Network Analysis:

The Hidden Structures behind the Webs We Weave

17-213 / 17-668

Homophily and Degree Correlation (Part 1)

Thursday, September 14, 2023

Patrick Park & Bogdan Vasilescu

Carnegie Mellon University
School of Computer Science

S3D
Software and Societal
Systems Department

2-min Quiz, on Canvas

Quick Recap – Last Tuesday’s Lecture

Graph signature of social ties

Social tie dynamics

Case Study: Clustering Coefficient

Measurement of Triadic Closure

Last Thursday

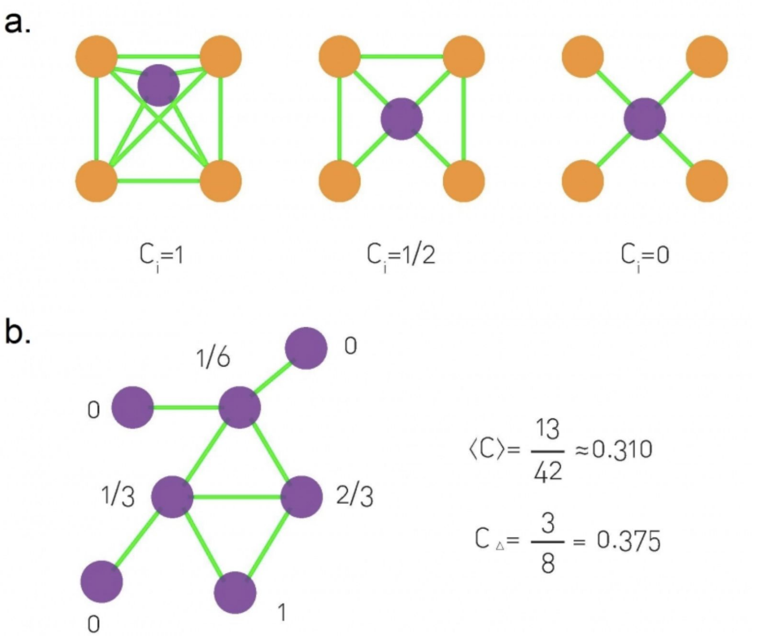
The extent of triadic closure in a network:

- Local clustering coefficient: The probability that two neighbors of a node are connected

Number of ties among *i*'s neighbors (excluding ties involving *i*)

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

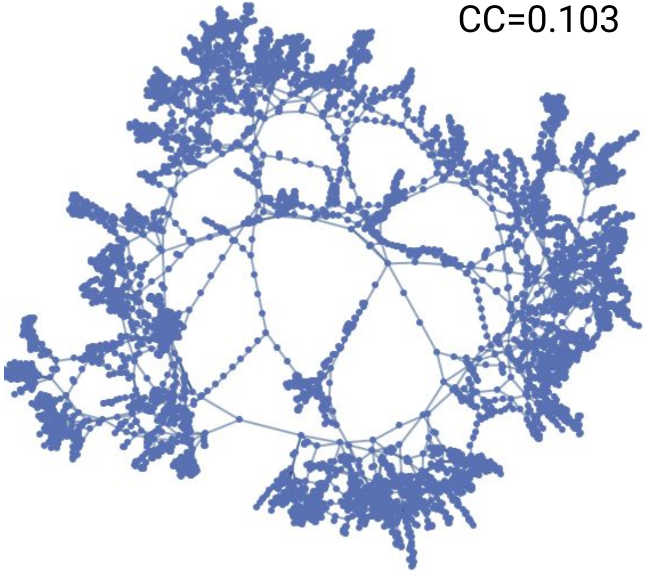
The average across *all* nodes is that network's "local" clustering coefficient



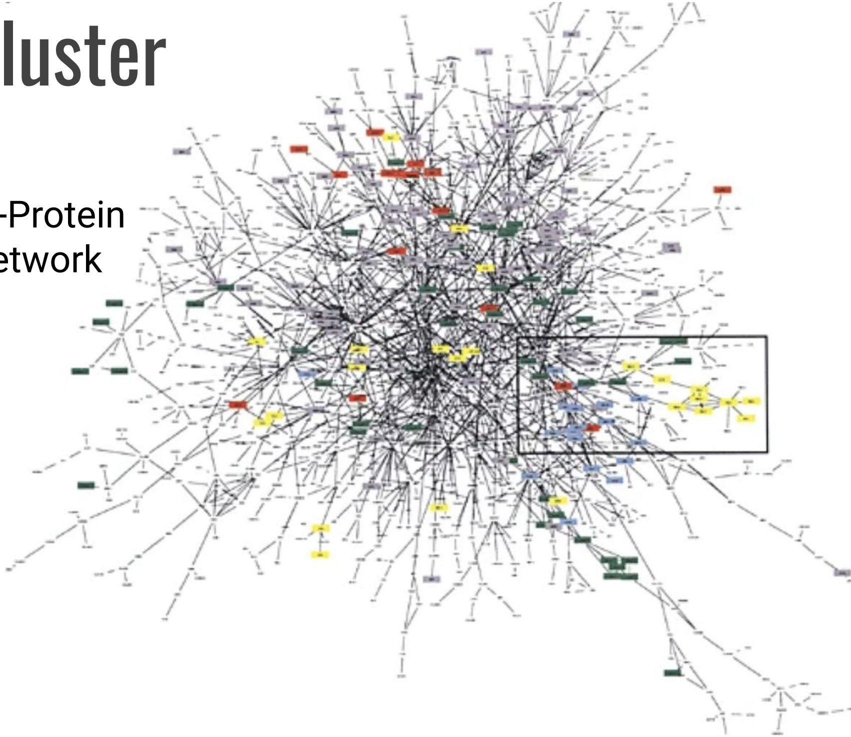
Many non-human networks don't cluster

U.S. Powergrid network

CC=0.103



Yeast Protein-Protein Interaction Network



Measure	TAP		HMS-PCI		Other data sets		
	"Small"	"Medium"	"Small"	"Medium"	Y2H	DIP	TP
Nodes n	193(15)	1,365(1,250)	99(7)	1,544(1,501)	1,870	1,788	434
Interactions I	191(38)	3,230(3,150)	67(7)	3,481(3,456)	2,240	3,003	868
Connectance C	0.01(0.36)	0.003(0.004)	0.01(0.33)	0.003(0.003)	0.001	0.002	0.009
Clustering cc	0.248(0.66)	0.216(0.233)	0.071(0)	0.048(0.049)	0.068	0.188	0.054
Diameter D	(1.94)	(4.93)	(1.81)	(4.41)			
Longest path	(4)	(12)	(3)	(11)			
Stretch parameter b	0.78	0.48	0.65	0.34	0.34	0.53	0.55

What do you see?

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

(Menczer et al, 2020)

High clustering in many human social networks

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

(Menczer et al, 2020)

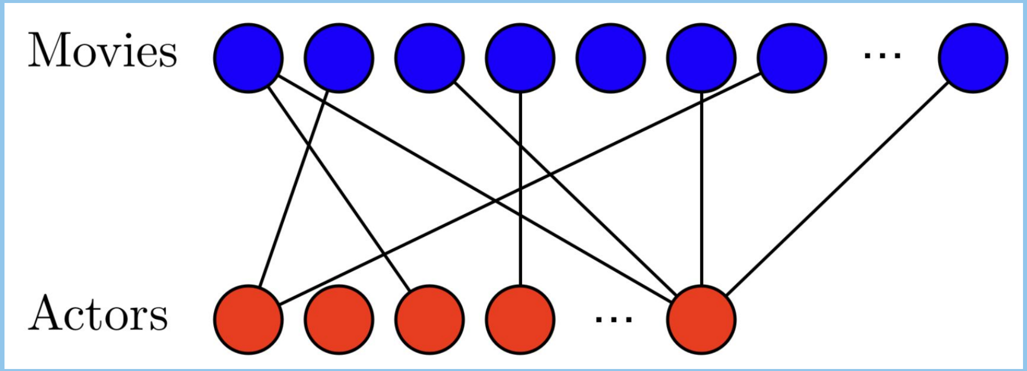
Huh???

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

(Menczer et al, 2020)

Bipartite network: links only between movies and stars

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				0
IMDB co-stars				0.67
Twitter US politics				0.03
Enron email				0.12
Wikipedia math				0.31
Internet routers				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



(Menczer et al, 2020)

Huh???

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

(Menczer et al, 2020)

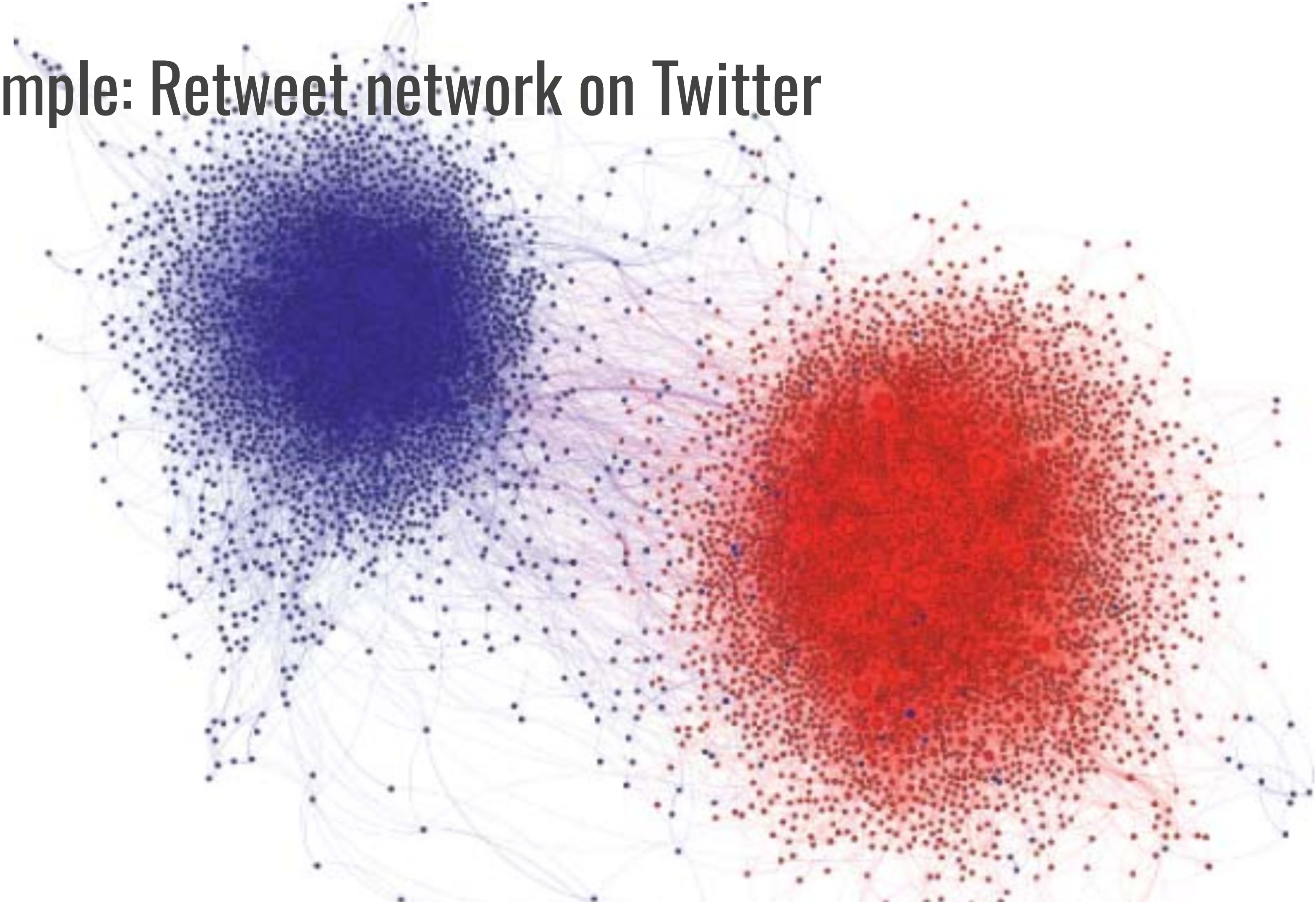
Retweet cascade trees look like stars (B rt A, C rt B → C rt A)

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	[illegible]	[illegible]	0.03
Enron email	87,273	[illegible]	[illegible]	0.12
Wikipedia math	15,220	[illegible]	[illegible]	0.31
Internet routers	190,914	[illegible]	[illegible]	0.16
US air transportation	546	[illegible]	[illegible]	0.49
World air transportation	3,179	[illegible]	[illegible]	0.49
Yeast protein interactions	1,870	[illegible]	[illegible]	0.07
C. elegans brain	297	[illegible]	[illegible]	0.29
Everglades ecological food web	69	916	2.2	0.55

(Menczer et al, 2020)

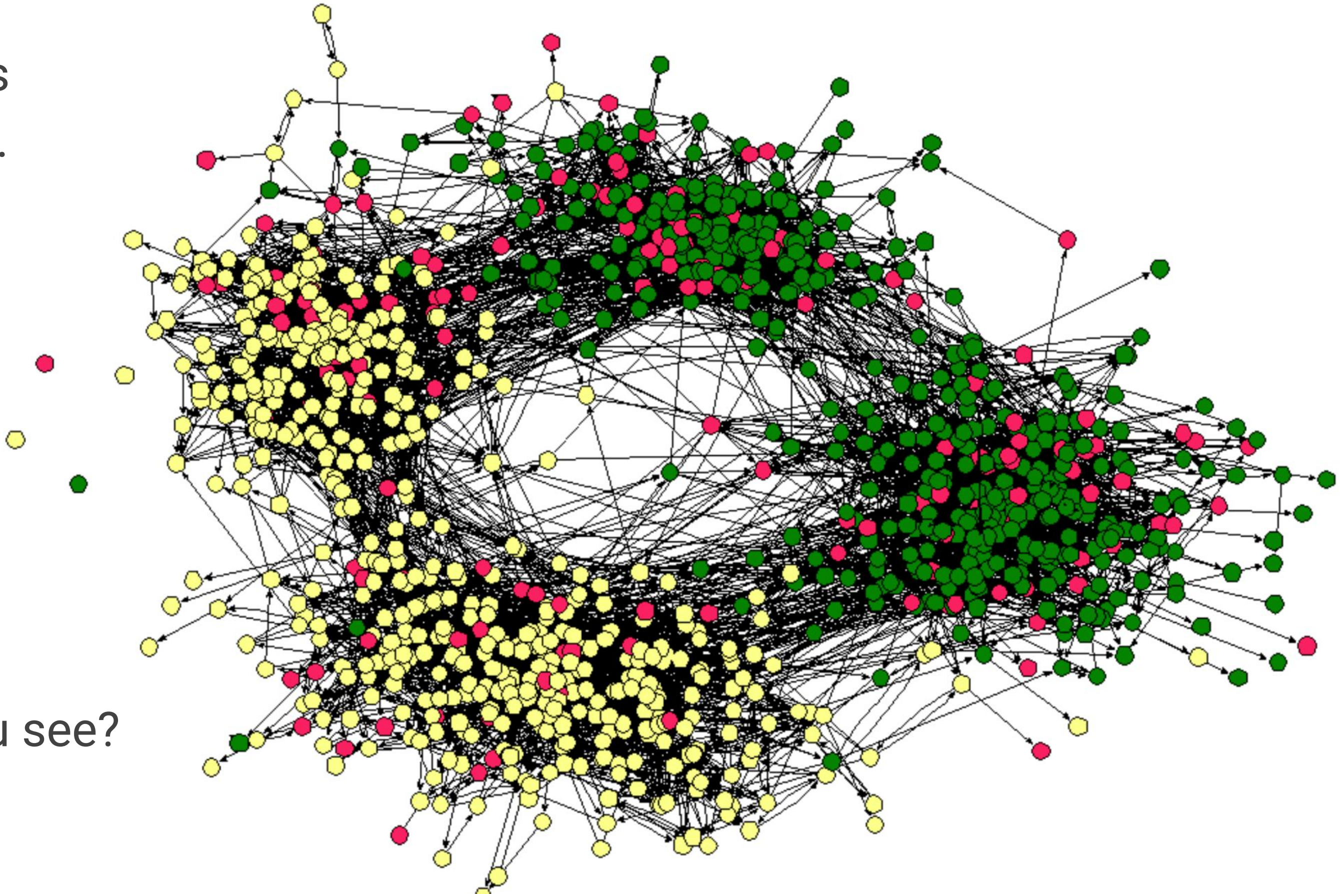
Birds of a Feather

Example: Retweet network on Twitter



Example: Social network from a town's middle school and high school

Circle colors denote race.

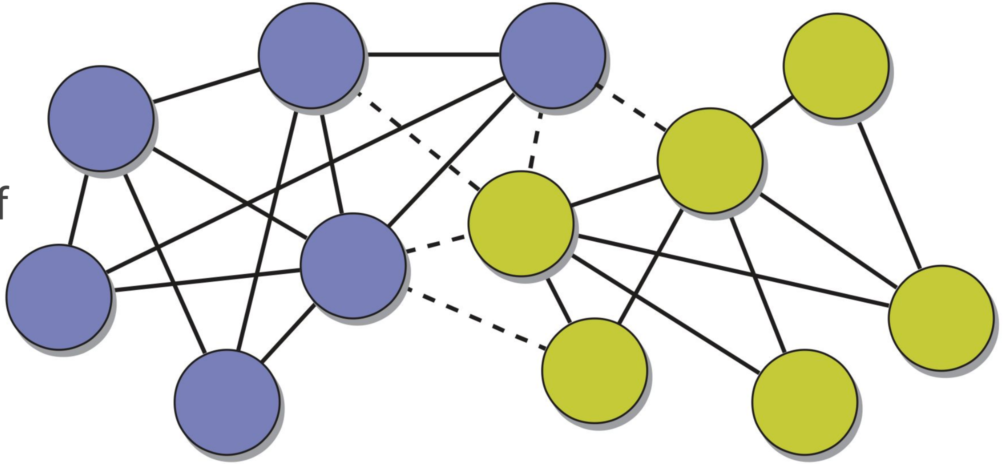


What do you see?

Homophily: Often, nodes that are connected to each other in a social network tend to have similar characteristics

The majority of links for each node go to nodes of the same color.

The majority of links connect nodes of the same color.



“People love those who are like themselves.” - Aristotle

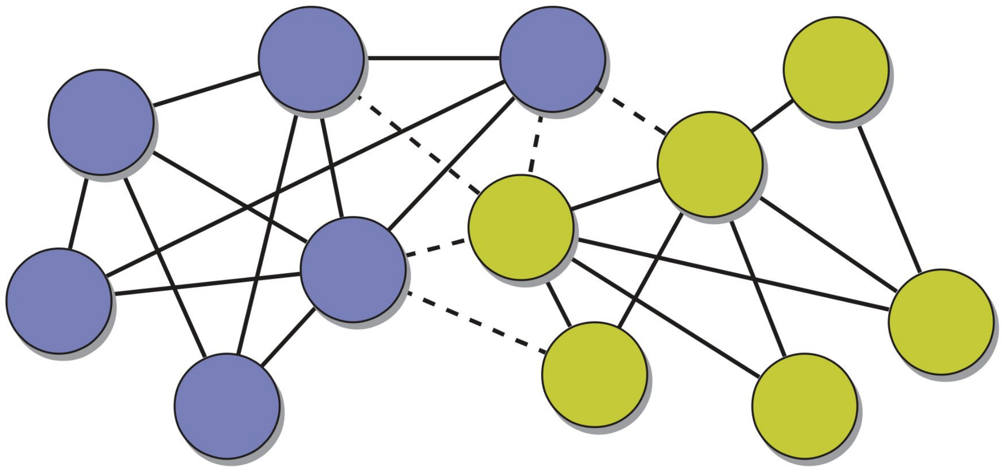
“Similarity begets friendship.” -Plato

(homo: same, phil: love → love for something that is the same)

Homophily: Often, nodes that are connected to each other in a social network tend to have similar characteristics

Salient dimensions:

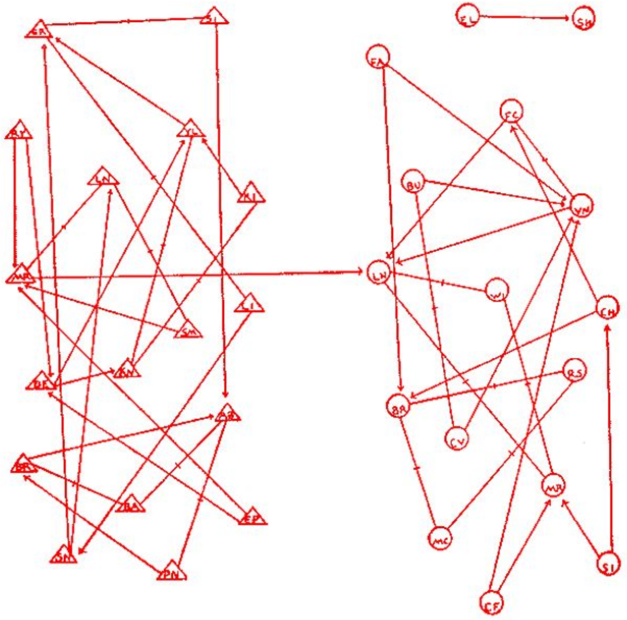
- Race, ethnicity
- Gender, sex
- Age
- Religion
- Occupation/education



Homophily: Gender

Salient dimensions:

- Race, ethnicity
- Gender, sex
- Age
- Religion
- Occupation/education

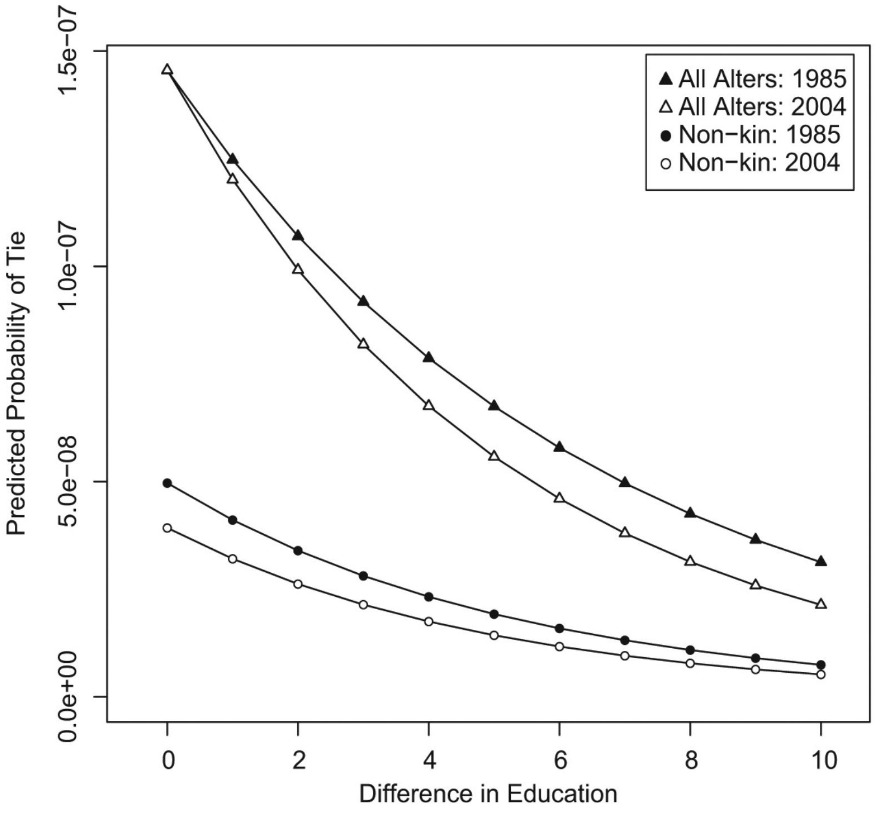


Homophily: Education

Tie probability decreases as the difference in education increases between two people

Tie probability is lower for non-kin

The effect is stronger in more recent years



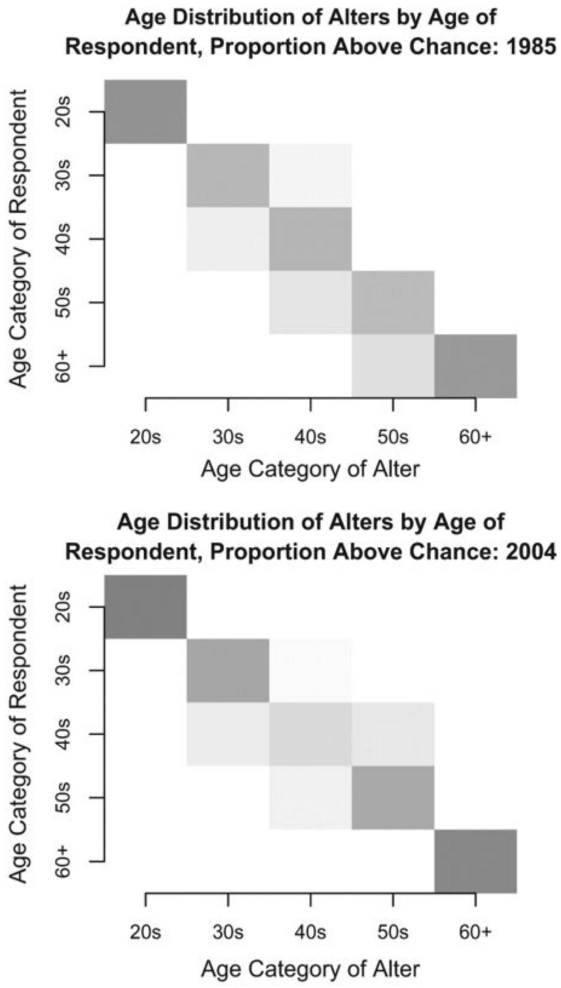
Smith et al. 2014

Homophily: Age

Age homophily slightly increased over time

Higher levels of homophily at 20s and 60s:

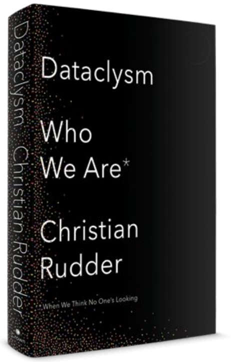
Why?

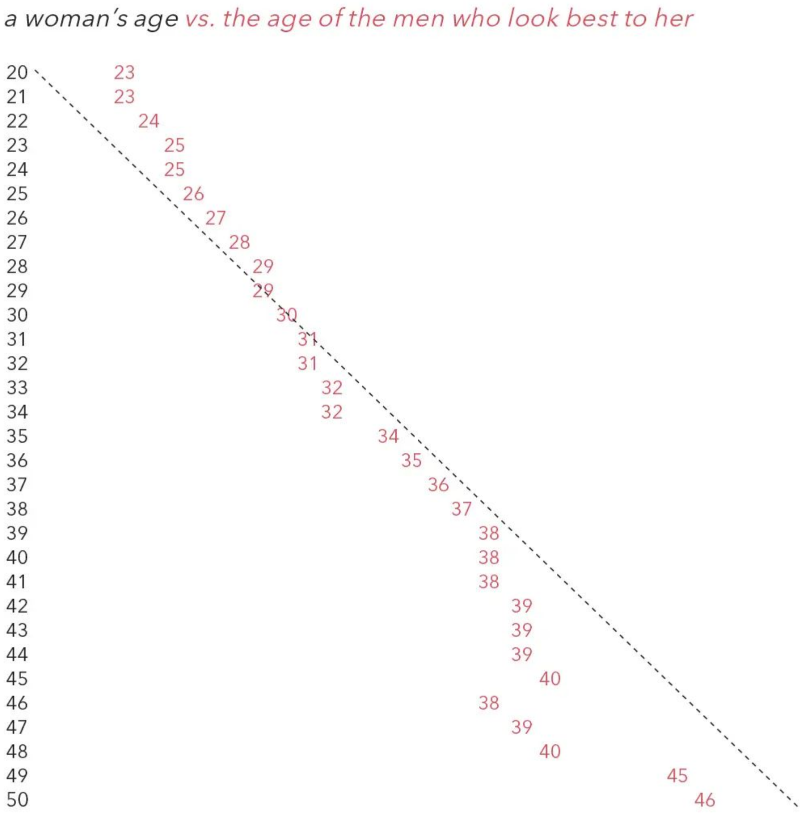


Smith et al. 2014

Homophily: Age

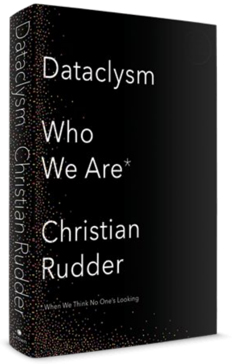
OkCupid data: Women are most interested in men their own age.

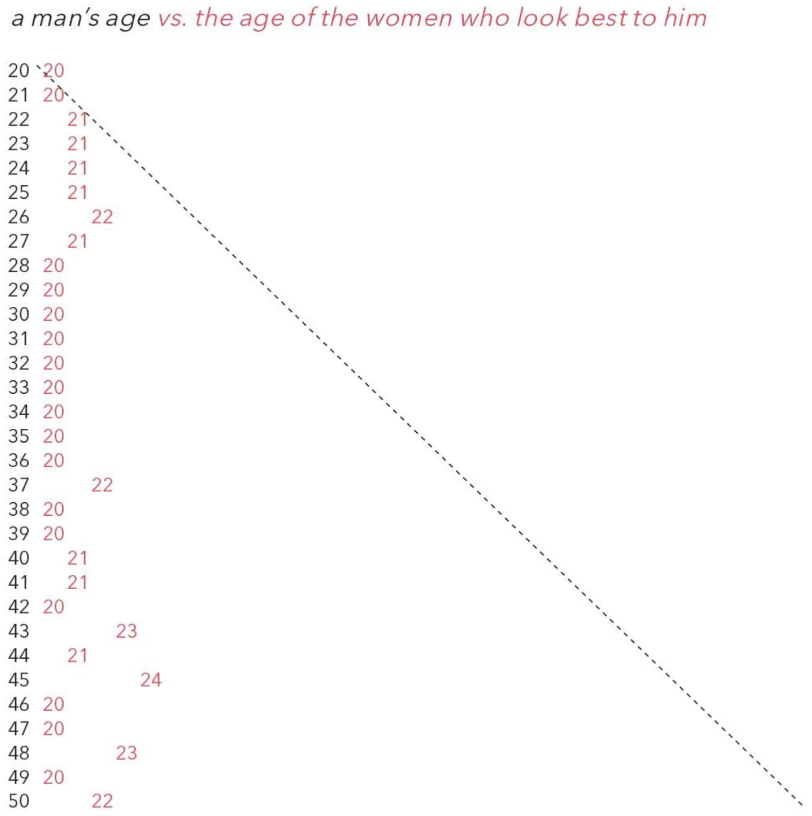




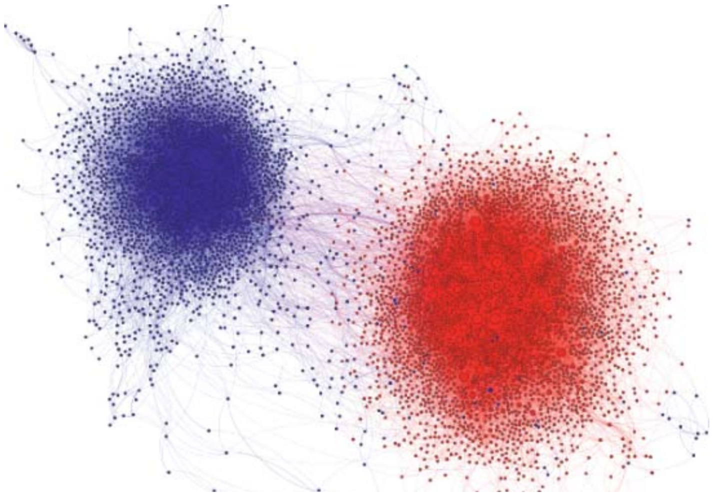
Homophily: Age

OkCupid data: Men are most interested in women in their early 20s.





Aside: The dark side of homophily



Exceedingly easy to connect with people who share our worldviews and unfriend / unfollow people with different opinions.

Information can be shared and consumed in such a selective and efficient way as to influence our opinions very effectively.

Result: segregation and polarization of our online communities.

High risk of manipulation by misinformation and social bots.

Competing mechanisms

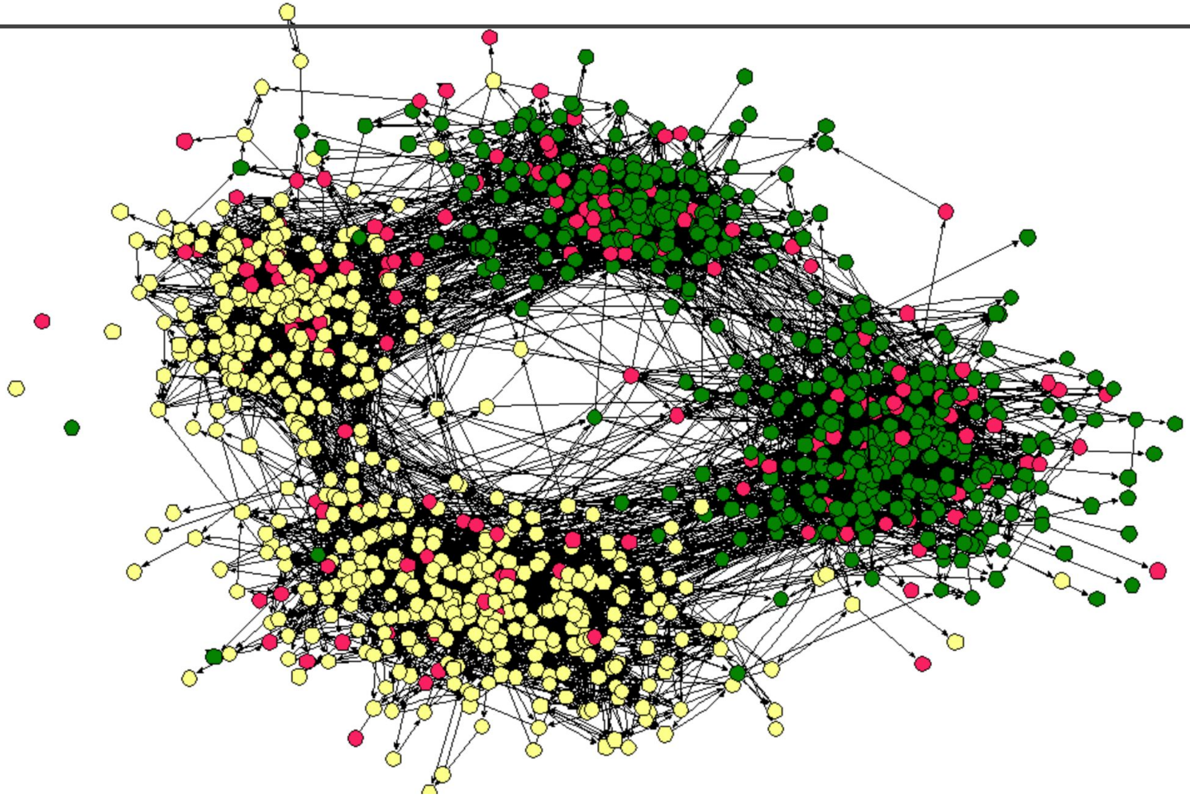
Selection (“homophily”): If people are similar in some way, they are more likely to select each other and become connected.

Social influence: People who are friends become more similar over time.

Homophily: Intrinsic vs contextual effects

Race (intrinsic)

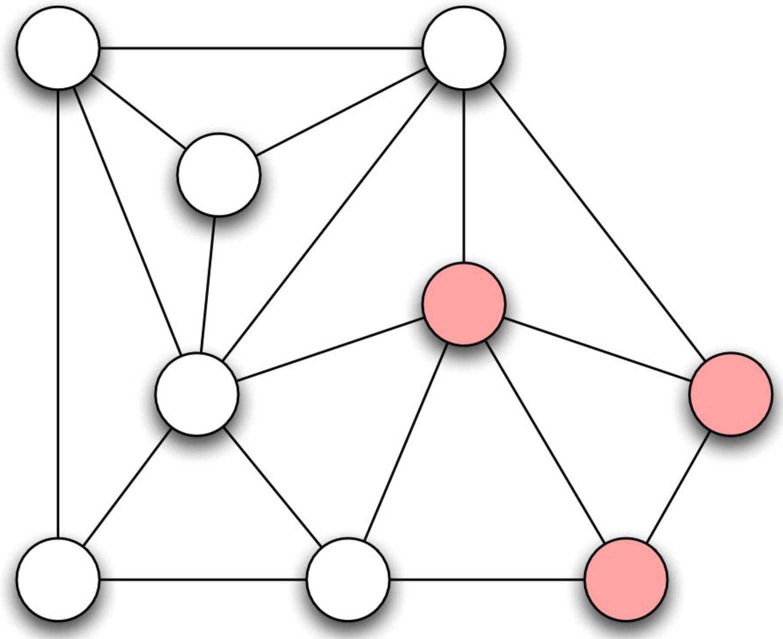
Middle vs
high school
(context)



Measuring homophily

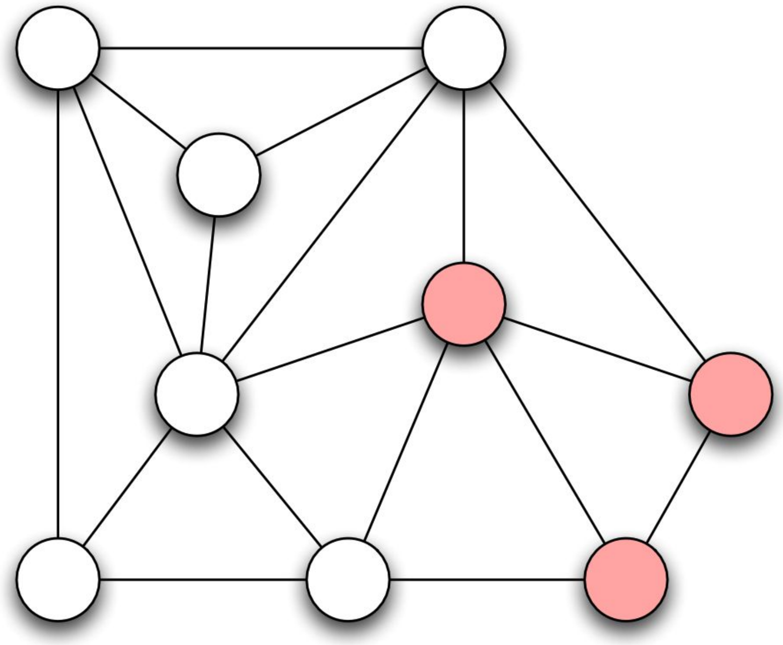
Given a particular characteristic of interest (like race, or age), is there a simple test we can apply to a network to estimate whether it exhibits homophily according to this characteristic?

Imagine this is the friendship network of an elementary-school classroom, with colors representing different genders.



Measuring homophily

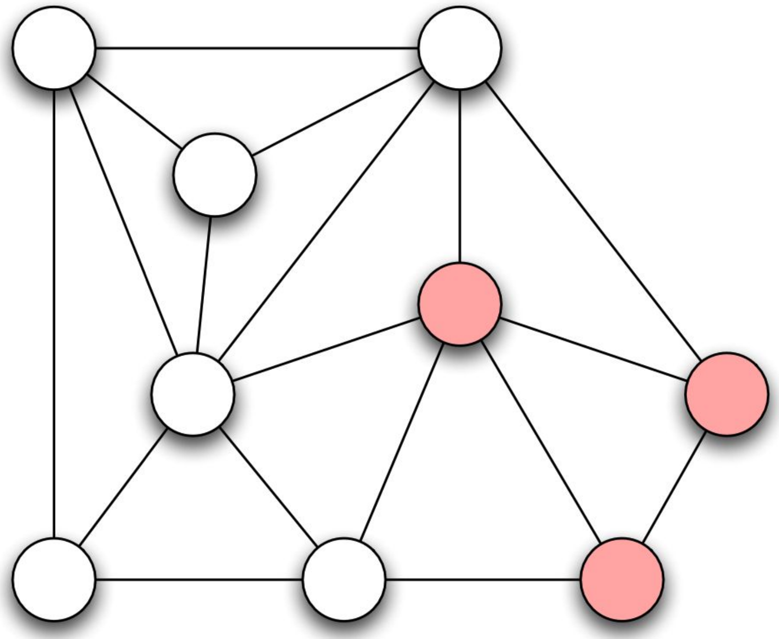
What would it mean for the network not to exhibit homophily by gender?



Measuring homophily

What would it mean for the network <u>not</u> to exhibit homophily by gender?

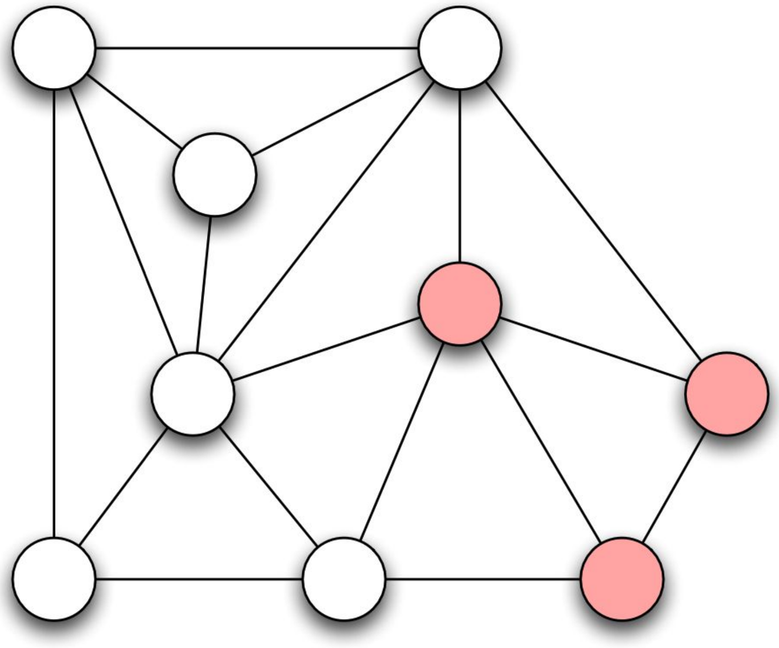
The proportion of male and female friends a person has should look like the background male/female distribution in the full population.



Measuring homophily

What would it mean for the network not to exhibit homophily by gender?

If we were to randomly assign each node a gender according to the gender balance in the real network, then the number of cross-gender edges should not change significantly relative to what we see in the real network.

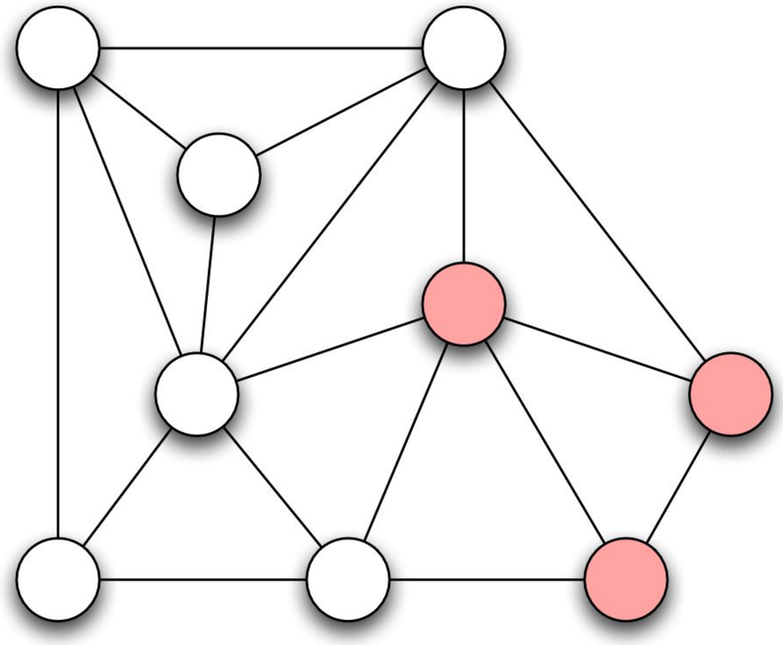


Measuring homophily

Suppose a p fraction of all individuals are male, and a q fraction are female.

Consider a given edge in this network:

- both ends of the edge will be male with probability … ?
- both ends will be female with probability …?
- if one end is male and the other is female, or vice versa, then we have a cross-gender edge with probability …?

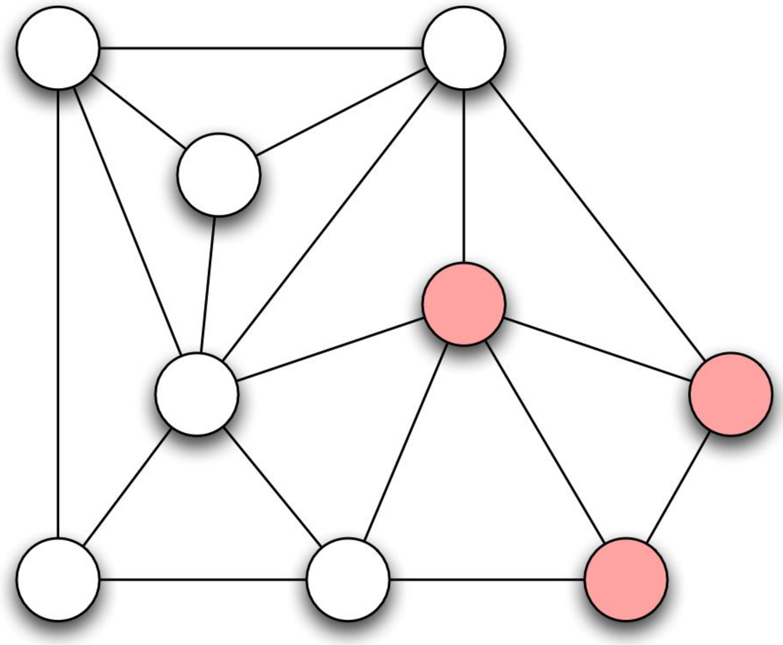


Measuring homophily

Suppose a p fraction of all individuals are male, and a q fraction are female.

Consider a given edge in this network:

- both ends of the edge will be male with probability p^2
- both ends will be female with probability q^2
- if one end is male and the other is female, or vice versa, then we have a cross-gender edge with probability $2pq$



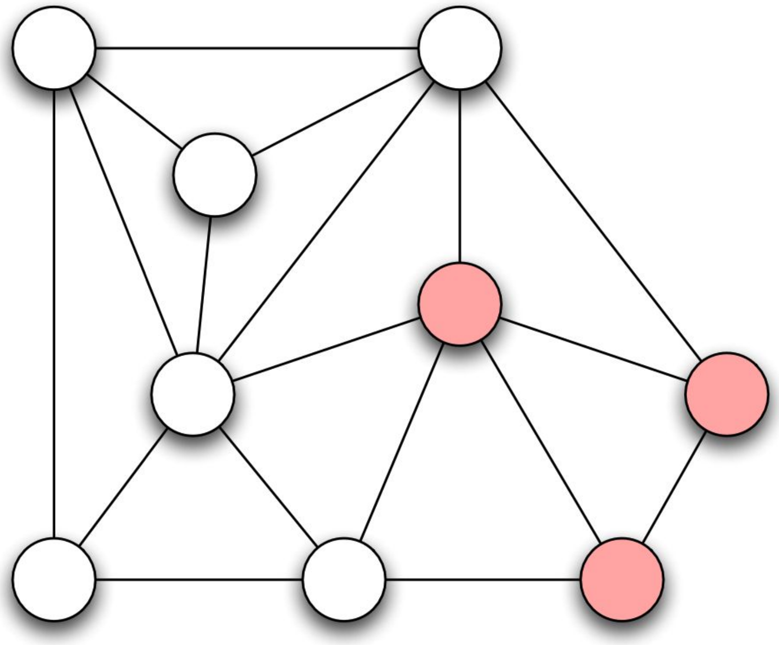
Measuring homophily

Homophily test:

If the fraction of cross-gender edges is significantly less than 2pq, then there is evidence for homophily.

p = 2/3 and q = 1/3 in our example
2pq = 4/9 = 8/18
5 / 18 edges are cross-gender

With no homophily, one should expect to see 8 cross-gender edges rather than than 5, so this example shows some evidence of homophily.



Aside: Networks can also exhibit inverse homophily

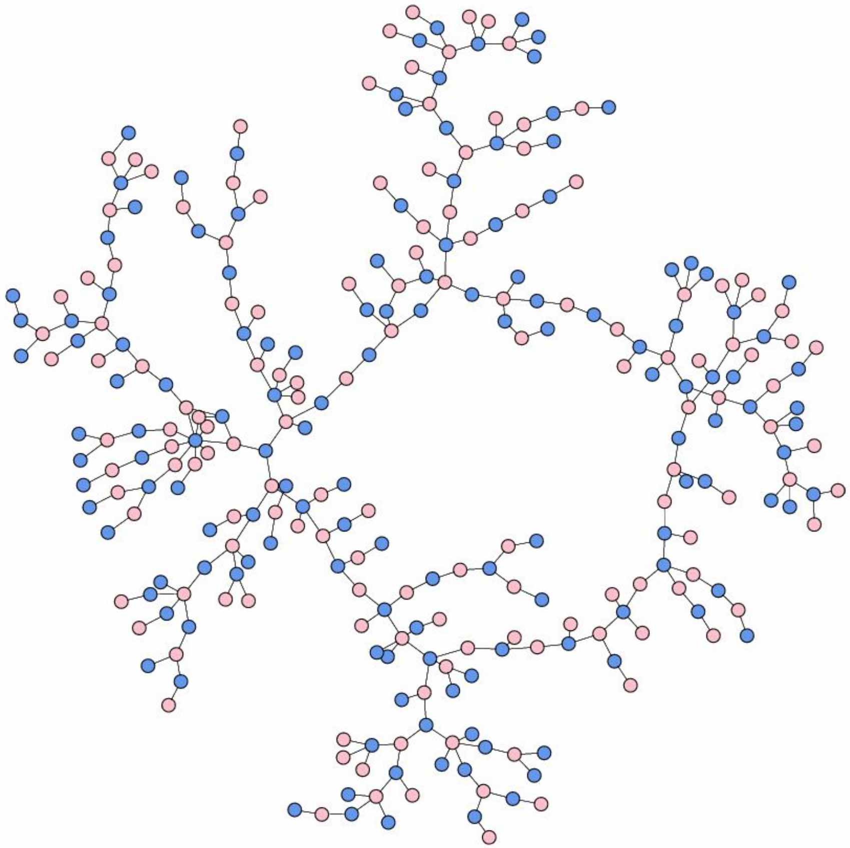
If the fraction of cross-gender edges is significantly more than 2pq.

Do you remember any example?

Aside: Networks can also exhibit heterophily

If the fraction of cross-gender edges is significantly more than 2pq.

Yes! The high school dating network



Summary

We’ve seen another fundamental property of networks: similarity between neighbors

(Recall short paths connecting nodes and triangles formed by common neighbors)

One <u>extremely</u> powerful analysis technique: comparison to a random (shuffled) network. We’ll see another one (longitudinal analysis) next time.