Network Analysis:

The Hidden Structures behind the Webs We Weave

17-213 / 17-668

Homophily and Degree Correlation

Tuesday, September 19, 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 Thursday's Lecture

Clustering coefficient case study

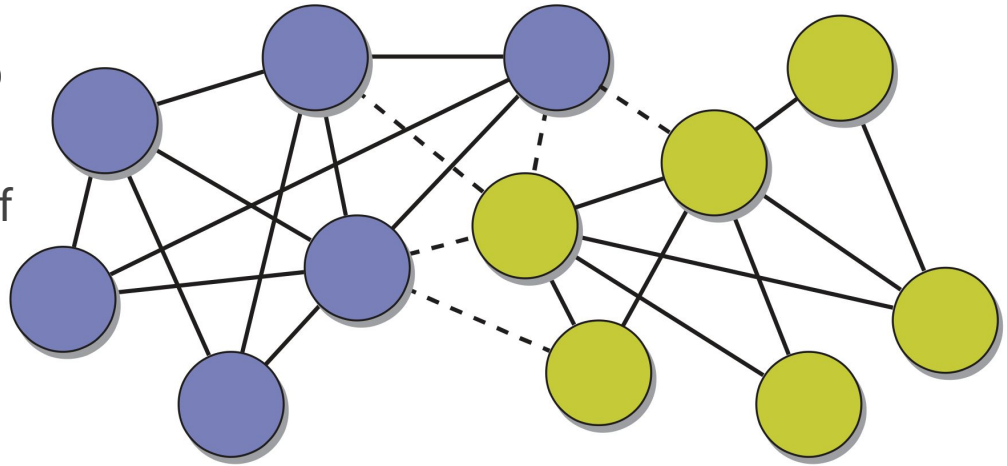
Homophily and how to measure

Birds of a Feather

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)

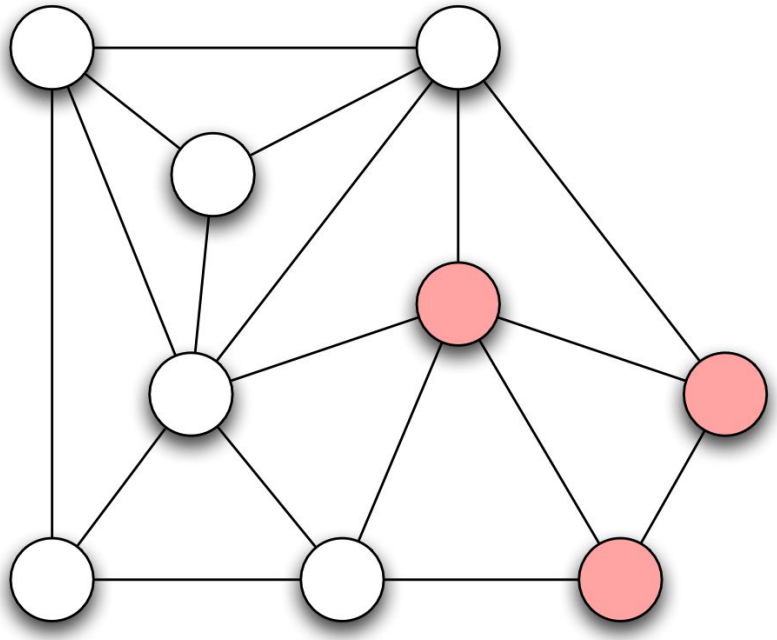
Measuring homophily

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

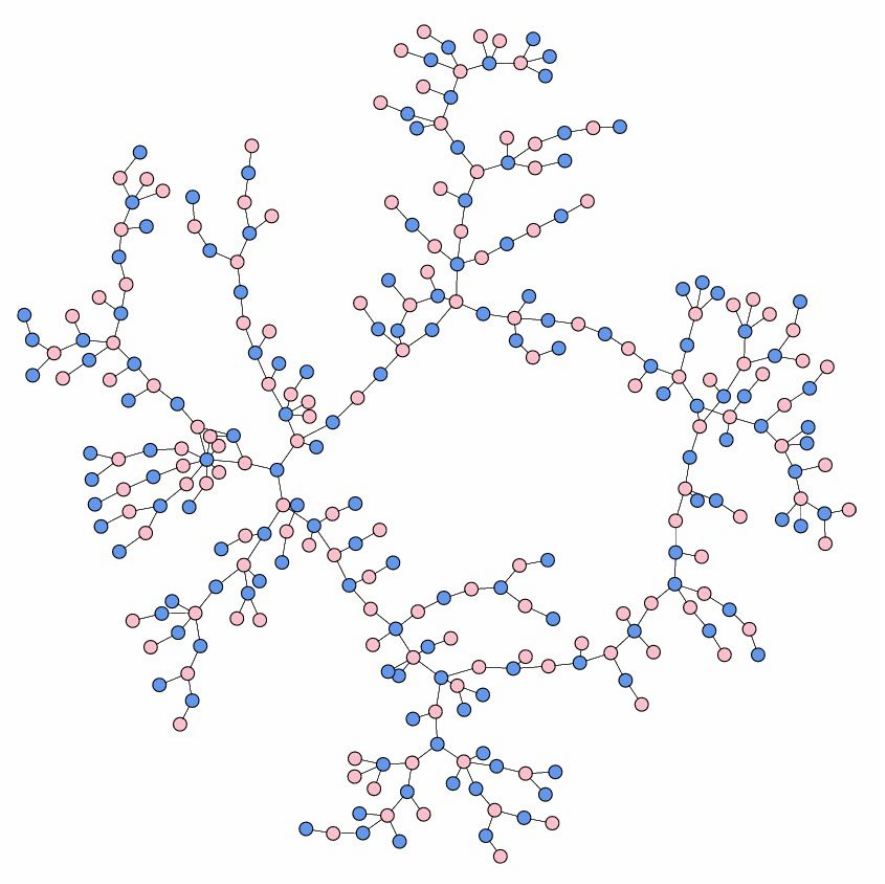
If the fraction of cross-gender edges is significantly more than 2pq.

Do you remember any example related to gender?

Aside: Networks can also exhibit heterophily

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

Yes! The high school dating network



The natural sciences perspective

Homophily: Status & Power

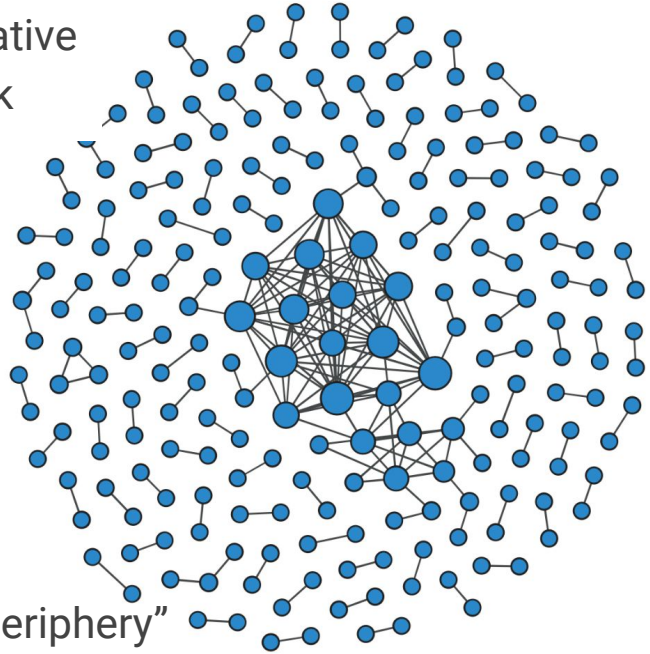
Degree homophily: “degree assortativity” or “degree correlation” – high-degree nodes tend to be connected to other high-degree nodes and vice versa.

Extensively studied from a graph-theoretic perspective.

Degree Assortativity / Disassortativity

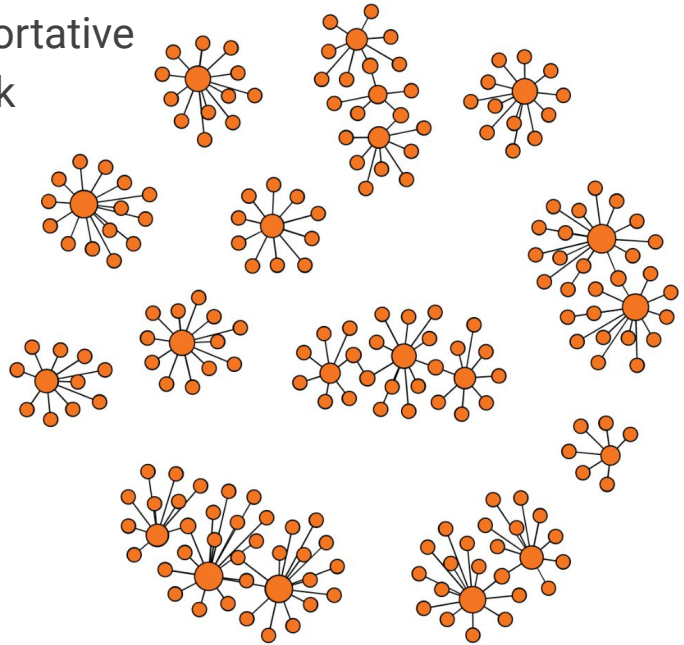
Example:

Assortative network



“Core-periphery”

Disassortative network

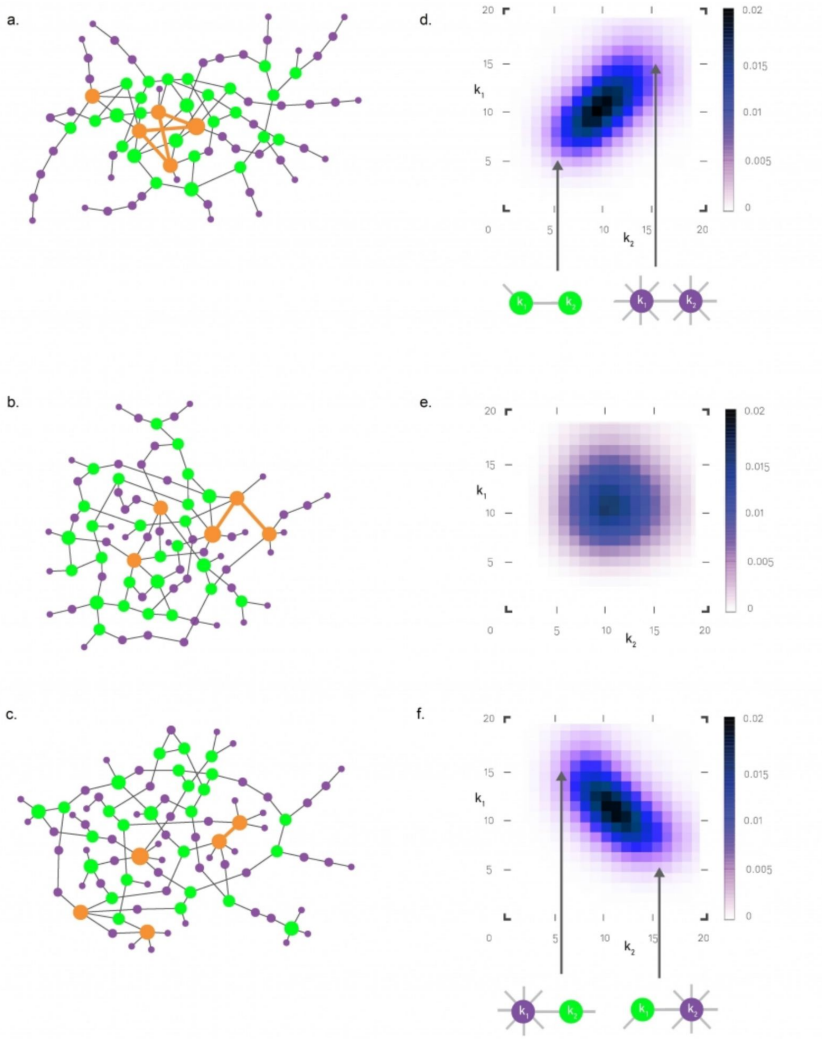


Degree Assortativity / Disassortativity

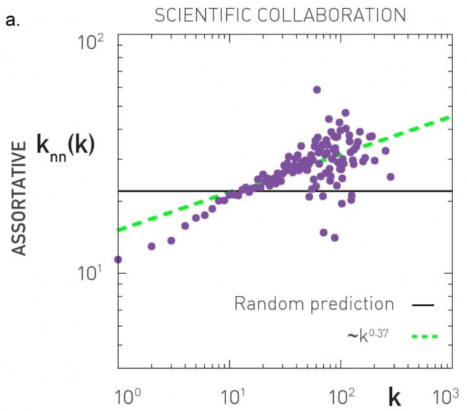
(a) **Positive** degree correlation: Connected nodes have similar degree

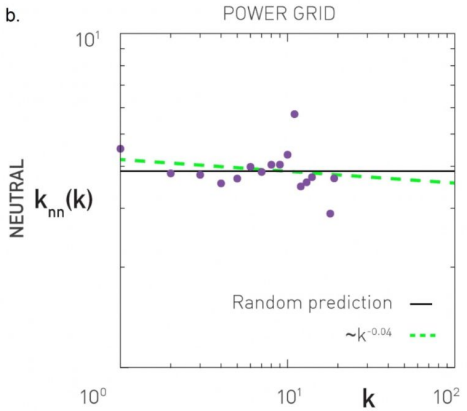
(b) **Neutral**: The degree of connected nodes have no correlation

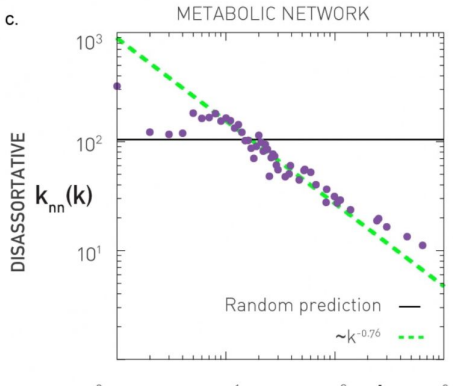
(c) **Negative** degree correlation: Connected nodes have dissimilar degree



Measuring degree correlation: Average degree of the neighbors of a node of degree k







Average degree of neighbors increases as k increases → assortative network

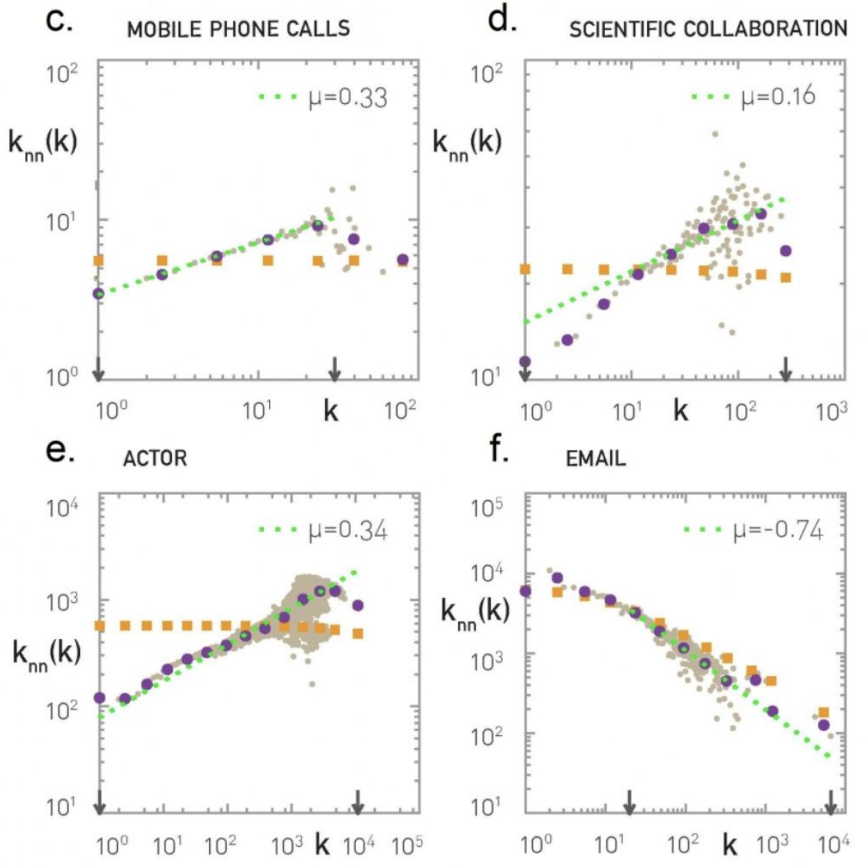
Average degree of neighbors neither increases nor decreases as k increases → degree neutral network

Average degree of neighbors decreases as k increases → disassortative network

Human social networks tend to exhibit positive degree correlations

Why positive?

Why is the email network negative?



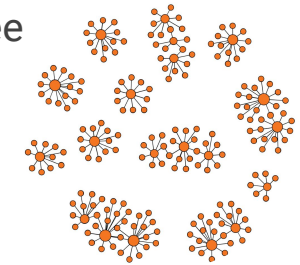
Human social networks tend to exhibit positive degree correlations

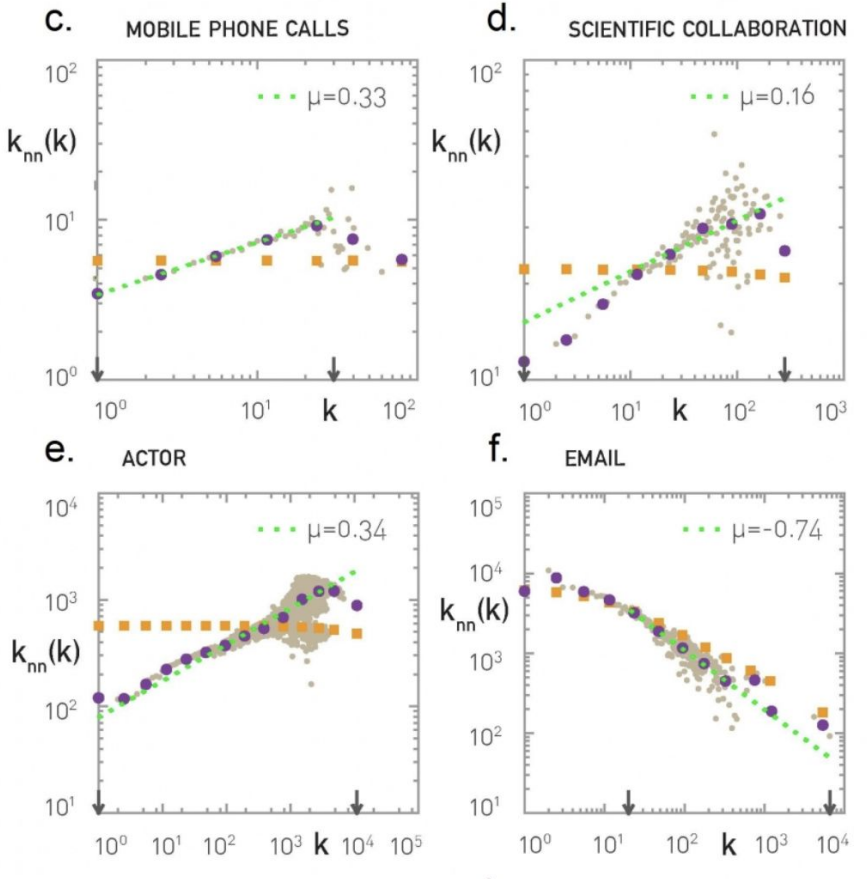
Why positive?
→ Open question. Several studies argue that it is related to the fact that humans form groups

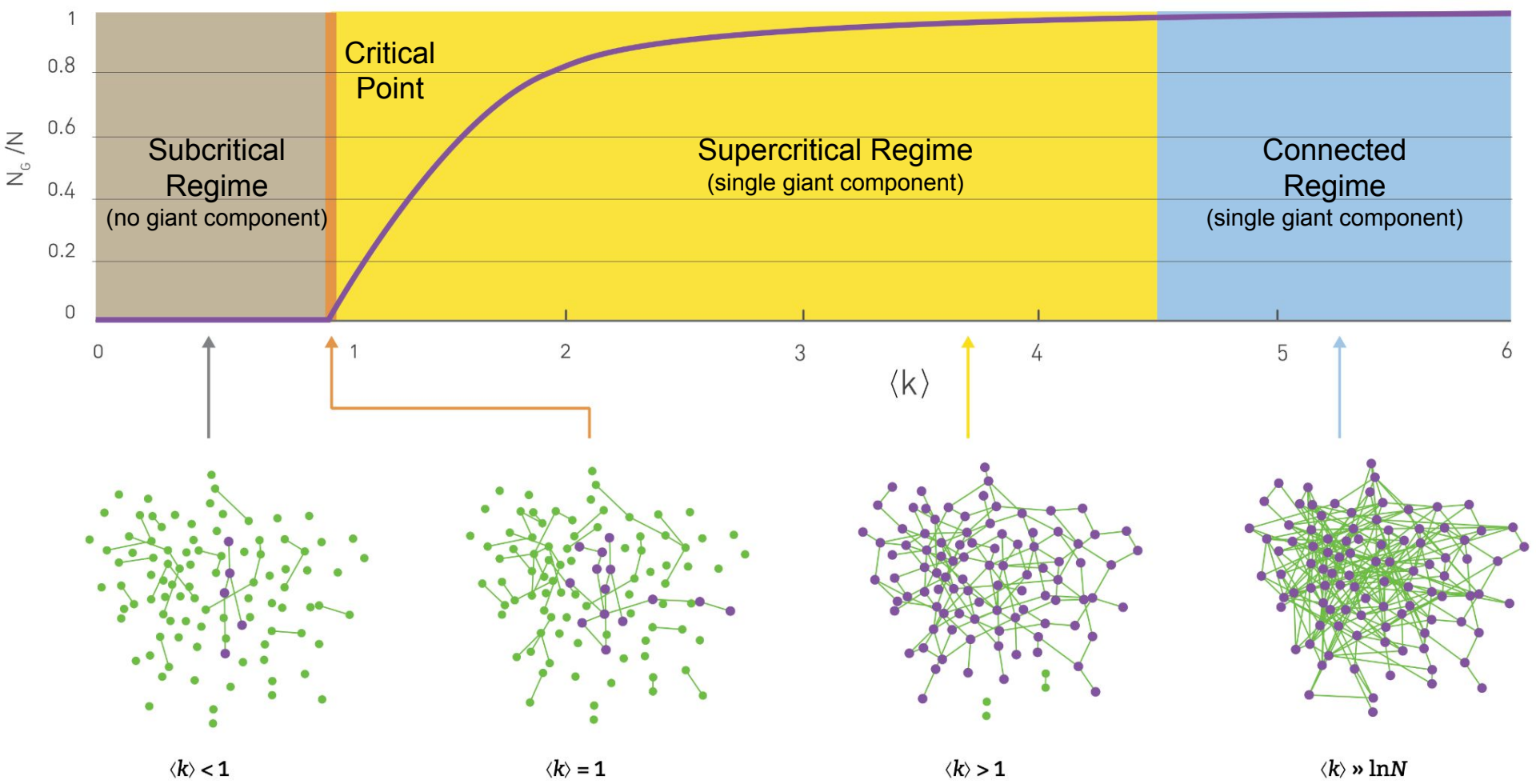
→ People in large groups tend to have high degree (more group members to connect with) and those in small groups are constrained in forming ties - hence low degree

Why is the email network negative?

→ Networks with skewed degree distributions tend to exhibit negative degree correlations





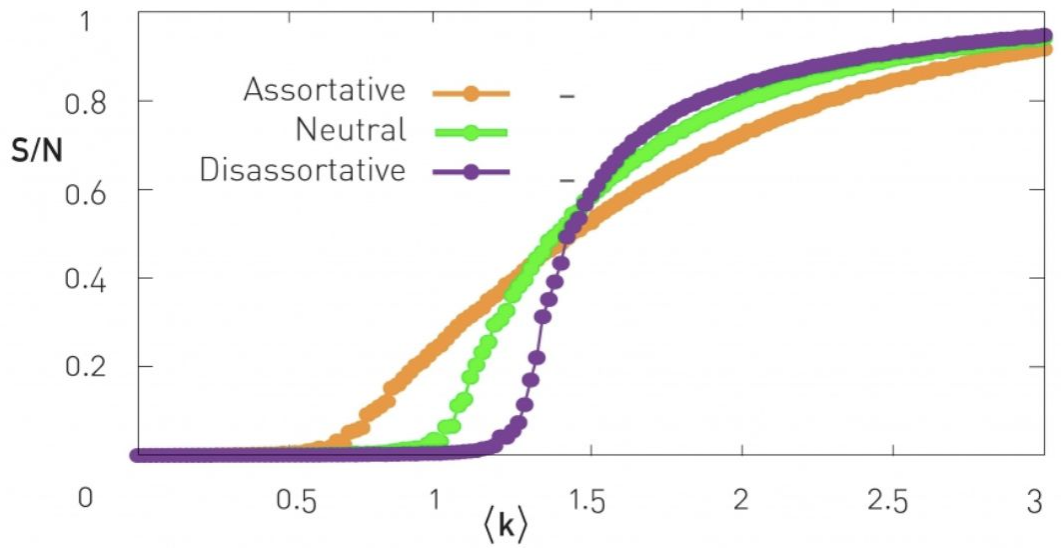


(Barabasi Ch. 3.6; Erdős & Rényi, 1959)

Impact of Assortativity: Higher connectivity

Giant component can emerge at lower mean degree <k>

Size of largest component / Size of entire network

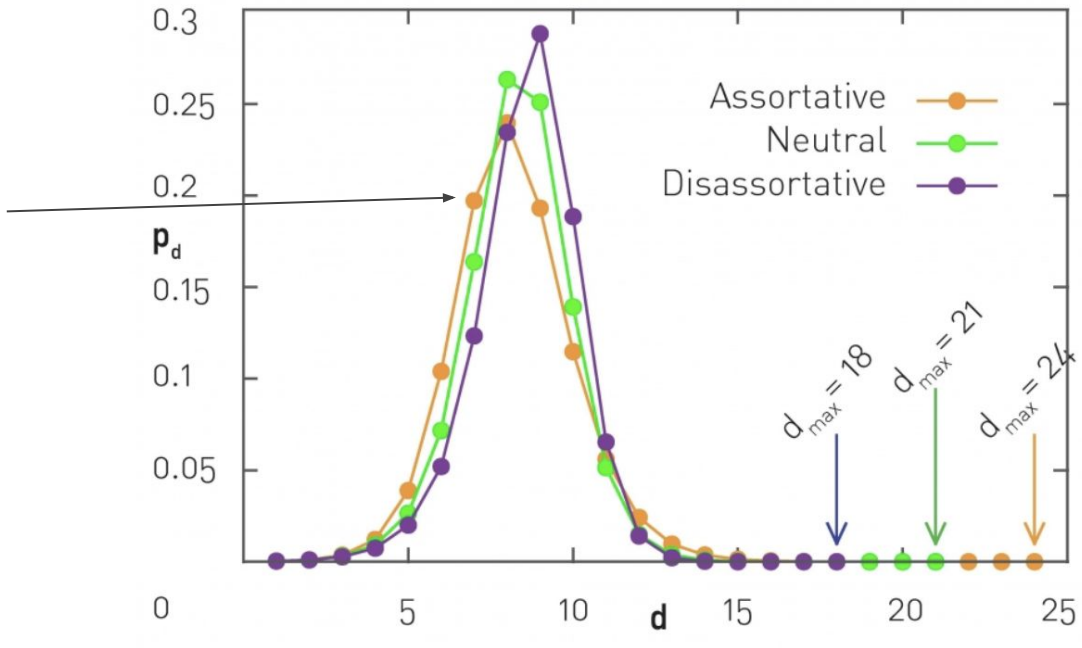


This means connectivity increases even if people do not have many connections

Impact of Assortativity: Higher connectivity

Giant component can emerge at lower mean degree <k>

Assortative networks have shorter average path length



Back to interpreting homophily

Million dollar question: Why does homophily happen?

Recall the two competing mechanisms:

Selection: 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.

Does similarity induce links, or do links induce similarity?

Important for reasoning about the effect of possible interventions

Consider an adolescent drug use network:

If drug use displays social influence — with students showing a greater likelihood to use drugs when their friends do — then target certain high-school students and influence them to stop using drugs; their social influence could cause their friends to stop using drugs as well.

If illicit drug arises almost entirely from selection effects, then as targeted students stop using drugs, they change their social circles and form new friendships with students who don't use drugs, but the drug-using behavior of other students is not strongly affected.

Selection may operate at several different scales, and with different levels of intentionality

In a small group, when people choose friends who are most similar from among a clearly delineated pool of contacts, there is clearly active choice going on.

In other cases, and at more global levels, selection can be more implicit and a result of the social environment.

For example, when people live in neighborhoods, attend schools, or work for companies that are relatively homogeneous compared to the population at large.

Million dollar question: Why does homophily happen?

Recall the two competing mechanisms:

Selection: 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.

Does similarity induce links, or do links induce similarity?

Million dollar question: Why does homophily happen?

Recall the two competing mechanisms:

Selection: 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.

Does similarity induce links, or do links induce similarity?

We need longitudinal studies: Have the people in the network adapted their behaviors to become more like their friends, or have they sought out people who were already like them?

Case Study: Christakis and Fowler obesity study showing evidence of social influence

Framingham heart study network

Red borders: women

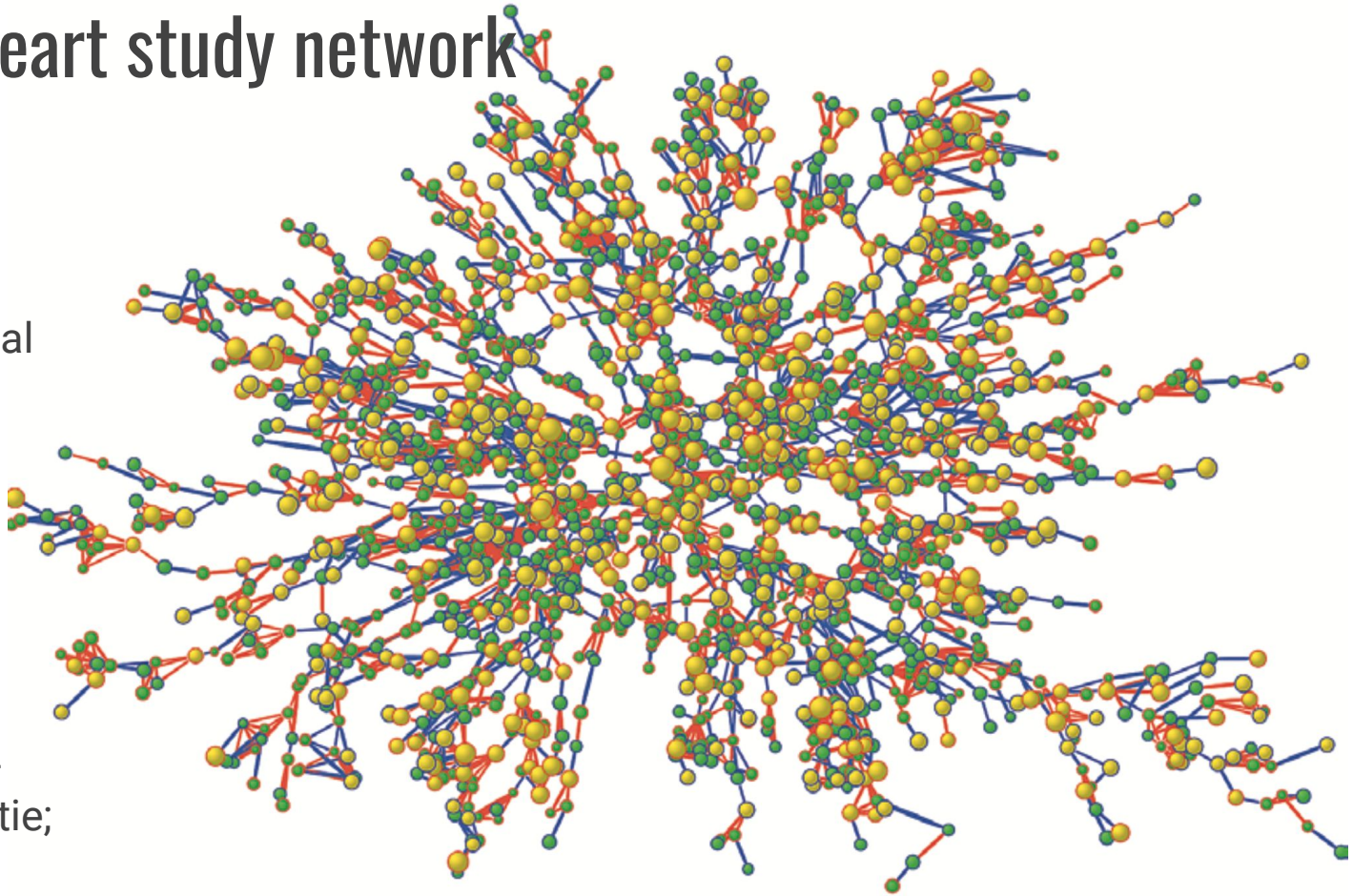
Blue borders: men.

Node size proportional to the person's body-mass index.

Yellow: body-mass index ≥30 ("obese")

Green: nonobese.

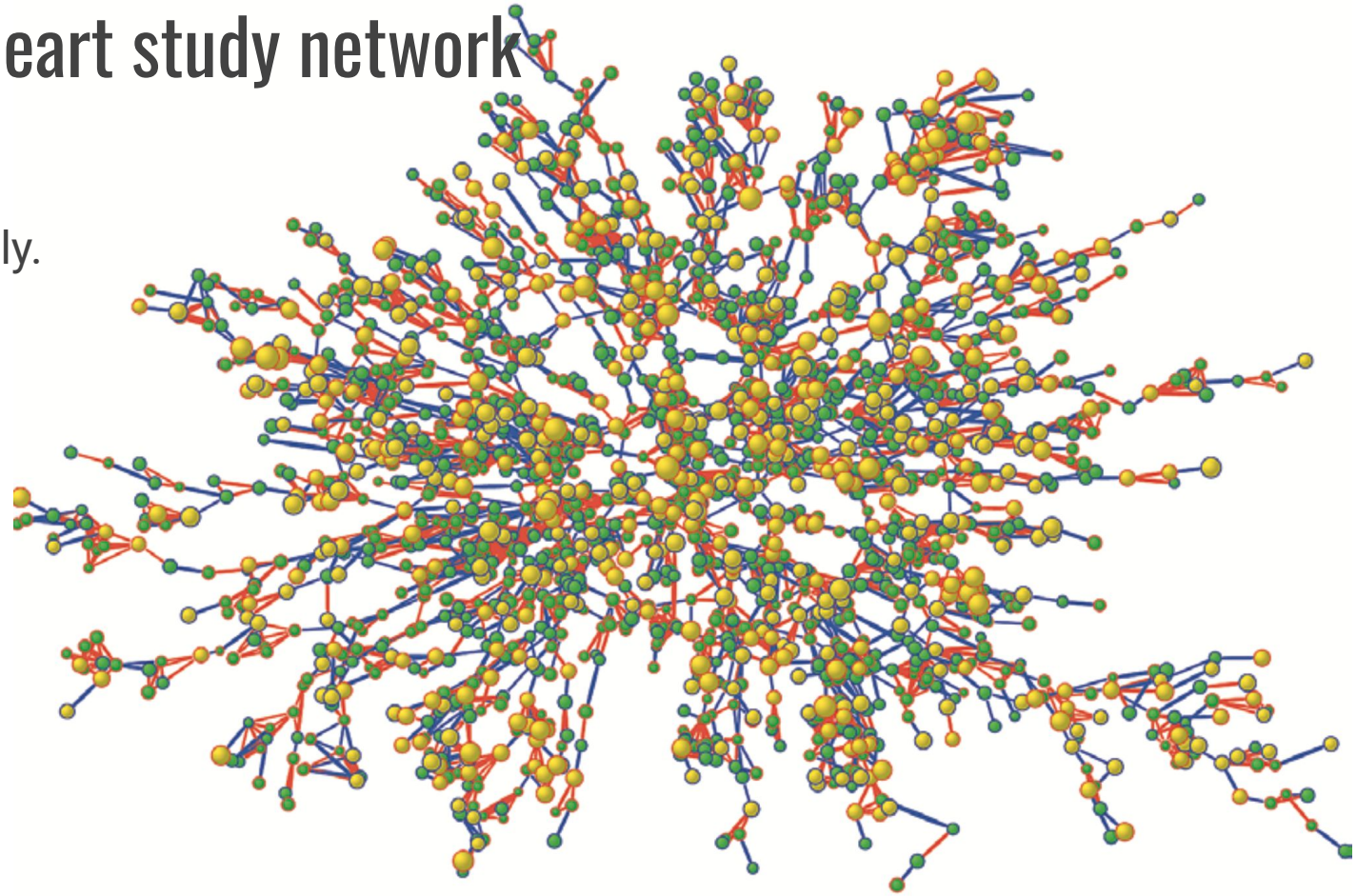
Tie colors indicate relationship: purple – friendship or marital tie; orange – familial tie.



Framingham heart study network

The researchers tested for homophily.

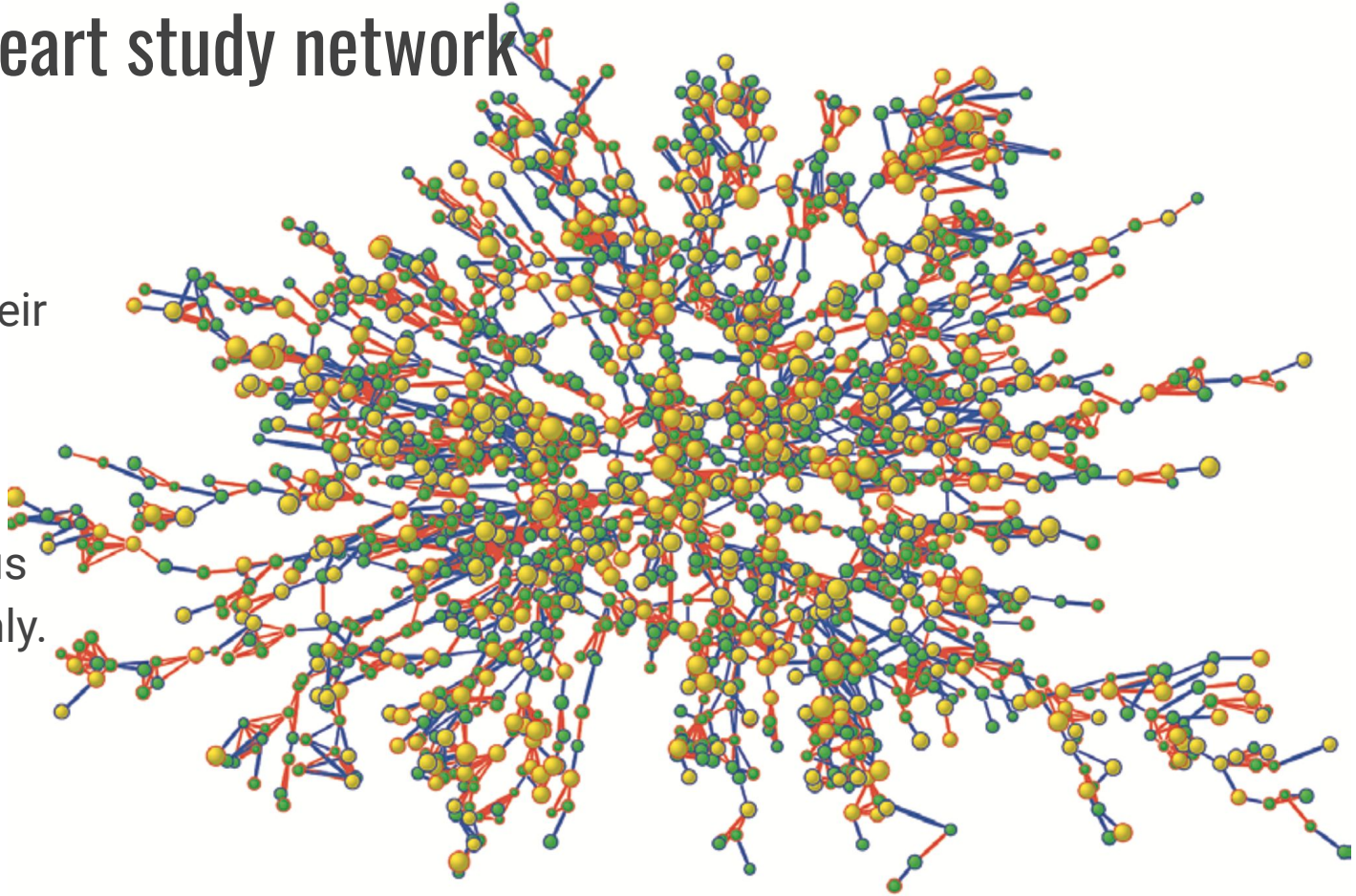
How?



Framingham heart study network

People tend to be more similar in obesity status to their network neighbors than in a version of the same network where obesity status is assigned randomly.

Now, why?



Hypotheses

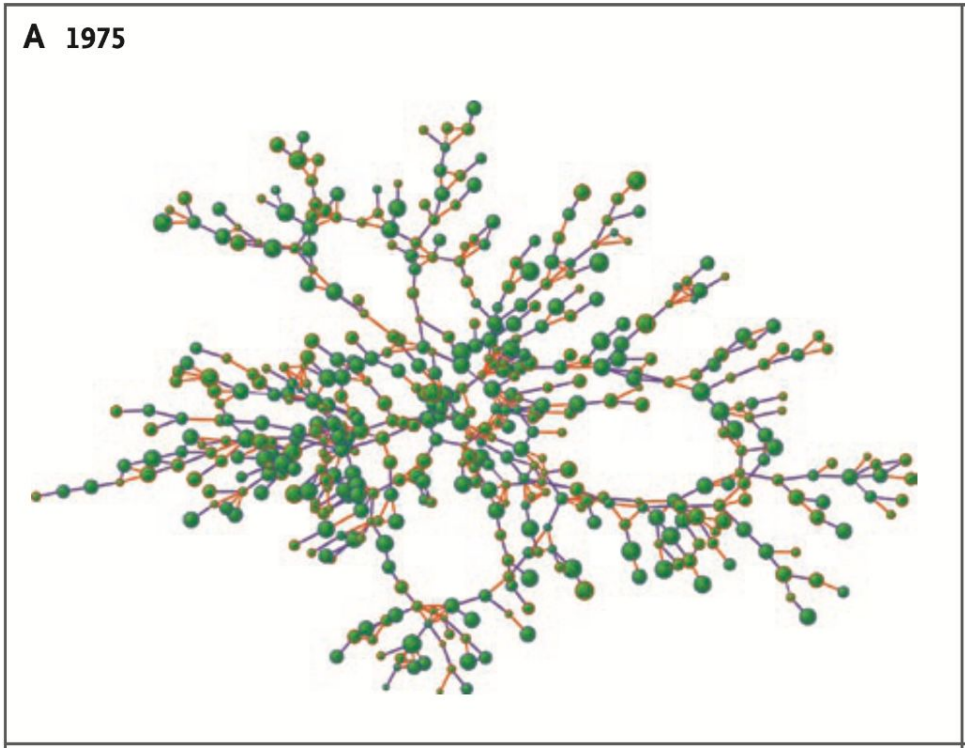
This clustering is present:

(1) because of selection effects, in which people are choosing to form friendships with others of similar obesity status?

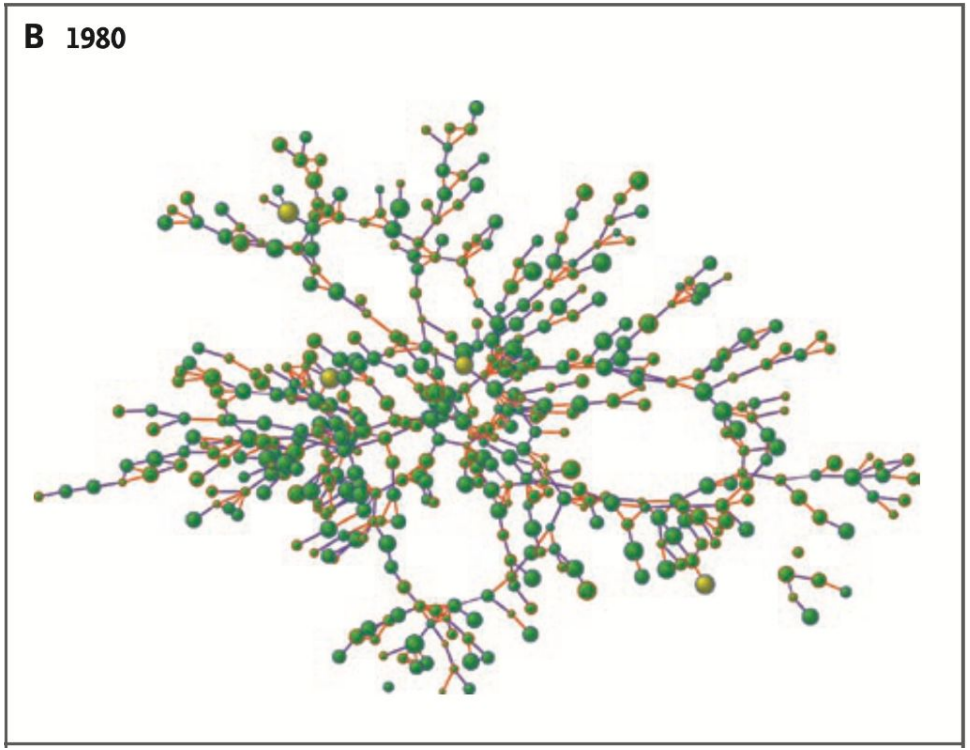
(2) because of the confounding effects of homophily according to other characteristics, in which the network structure indicates existing patterns of similarity in other dimensions that correlate with obesity status? or

(3) because changes in the obesity status of a person's friends was exerting a (presumably behavioral) influence that affected his or her future obesity status?

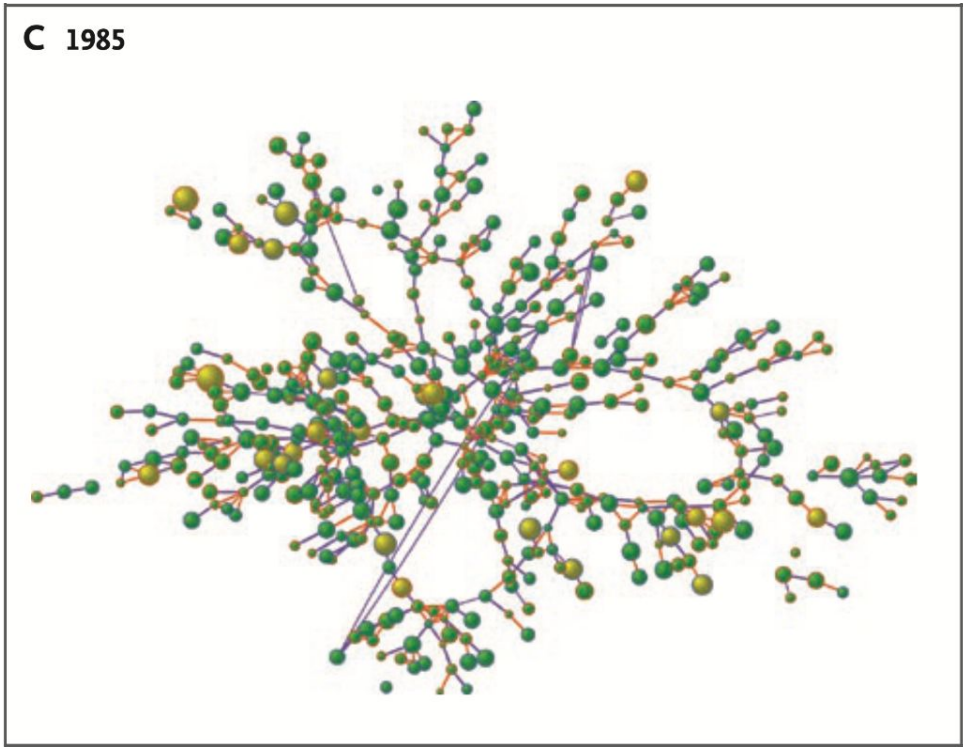
Key idea: Study the network longitudinally



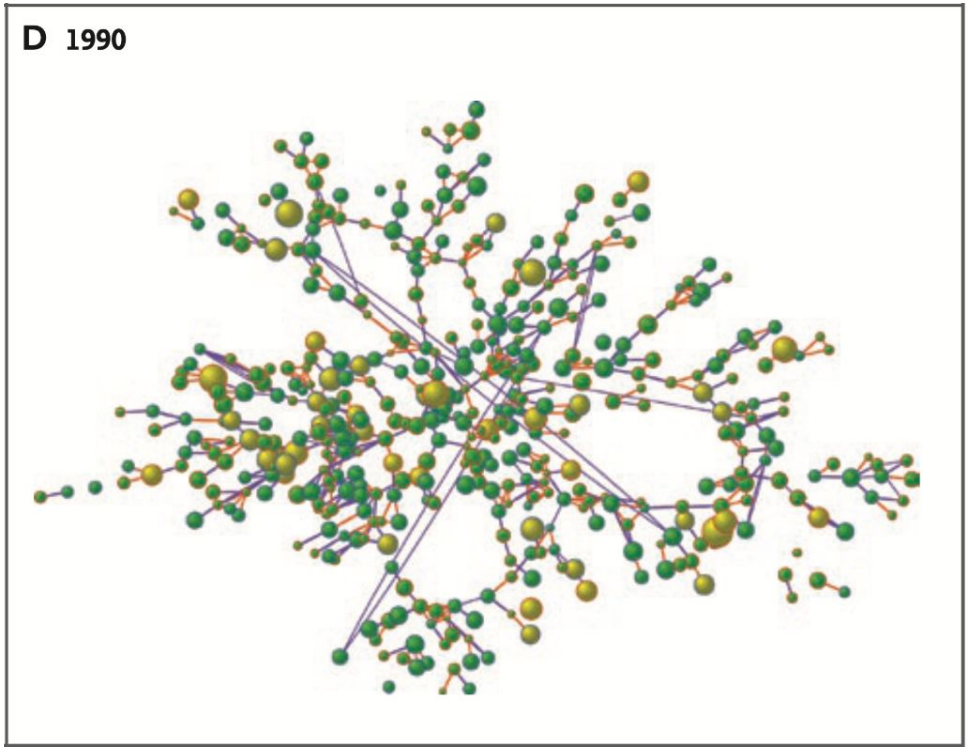
Key idea: Study the network longitudinally



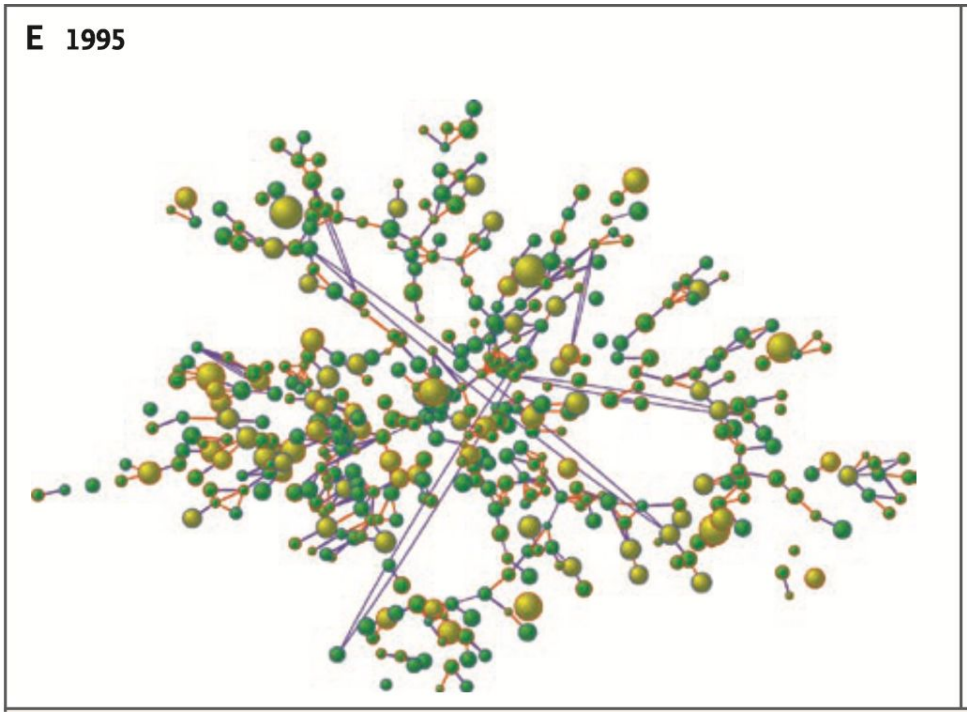
Key idea: Study the network longitudinally



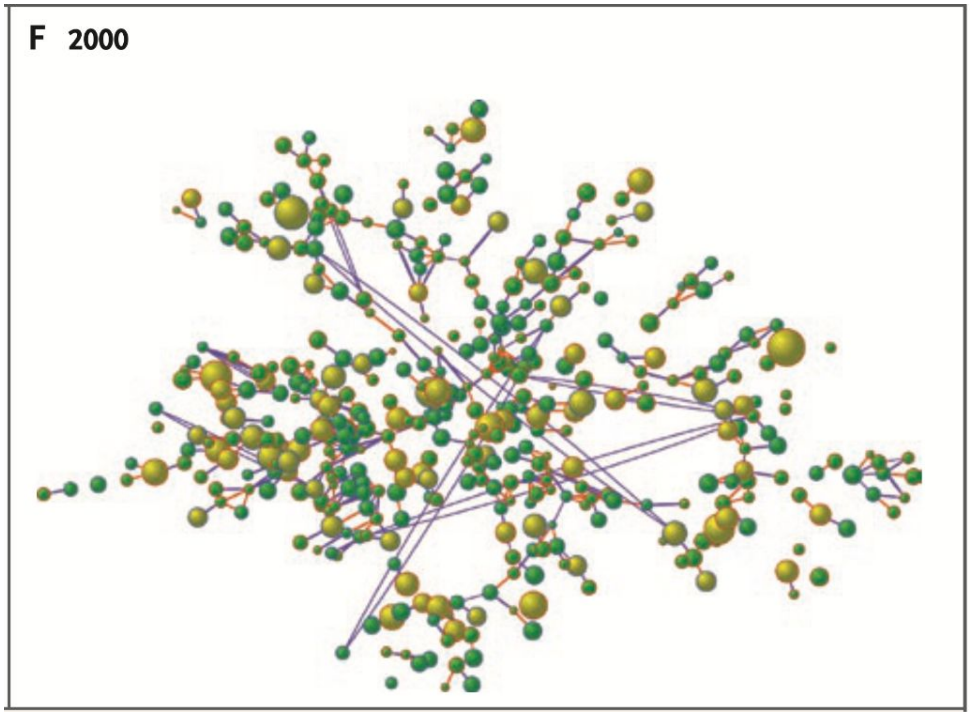
Key idea: Study the network longitudinally



Key idea: Study the network longitudinally



Key idea: Study the network longitudinally



Statistical modeling intuition

Model one's obesity status at time point t+1 as a function of

- their age, sex, and educational level;
- their obesity status at the previous time point (t); and
- their neighbors' obesity status at times t and t+1

Statistical modeling intuition

Model one's obesity status at time point t+1 as a function of

- their age, sex, and educational level;
- their obesity status at the previous time point (t); and
- their neighbors' obesity status at times t and t+1

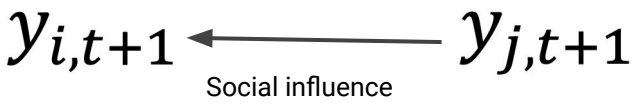
confounding factors (H2)

genetics plus intrinsic, stable predisposition to obesity (H2)

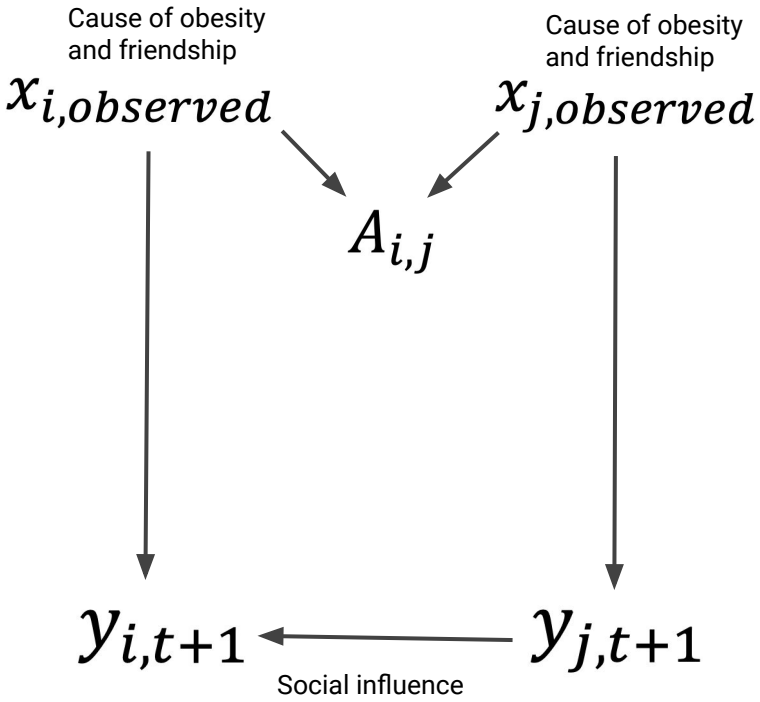
H1 – homophily (people choosing to form friendships with others of similar obesity status)

H3 – influence (a neighbor's weight affected the person's weight)

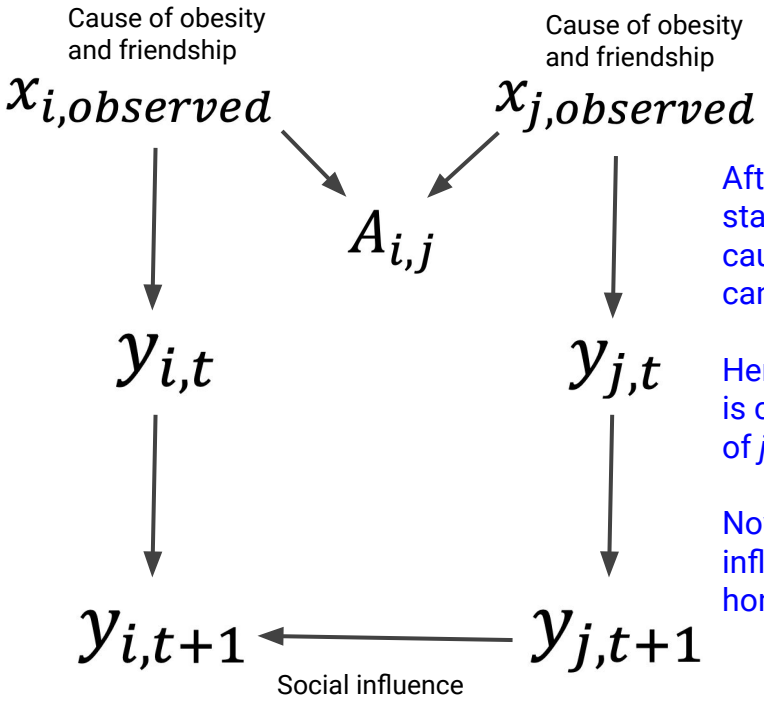
Causal Diagram



Causal Diagram



Causal Diagram

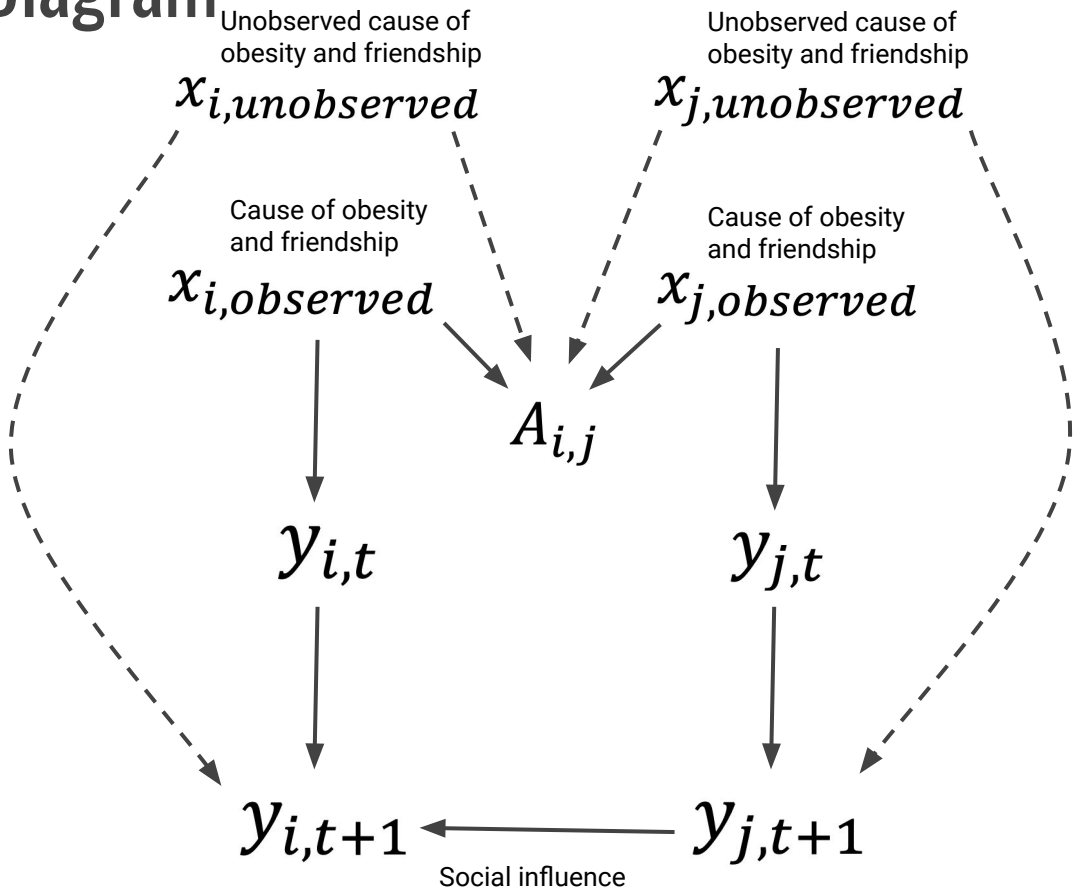


After controlling for *j*'s obesity status at *t*, the variable (*x*) that caused *i* and *j*'s friendship cannot affect *j*'s obesity at t+1

Hence, the effect of homophily is controlled for by the inclusion of *j*'s obesity at *t*

Now, the effect of social influence is not confounded by homophily

Causal Diagram



But, wait! It's a million dollar question for a reason

Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis

BMJ 2008 ; 337 doi: https://doi.org/10.1136/bmj.a2533 (Published 05 December 2008)

Longitudinal statistical analysis cannot always differentiate the effect of influence from selection.

Using the same longitudinal analysis, one might conclude that height is contagious!

We will explore why this is a hard problem in future lectures.

But, wait! It's a million dollar question for a reason

Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis

BMJ 2008 ; 337 doi: https://doi.org/10.1136/bmj.a2533 (Published 05 December 2008)

Results Significant network effects were observed in the acquisition of acne, headaches, and height. A friend's acne problems increased an individual's odds of acne problems (odds ratio 1.62, 95% confidence interval 0.91 to 2.89). The likelihood that an individual had headaches also increased with the presence of a friend with headaches (1.47, 0.93 to 2.33); and an individual's height increased by 20% of his or her friend's height (0.18, 0.15 to 0.26). Each of these results was estimated by using standard methods found in several publications. After adjustment for environmental confounders, however, the results become uniformly smaller and insignificant.

Conclusions Researchers should be cautious in attributing correlations in health outcomes of close friends to social network effects, especially when environmental confounders are not adequately controlled for in the analysis.

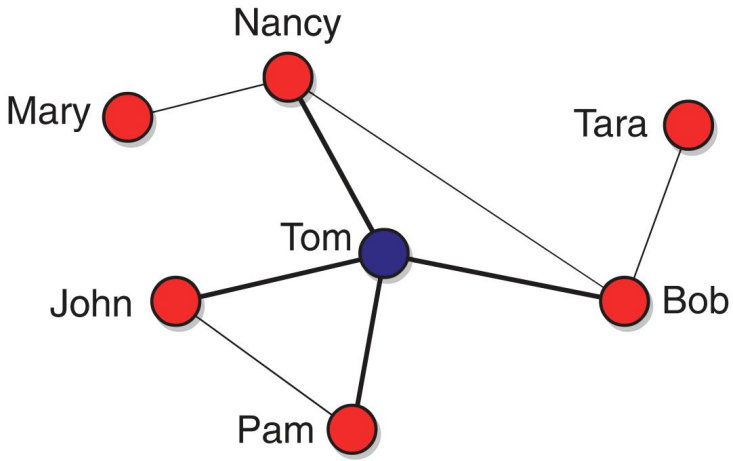
Case Study: The Friendship Paradox

Suppose you are looking for the person with the most friends

You only have a directory of phone numbers

Option 1: Call a person randomly

The chance that you pick Tom is … ?

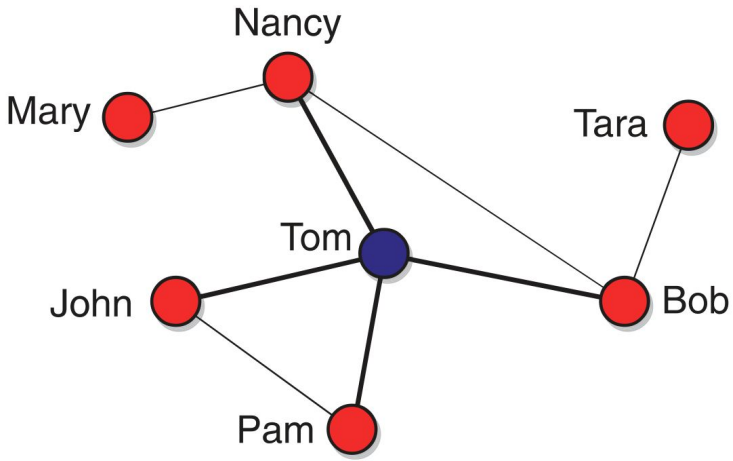


Suppose you are looking for the person with the most friends

You only have a directory of phone numbers

Option 1: Call a person randomly

The chance that you pick Tom is 1/7 ~ 14%

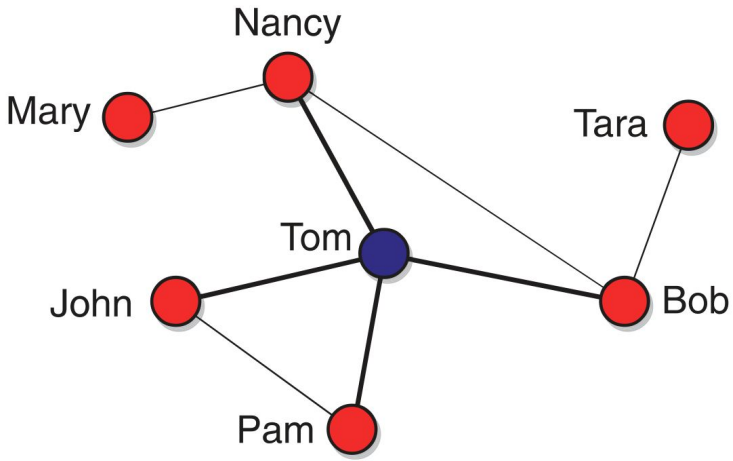


Suppose you are looking for the person with the most friends

You only have a directory of phone numbers

Option 2: Call a person randomly, and ask them about a random friend

The chance that you pick Tom is …?

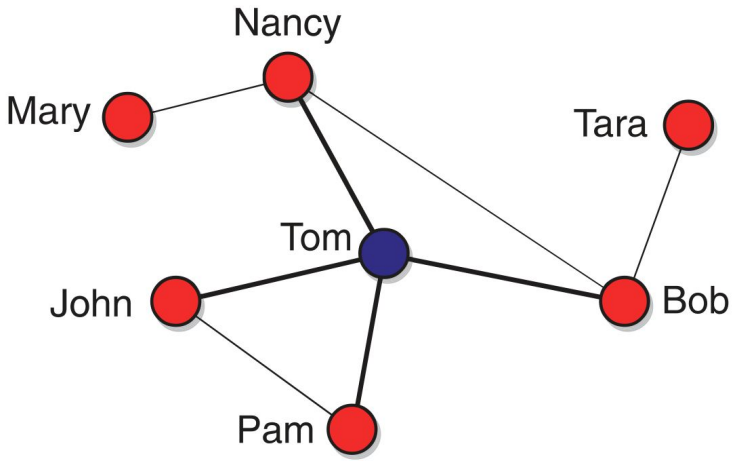


Suppose you are looking for the person with the most friends

You only have a directory of phone numbers

Option 2: Call a person randomly, and ask them about a random friend

The chance that you pick Tom is 5/21 ~ 24%



Mary: 0/1, Nancy: ⅓, John: ½, Pam: ½, Bob: ⅓, Tara: 0/1, Tom: 0/4

Probability of being called: 1/7

Therefore: (0/1+⅓+½+½+⅓+0/1+0/4)*1/7 = 5/21

Now, the paradox:

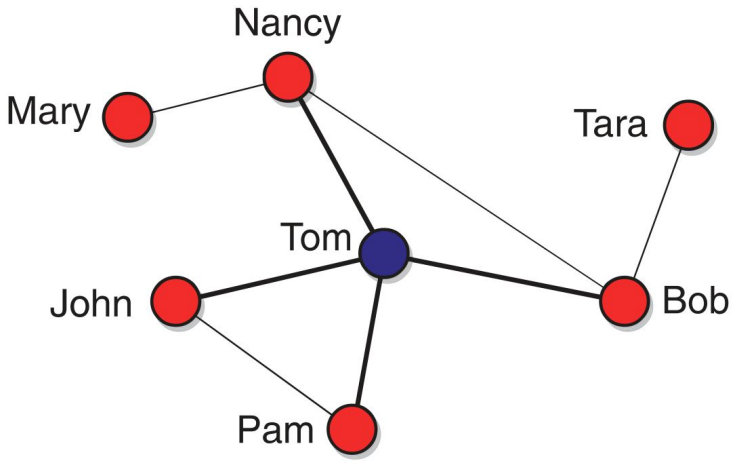
Nancy has 3 friends: Mary, Tom, Bob

They have in total 1 + 4 + 3 = 8 friends

→ Nancy's friends have on average 8/3 friends

Average degree: (1+3+4+2+2+3+1)/7
= 16 / 7 = 2.29

Average degree of neighbors: 2.83



Now, the paradox:

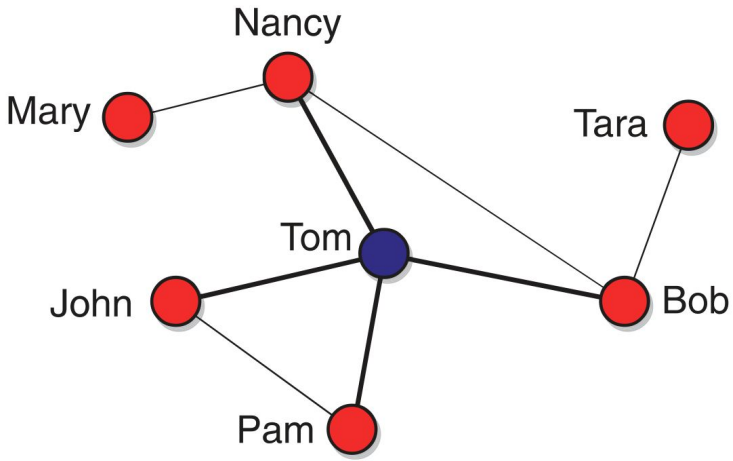
Nancy has 3 friends: Mary, Tom, Bob

They have in total 1 + 4 + 3 = 8 friends

→ Nancy's friends have on average 8/3 friends

Average degree: 16 / 7 = 2.29

Average degree of neighbors: 2.83



Your friends have more friends than you, on average!

Summary

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

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

Two extremely powerful analysis techniques: comparison to a random (shuffled) network and longitudinal analysis!