Network Analysis:

The Hidden Structures behind the Webs We Weave

17-338 / 17-668

Homophily and Degree Correlation (Part 1)

Tuesday, September 16, 2025

Patrick Park

Carnegie Mellon University
School of Computer Science

S3D
Software and Societal Systems Department

Quick Recap – Last Tuesday's Lecture

Graph signature of social ties

Social tie dynamics

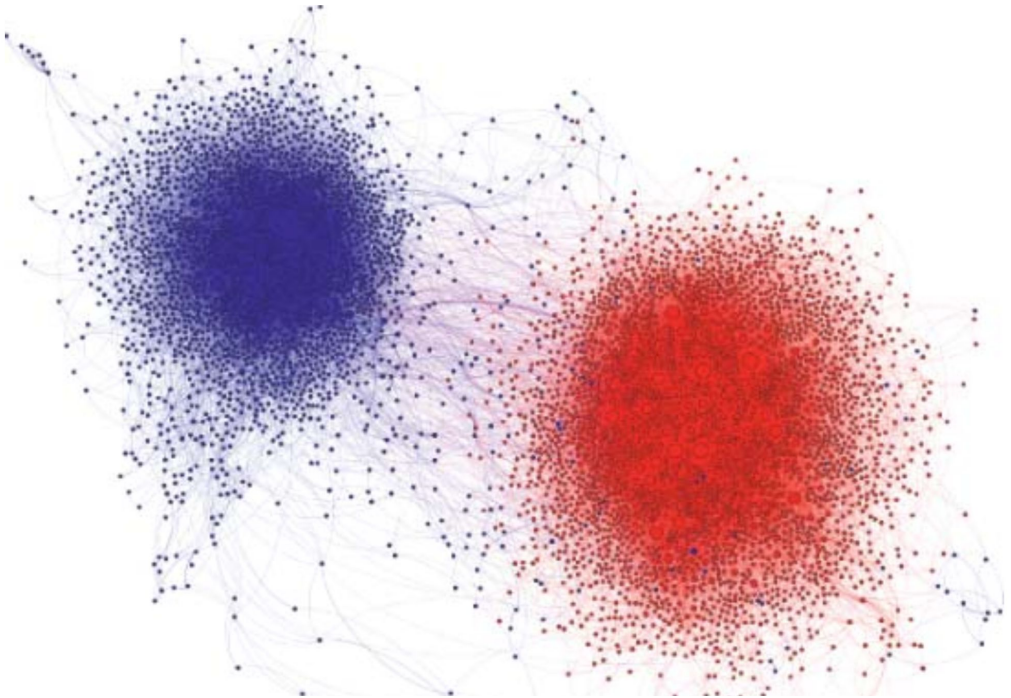
Birds of a Feather

Birds of a Feather Flock Together

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

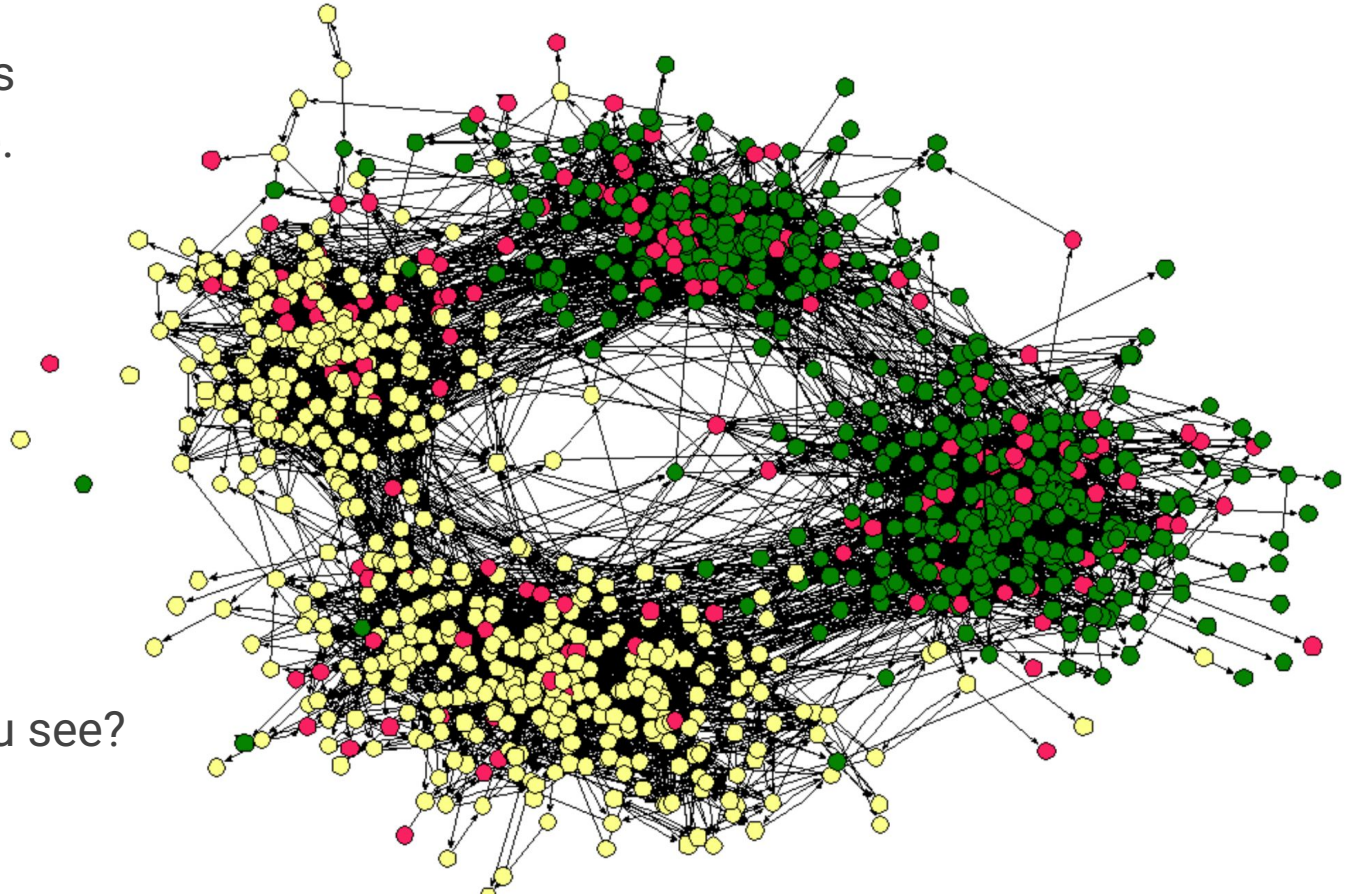
“Similarity begets friendship.” -Plato

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



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

Circle colors denote race.

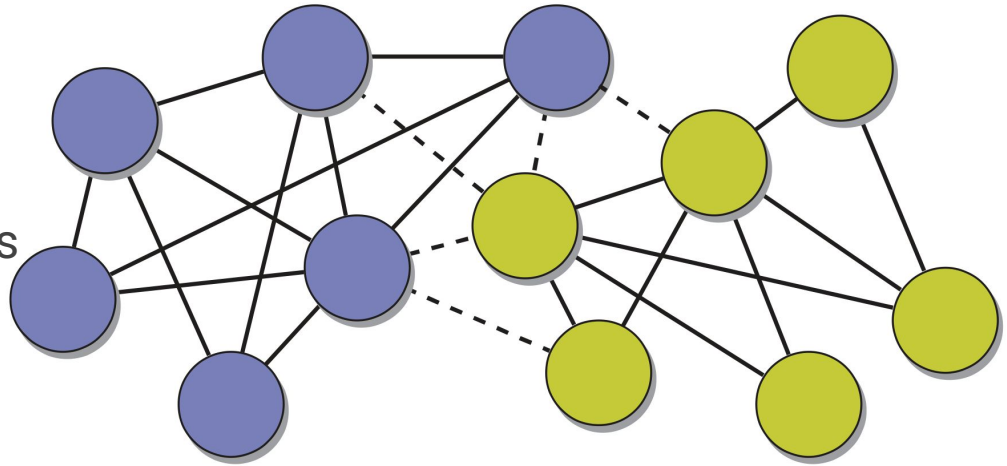


What do you see?

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.

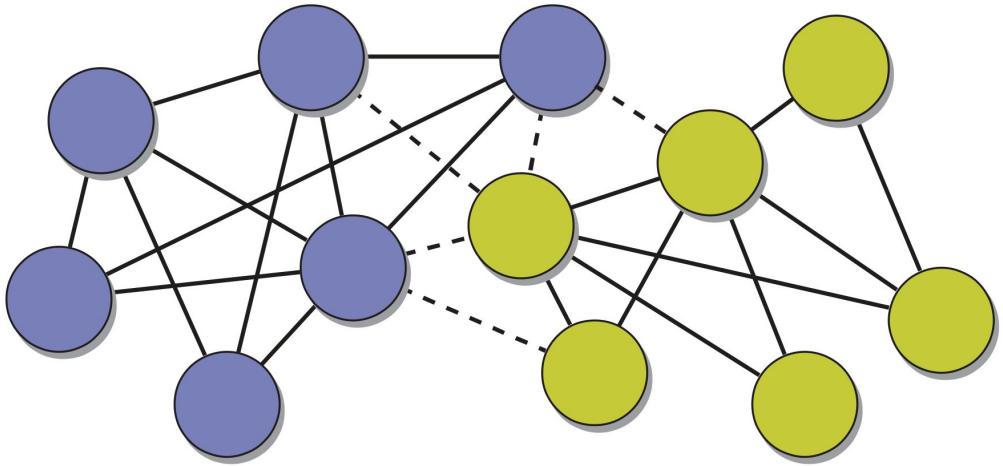
The majority of links connect nodes of the same color.



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

Salient dimensions:

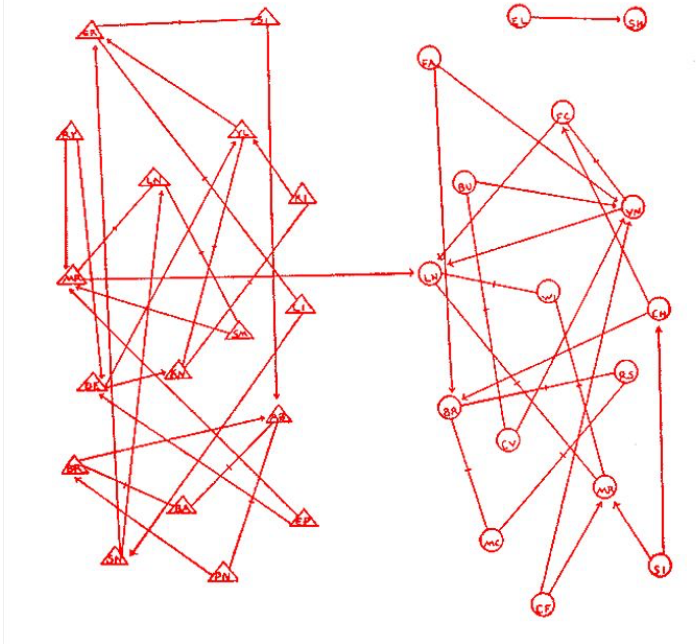
- Race, ethnicity
- Gender, sex
- Age
- Religion
- Occupation/education



Homophily: Gender

Salient dimensions:

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

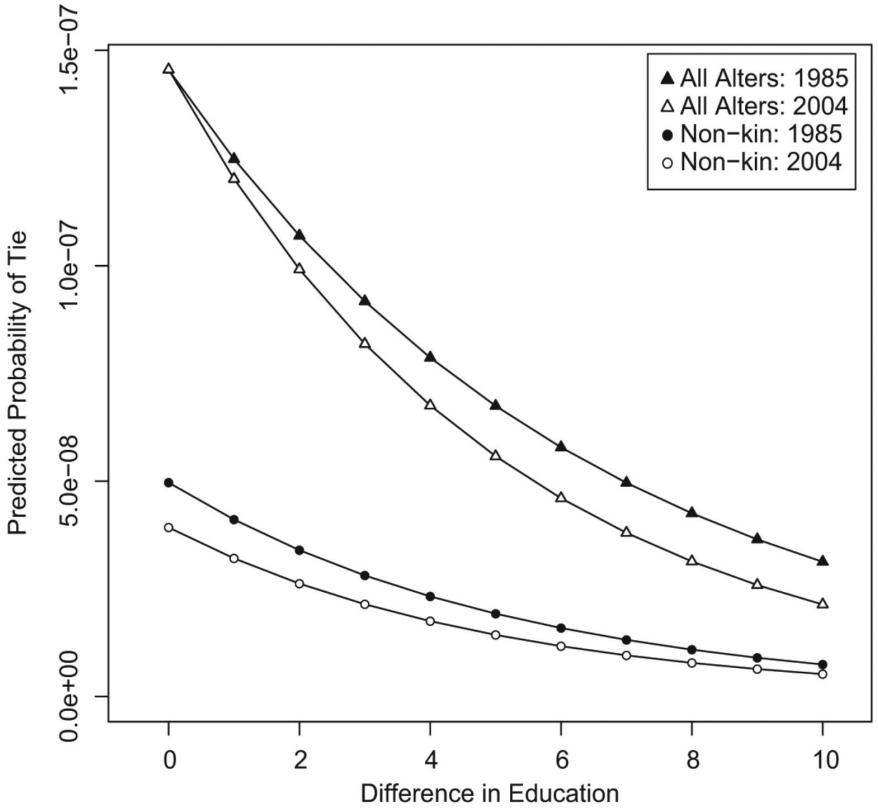


Homophily: Education

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

Tie probability is lower for non-kin.

Educational homophily weakened in more recent years.



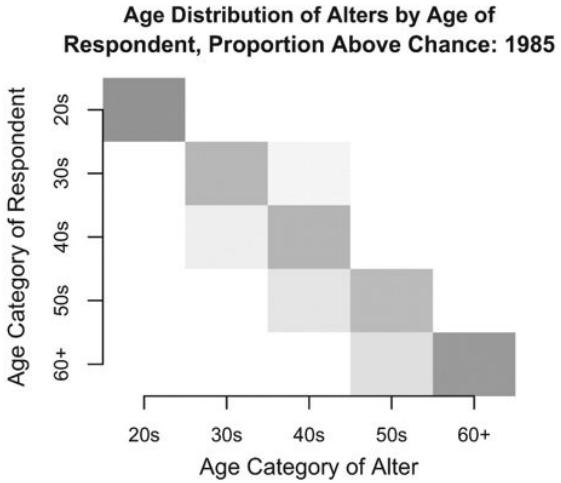
Smith et al. 2014

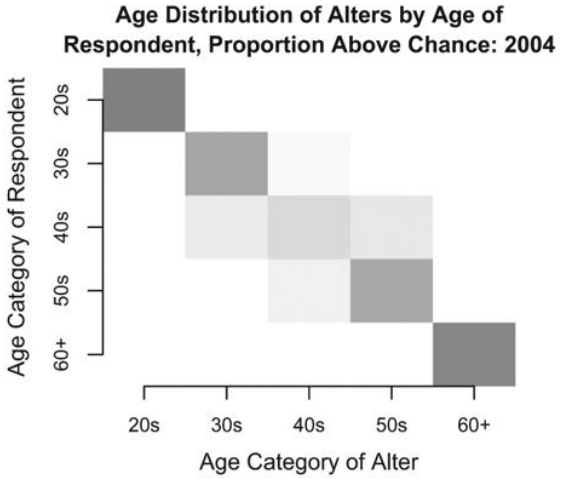
Homophily: Age

Age homophily slightly increased over time.

Higher levels of homophily at 20s and 60s:

Why?

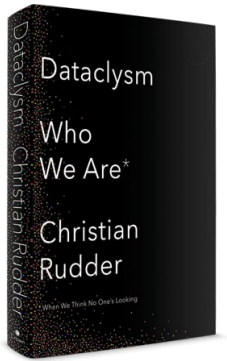


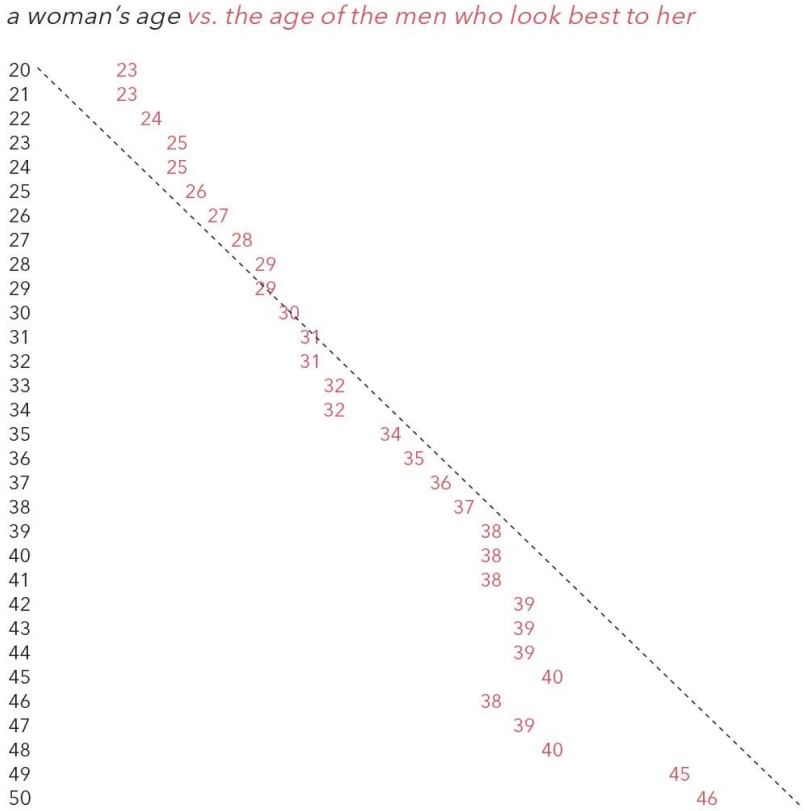


Smith et al. 2014

Homophily: Age

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

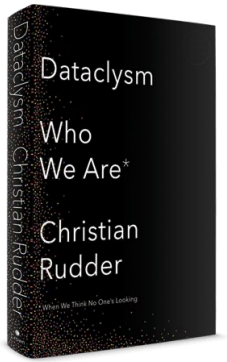




Homophily: Age

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

Homophily can be asymmetric



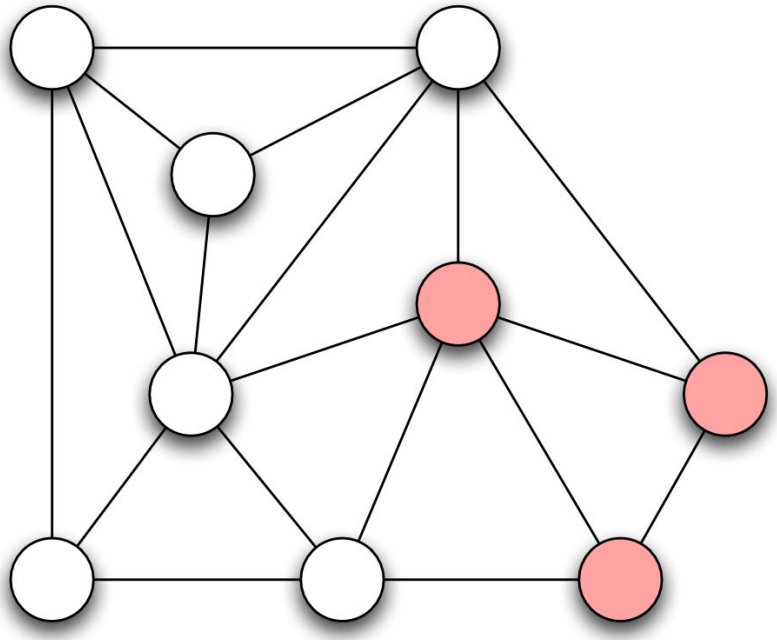
a man's age vs. the age of the women who look best to him

20	20
21	20
22	21
23	21
24	21
25	21
26	22
27	21
28	20
29	20
30	20
31	20
32	20
33	20
34	20
35	20
36	20
37	22
38	20
39	20
40	21
41	21
42	20
43	23
44	21
45	24
46	20
47	20
48	23
49	20
50	22

Measuring homophily

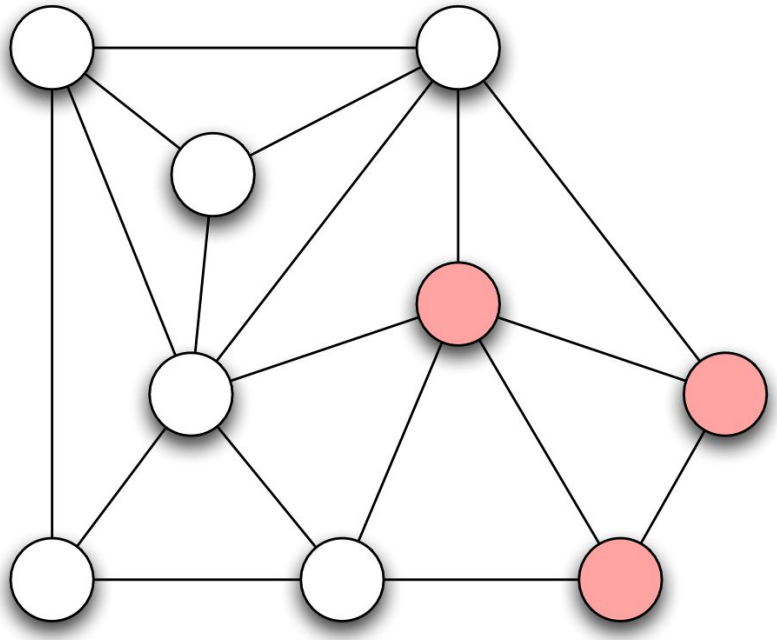
Given a particular characteristic of interest (like race, or age), is there a simple test we can apply to a network to estimate whether it exhibits homophily according to this characteristic?

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



Measuring homophily

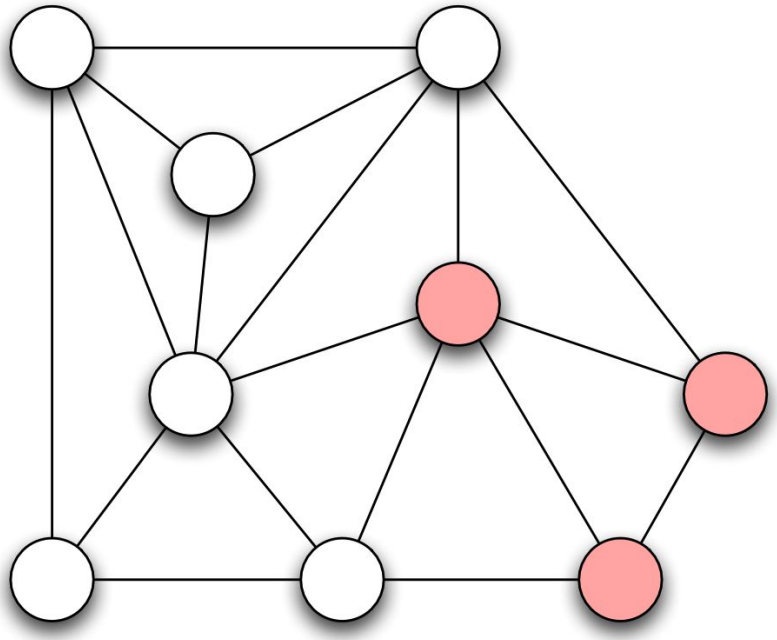
What would it mean for the network not to exhibit homophily by gender?



Measuring homophily

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

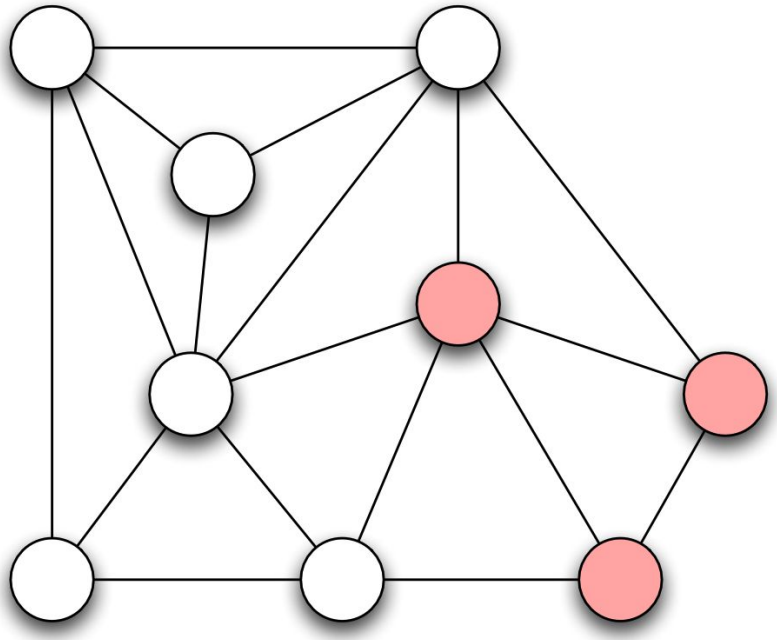
The proportion of male and female friends a person has should look like the background male/female distribution in the full population.



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.

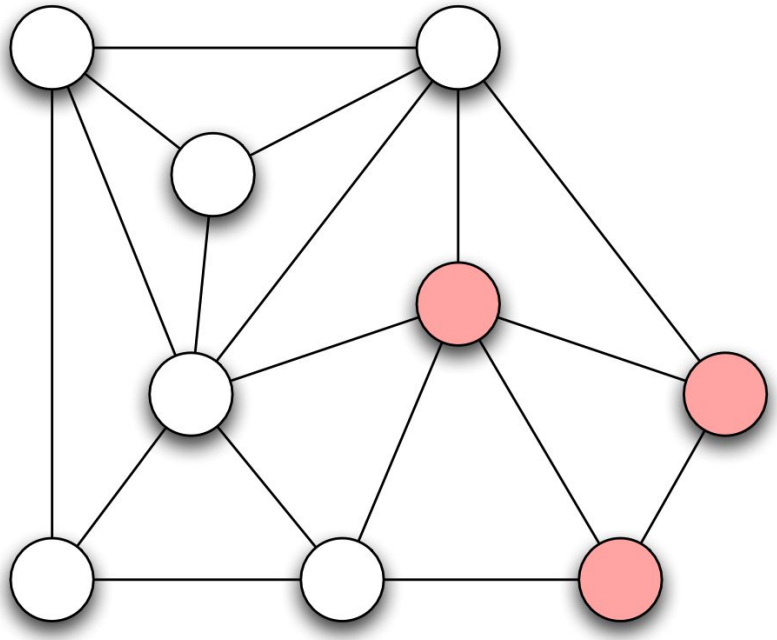


Measuring homophily

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 …?

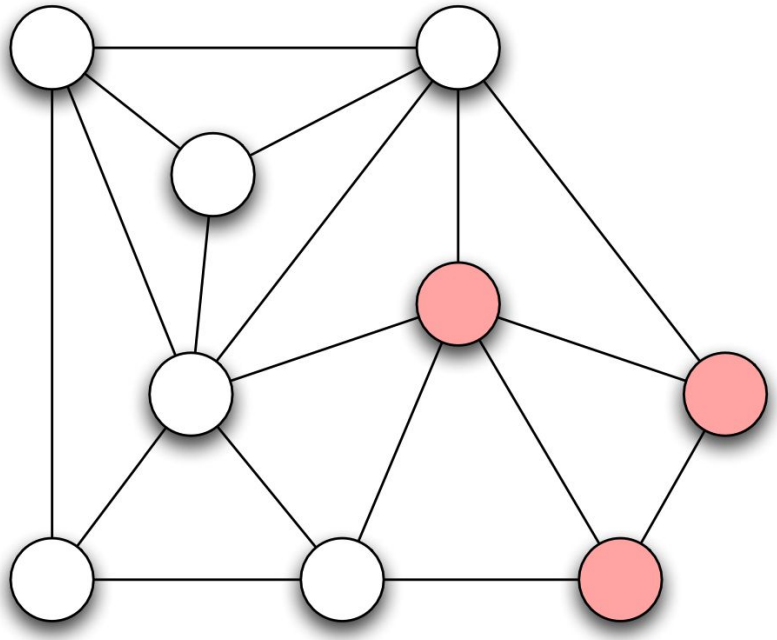


Measuring homophily

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 $2pq$



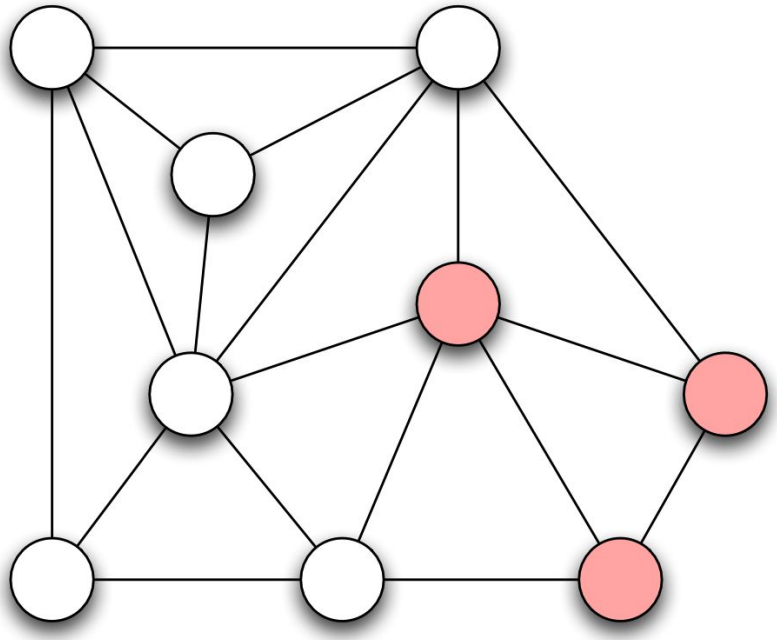
Measuring homophily

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.



Back to interpreting homophily

Competing mechanisms

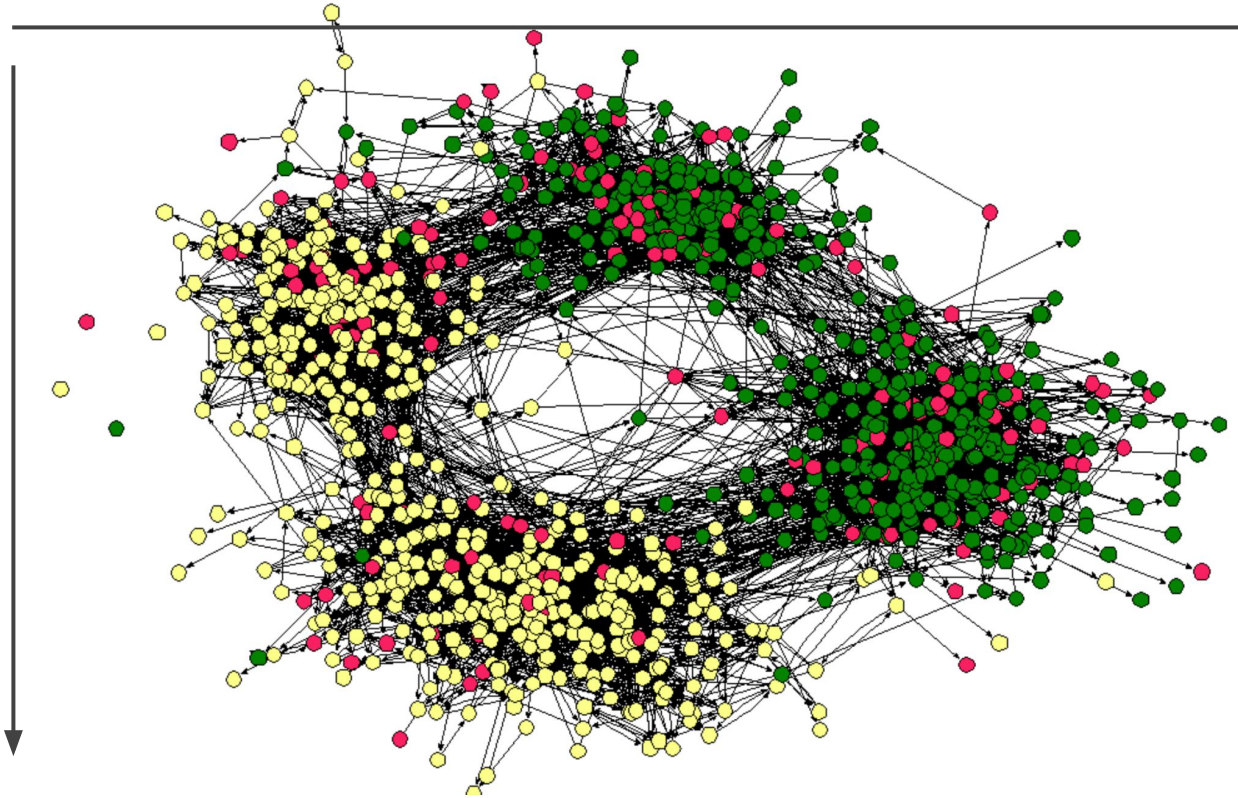
Selection (“homophily”): If people are similar in some way, they are more likely to select each other and become connected.

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

Homophily: Intrinsic vs contextual effects

Race (intrinsic)

Middle vs high school (context)



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?

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

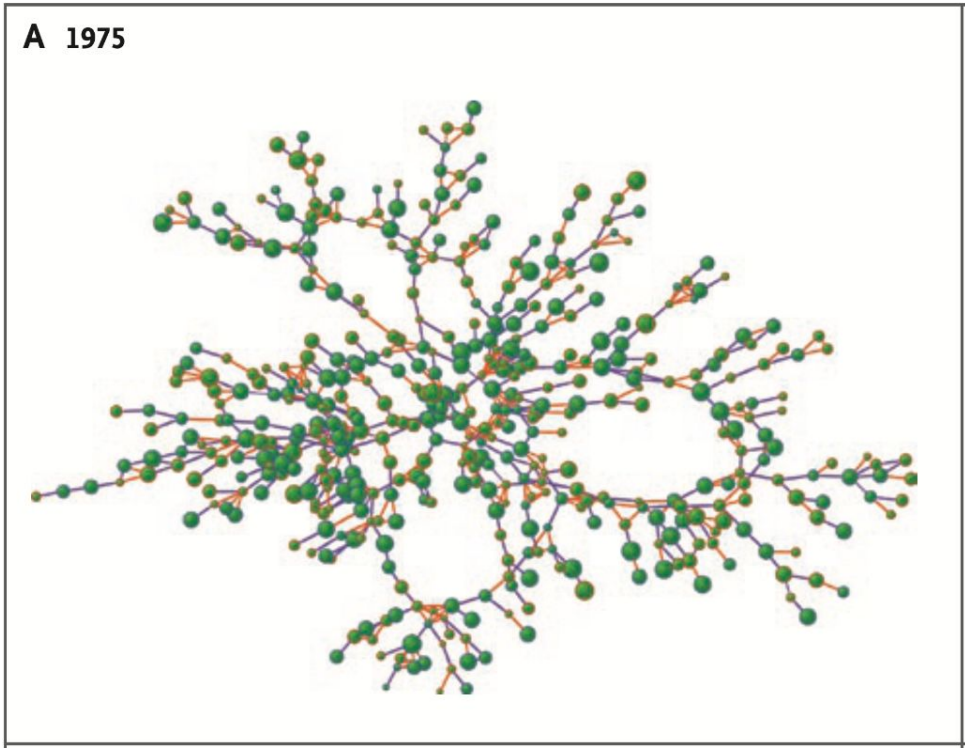
In a small group, when people choose friends who are most similar from among a clearly delineated pool of contacts, there is clearly active choice going on.

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

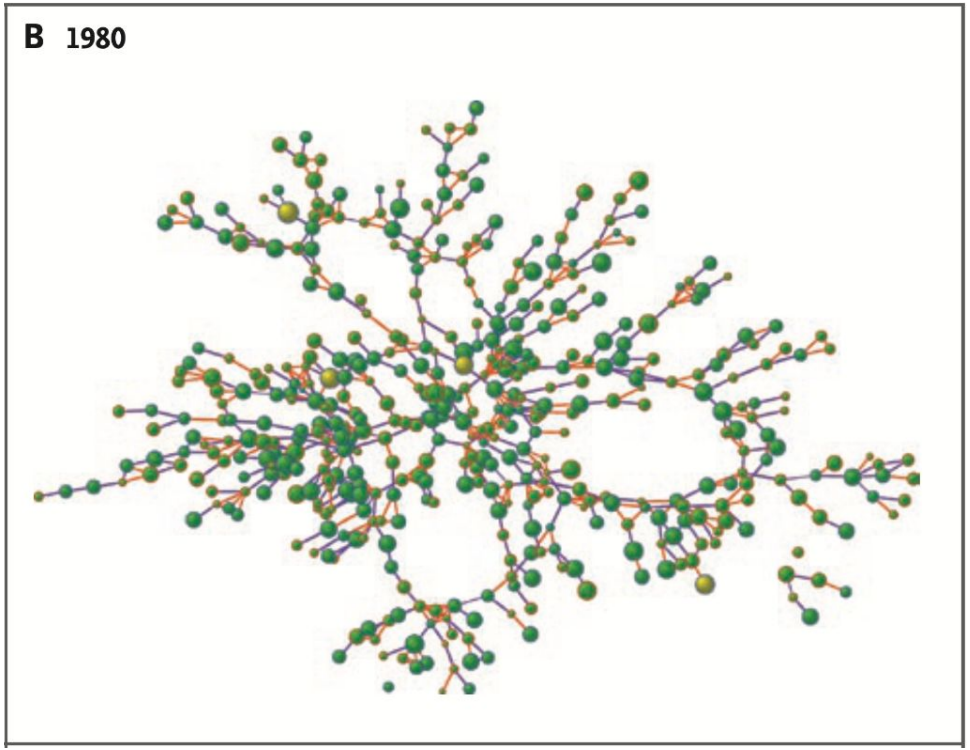
For example, when people live in neighborhoods, attend schools, or work for companies that are relatively homogeneous compared to the population at large.

Case Study: obesity study showing evidence of social influence

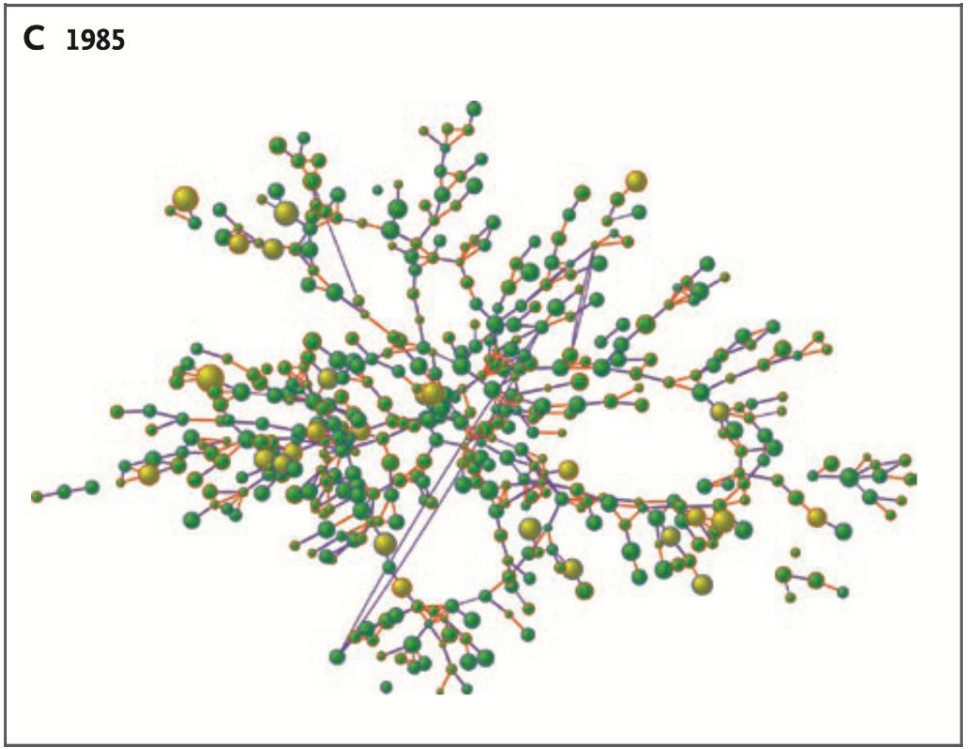
Framingham heart study



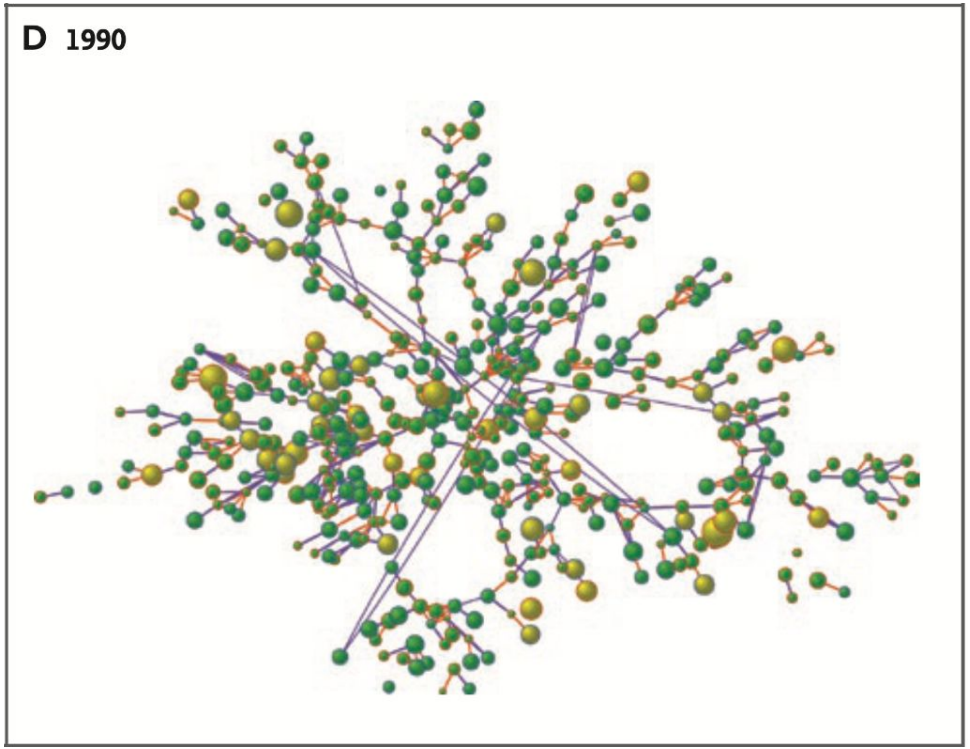
Framingham heart study



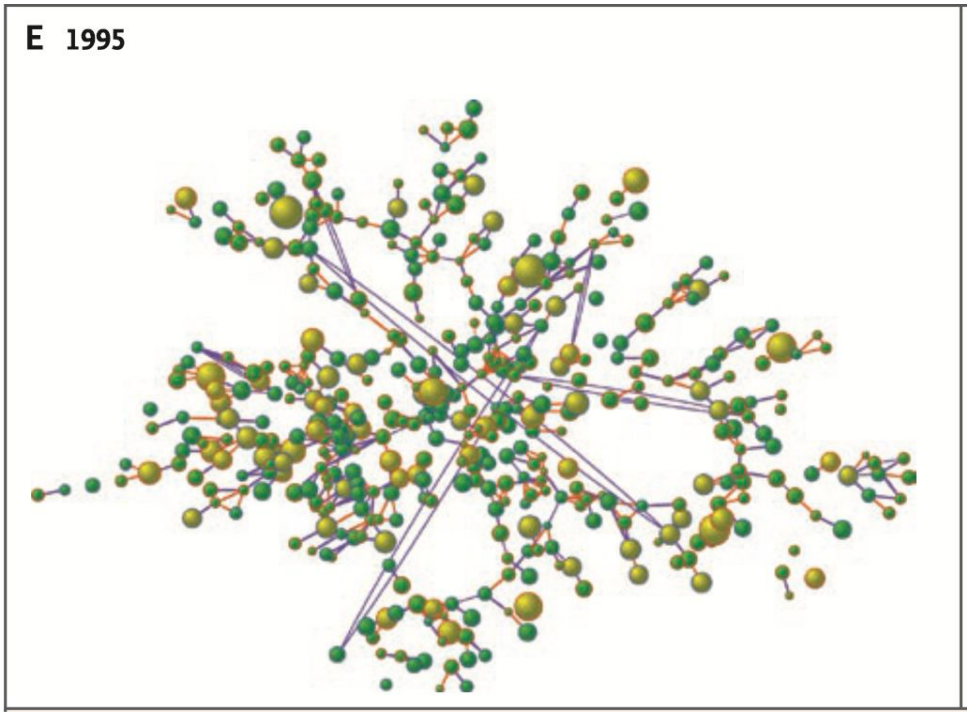
Framingham heart study



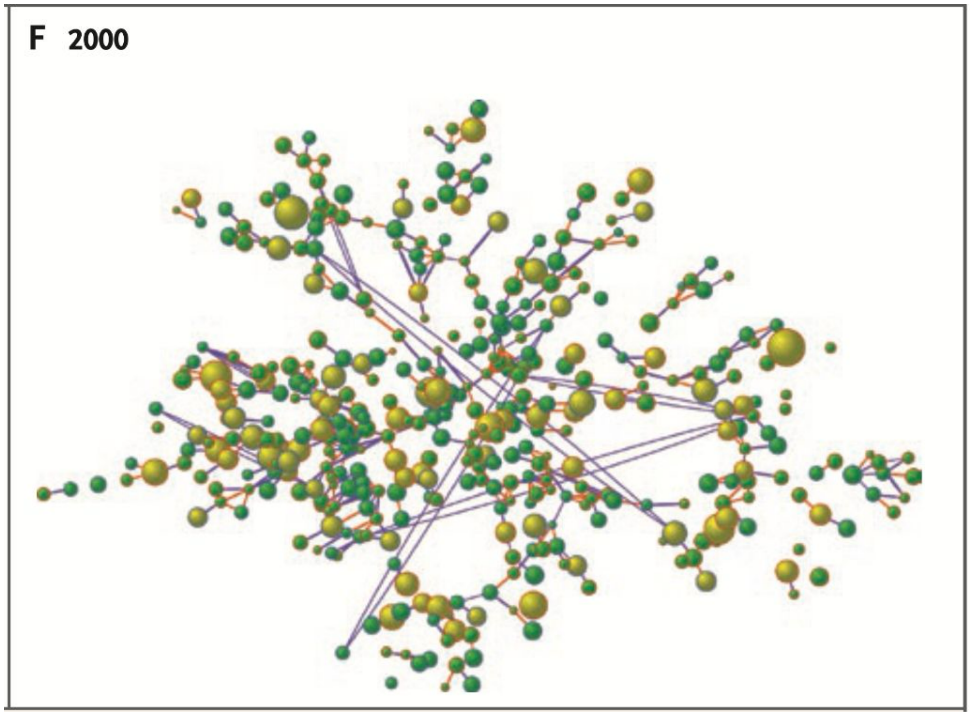
Framingham heart study



Framingham heart study



Framingham heart study



Framingham heart study network

Red borders: women

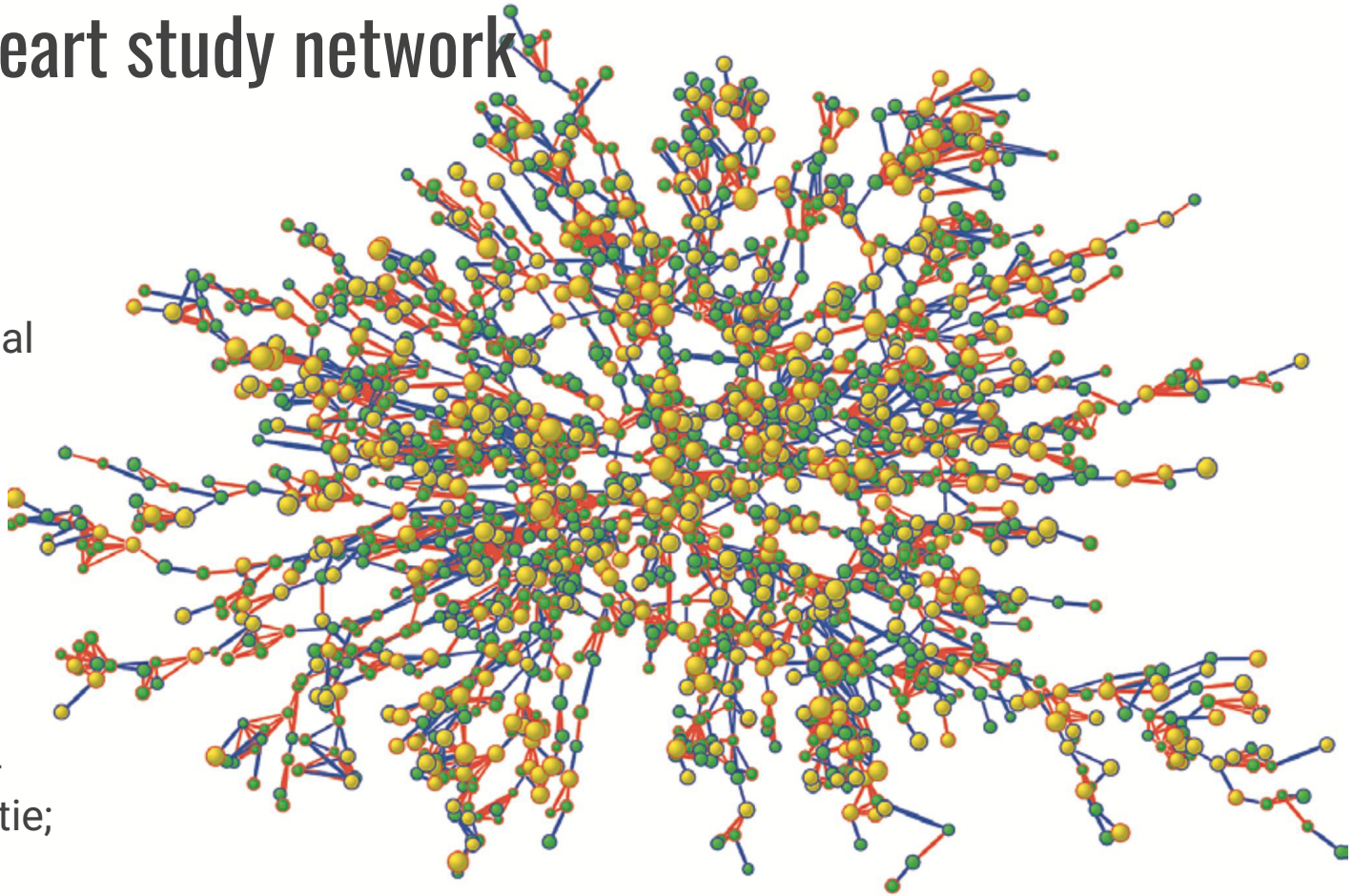
Blue borders: men.

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

Yellow: body-mass index ≥30 ("obese")

Green: nonobese.

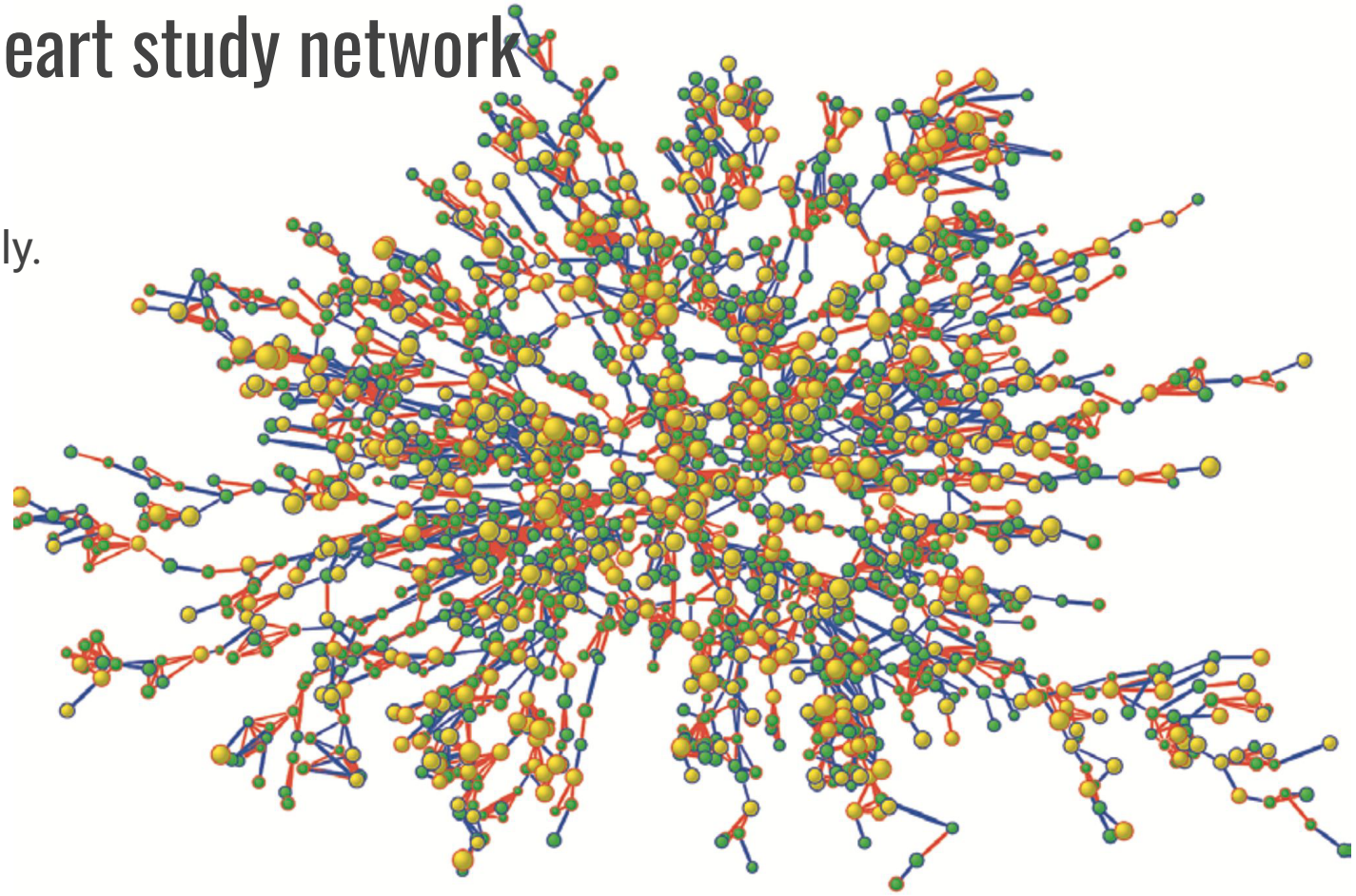
Tie colors indicate relationship: purple – friendship or marital tie; orange – familial tie.



Framingham heart study network

The researchers
tested for homophily.

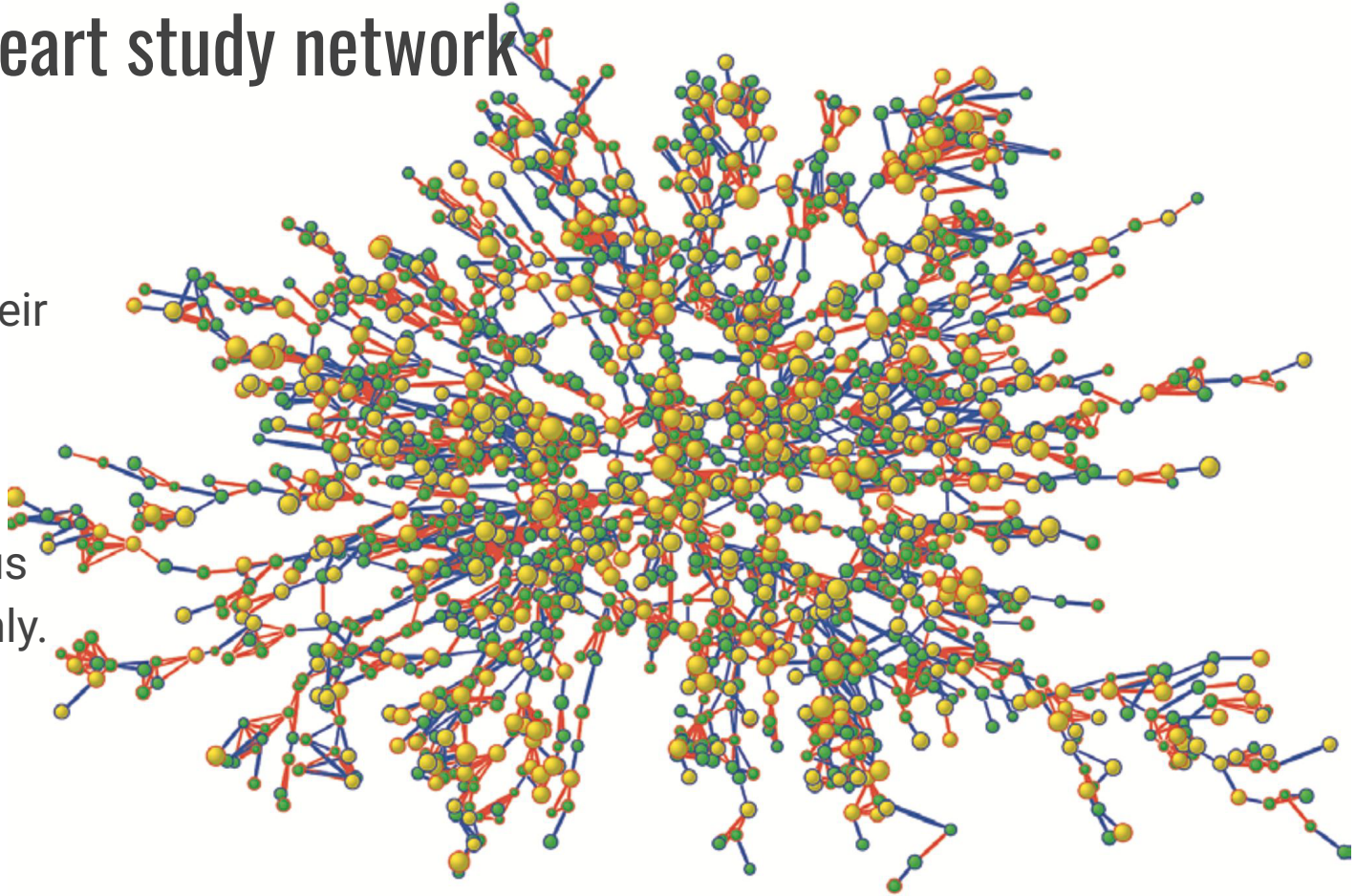
How?



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 on other characteristics, in which the network structure indicates existing patterns of similarity in other dimensions that correlate with obesity status

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

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

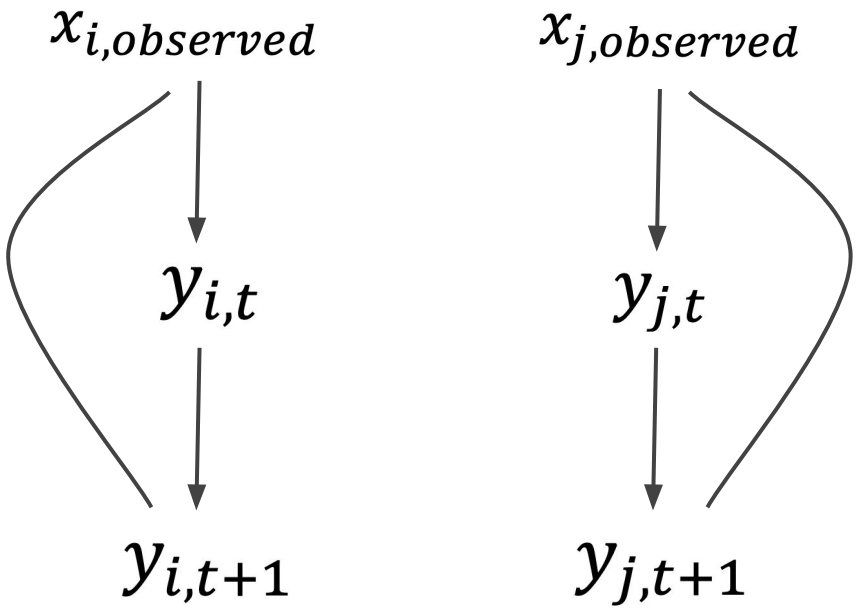
confounding factors (H2)

genetics plus intrinsic, stable predisposition to obesity (H2)

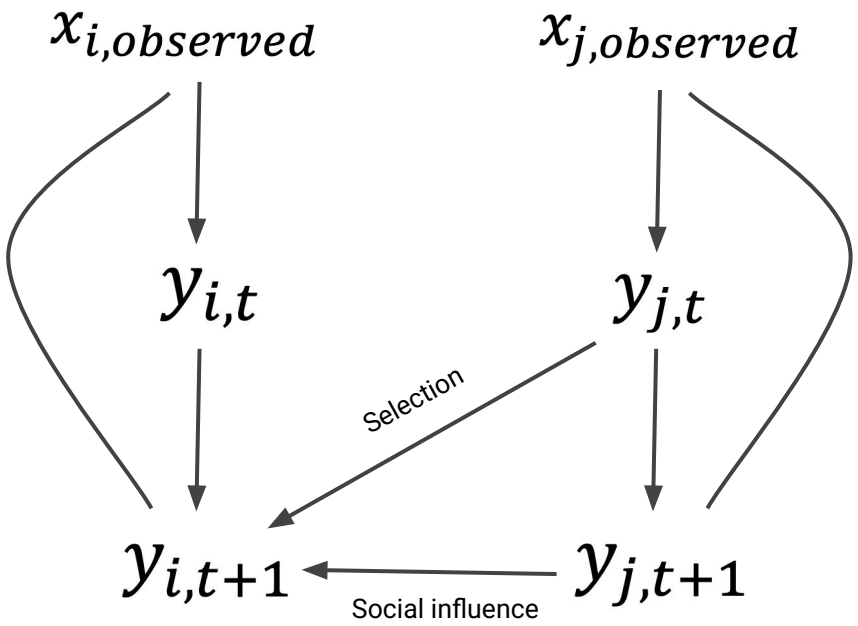
H1 – homophily (people choosing to form friendships with others of similar obesity status)

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

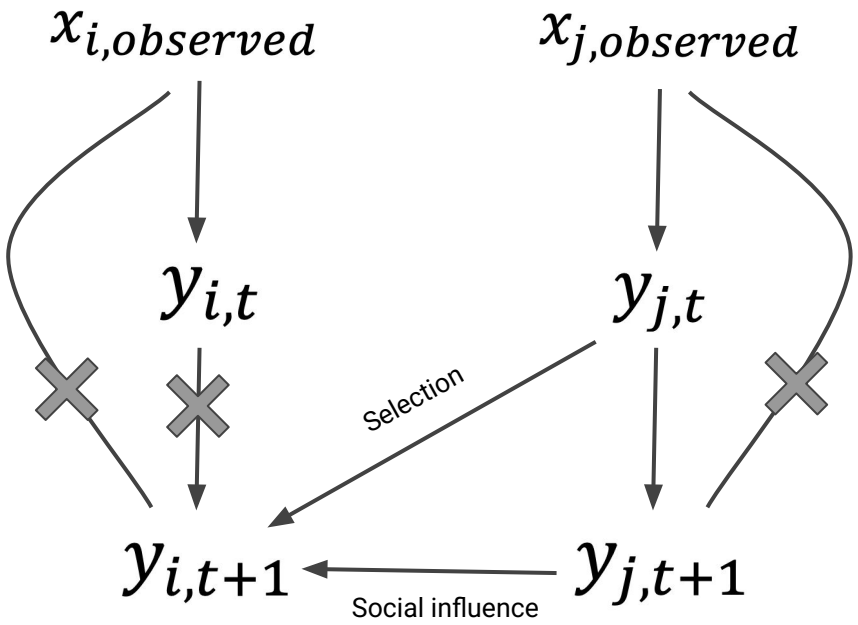
Causal diagram



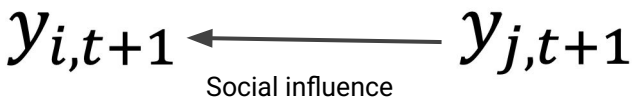
Causal diagram



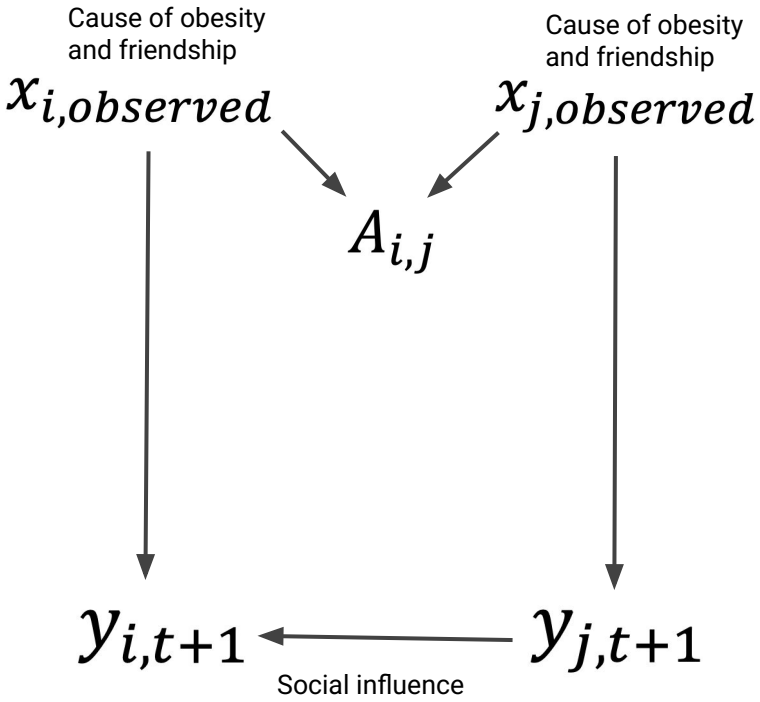
Causal diagram



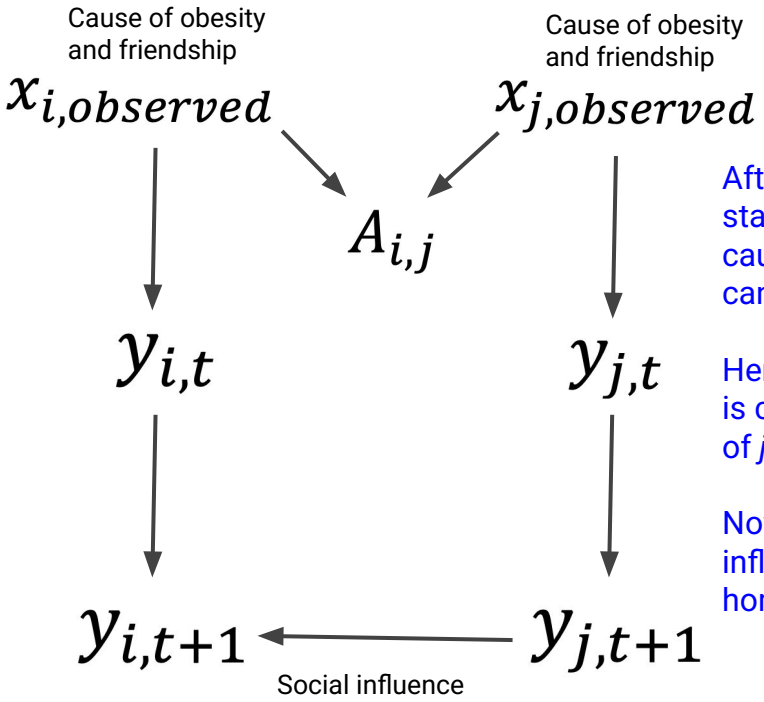
Causal Diagram



Causal Diagram



Causal Diagram

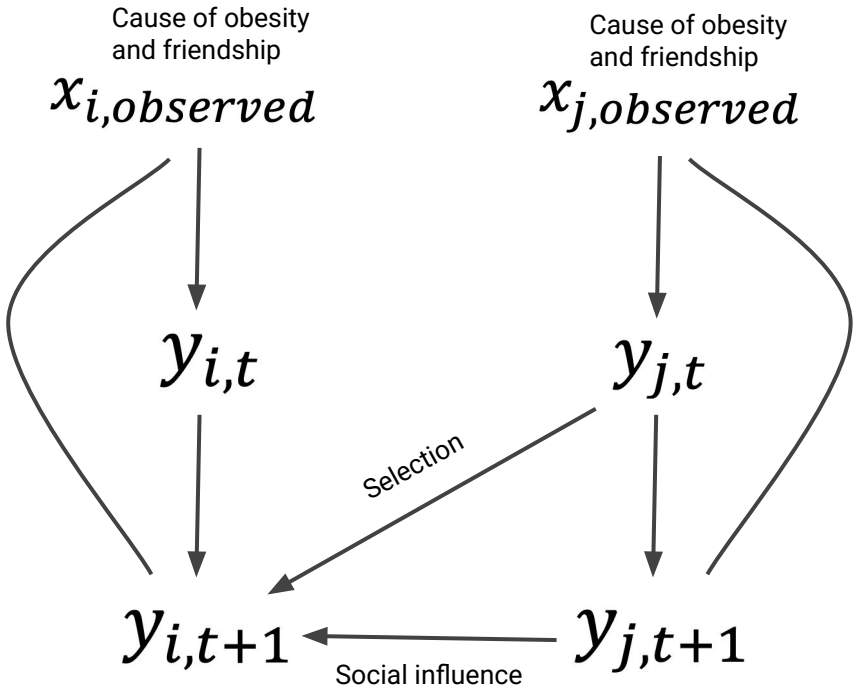


After controlling for *j*'s obesity status at *t*, the variable (*x*) that caused *i* and *j*'s friendship cannot affect *j*'s obesity at t+1

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

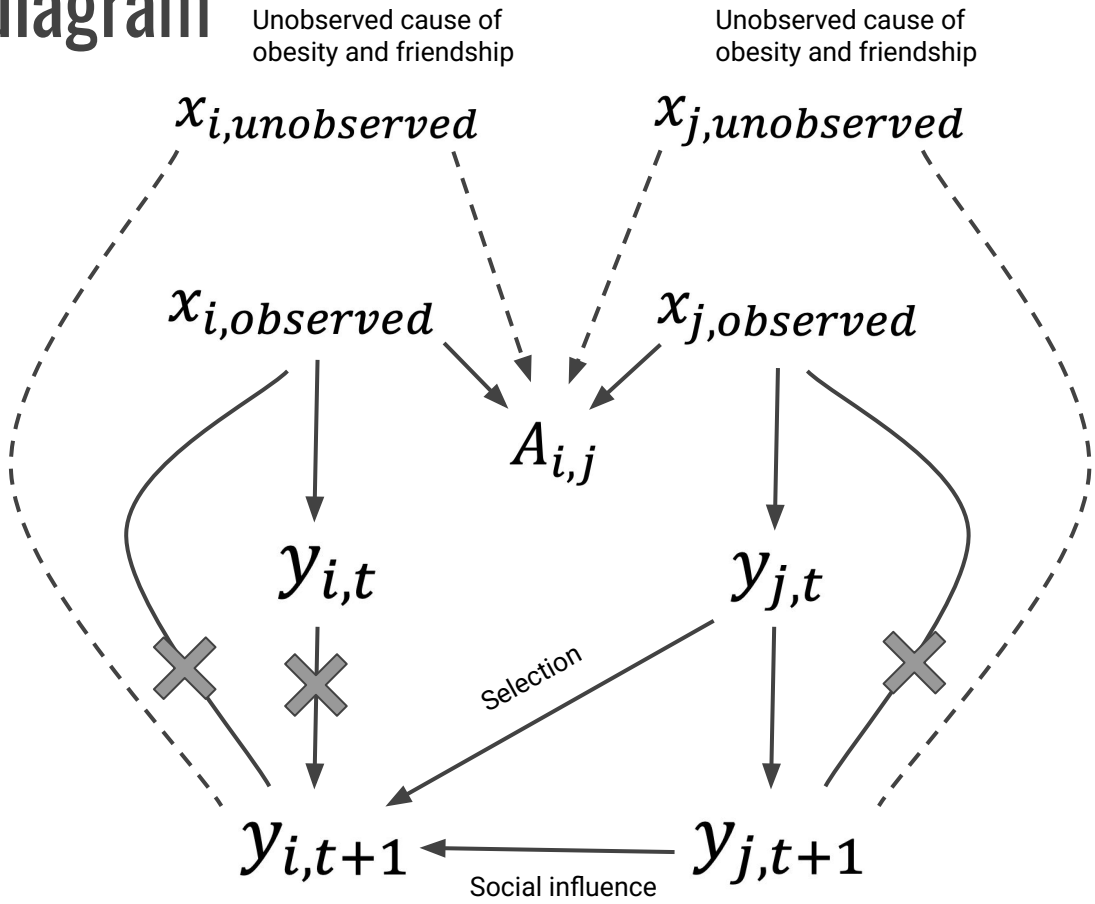
Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis

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

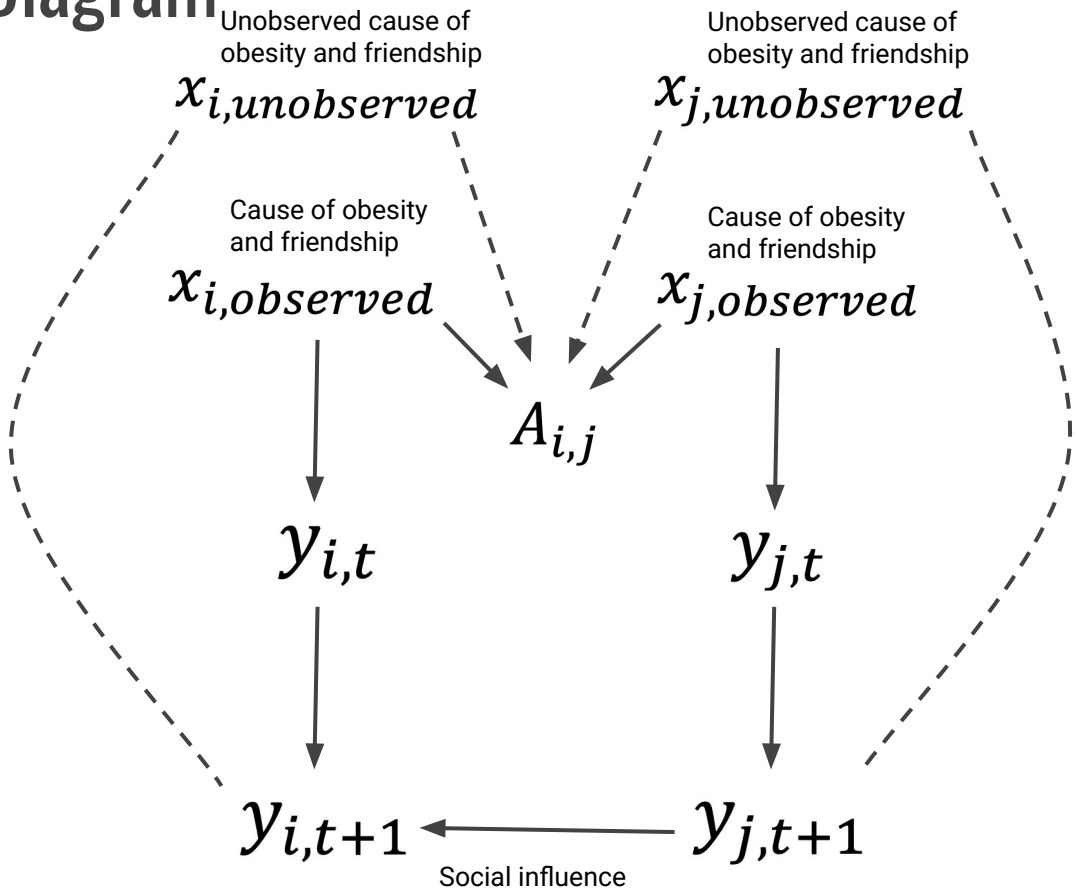
Longitudinal statistical analysis cannot always differentiate the effect of social influence from homophily-based selection.

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

Causal diagram



Causal Diagram



Summary

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

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

One extremely powerful analysis technique: comparison to a random (shuffled) network.