http://www.di.unito.it/~tizzani

michele.tizzani@unito.it

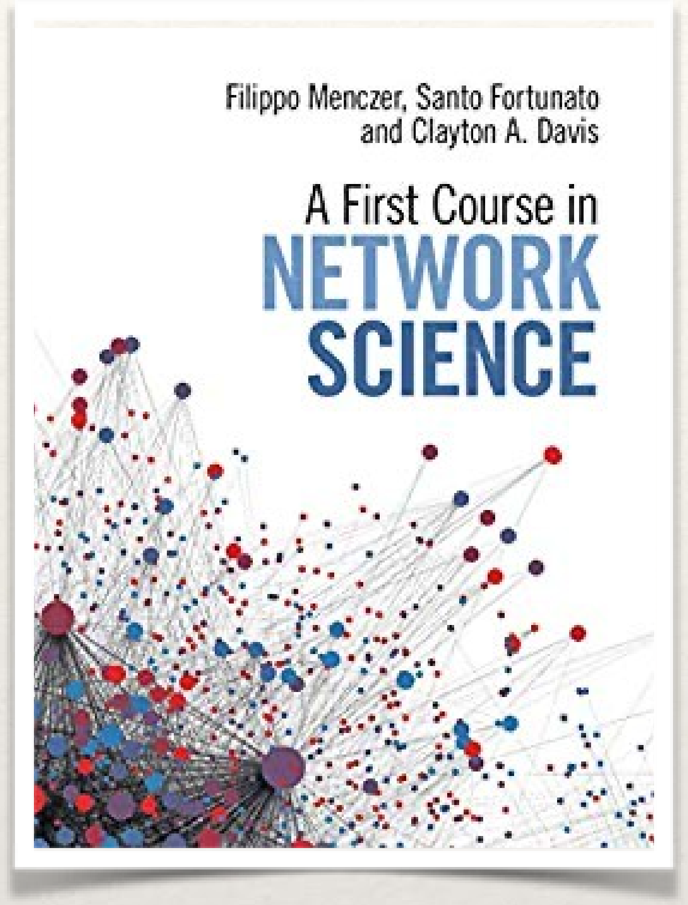
Michele Tizzani - Università degli Studi di Torino (Italy)

Lecture 06.ns04

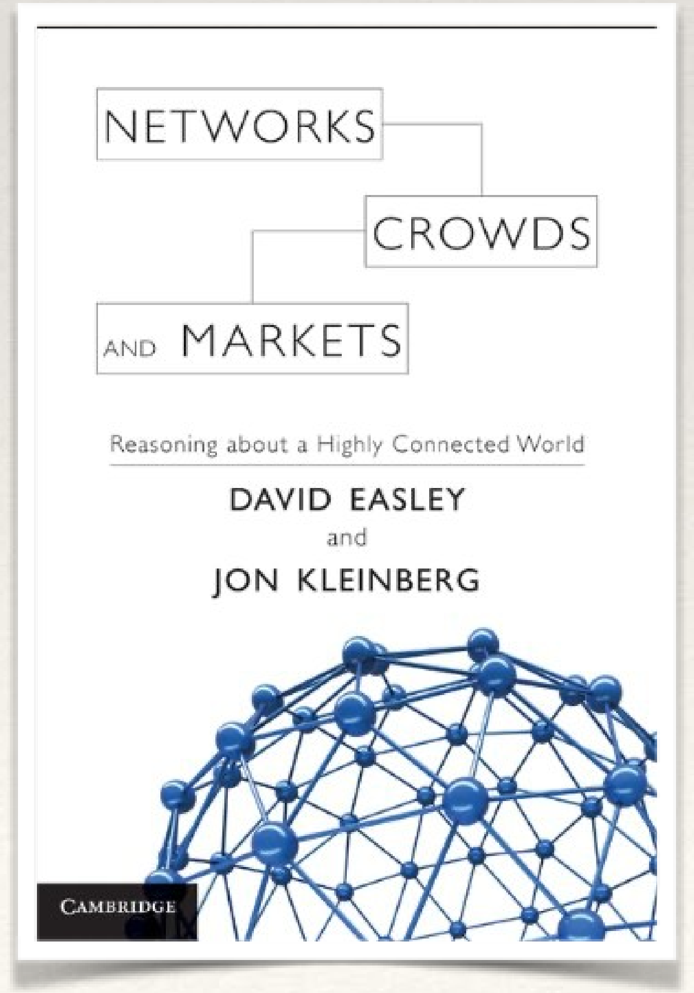
Homophily

Course: **Complex Networks Analysis and Visualization**
Sub-Module: **NetSci**

References



[ns1] Chapter 2 (2.1)



[ns2] Chapter 4 (4.1-4.3, 4.5) —->

https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch04.pdf

Agenda

Homophily or Assortative mixing

Homophily by Degree

Mechanisms Underlying Homophily: Selection and Social Influence

Affiliation

Homophily or "Birds of a feather"

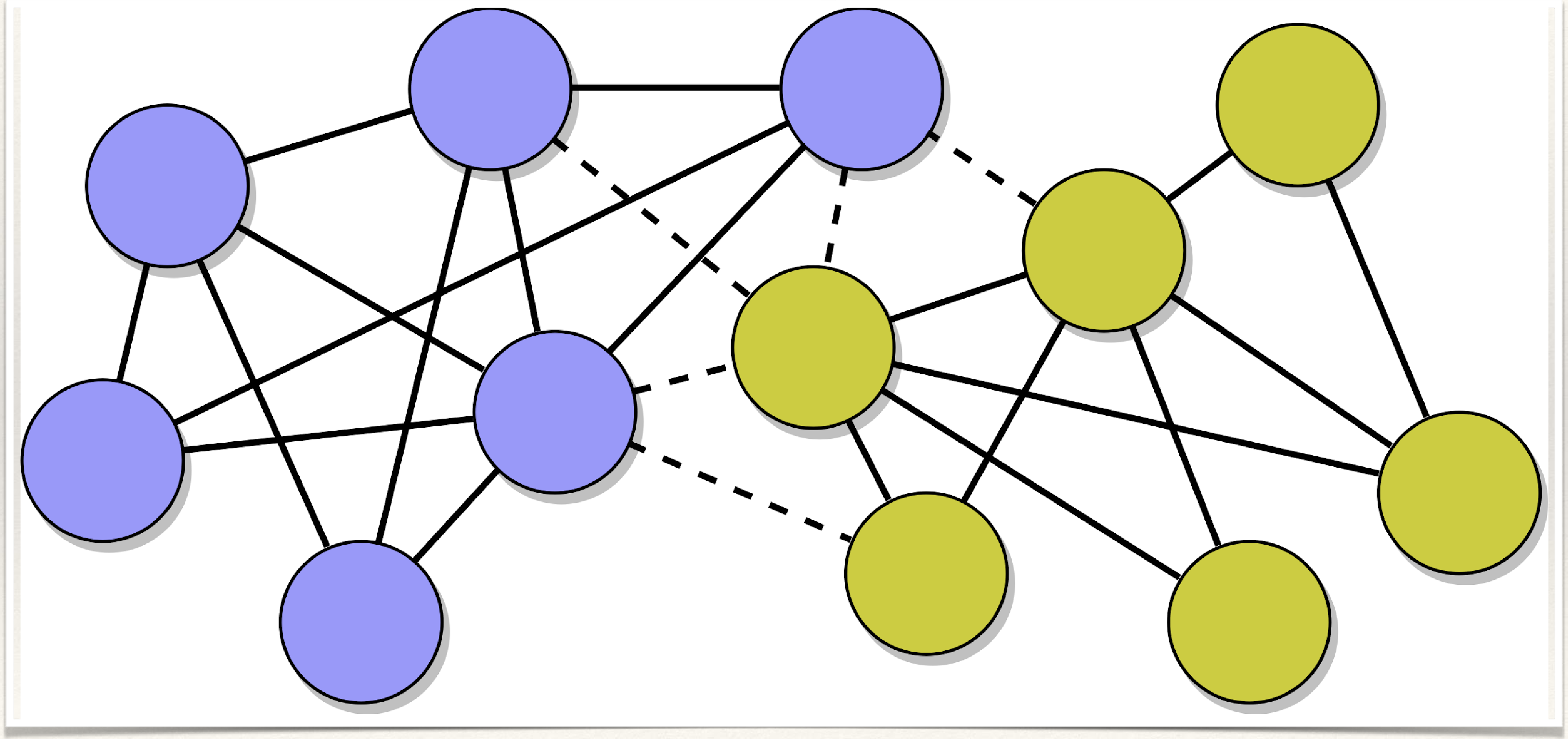
Surrounding contexts

- We focus on
  - structural properties
  - link formation processes
  - personal characteristics
  - similarities between individuals
- **surrounding contexts**: factors that exist outside the nodes and the edges of a network but that can have an effect on the evolution of the system

Homophily

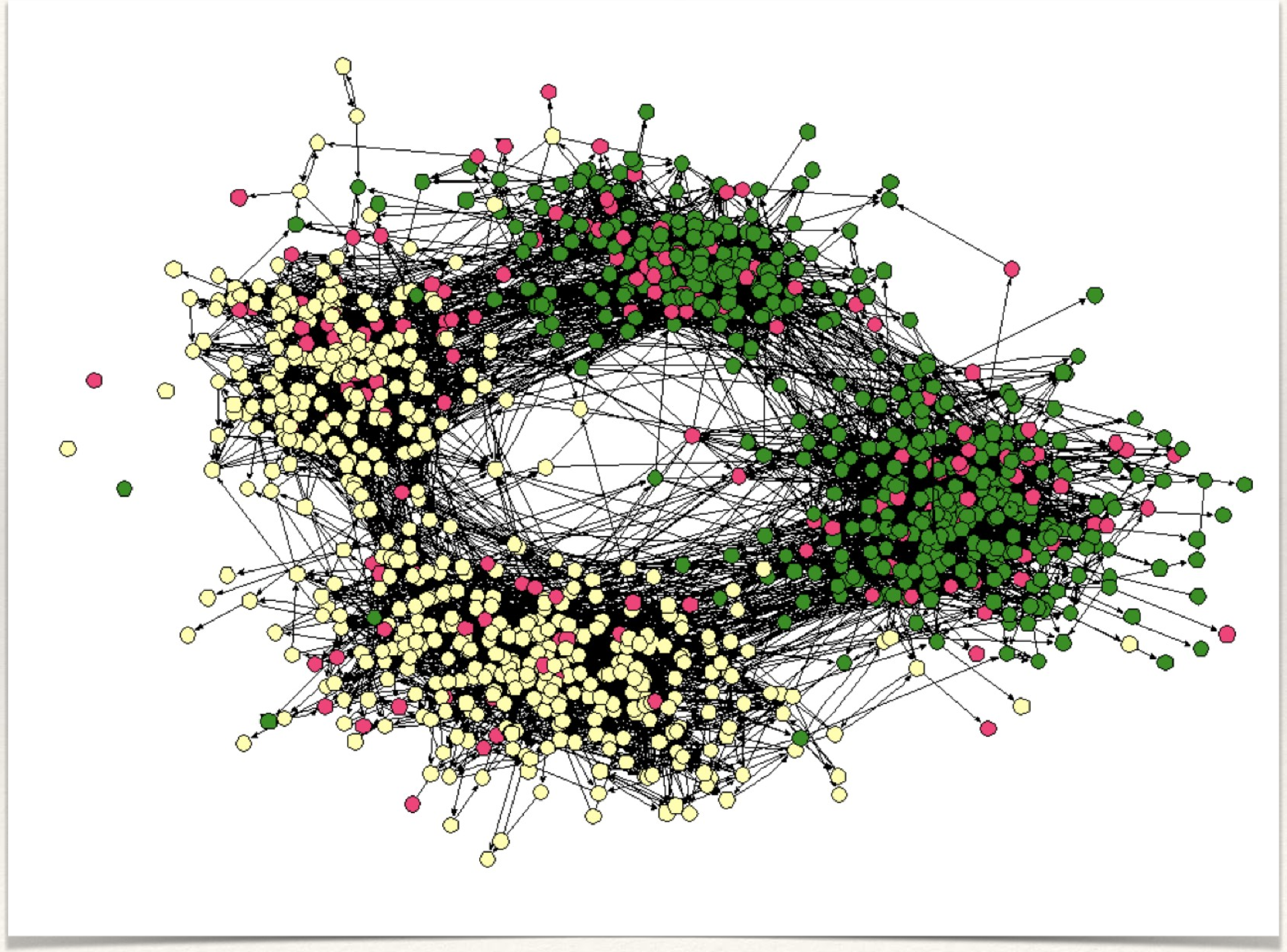
- The principle that we tend to be similar to our friends
- This makes your friends not statistically significant as a random sample of the population
- similarities
  - immutable characteristics
  - mutable characteristics

"Birds of a feather flock together"



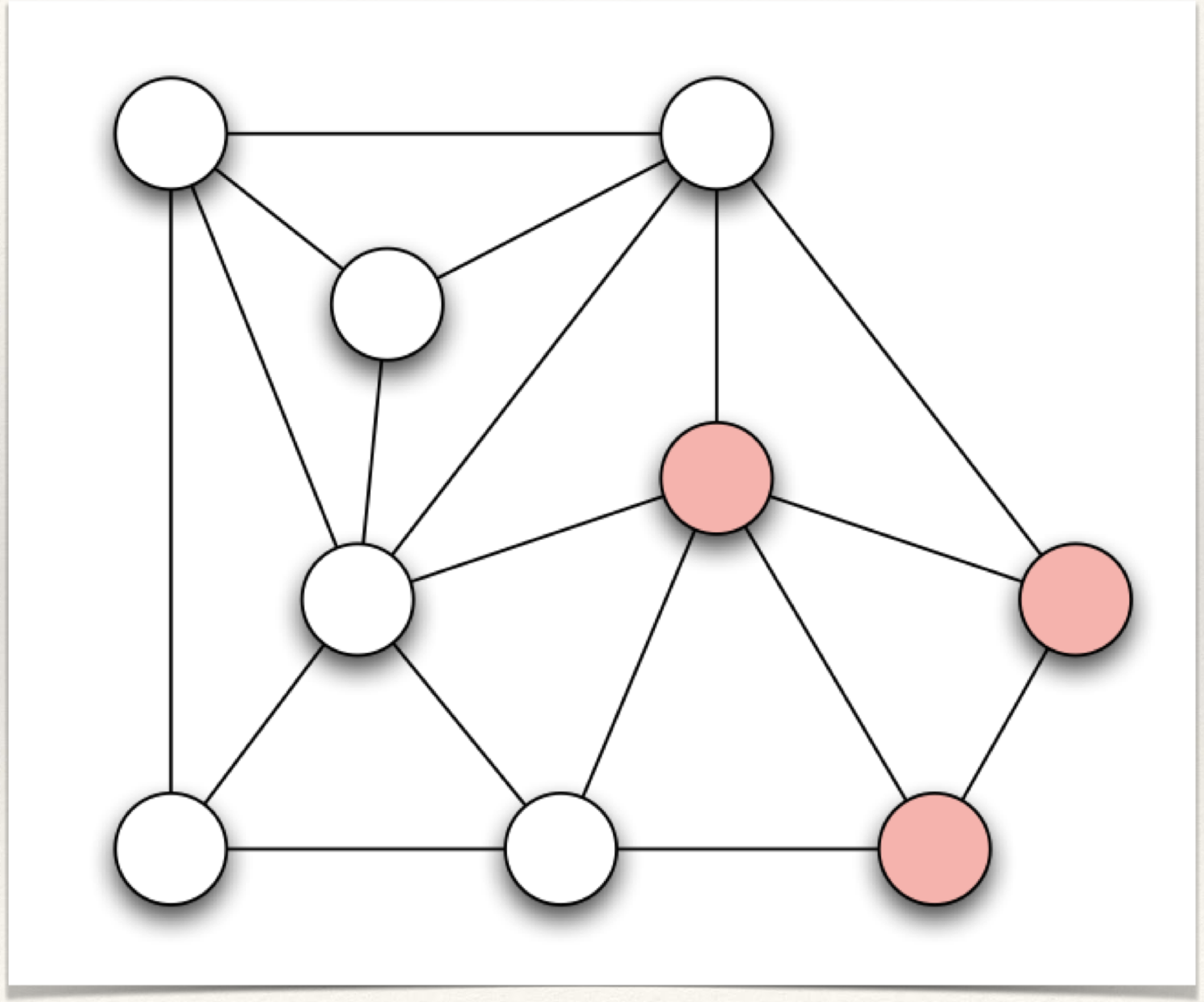
Example: middle and high school

- In real networks we find:
  - intrinsic triadic closure
  - contextual characteristics, that influence similarities and that shape the network



Measuring homophily

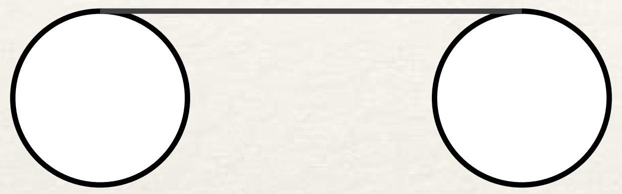
- Simple test:
  1. let's assign randomly a color to each node
  2. count number of cross-colors edges
  3. compare numbers with actual network



$p = \frac{6}{9} = \frac{2}{3}$

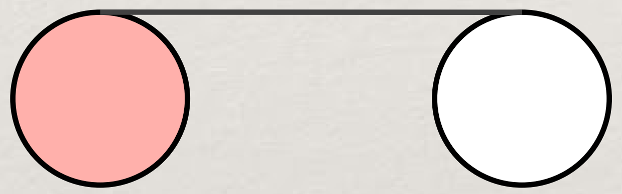
$q = \frac{3}{9} = \frac{1}{3}$

fraction of white nodes: $p = \frac{2}{3}$ fraction of pink nodes: $q = \frac{1}{3}$



$$p \cdot p = p^2 = \frac{4}{9}$$

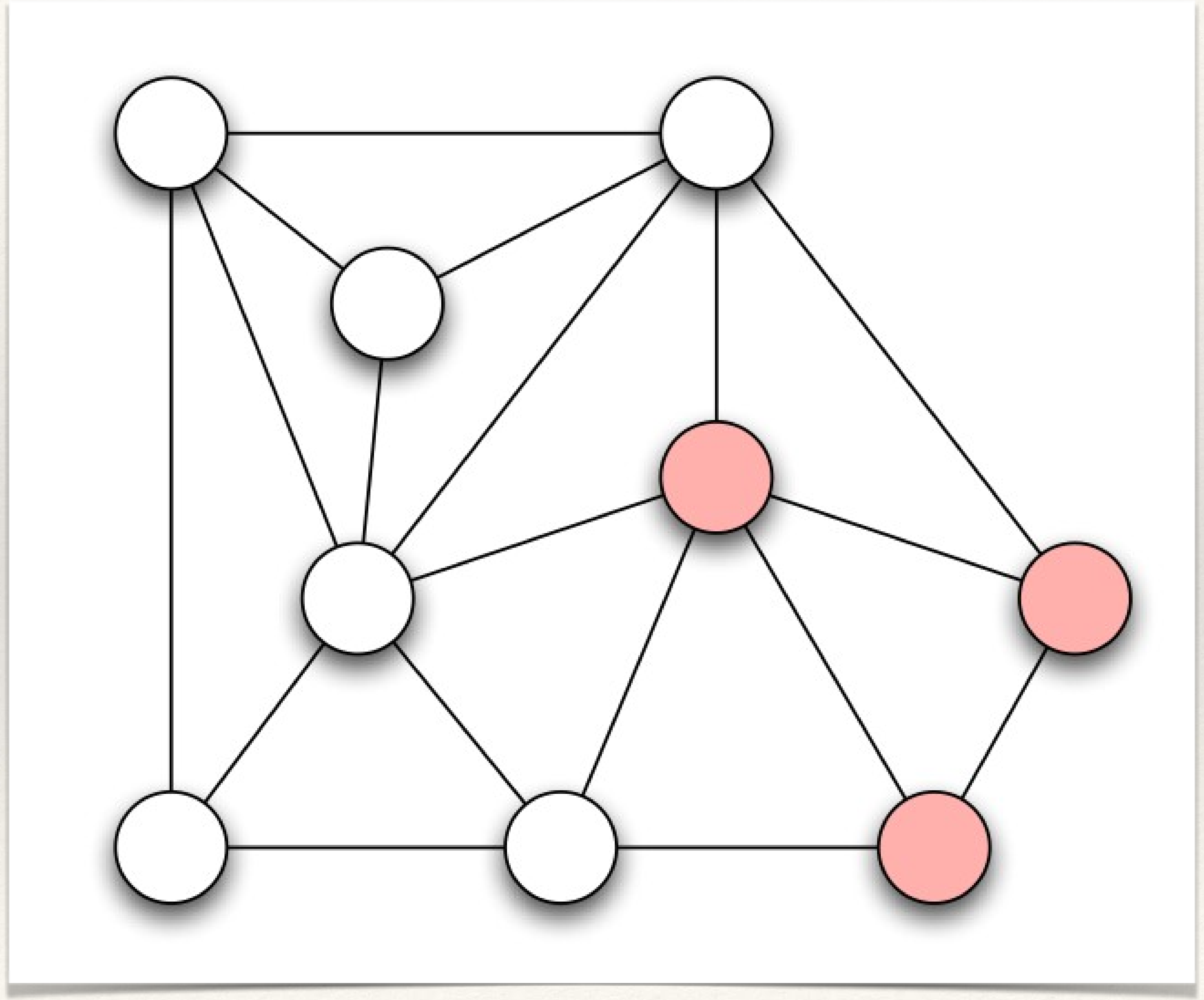
$$q \cdot q = q^2 = \frac{1}{9}$$



$$2p \cdot q = \frac{4}{9}$$

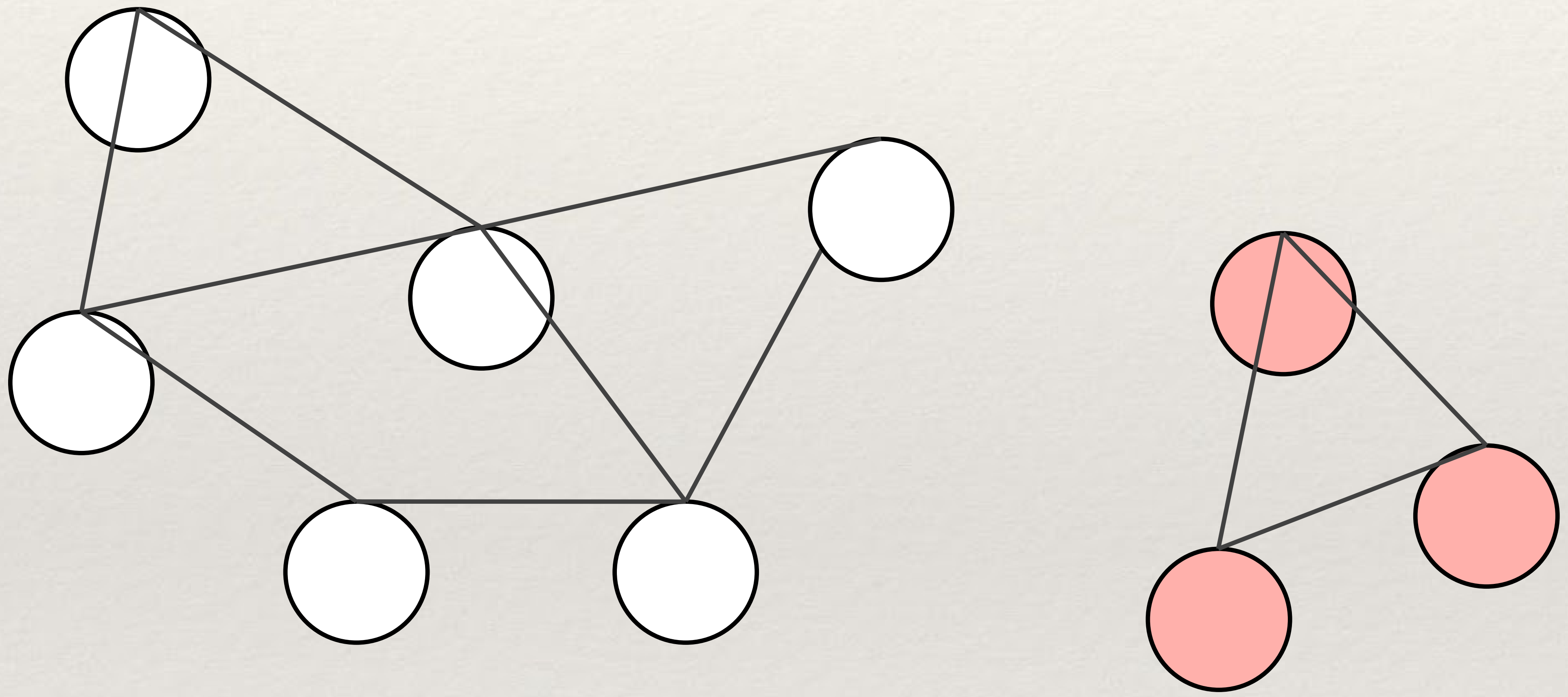
$$\frac{5}{18} < \frac{4}{9}$$

Cross-characteristic edges = number of non homophilic edges over all edges



homophily test: check if *#actual cross groups edges < 2pq*

$$\frac{5}{18} < \frac{4}{9}$$ **homophily!**



perfect **homophily:**

$$0 < \frac{4}{9}$$

More precisely

homophily test: if the fraction of cross-types edges is *significantly less* then *2pq*, then there is a *signal* of homophily

Measuring homophily: Modularity

The network is assortative if a significant fraction of the edges in the network run between nodes of the same type.

Problem: This fraction is 1 if all nodes belong to the same single type and 0 if there is a complete separation. We would like a measure that is large in non-trivial cases and small in trivial ones.

A better measure quantifies the level on non-randomness in placement of edges in the network.

This is the difference between the fraction of edges between nodes of the same type and the expected number of edges between all pairs of node of the same type.

Modularity

$$g_i = 1 \ldots N$$ → Group, class or type of node *i*

Fraction of edges between nodes of the same type → $$\sum_{edges(i,j)} \delta_{g_i g_j} = \tfrac{1}{2} \sum_{ij} a_{ij} \delta_{g_i g_j}$$

Expected number of edges between all pairs of node of the same type → $$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta_{g_i g_j}$$

Modularity

Modularity = Fraction of edges between nodes of the same type - Expected number of edges between all pairs of node of the same type

$$Q = \frac{1}{2} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta_{g_i g_j}$$

Modularity is a measure of the extent to which similar nodes connect in a ntwork. It is strictly less than 1 and takes positive values if there are more edges between nodes of the same type than we would expect by random chance.

It can also take negative values if there are fewer such edges than we would expect by chance.

Assortativity coefficient

Now we consider a scalar characteristics x_i for the node i, i.e. age, income, etc.

Then consider the pairs of values $x_{i,}$ x_j for the nodes i and j at the ends of each edge in the network and let us calculate their covariance over all edges, i.e. the joint variability of the values over the network.

$$cov(x_i, x_j) = \frac{\sum_{ij} a_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{ij} a_{ij}}$$

$$\mu = \frac{1}{2m} \sum_i k_i x_i$$

Mean value of x_i at the end of an edge

Assortativity coefficient

Then the covariance becomes:

$$cov(x_i, x_j) = \frac{1}{2}\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j$$

The covariance will be positive if, on balance, values at either end of an edge tend to be both large or both small, and negative if they tend to vary in opposite directions.
In other words, the covariance will be positive when we have assortative mixing and negative for disassortative mixing.

Assortativity coefficient

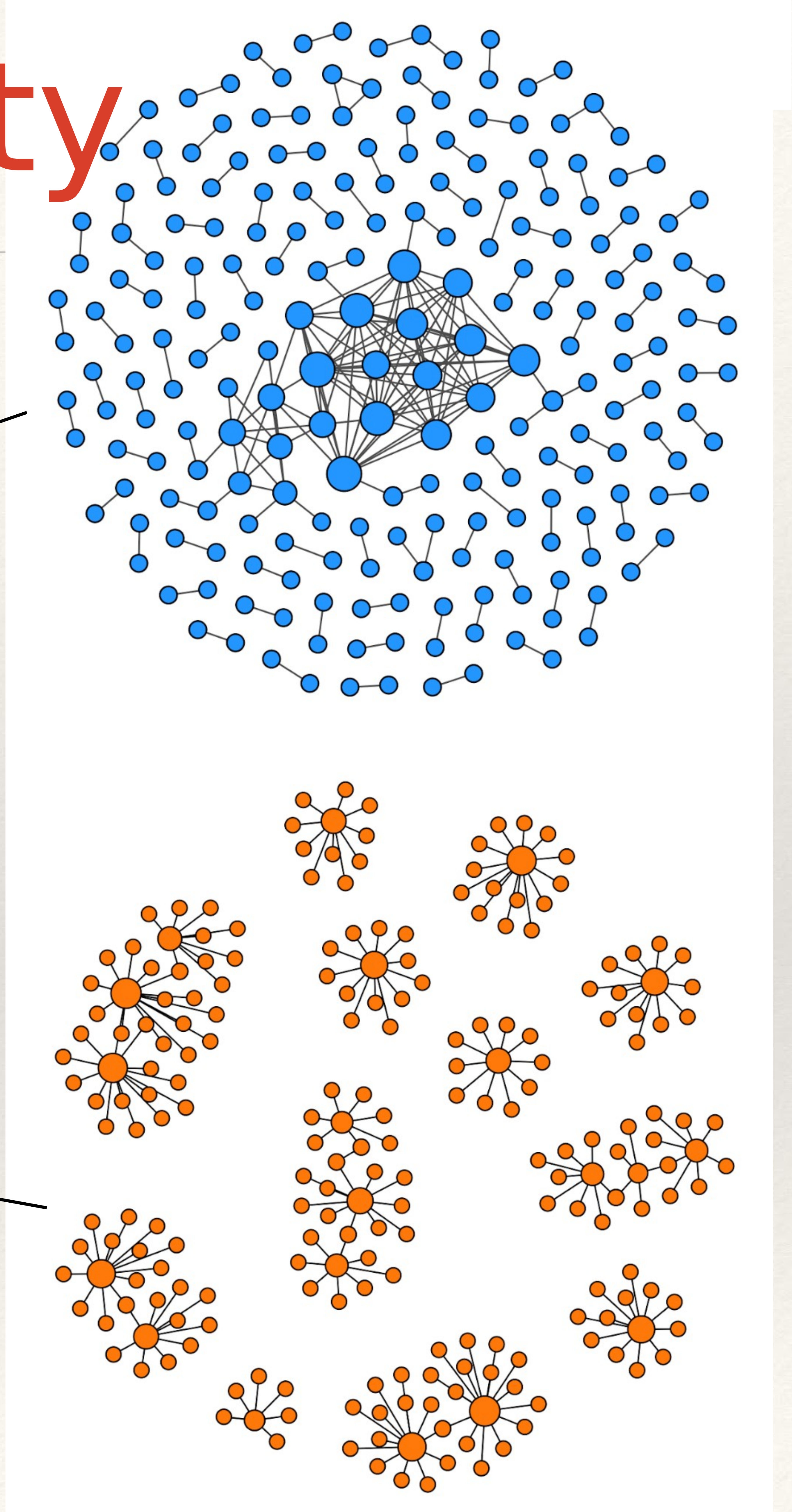
If we want that the covariance takes value 1 in a perfect assortative mixing, we can normalize by the variance to get the assortative coefficient r

$$r = \frac{\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j}{\sum_{ij}\left(a_{ij}\delta_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j}$$

Note: this is an example of correlation coefficient (Pearson)

The correlation coefficient varies in value between a maximum of 1 for a perfectly assortative network and a minimum of -1 for a perfectly disassortative one.

Degree assortativity



- A.k.a. **degree correlation**
- Assortative networks have a **core-periphery** structure with hubs in the core
  - Ex: social networks
- Disassortative networks have **hub-and-spoke** (or **star**) structure
  - Ex: Web, Internet, food webs, bio networks

Assortativity coefficient by degree

If we want to consider similar nodes with the same degree then we can substitute x_i with k_i and obtain the assortativity coefficient by degree.

$$r = \frac{\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}{\sum_{ij}\left(a_{ij}\delta_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}$$

$$cov(k_i, k_j) = \frac{1}{2}\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j$$

Assortativity coefficient by degree

If we want to consider similar nodes with the same degree then we can substitute x_i with k_i and obtain the assortativity coefficient by degree.

$$r = \frac{\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}{\sum_{ij}\left(a_{ij}\delta_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}$$

$$cov(k_i, k_j) = \frac{1}{2}\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j$$

Measuring assortativity by neighbors

Another way to compute the degree assortativity is by measuring the **degree correlation function**, i.e., the correlation between the degree and the average degree of the neighbors of nodes with that degree.

Let's calculate the average of *i*'s neighbors' degrees, first:

$$k_{nn}(i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j$$

With a little abuse of notation, let's define the degree correlation function as the average of all degree-k nodes.

$$k_{nn}(k) = \sum_{k'} k' P(k'|k)$$

to check visually network assortativity, we plot:

$$(k, k_{nn}(k))$$

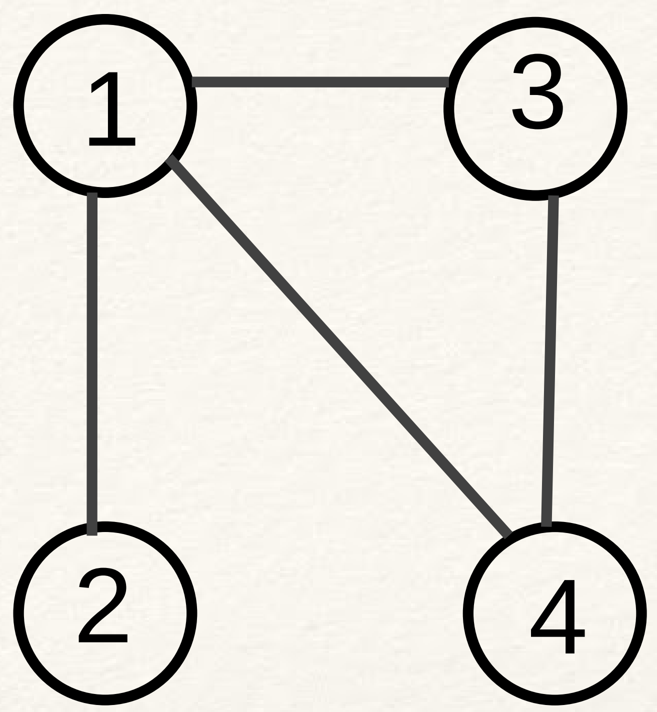
Assortativity in a random network

It is possible to prove that, in a neutral network (i.e., where there is no correlation between a node's degree and its neighbors' average degree), plotting result in a horizontal line, more precisely:

$$P(k'|k) = k' \frac{p(k')}{\langle k \rangle} \longrightarrow k_{nn}(k) = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

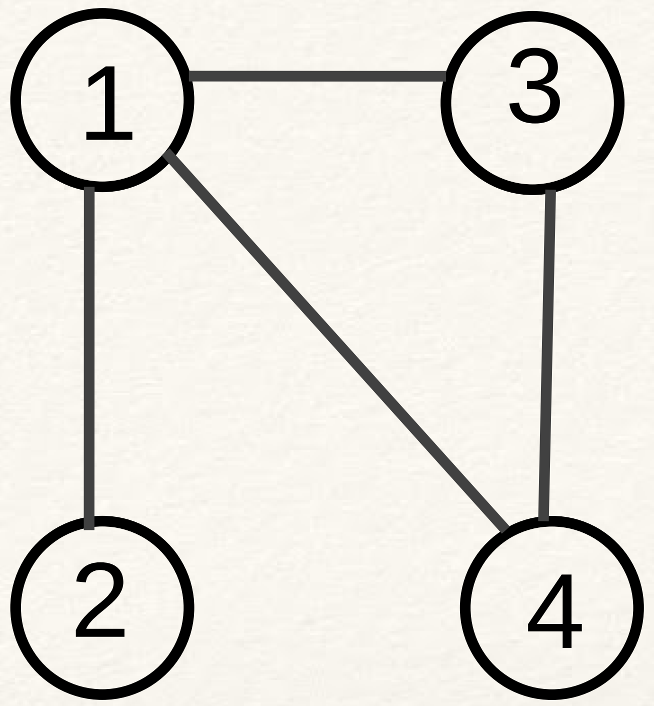
See chapter 7 of Barabasi's book for more mathematical insights: http://networksciencebook.com/chapter/7#measuring-degree

$$k_{nn}(i) = \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}k_j$$ average of i's neighbors' degrees



Let’s compute it!

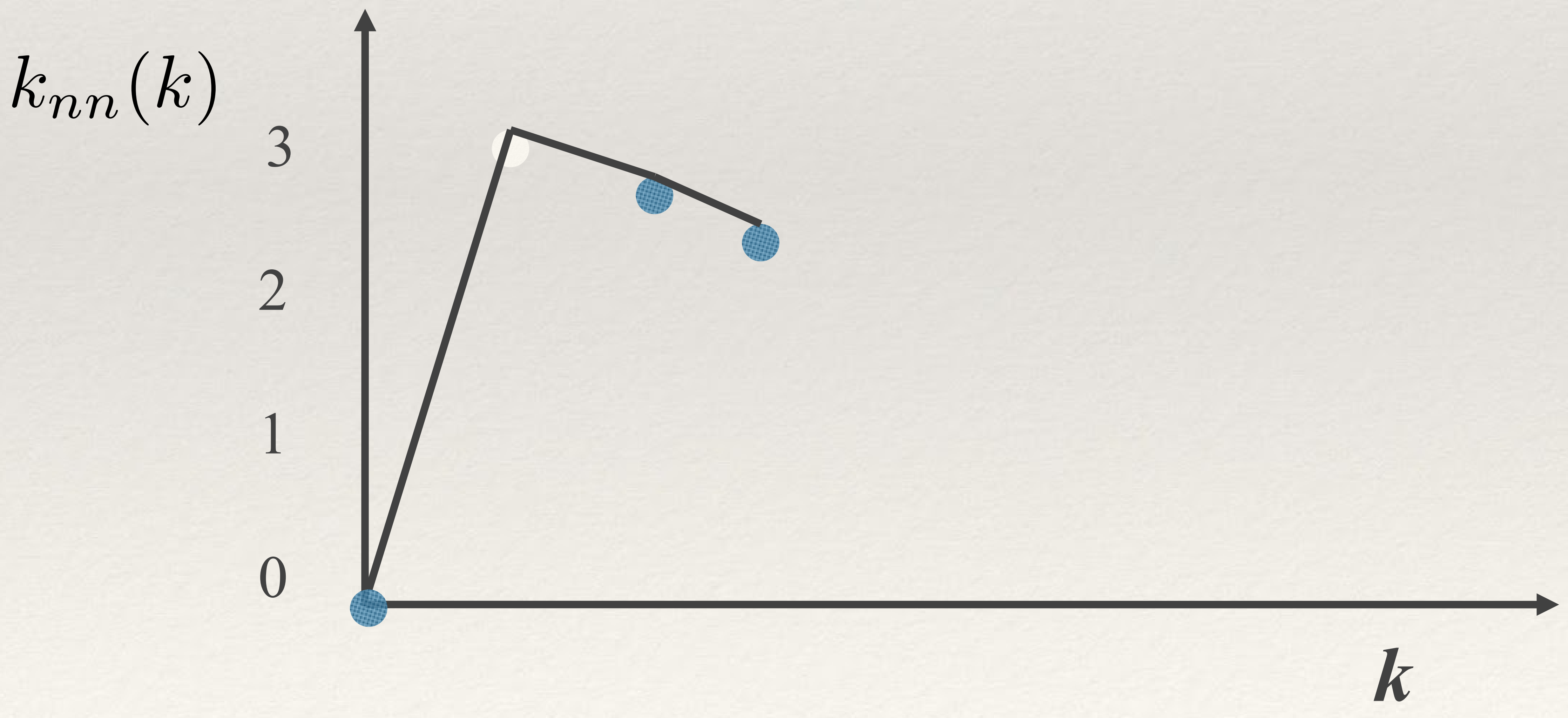
correlation degree function: $k_{nn}(k)$



is the average of all the neighbors' degree of k-degree nodes

$$k = (0, 1, 2, 3)$$

$$k_{nn}(k) = (0, 3, 5/2, 5/3)$$

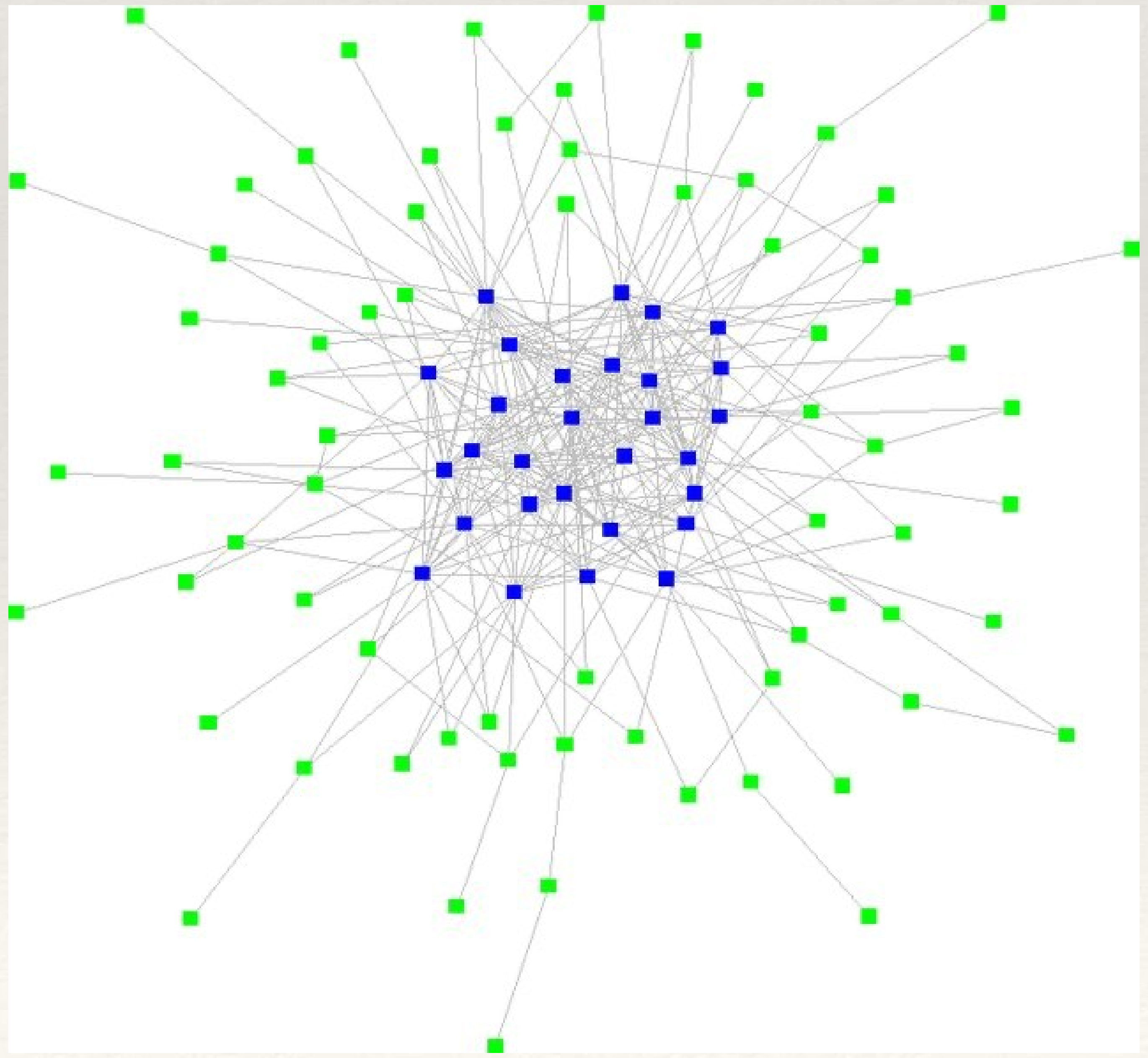


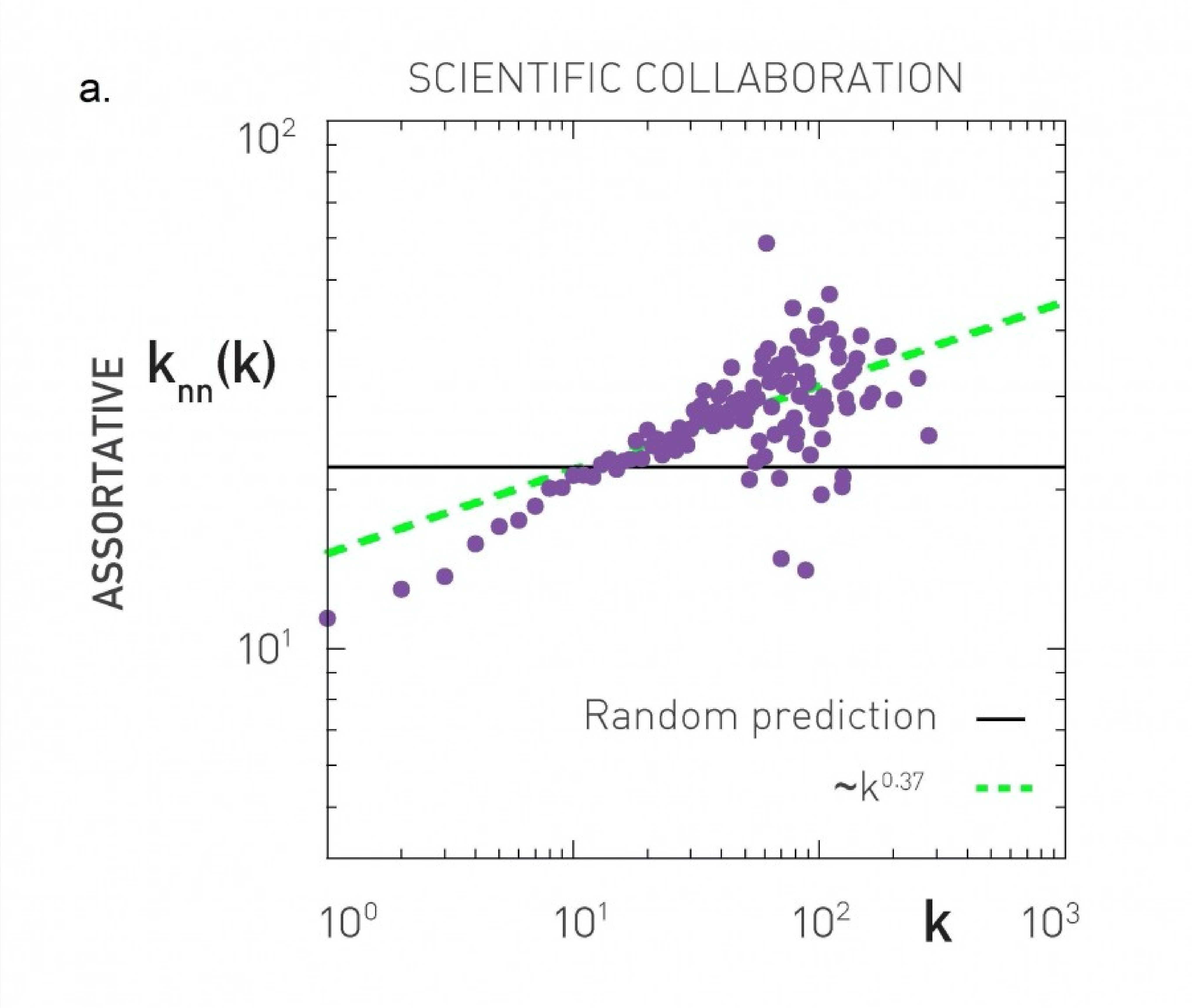
Positive assortativity

Scientific collaboration network

The increasing $k_{nn}(k)$ with k indicates that the network is assortative.

ex. core-periphery

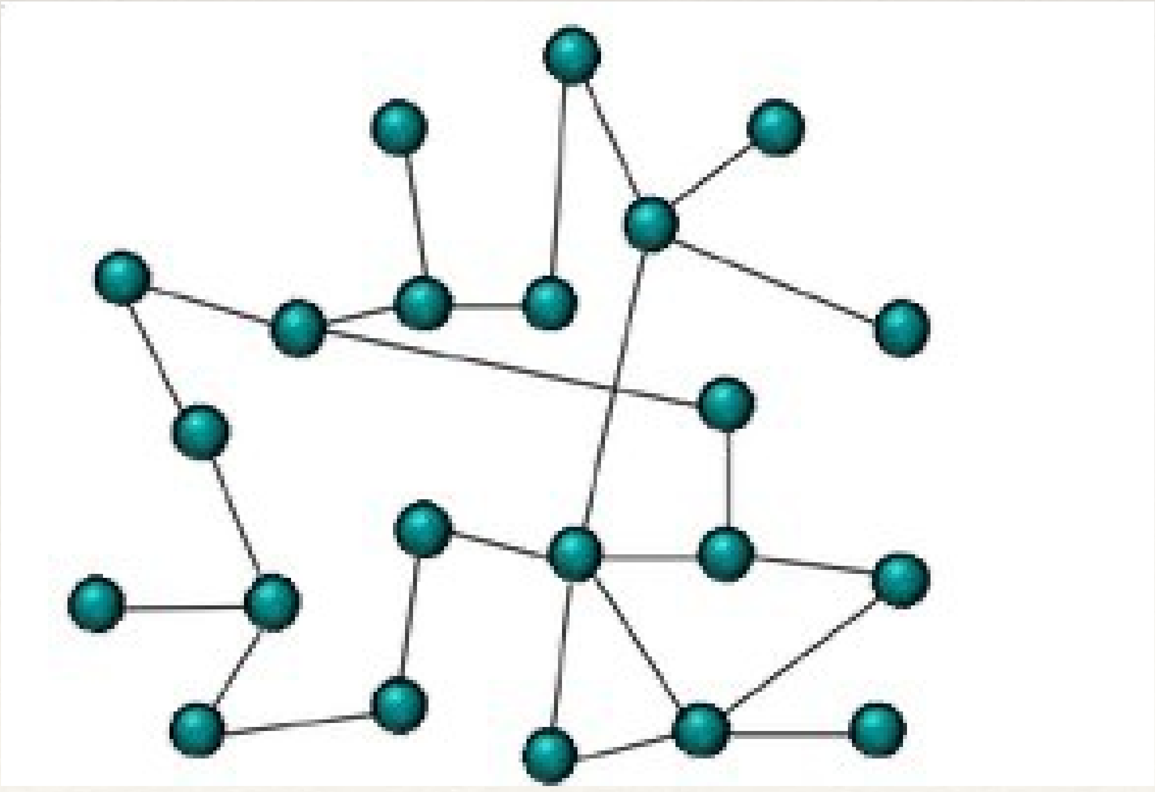


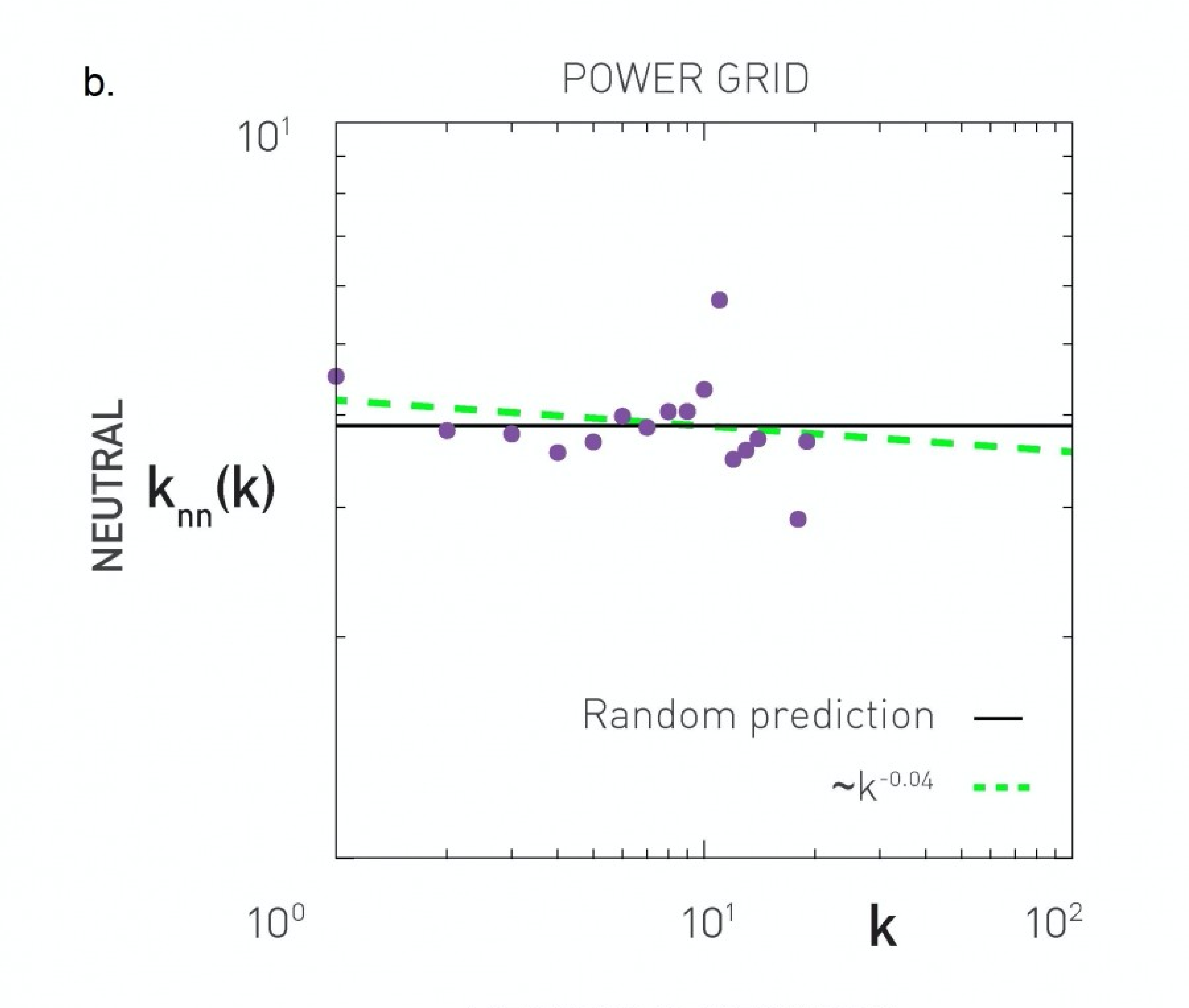


Neutral assortativity

Power grid

The horizontal $k_{nn}(k)$ indicates the lack of degree correlations, in line with our expectations for neutral networks.

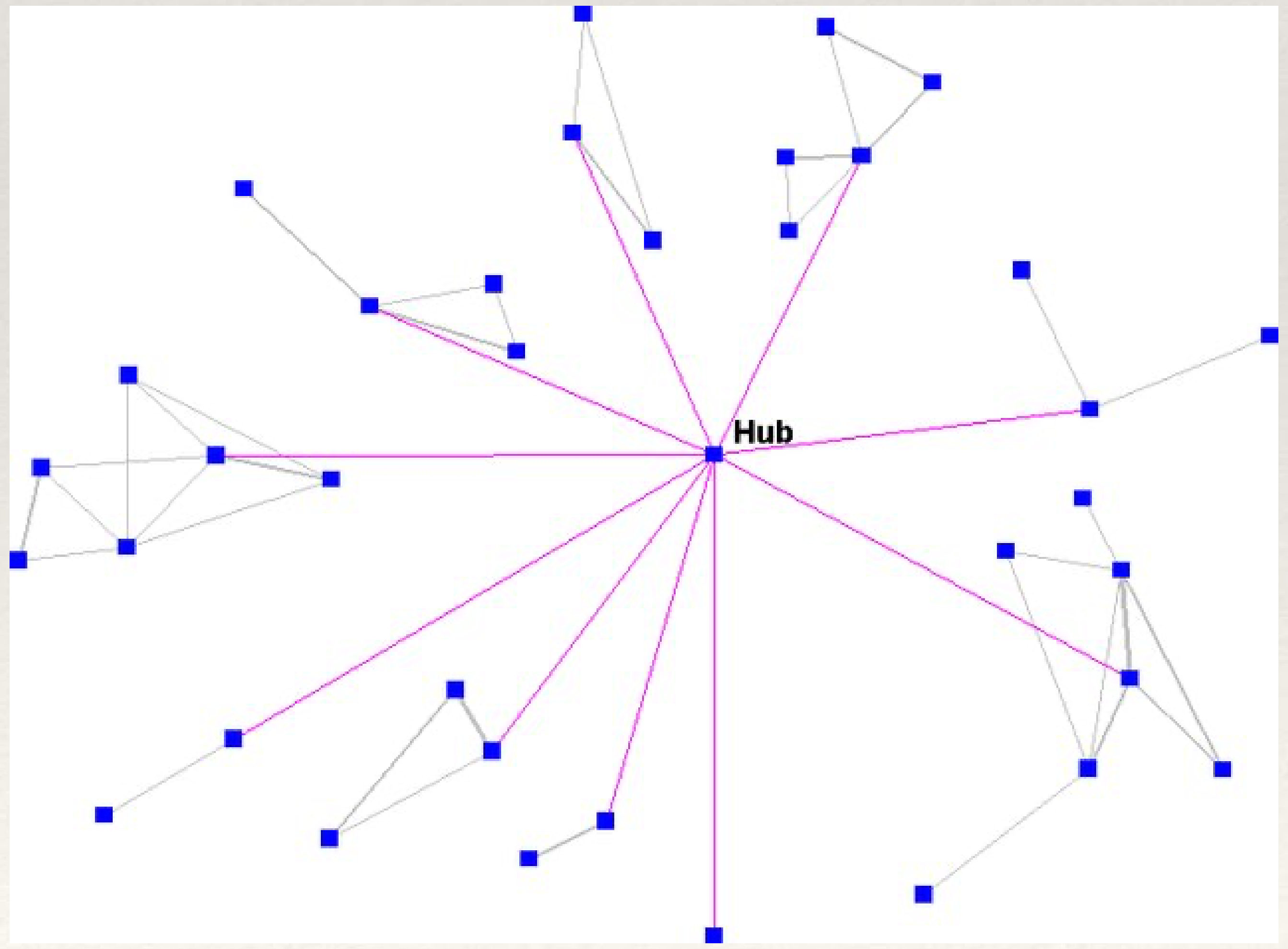


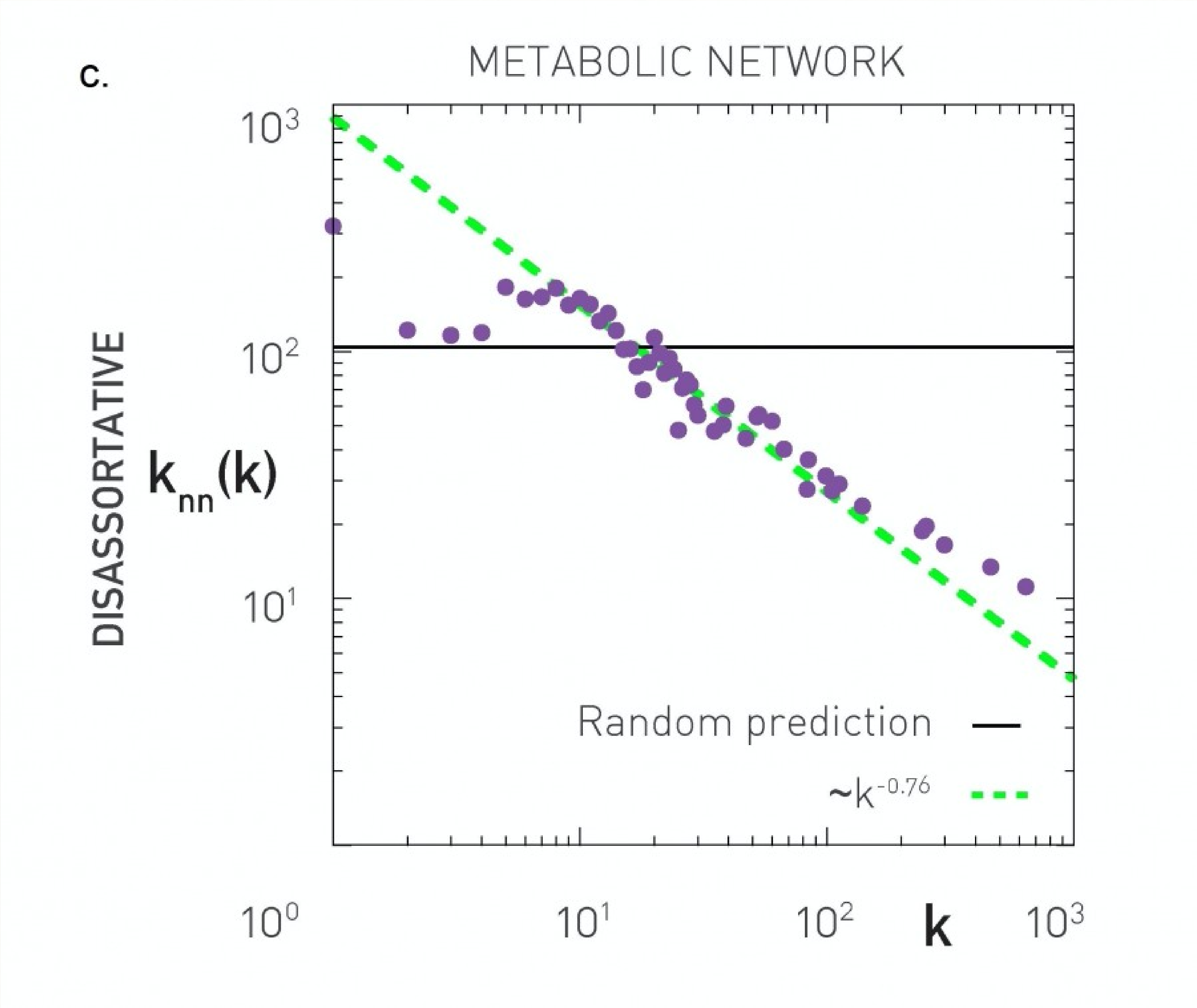


Negative assortativity

Metabolic Network

The decreasing $k_{nn}(k)$ documents the network's disassortative nature.



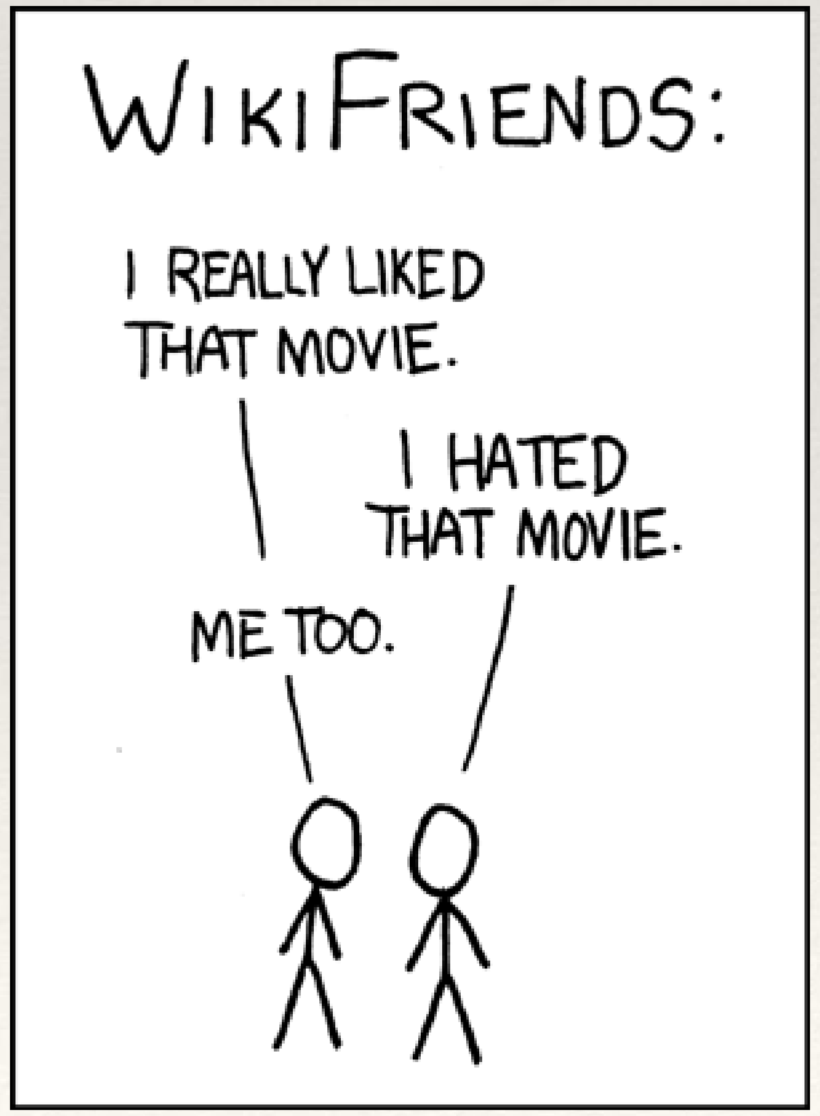


Mechanisms Underlying Homophily: Selection and Social Influence

Underlying mechanisms of homophily

- Two possible mechanisms by which homophily (also: assortativity) emerges naturally:

  1. **Selection**: similar nodes become connected
  2. (Social) **influence**: connected nodes become more similar



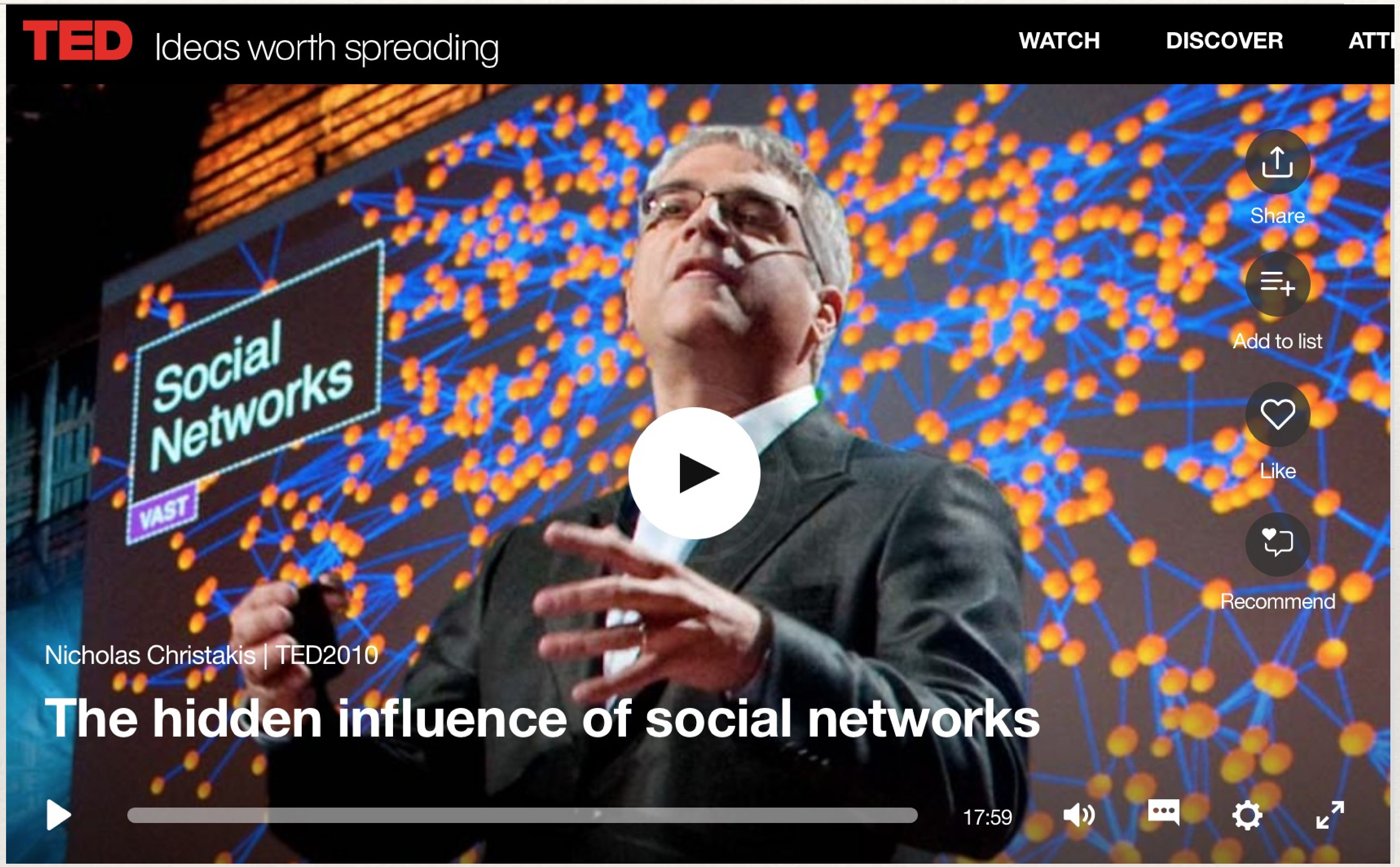
- It can also be a bad thing. For example "**echo chambers**" and "**groupthink**" are situations where your friends are like you, diversity is killed, and you are only exposed to opinions that reinforce your pre-existing beliefs…

The interplay of selection and social influence

- longitudinal methodology:
  - observe a network for a long period of time
  - observe both factors in action
  - how do we quantify the impact?
- ex1: drug dependency:
  - social circle?
  - peer-pressure

ex2: obesity "contagion"

- dataset: 12,000 people
- obesity status
- social network structure
- obese vs non obese: there is a tendency toward clustering
- homophily test: passed
- why?
    - selection?
    - homophily that correlates with something else?
    - social influence? —> contagion!



https://www.ted.com/talks/nicholas_christakis_the_hidden_influence_of_social_networks

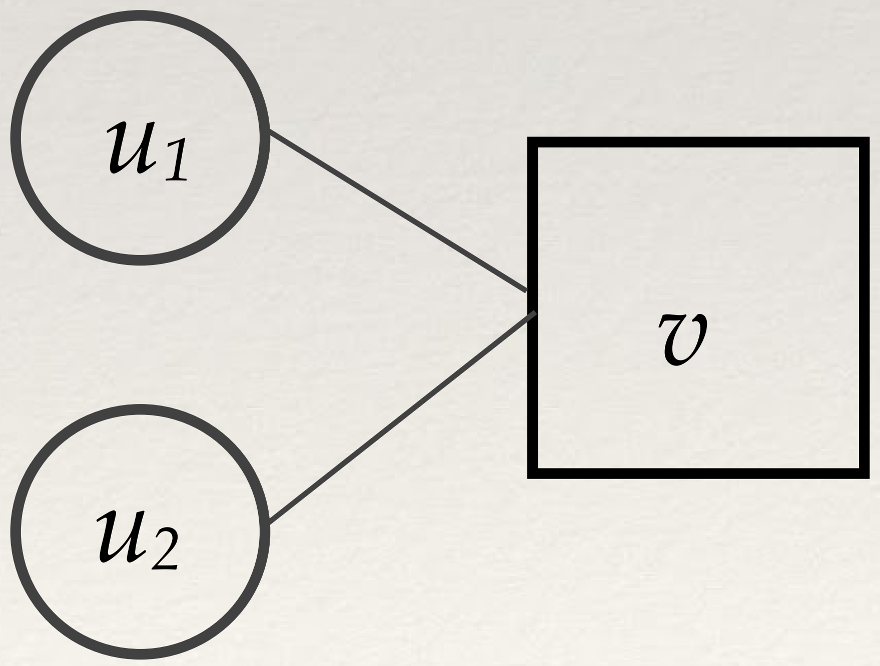
Affiliation

Affiliation

- Can we represent surrounding context by means of networks, too?
- An affiliation network is a network in which actors are connected via their membership in groups of some kind.
- FOCI (or groups): focal points
- Focus —> activity
- We can use **bipartite graphs**

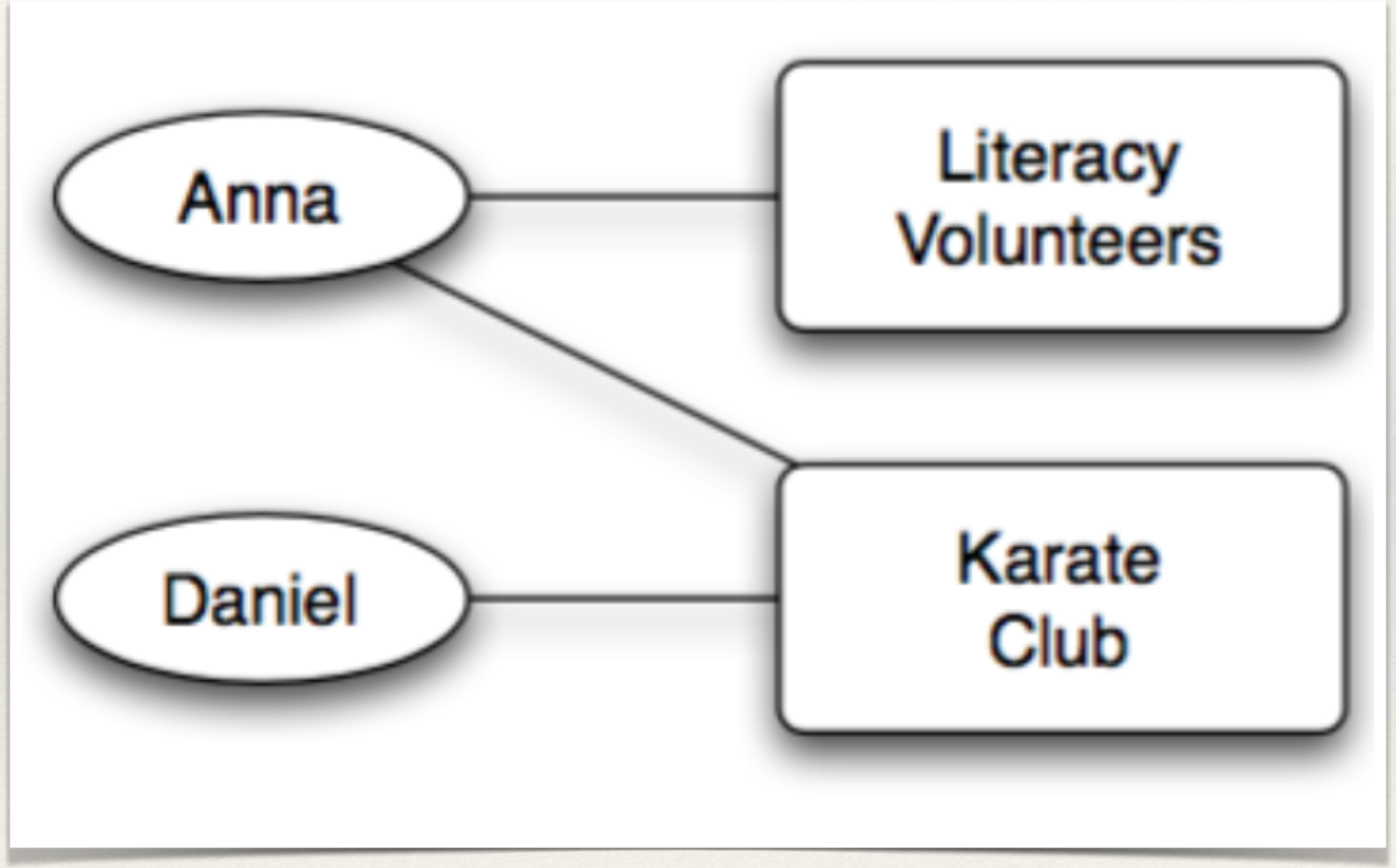
$$G = (\overset{\text{persons}}{U}, \overset{\text{foci}}{V}, E)$$

$$\forall (u, v) \in E : u \in U \wedge v \in V$$



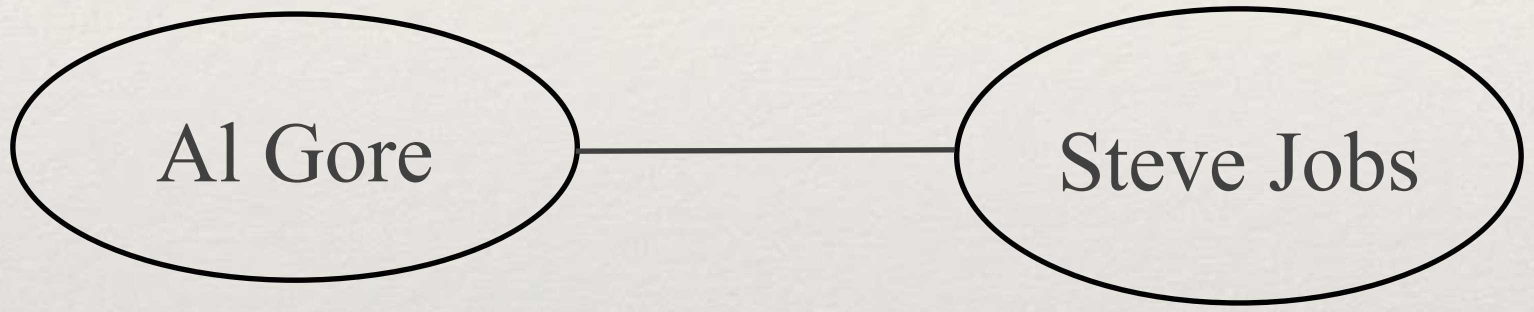
Affiliation networks

- Can I complement my affiliation network with a social network?

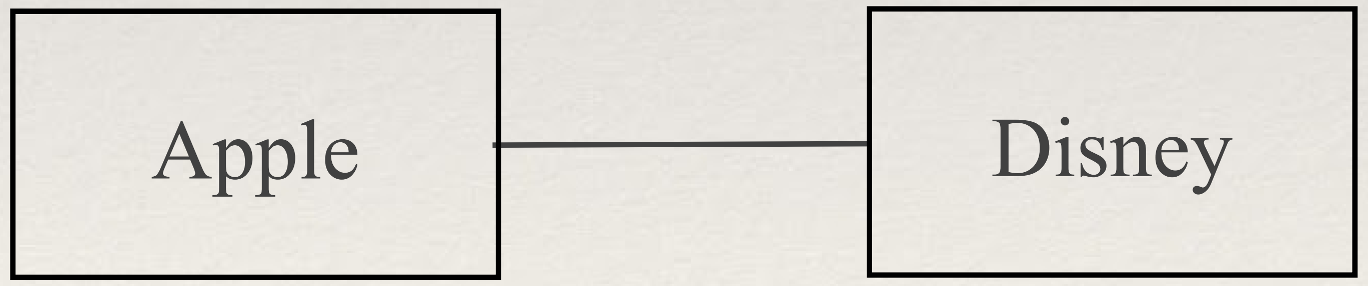


Using projections

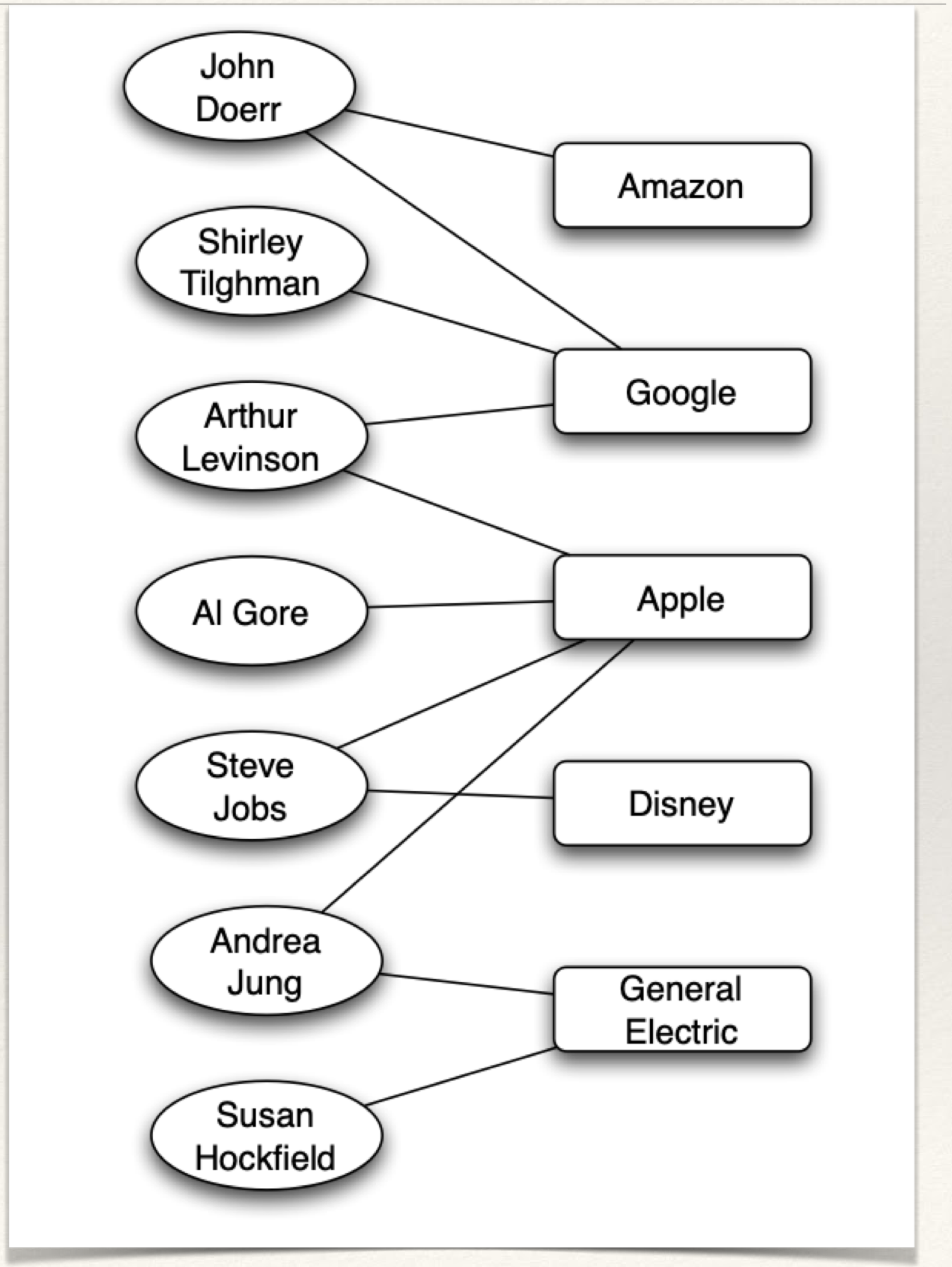
- ex. memberships of people on corporate boards of directors
- projection1: Network of interaction among board members



- projection2: Network of interaction among companies

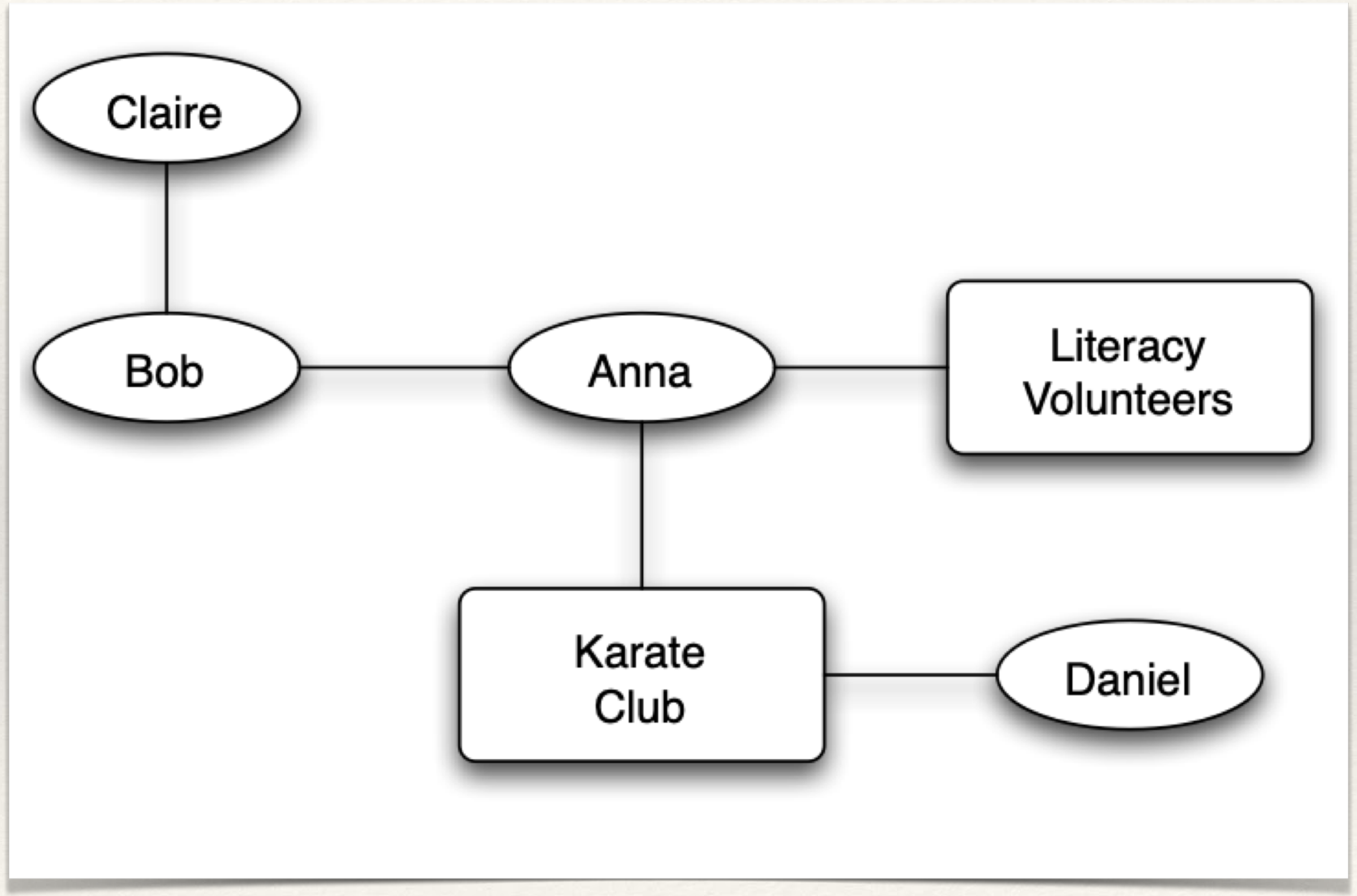


- Can you do the same with the movie networks?



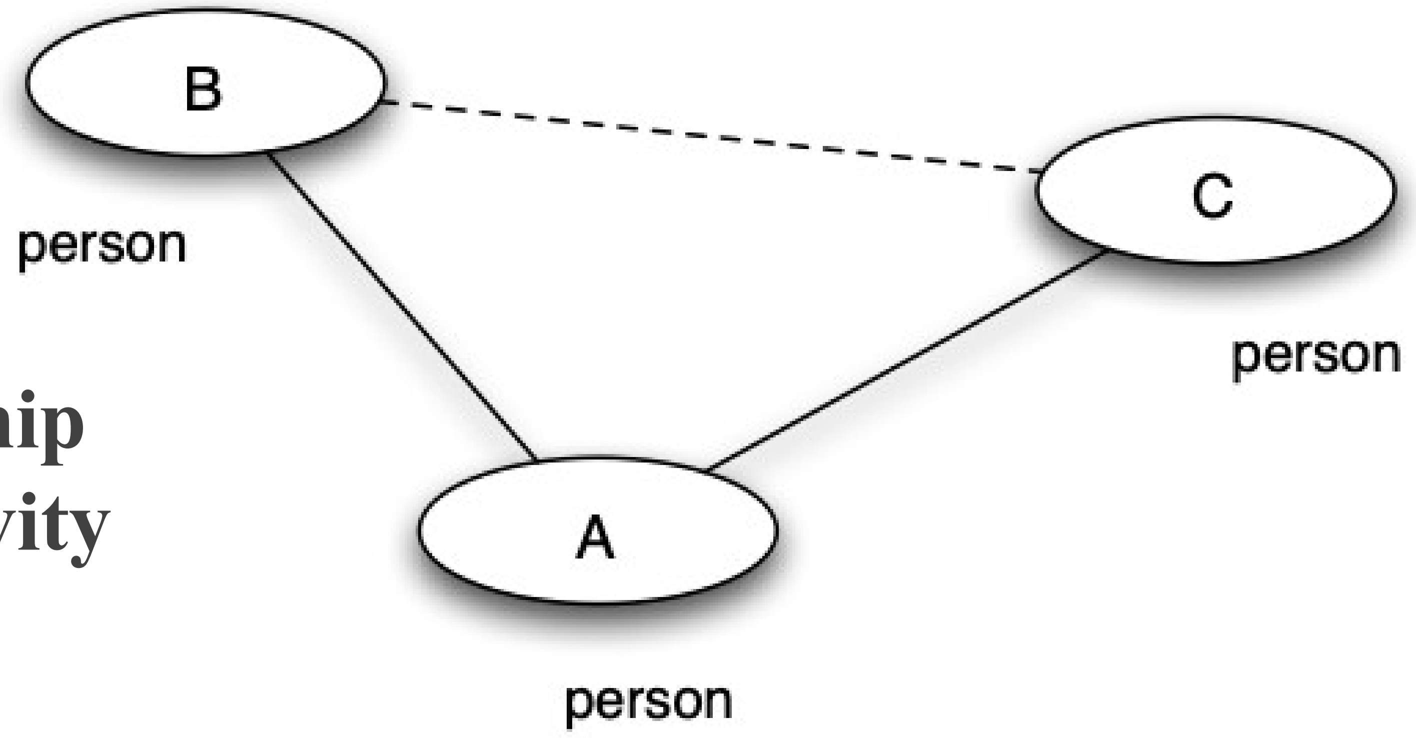
Co-evolution of social and affiliation nets

- Both social and affiliation networks **change over time**
- We want to merge the information from both networks -> **social-affiliation networks**
- Can we predict **new links** formation?

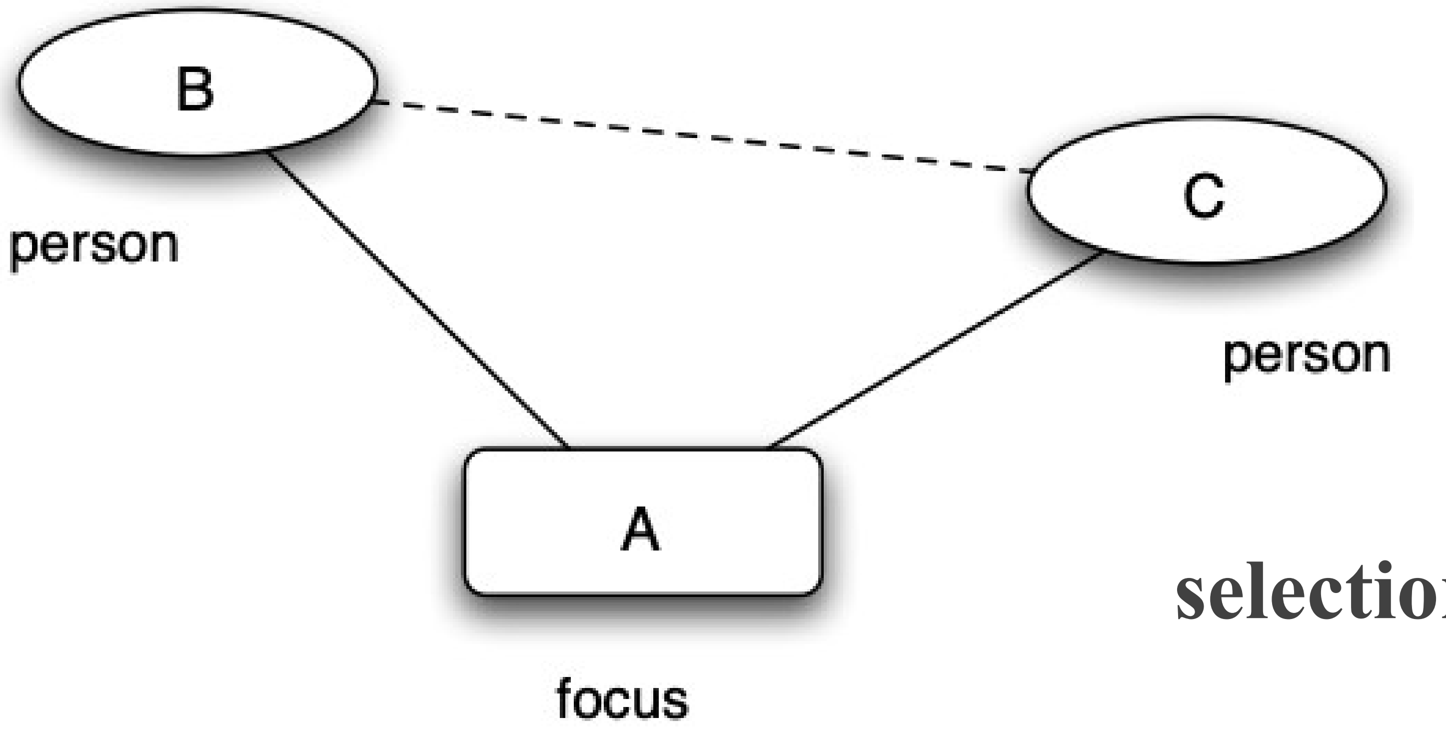


Closures

friendship transitivity

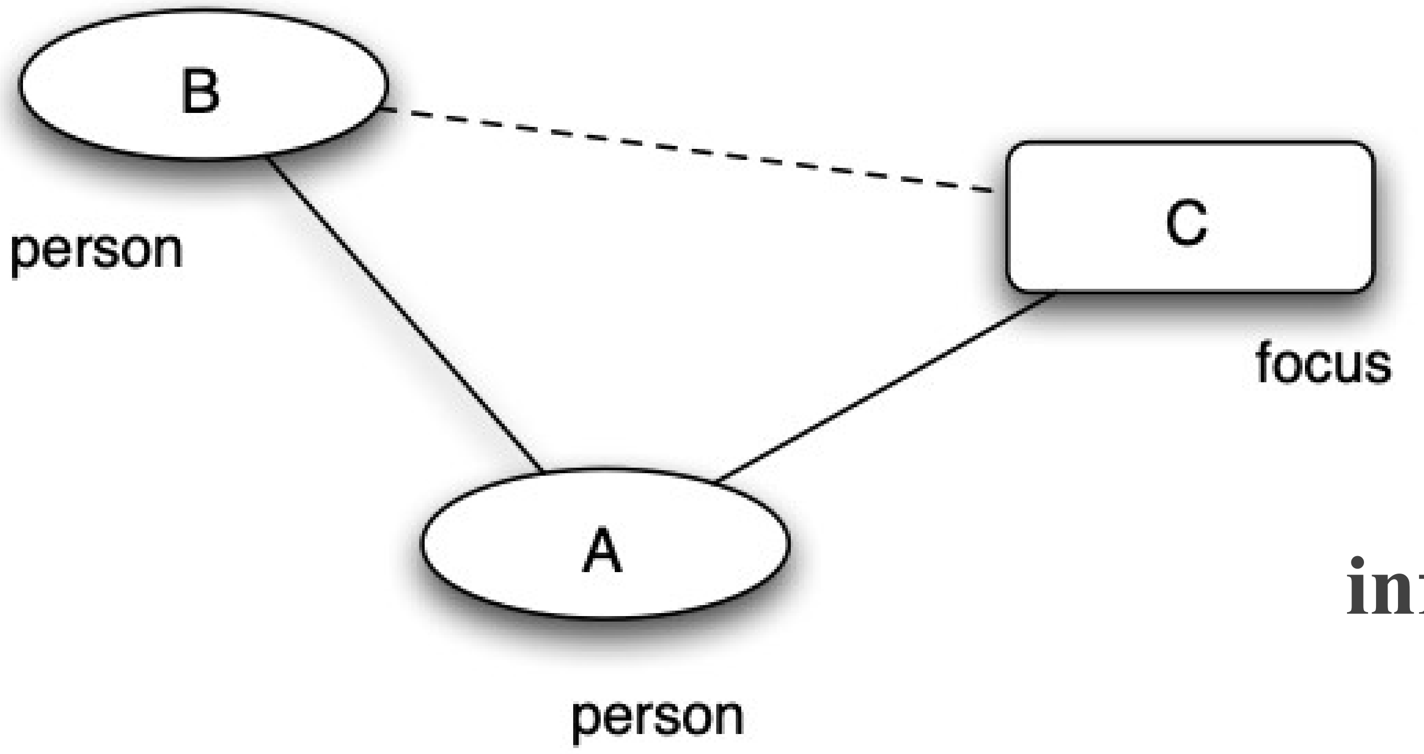


(a) *Triadic closure*



selection

(b) *Focal closure*



influence

(c) *Membership closure*

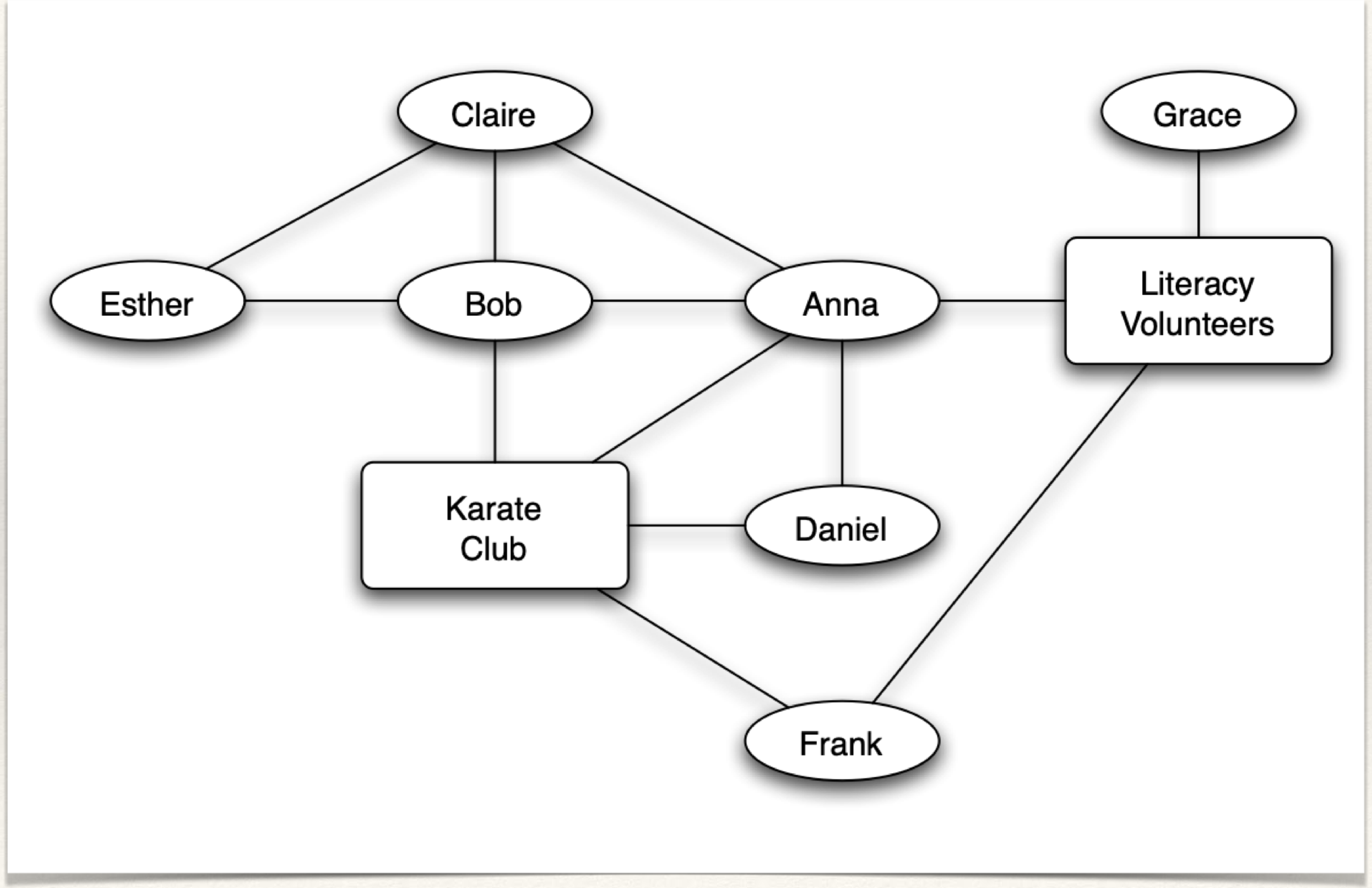
Tracking Link Formation in Online Data

Data from online platforms

- caveat: we can extrapolate from digital interactions information to other interactions that are not computer mediated

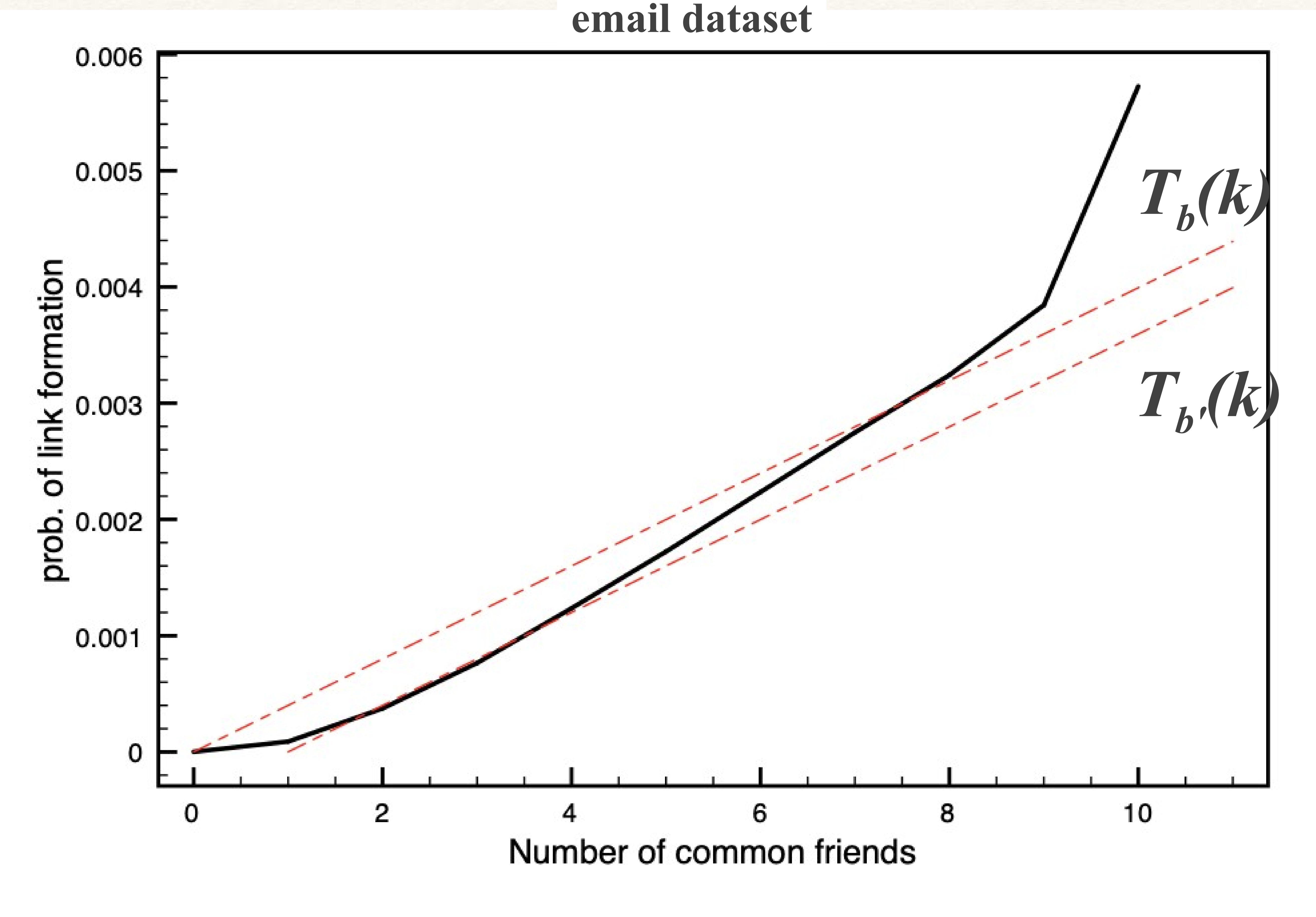
Triadic closure: numerical questions

- how much is likely for a link to form if it has the effect of **closing a triangle**?
- how much more likely is a link to form if it closes **many** triangles?



- Two snapshots of the network at time ***t*** and ***t'***
- For each ***k***, identify all pairs of nodes who have exactly ***k*** friends in common at time ***t***, but who are not directly connected by an edge
- ***T(k)*** = the fraction of these pairs that have formed an edge by ***t'***
  - empirical probability that a link will form between two people with ***k*** friends in common

email dataset



- 2 people having a friend in common, have an independent small probability p to connect each other, $(1-p)^k$ probability the link fails to form
- Two baselines:
  - $T_b(k) = 1-(1-p)^k$
  - $T_{b'}(k) = 1-(1-p)^{k-1}$

- Common friends assumptions is **too simple**!
- There is "something" more than this

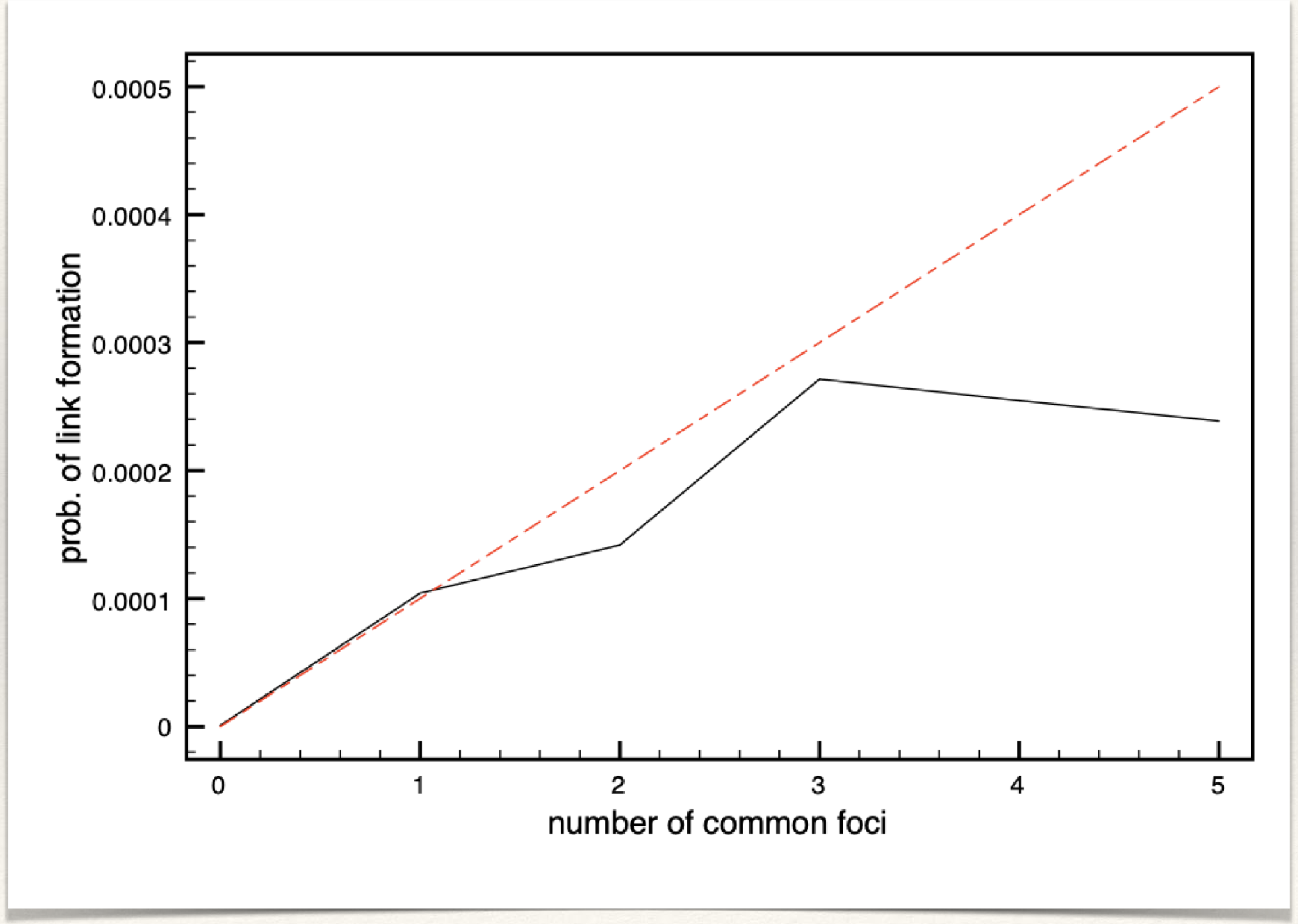
Focal closure: looking for evidences

focal closure: what is the probability that two people form a link as a function of the number of foci they are jointly affiliated with?

- email dataset + information on class schedules for each student

Empirical evidence: from to

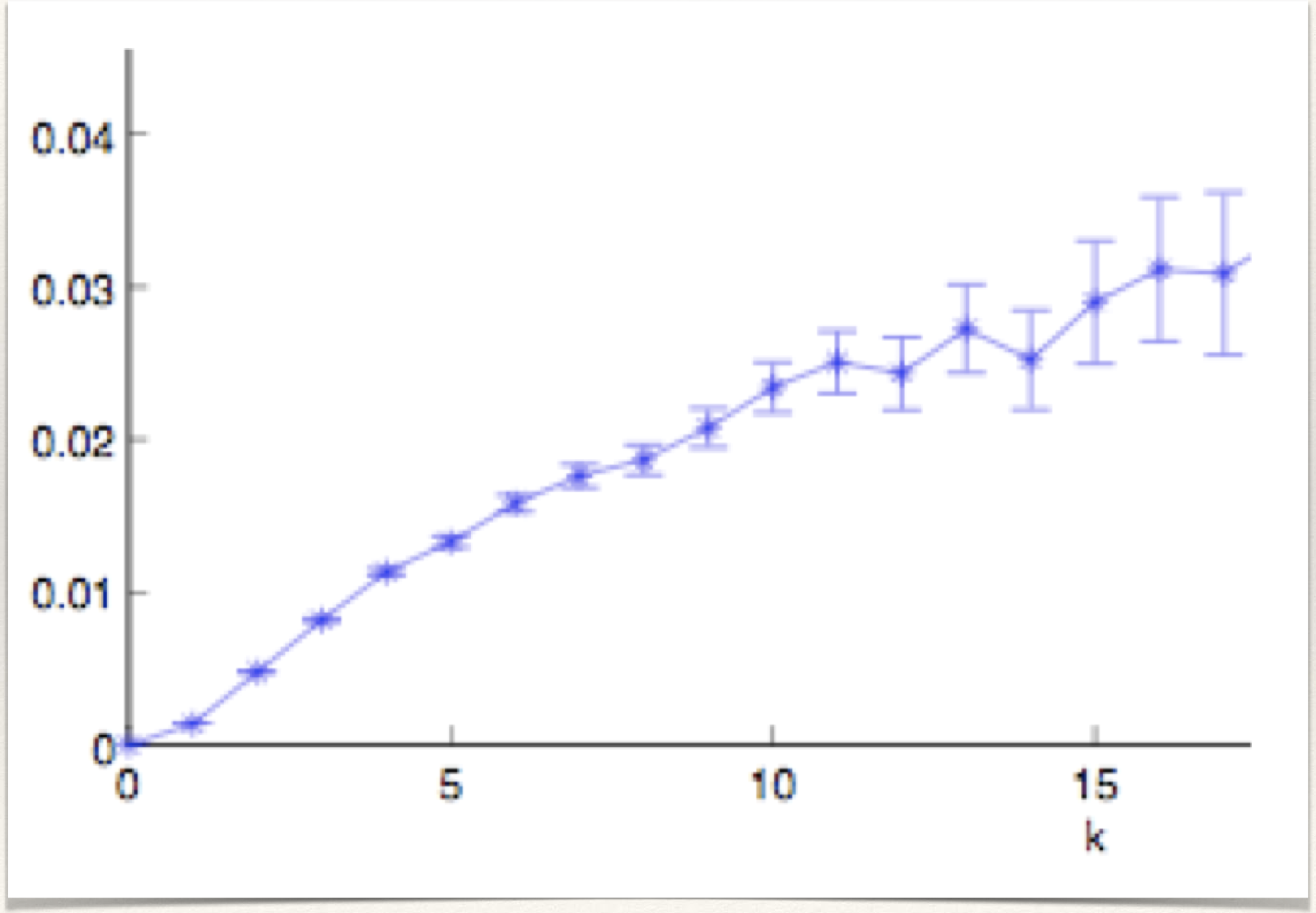
What for when ?



Membership closure: wikipedia case study

membership closure: what is the probability that a person becomes involved with a particular focus as a function of the number of friends who are already involved in it?

- Probability of editing Wikipedia articles as a function of the number of friends that already have done so



Discussion

- Diminishing effect over k
- Multiple effects that operate simultaneously on the formation of a link
- Homophily suggests that friends tend to have similar characteristics
    - triadic closure
    - focal closure
    - membership closure
- -> Correlation is not causation!
- -> forget linear explanations of cause-effects!