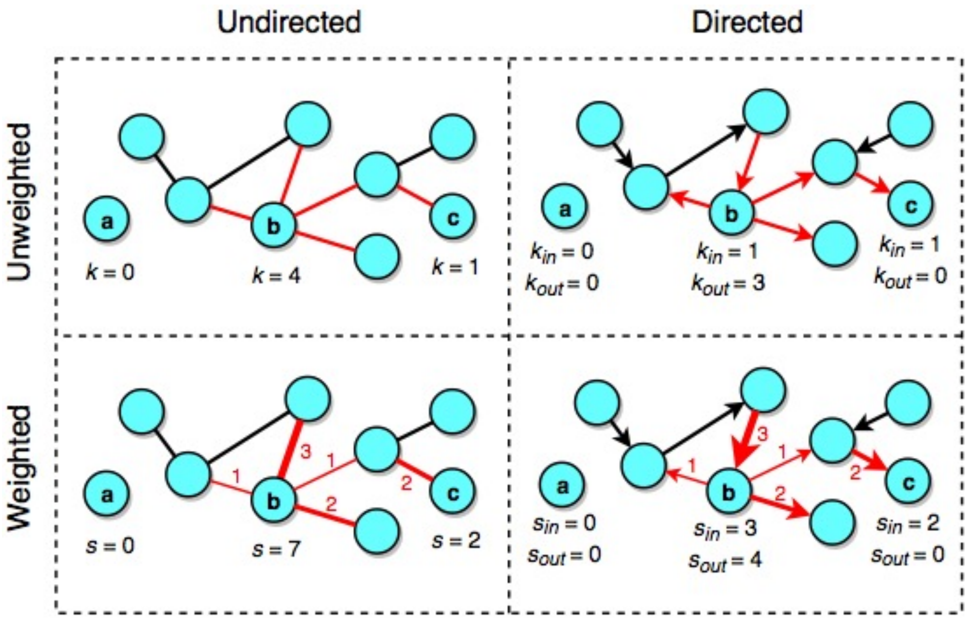
Network Analysis

AN INTRODUCTION FOR HUMANISTS

Dr Katarzyna Anna Kapitan
13 February 2025

Recap



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

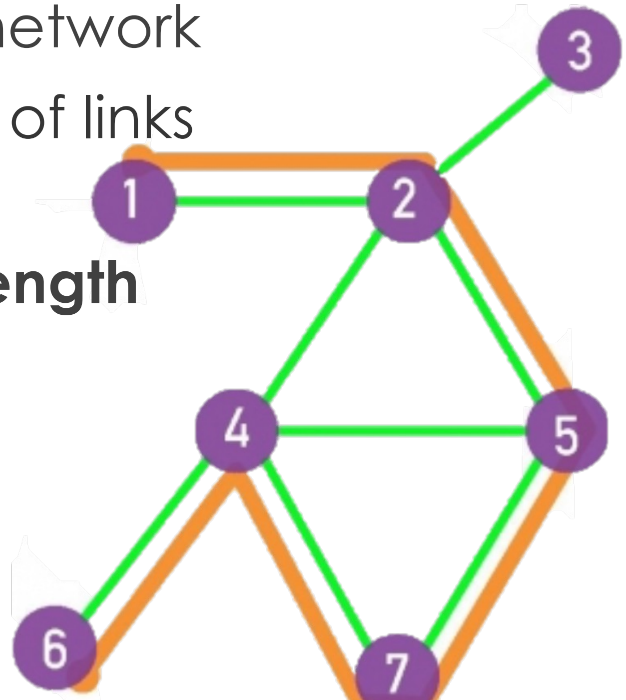
Degree
- In-degree
- Out-degree
- Total degree

Strength
- In-strength
- Out-strength
- Total strength

Paths

Paths

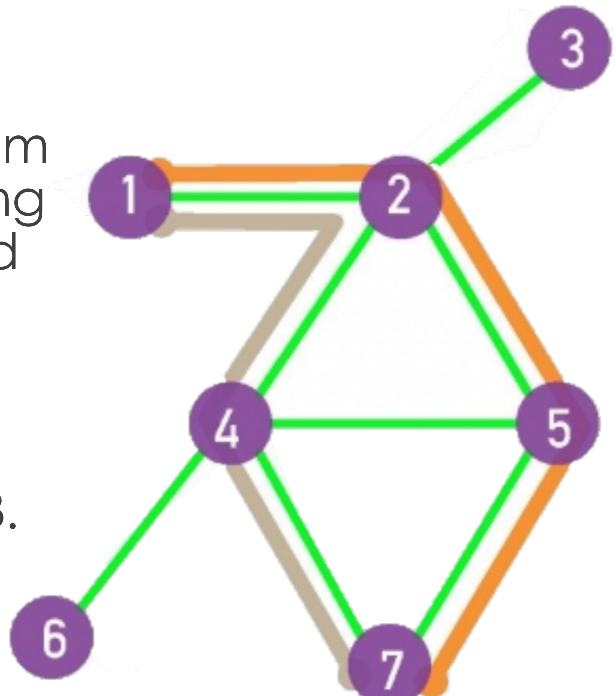
- A **path** is a route that runs along the links of the network
- A **path** between nodes i_0 and i_n is an ordered list of links $P = \{(i_0, i_1), (i_1, i_2), \ldots, (i_{n-1}, i_n)\}$.
- The number of links in a path is called the **path length**
- The **path** in orange between ***node 1*** and ***node 6*** follows the route 1→2→5→7→4→6 with links {(1,2), (2,5), (5,7), (7,4), (4,6)}, so its **length is 5**.



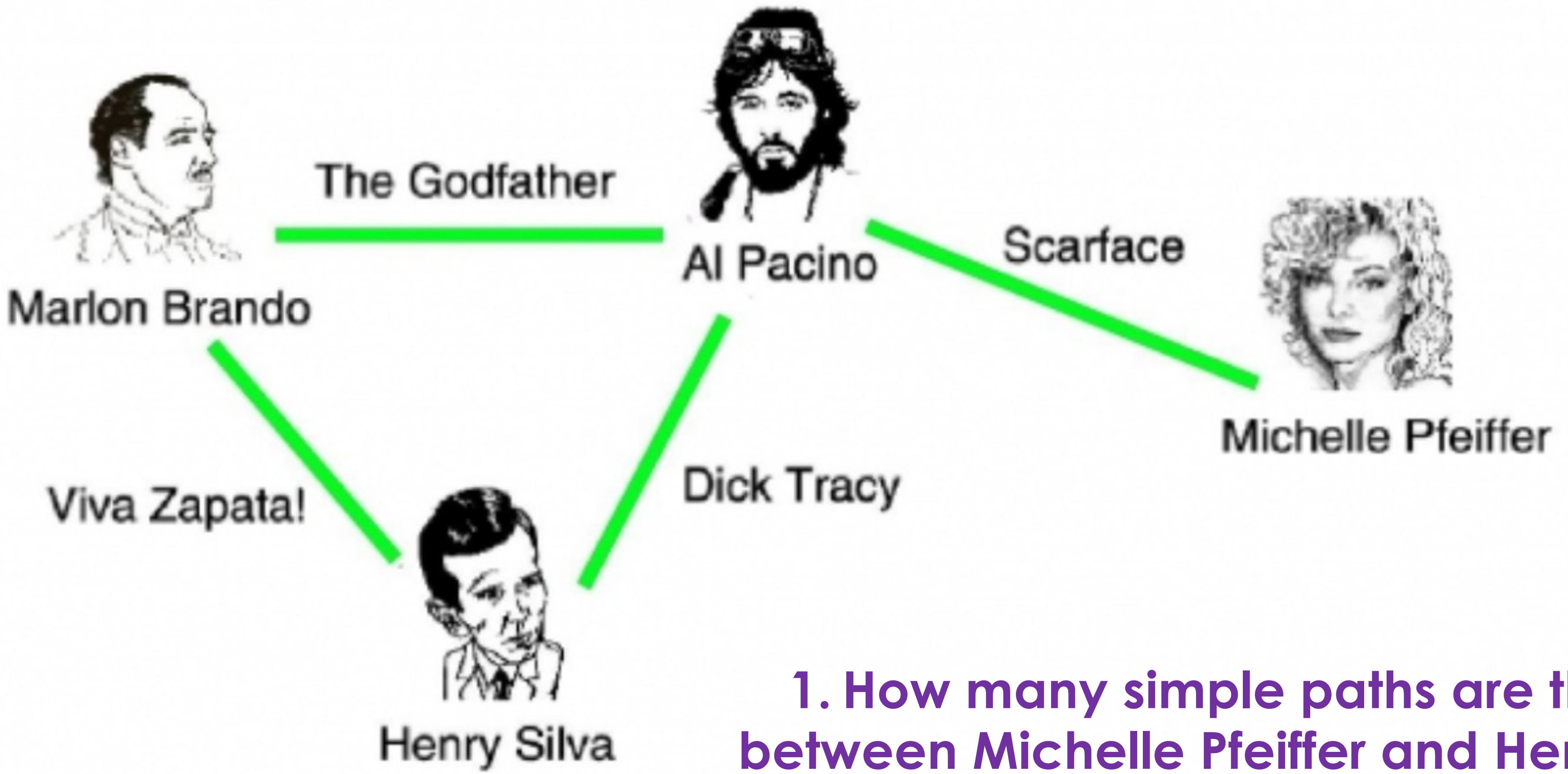
Source: Barabási, Network Science (https://networksciencebook.com)

The Shortest Path (Distance)

- The concept of a **path** is the basis of the definition of **distance** among nodes in a network.
- The **distance** between two nodes is defined as the minimum number of links that must be traversed in a path connecting the two nodes. Such a path is called **the shortest path**, and its length is called **the shortest-path length**.
- The distance between nodes *i* and *j* is denoted by $\boldsymbol{d_{ij}}$
- The shortest path between ***node 1*** and ***node 7*** has the **length of 3, so the distance between these nodes ($d_{1,7}$) is 3**. **Two possible shortest paths exist, one marked in grey 1→2→4→7 and one in orange 1→2→5→7.*



Source: Barabási, Network Science (https://networksciencebook.com)

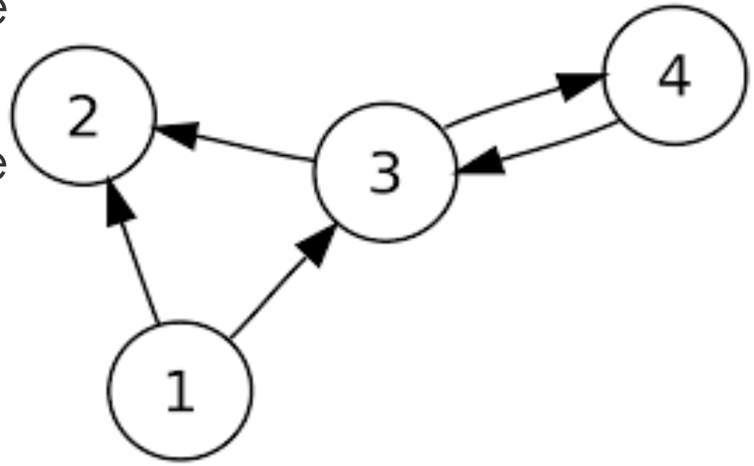


1. How many simple paths are there between Michelle Pfeiffer and Henry Silva?
2. What's the length of each of these paths?
3. What is the distance between Michelle Pfeiffer and Henry Silva?

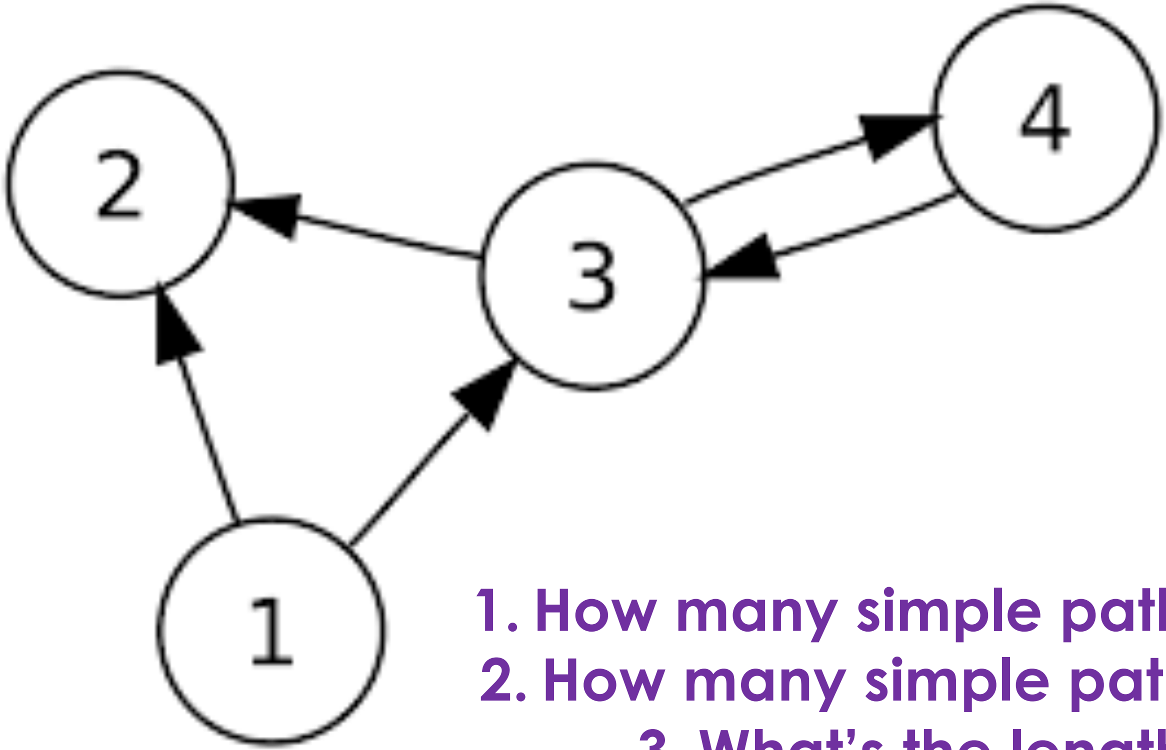
Source: Barabási, Network Science (https://networksciencebook.com)

The Shortest Path (Distance)

- In an undirected network $\mathbf{d_{ij} = d_{ji}}$, (the distance between node *i* and *j* is the same as the distance between node *j* and *i*).
- In directed graphs each path needs to follow the direction of the relationships.
- In a directed network the existence of a path from node *i* to node *j* does not guarantee the existence of a path from *j* to *i*. In a directed network often $\mathbf{d_{ij} \neq d_{ji}}$.
- **If the two nodes are disconnected, the distance is infinity.*



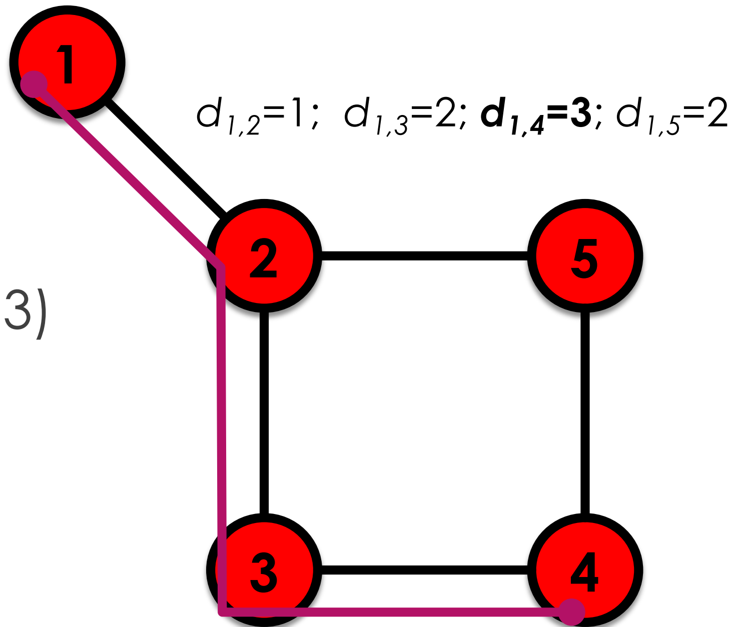
Source: Wikipedia, Directed Graph.



1. How many simple paths are there between 1 and 4?
2. How many simple paths are there between 4 and 1?
3. What's the length of each of these paths?
4. What is the distance between 1 and 4?
5. What is the distance between 4 and 1?

Diameter

- The **diameter** of the network (***d**$_{max}$*) is the maximum shortest-path length across all pairs of nodes (or the length of the longest shortest path in the network, or the **maximum distance**).
- The diameter of this network is equal to the distance from ***node 1*** to ***node 4***, which is 3 ($d_{1,4} = 3$)
- $d_{max} = 3$



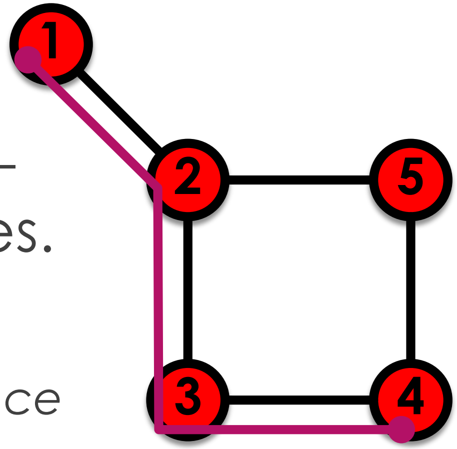
Source: Barabási, Network Science (https://networksciencebook.com)

Average Path Length

- The average shortest-path length (**average path length | average distance**) is denoted as **⟨*d*⟩** and is obtained by averaging the shortest-path lengths across all pairs of nodes.

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j)$$

- **In undirected networks we count each distance twice, so the distance of the path $d_{1,2}$ and $d_{2,1}$.*



$\boldsymbol{d_{1,2}=1}$; $d_{1,3}=2$; $d_{1,4} = 3$; $\boldsymbol{d_{1,5}=2}$
$\boldsymbol{d_{2,1}=1}$; $d_{2,3}=1$; $d_{2,4}=2$; $d_{2,5}=1$
$d_{3,1}=2$; $d_{3,2}=1$; $d_{3,4}=1$; $d_{3,5}=2$
$d_{4,1}=3$; $d_{4,2}=2$; $d_{4,3}=1$; $d_{4,5}=1$
$\boldsymbol{d_{5,1}=2}$; $d_{5,2}=1$; $d_{5,3}=2$; $d_{5,4}=1$

N = 5

$\langle d \rangle$ = $(d_{1,2}$ + … + $d_{ij})$/ N(N-1)
$\langle d \rangle$ = (1+2+3+2) + (1+1+2+1) + (2+1+1+2) + (3+2+1+1) + (2+1+2+1) /5*4
$\langle d \rangle$ = 8 + 5 + 6 + 7 + 6 / 20
$\langle d \rangle$ = 32/20
$\langle d \rangle$ = 1.6

Why should we care about path lengths?

Small Worlds

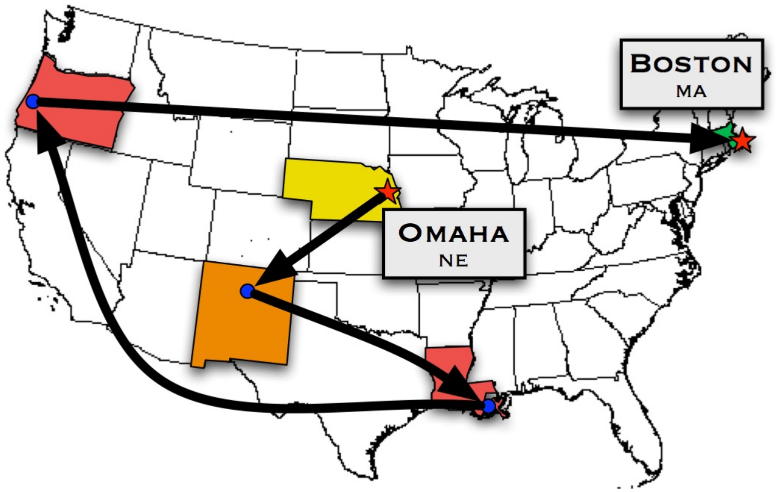
Small Worlds

- **Social networks** (and many other real-life networks) have **short distances** (short average shortest-path length) between any pair of nodes.
- The is known as **small world phenomenon**, or **six degrees of separation**.
- If you choose any two individuals anywhere on earth, you will find a path of at most six acquaintances between them.

Small Worlds & Frigyes Karinthy

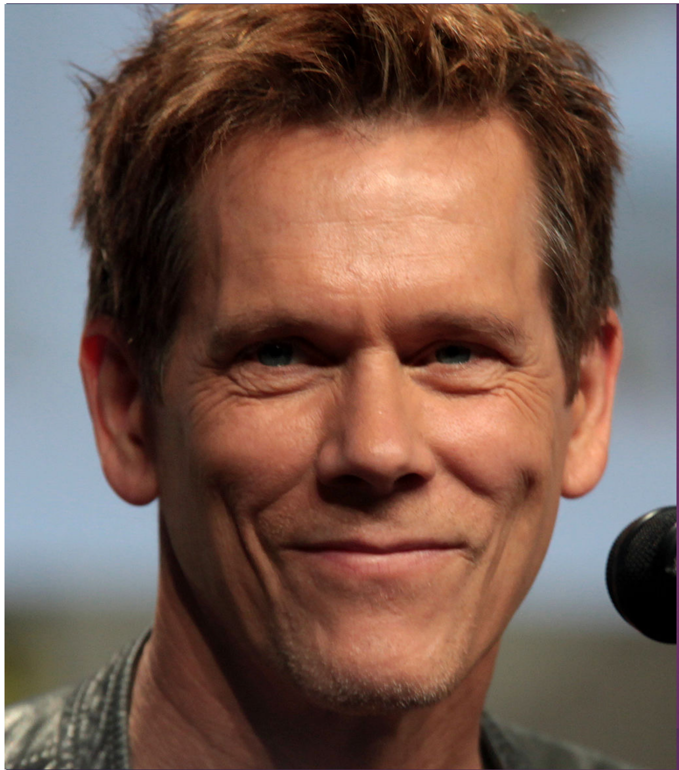
- The idea was first introduced by a Hungarian writer names Frigyes Karinthy in his short story “Chains” from 1929:
- "Planet Earth has never been as tiny as it is now. It shrunk […] due to the quickening pulse of both physical and verbal communication. […] [T]he population of the Earth is closer together now than they have been before. We should select any person from the 1.5 billion inhabitants of the Earth – anyone, anywhere at all. He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances." (Source: http://vadeker.net/articles/Karinthy-Chain-Links_1929.pdf)

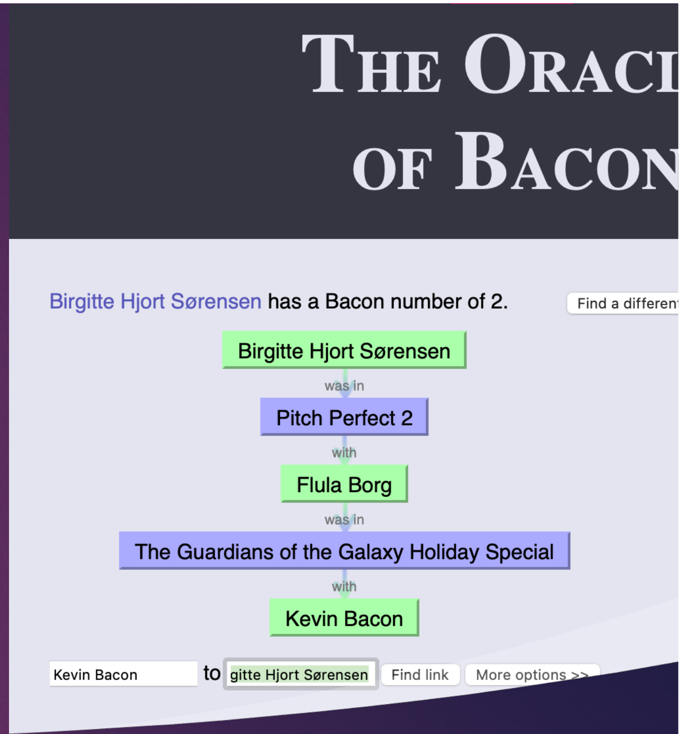
Small Worlds & Stanley Milgram



The path of one of the letters: Omaha, Nebraska -> Santa Fe, New Mexico-> New Orleans, Louisiana -> Eugene, Oregon-> Boston, Massachusetts. Source: Menczer et al.

- **Stanley Milgram's**1967 experiment measuring the **social distance** between people in the US.
- 160 letters sent to random people in Nebraska & Kansas.
- Instructions to forward it to a personal acquaintance who is likely to know the target
- 2 targets in Massachusetts
- 42 letters made it to the targets (26%)
- Average 6.5 steps (range: 3-12 steps)





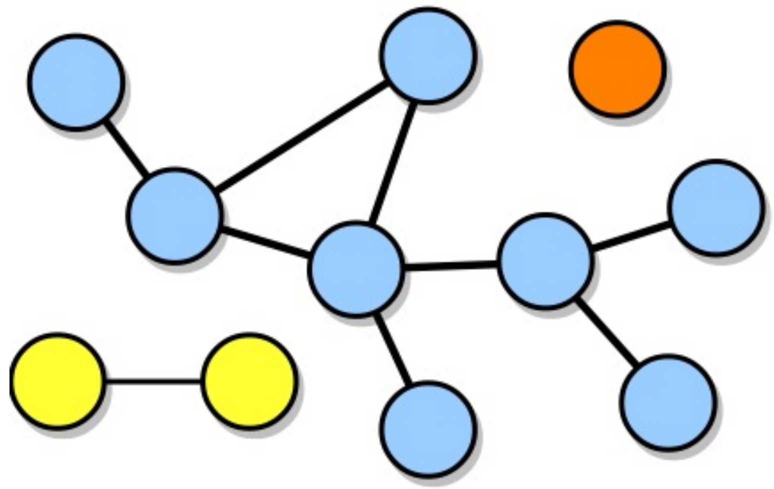
Six degrees of Kavin Bacon

Kevin Bacon (1958–), photo by Gage Skidmore, CC BY-SA 2.0 (source: wikipedia); Birgitte Hjort Sørensen (1982–) as on a poster for "The Passion of Marie", CC BY 3.0 (source: wikipedia); Sørensen's Becon score of 2 (source: https://oracleofbacon.org)

Connectedness

Connectedness in Undirected Networks

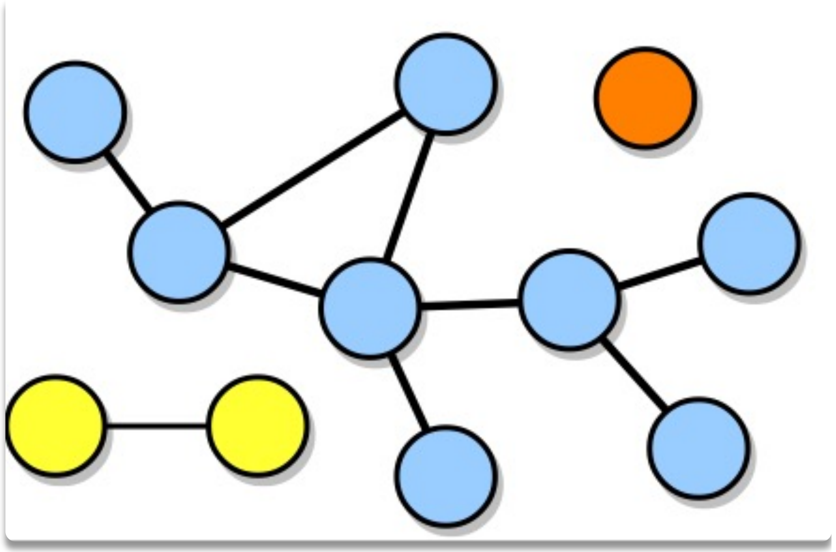
- In an undirected network nodes *i* and *j* are **connected** if there is a **path** between them. They are **disconnected** if such a path does not exist, in which case we have $d_{ij} = \infty$.
- A network is **connected** if there is a path between any two nodes.
- **A network is disconnected** if there is at least one pair with $d_{ij} = \infty$.
- Disconnected network is composed of more than one connected components, or simply components.



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Connectedness in Undirected Networks

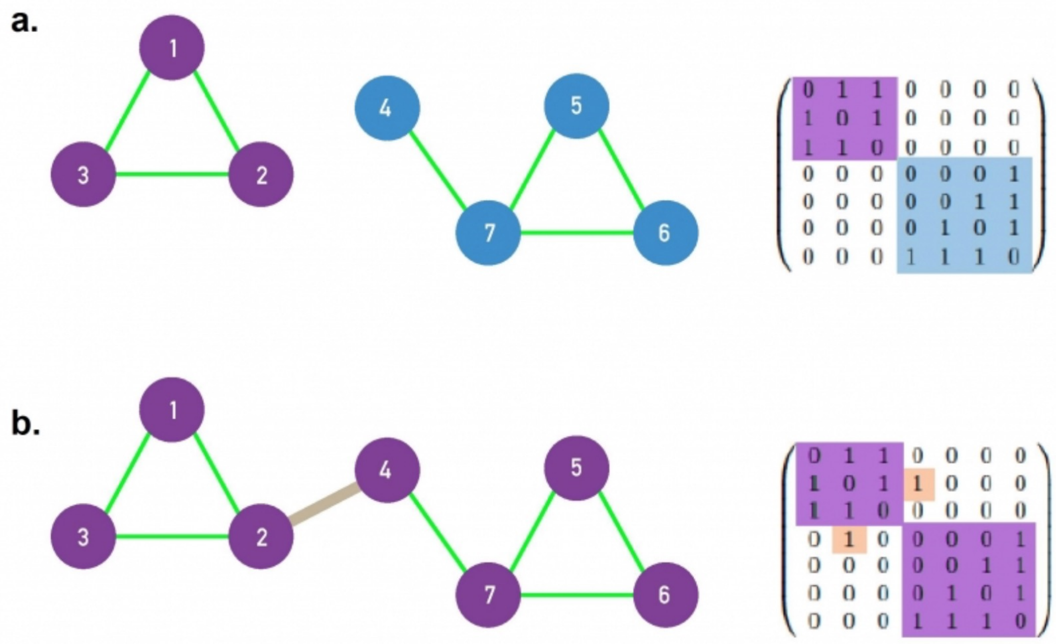
- **A component** is a subnetwork containing one or more nodes - there is a path connecting any pair of these nodes, but there is no path connecting them to other components.
- The largest connected component is called **the giant component**
- **A singleton** is the smallest-possible component (a node not connected to anything)



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Connectedness in Undirected Networks

- If a network consists of two components, a properly placed single link can connect them, making the network connected. Such a link is called a **bridge**.
- In general, a bridge is any link that, if cut, disconnects the network.



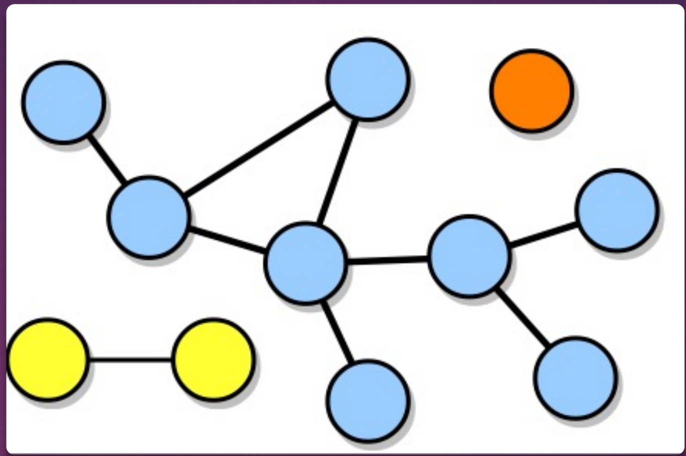
Source: Barabási, Network Science (https://networksciencebook.com)

Disconnected Network & Distance

- If the two nodes are disconnected, the distance is infinity $d_{ij} = \infty$.
- We cannot measure the average path length nor the diameter for the entire network, if the network is disconnected.

BUT

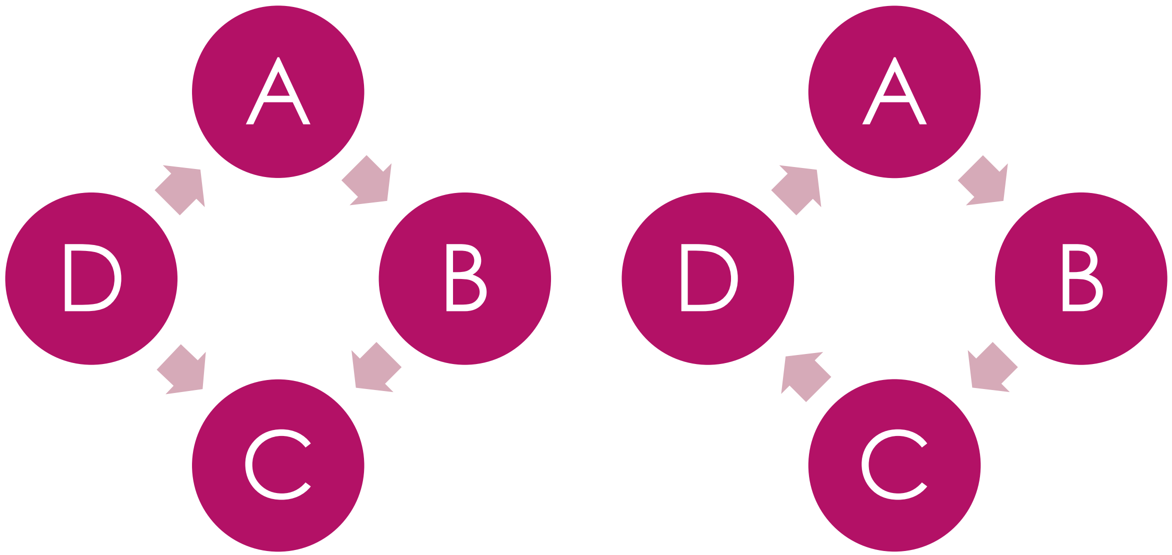
- We can measure all of them for each of the components (subnetworks).



What is the diameter of each of the components?

Connectedness in Directed Networks

- A directed network can be **strongly** or **weakly connected.**
- A network is **strongly connected** if there is a path between any two nodes respecting the link directions.
- A network is **weakly connected** if the is a path between any two nodes regardless of the link directions.



Network 1

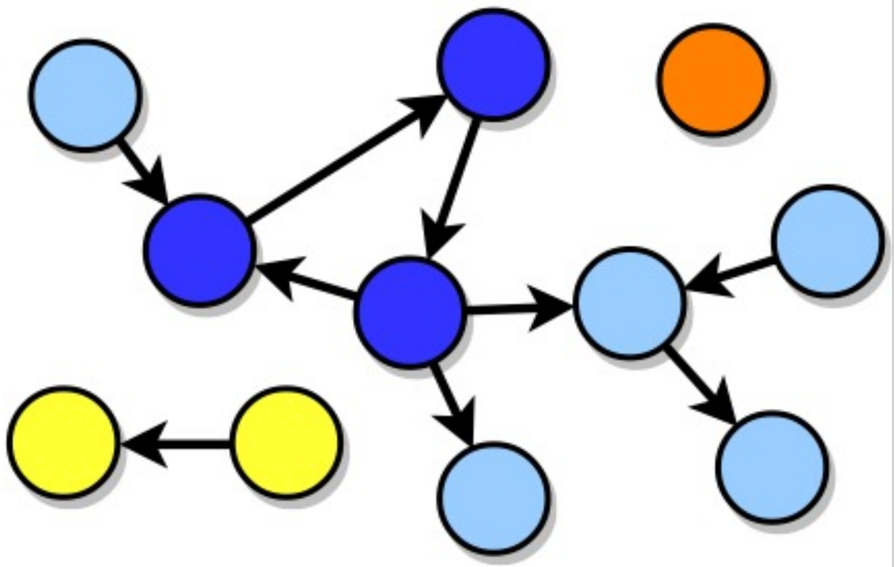
Network 2

Which one is which?

Connectedness in Directed Networks

- In disconnected networks we can talk about **strongly** or **weakly connected** components.

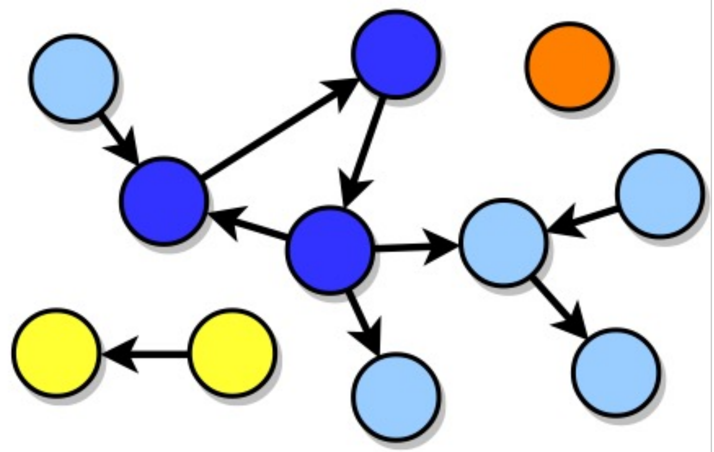
- The yellow component is: ?
- The orange component is: ?
- The dark blue component is: ?
- The blue (both light and dark) component is: ?



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Connectedness in Directed Graphs

- The **in-component** of a **strongly connected component** *S* is the set of nodes from which one can reach *S*, but that cannot be reached from *S*.
- The **out-component** of a **strongly connected component** *S* is the set of nodes that can be reached from *S*, but from which one cannot reach *S*.

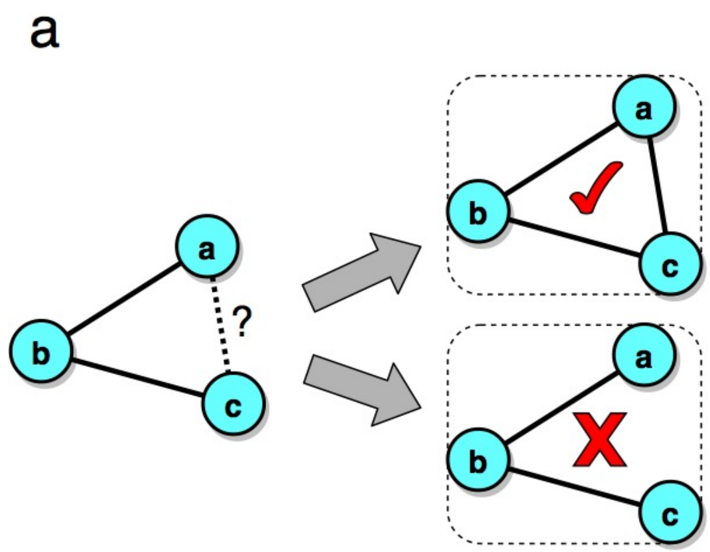


Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Clustering Coefficient

Clustering Coefficient

- The clustering coefficient captures the degree to which the neighbours of a given node link to each other.
- Think about as triangles of friendship
- If A(Alice) is friends with B (Bob) and C (Charlie) is also friends with B (Bob), then A (Alice) and C (Charlie) are also likely to be friends of each other.



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Clustering Coefficient

- We can measure the number of triangles that a node actually has relative to how many it could have
- Similarly to **density**, the **clustering coefficient** of a node is the fraction of pairs of the node's neighbours that are connected to each other:

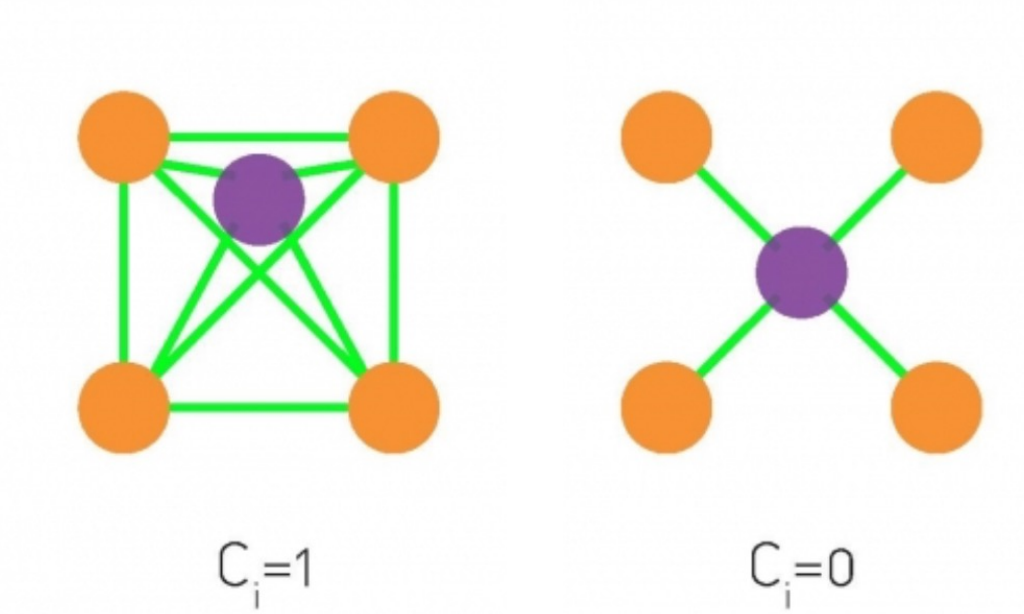
$$C_{i\,=\,2}Li\ /\ k_i\,(ki{-}1)$$

L_i - Number of edges between neighbours of node i
k_i - Degree of node i (number of neighbours)

The denominator k_i (k_i−1)/2 is the maximum possible number of edges between i's neighbours (the number of possible triangles).

Clustering Coefficient

- The value of C_i is always between 0 and 1 (it's a fraction)
- $\mathbf{C_i = 0}$ if none of the neighbours of node *i* (purple) link to each other.
- $\mathbf{C_i = 1}$ if the neighbours of node *i* (purple) form a complete graph, i.e. they all link to each other.



Source: Barabási, Network Science (https://networksciencebook.com)

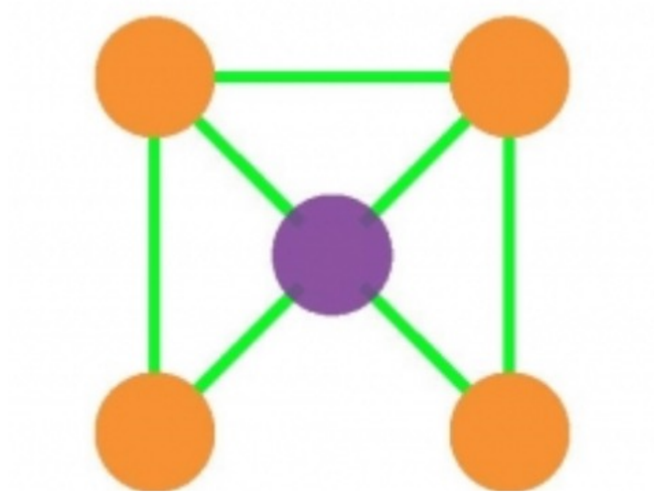
Clustering Coefficient

$$C_{i} = {}_{2}Li \;/\; k_{i}(ki-1)$$

L_i - Number of edges between neighbours of node *i* (purple)

k_i = Degree of node *i* (number of its neighbours)

Ci = ?



Source: Barabási, Network Science (https://networksciencebook.com)

Clustering Coefficient

$$C_{i} = 2L_i / k_i(k_i-1)$$

L_i - Number of edges between neighbours of node *i* (purple)

$L_i = 3$

k_i = Degree of node *i* (number of its neighbours)

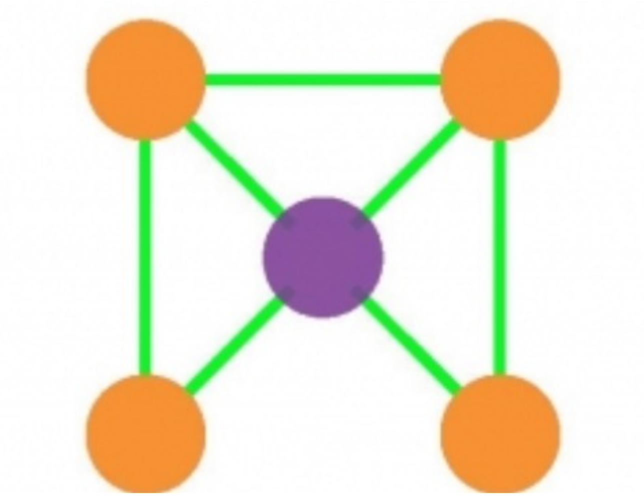
$k_i = 4$

Ci = 2 * 3 / 4 (4-1)

Ci = 6 / 4*3

Ci = 6/12

Ci = 0.5



Clustering Coefficient

- The **global clustering coefficient** measures the overall tendency of nodes to form triangles.
- It is defined as:

$$C = \frac{\sum_i C_i}{N}$$

- C_i = Local clustering coefficient of node i
- N = Total number of nodes

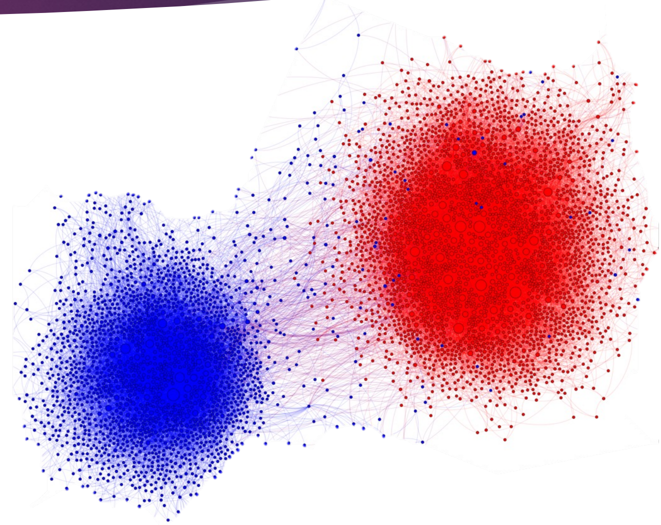
Why should we care about Clustering Coefficient?

Assortativity

Katarzyna Anna Kapitan, Network Analysis for Humanists,
Paris 2025

Assortativity

- Assortativity (positive assortativity) – the tendency for nodes to connect to other nodes with similar properties within a network.
- Disassortativity (negative assortativity) – the tendency for nodes to connect to other nodes with dissimilar properties within a network



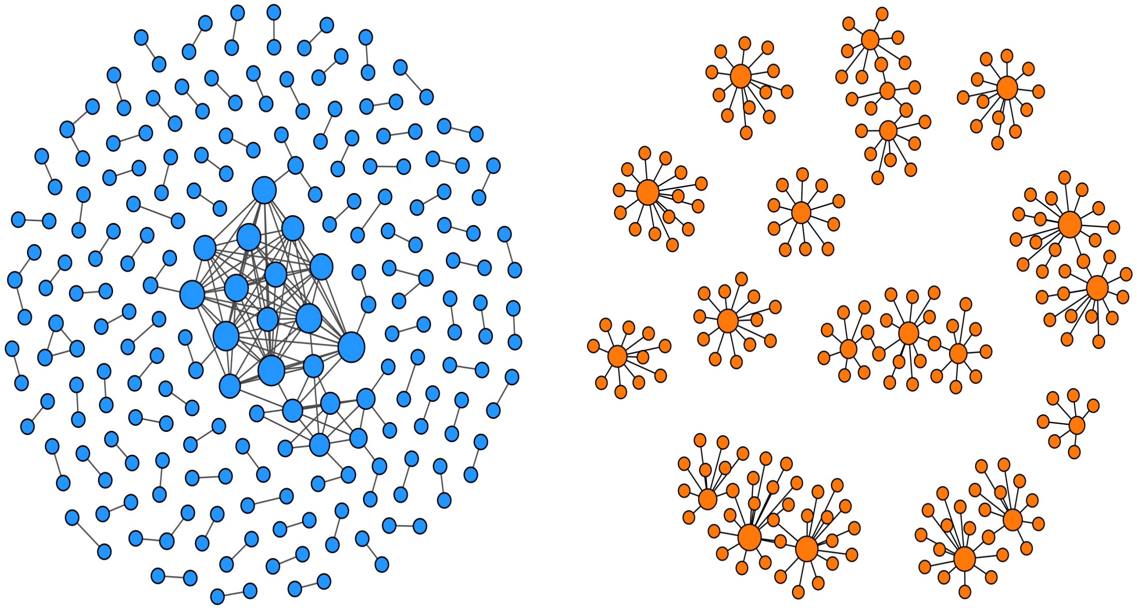
Retweet network on Twitter, based on political posts during 2010 US election. Links represent retweets of posts that used hashtags such as #tcot (top conservatives on Twitter) and #p2 (Progressives 2.0) associated with conservative (red) and progressive (blue) messages.
Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Degree Assortativity (degree correlation)

- Assortativity based on degree is called **degree assortativity** or **degree correlation.**
- Networks where high-degree nodes tend to be connected to other high-degree nodes (and low-degree to low-degree) are called **assortative**.
- Networks where high-degree nodes tend to be connected to low-degree nodes and vice versa are called **disassortative**.

Degree Assortativity (degree correlation)

If node's degree is represented by its size, which types of networks are the blue network and the orange network?



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*,(2023).

Why should we care about degree assortativity?