Friendly Graph Theory: Clustering & Homophily

Introduction to Network Science

Instructor: Michele Starnini — https://github.com/chatox/networks-science-course

Universitat Pompeu Fabra Barcelona

Contents:

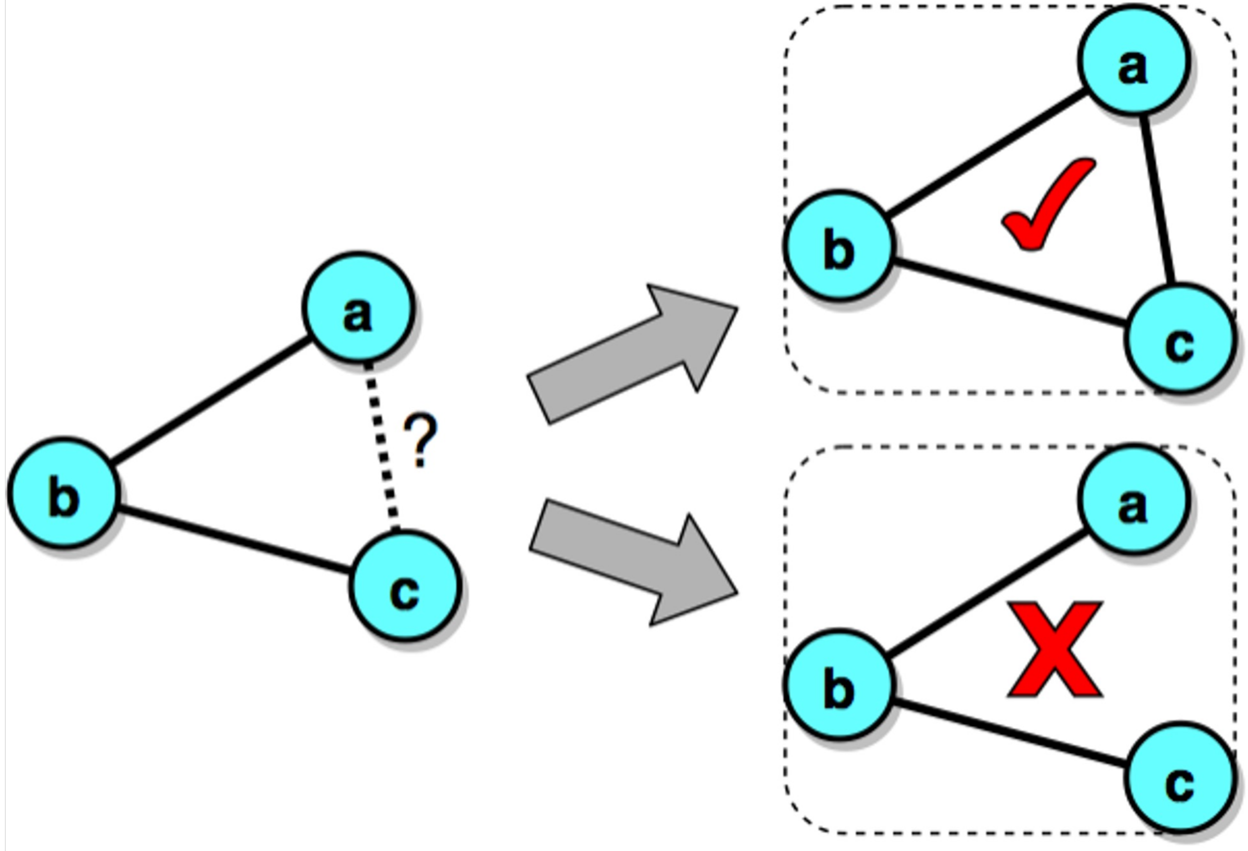
- Clustering
- Homophily

all related to friendship in social networks!

Clustering

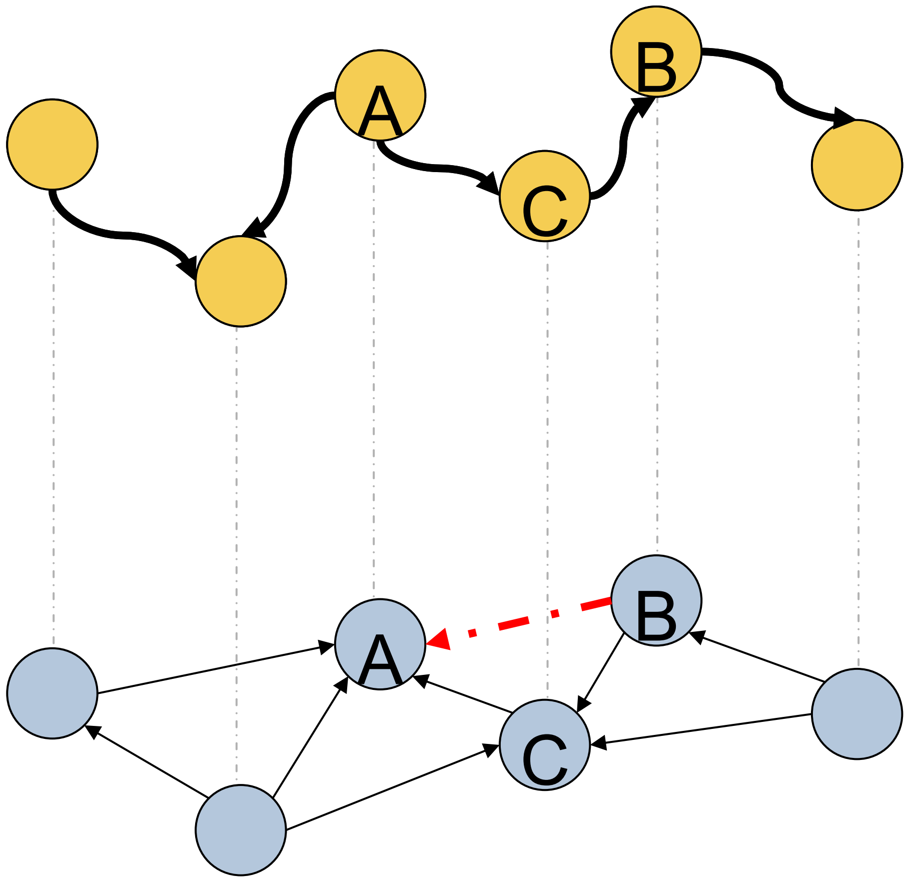
Who is a friend? [Triangle closure]

A prevalent way in which we form friendships is by befriending **friends of friends**



Source: Menczer, Fortunato, David: A First Course on Networks Science. Cambridge, 2020.

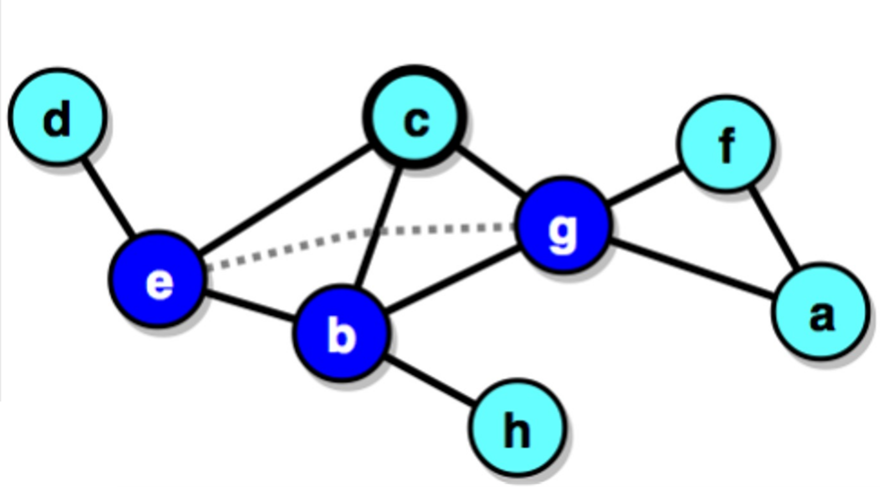
Tendency to form triangles



- The dynamics **on** the network, i.e., information diffusion, affect the dynamics **of** the network, i.e., the creation of links
- *B* is more likely to start following *A* after seeing content posted by *A* and re-posted by an account *C* that *B* already follows

https://doi.org/10.1145/2487575.2487607

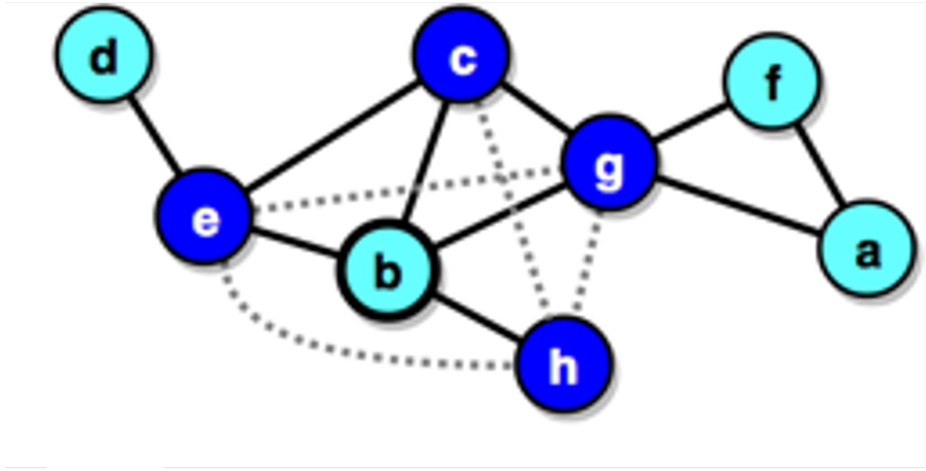
Example 1



Node c has 3 neighbors: e, b, g
They form two triangles out of the possible 3 (the missing one is drawn with a dotted line)

Source: Menczer, Fortunato, David: A First Course on Networks Science. Cambridge, 2020.

Example 2



Node b has 4 neighbors: e, c, g, h
They form two triangles out of the possible 6 (the missing ones are drawn with a dotted line)

Source: Menczer, Fortunato, David: A First Course on Networks Science. Cambridge, 2020.

Remember

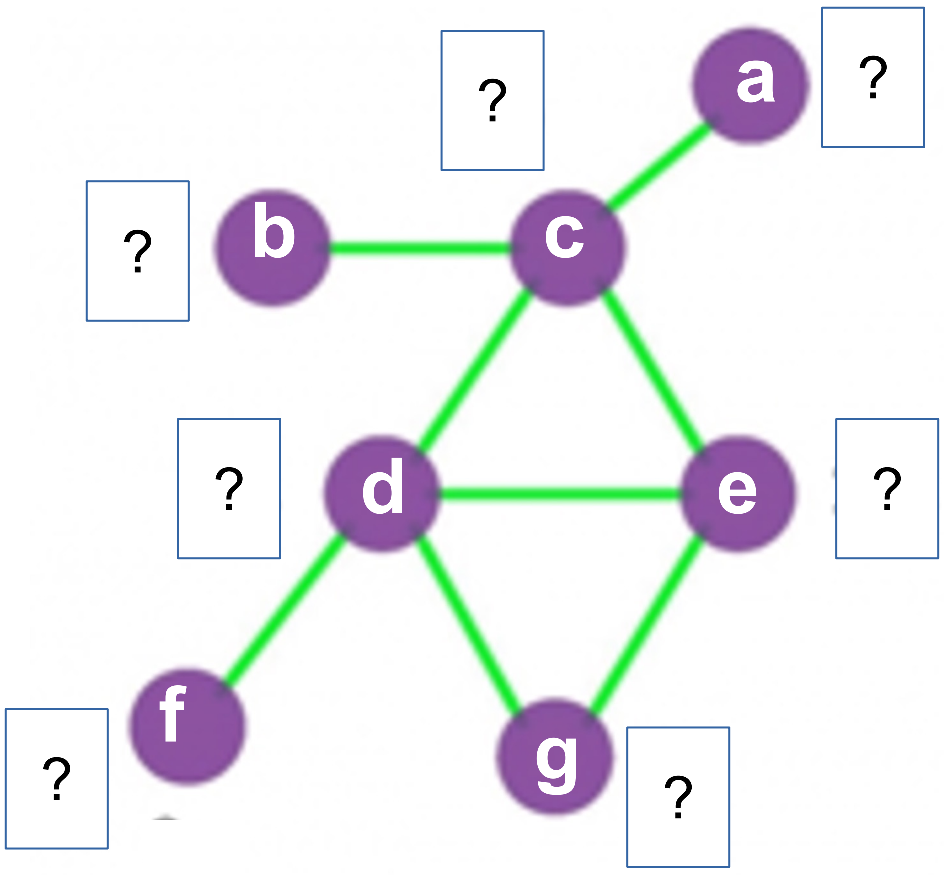
- The maximum number of links between *k* nodes is

$$\frac{k(k-1)}{2}$$

Local clustering coefficient

- The **local clustering coefficient** C_i is a property of a node i
- Let L_i represent the number of links among neighbors of node i

$$C_i = \frac{L_i}{\frac{k_i(k_i-1)}{2}} = \frac{2L_i}{k_i(k_i-1)} \quad C_i \triangleq 0 \text{ if } k_i \leq 1$$



Exercise

What is the local clustering coefficient of each node?

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

$$C_i \triangleq 0 \text{ if } k_i \leq 1$$

Pin board: https://upfbarcelona.padlet.org/chato/v0apheshv2l4hbot

Local clustering coefficient

- Degree correlations (assortativity) are related to the **two-point (nodes) correlations**
- Clustering is related to the **three-point (nodes) correlations**

Prob. that a k'-node is connected to a k″-node

Average clustering coefficient of k-nodes

$$\bar{c}(k) = \sum_{k',k''} p(k'', k'|k)p(k'', k')$$

Conditional prob. that a k-node is connected to a k'-node & a k″-node

Average clustering coefficient
(*“global clustering coefficient”*)

The **average clustering coefficient** is a property of the entire graph

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

Sometimes this is called the *curvature* of a graph

Network clustering coefficient

- Social networks tend to have high clustering coefficients because of **triadic closure**: we meet through common friends

- Other networks, e.g., bipartite and tree-like networks, have low clustering coefficient

Table 2.1 Average path length and clustering coefficient of various network examples. The networks are the same as in Table 1.1, their numbers of nodes and links are listed as well. Link weights are ignored. The average path length is measured only on the giant component; for directed networks we consider directed paths in the giant strongly connected component. To measure the clustering coefficient in directed networks, we ignore link directions.

Network	Nodes (N)	Links (L)	Average path length ($\langle \ell \rangle$)	Clustering coefficient (C)
Facebook Northwestern Univ.	10,567	488,337	2.7	0.24
IMDB movies and stars	563,443	921,160	12.1	0
IMDB co-stars	252,999	1,015,187	6.8	0.67
Twitter US politics	18,470	48,365	5.6	0.03
Enron Email	87,273	321,918	3.6	0.12
Wikipedia math	15,220	194,103	3.9	0.31
Internet routers	190,914	607,610	7.0	0.16
US air transportation	546	2,781	3.2	0.49
World air transportation	3,179	18,617	4.0	0.49
Yeast protein interactions	1,870	2,277	6.8	0.07
C. elegans brain	297	2,345	4.0	0.29
Everglades ecological food web	69	916	2.2	0.55

Real networks

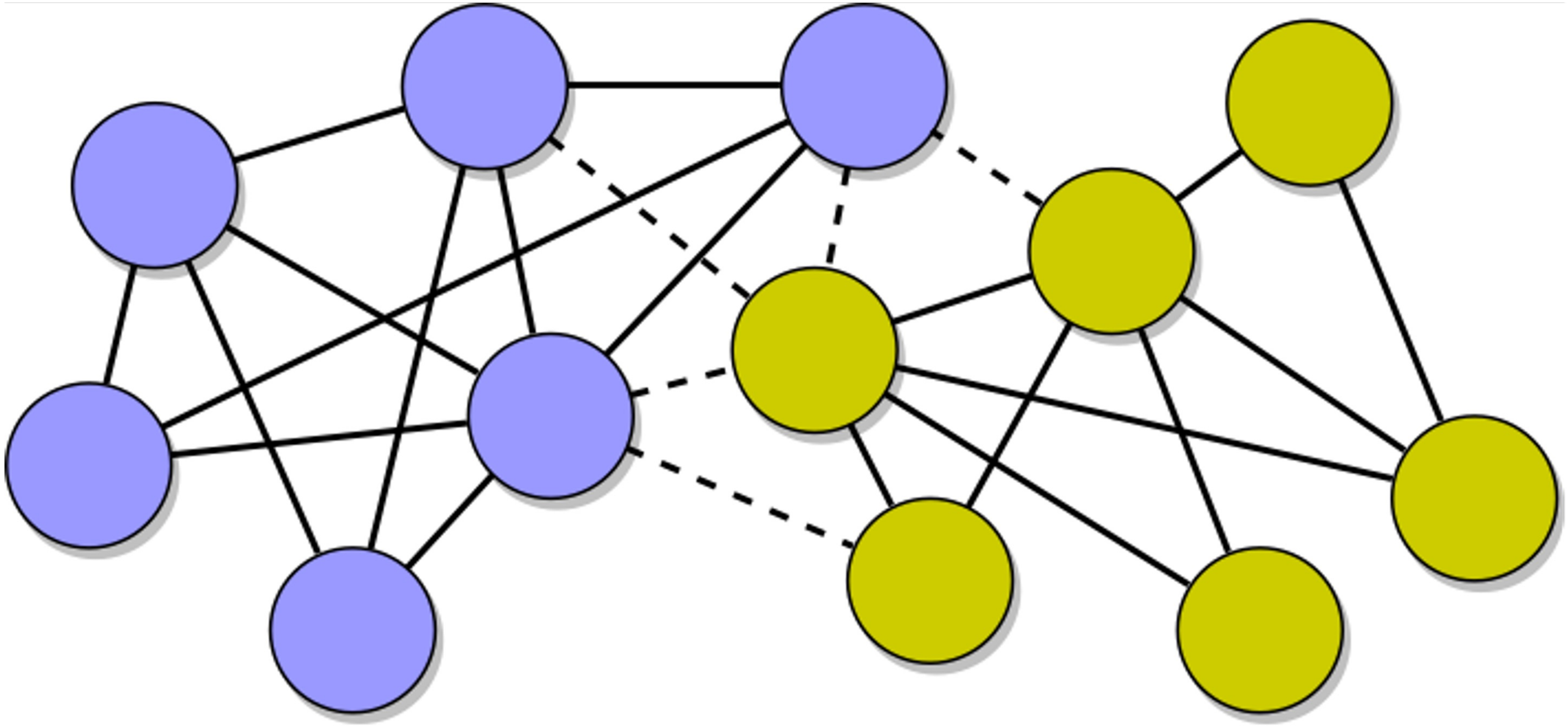
- Most real, complex networks show high clustering
- Clustering is particularly high in social networks
- Example: triangle closure is very successful recommending system

Homophily

Who is a friend? [Homophily]

- In social networks, nodes have **features** that influence their connectivity preferences
  - Age, gender identity, ethnicity, sexual preference, location, topics of interest, artistic sensitivities, …
- People tend to befriend those who are like them: that is called **homophily**

“Birds of a feather flock together”

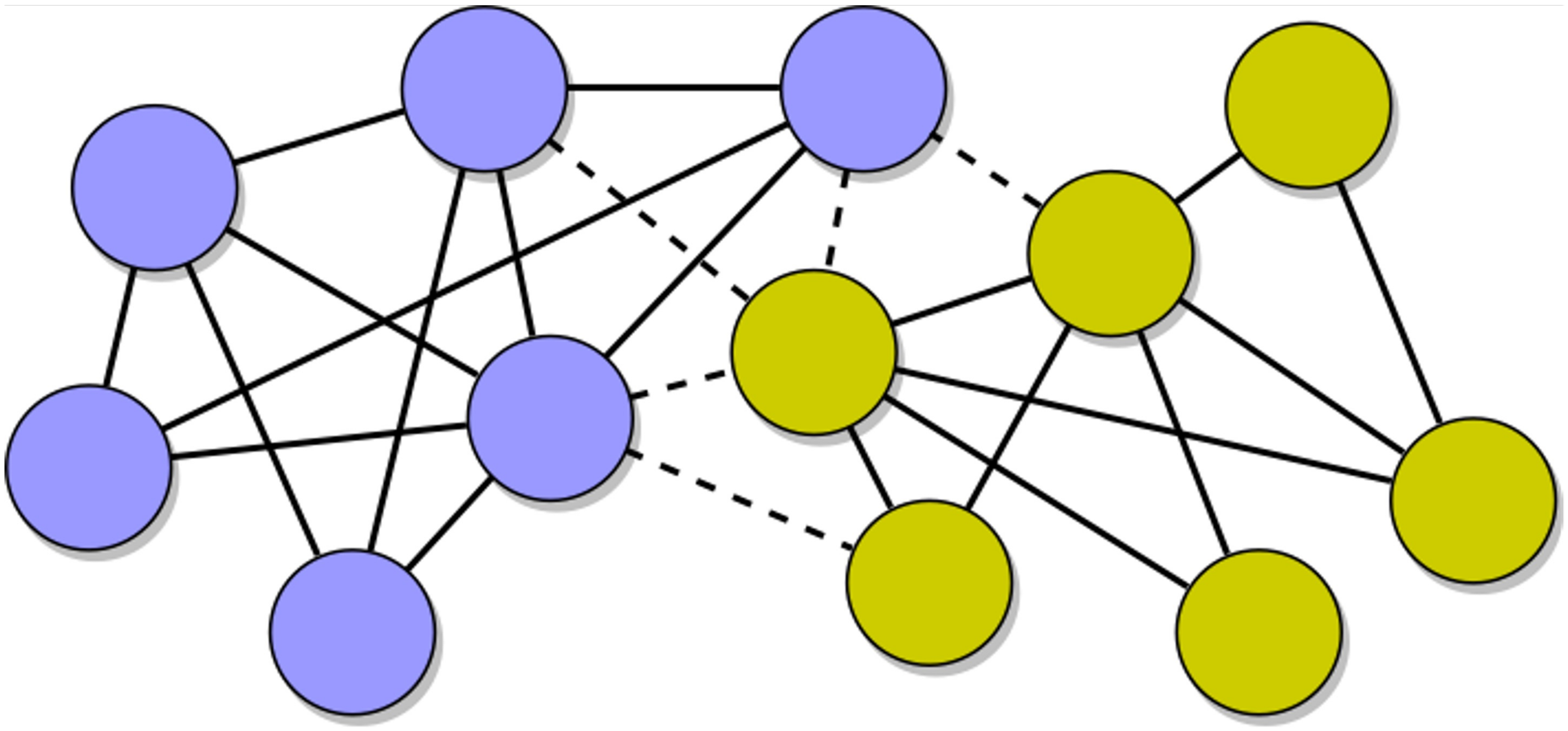


Source: Menczer, Fortunato, David: A First Course on Networks Science. Cambridge, 2020.

Quantifying homophily

- Let G be a graph of N nodes: N_a "yellow" and N_b "blue"
  - $N = N_a + N_b$
- Let G have L undirected links (including self loops), of which L_{aa} connect yellow to yellow, L_{ab} connect yellow to blue, and L_{bb} connect blue to blue
  - $L = L_{aa} + L_{ab} + L_{bb}$ $L_a = L_{aa} + L_{ab}$ $L_b = L_{bb} + L_{ab}$

$N_a = 6, N_b = 6, L_a = 14, L_b = 16, L_{ab} = 5,$



Source: Menczer, Fortunato, David: A First Course on Networks Science. Cambridge, 2020.

Expected links across groups

If yellow nodes have L_a links placed at random (incl. self loops), how many should go to a blue node?

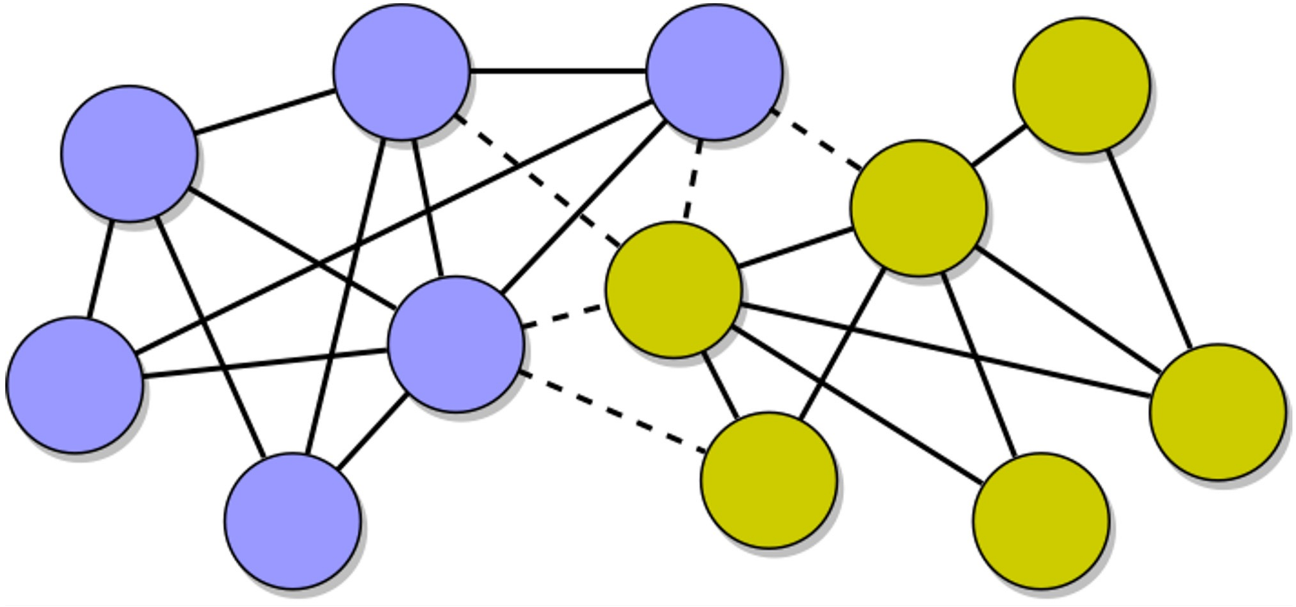
$$L_a \left(\frac{N_b}{N} \right)$$

Quantifying homophily of a group

- We compare observed against the **expected number of links crossing to the other group**
- <1 ⇒ homophily 1 ⇒ neutral >1 ⇒ heterophily

$$\text{Homophily}(a) = \frac{L_{ab}}{L_a\left(\frac{N_b}{N}\right)} \qquad \text{Homophily}(b) = \frac{L_{ab}}{L_b\left(\frac{N_a}{N}\right)}$$

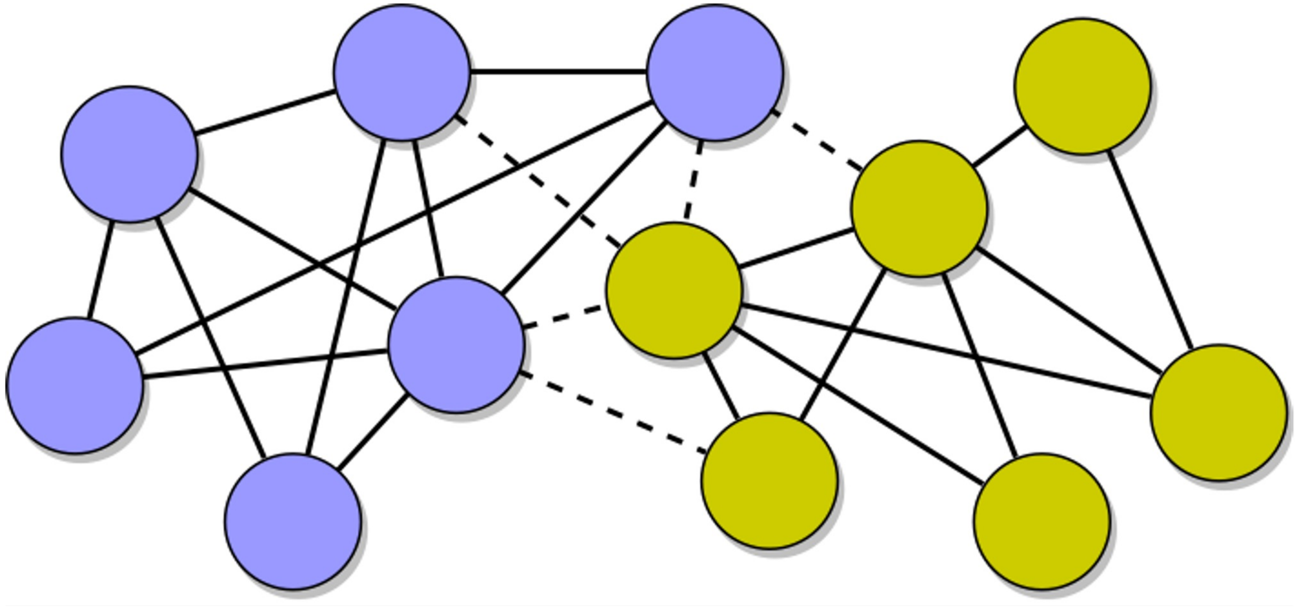
$$\text{Homophily}(a) = \frac{L_{ab}}{L_a\left(\frac{N_b}{N}\right)} = \frac{5}{14\left(\frac{6}{12}\right)} = \frac{5}{7}$$



Yellow nodes are homophilic

$N_a = 6$, $N_b = 6$, $L_a = 14$, $L_b = 16$, $L_{ab} = 5$,

$$\text{Homophily}(b) = \frac{L_{ab}}{L_b\left(\frac{N_a}{N}\right)} = \frac{5}{16\left(\frac{6}{12}\right)} = \frac{5}{8}$$

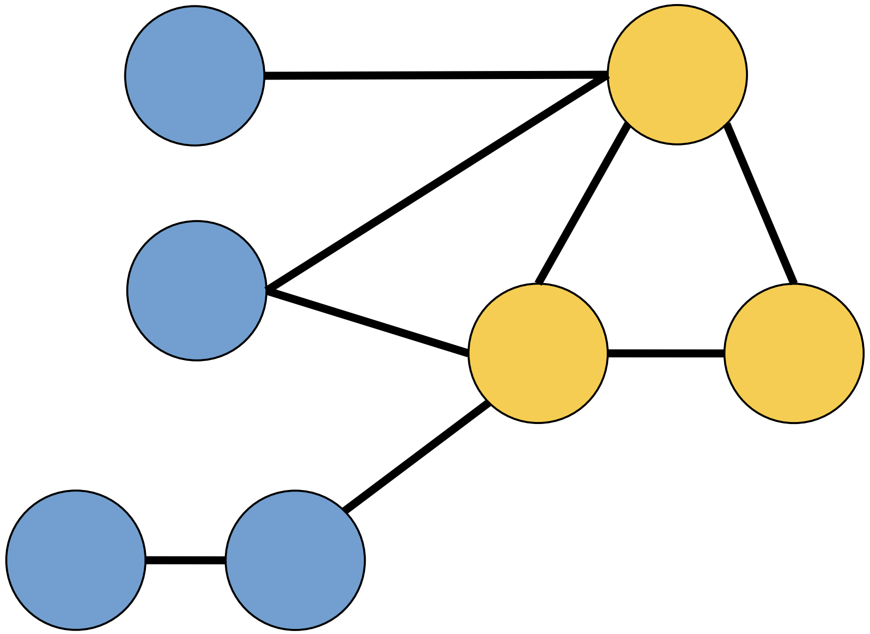


Blue nodes are homophilic

$N_a = 6$, $N_b = 6$, $L_a = 14$, $L_b = 16$, $L_{ab} = 5$,

Exercise

Compute homophily of both groups and indicate if each group is homophilic, heterophilic, or neutral



$$\text{Homophily}(a) = \frac{L_{ab}}{L_a\left(\frac{N_b}{N}\right)}$$

$$\text{Homophily}(b) = \frac{L_{ab}}{L_b\left(\frac{N_a}{N}\right)}$$

Pin board: https://upfbarcelona.padlet.org/chato/iig2u83qdzk4y7xc

Quantifying homophily of a group

- Preference to interact with similar individuals
- In networks: More in-group than out-group links

Social influence

Homophily:
Similar nodes become connected

Features → Connections

The opposite mechanism may also happen!

Social influence:
Connected nodes become more similar

Connections → Features

Social network structure (ie, who our friends are) can determine our thinking!

Echo-chambers: like-minded people tightly connected

- No diversity of opinions
- Confirmation bias
- Reinforcement of prejudices

Summary

Things to remember

- How to compute local and global clustering coefficients
- How to compute Homophily

Structural properties of real, complex networks

- Sparsity
- Heterogeneous degree distribution (hubs)
- Small world property (small average distance)
- Degree correlations
- Clustering

Sources

- A. L. Barabási (2016). Network Science – Chapter 02
- F. Menczer, S. Fortunato, C. A. Davis (2020). A First Course in Network Science – Chapter 02
- URLs cited in the footer of specific slides

Practice on your own

- Determine if the set {C, D, G} is homophilic or heterophilic
- Calculate local clustering coefficient of each node in this graph