



# Analisi e Visualizzazione delle Reti Complesse

## NS04 - Homophily

Prof. Rossano Schifanella





# Homophily



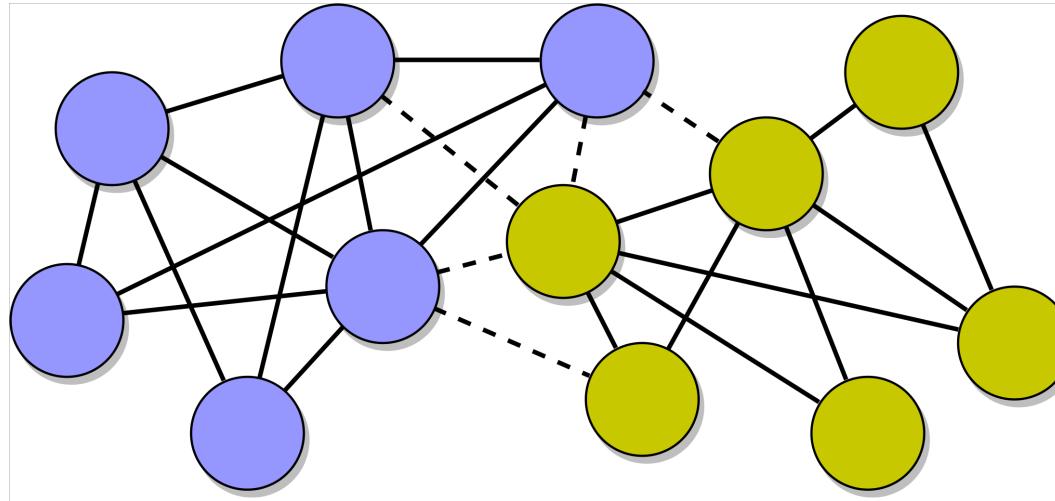
## Surrounding contexts

- 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 affect the evolution of the system

# Homophily

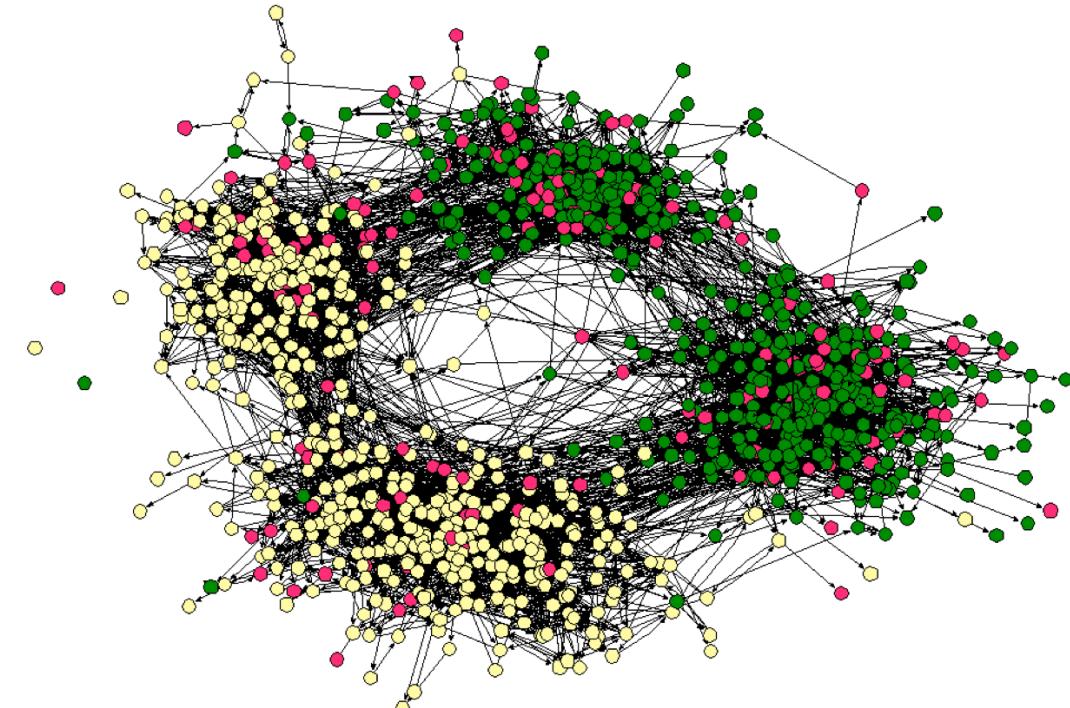
- The principle that **we tend to be similar to our friends**
- This makes your friends not statistically significant as a random sample of the population
- Similarities
  - immutable characteristics
    - e.g., race and ethnic dimensions
  - mutable characteristics
    - e.g., occupation, level of affluence, interests, beliefs, opinions, activities, place where they live
- We all have exceptions however, overall, this observation empirically holds

## Birds of a feather flock together



## Example: middle and high school

- In real social networks, we find:
  - intrinsic triadic closure
  - contextual characteristics that influence similarities, and that shape the network
- In this example, the circle's color represents students of different races
- **Important:** the formation of a link is likely to be the effect of a **combination of effects**
  - intrinsic structure + context



## Measuring homophily

Simple test:

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

$$\begin{array}{ll} \text{white circle} & p = \frac{6}{9} = \frac{2}{3} \\ \text{red circle} & q = \frac{3}{9} = \frac{1}{3} \end{array}$$



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

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



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

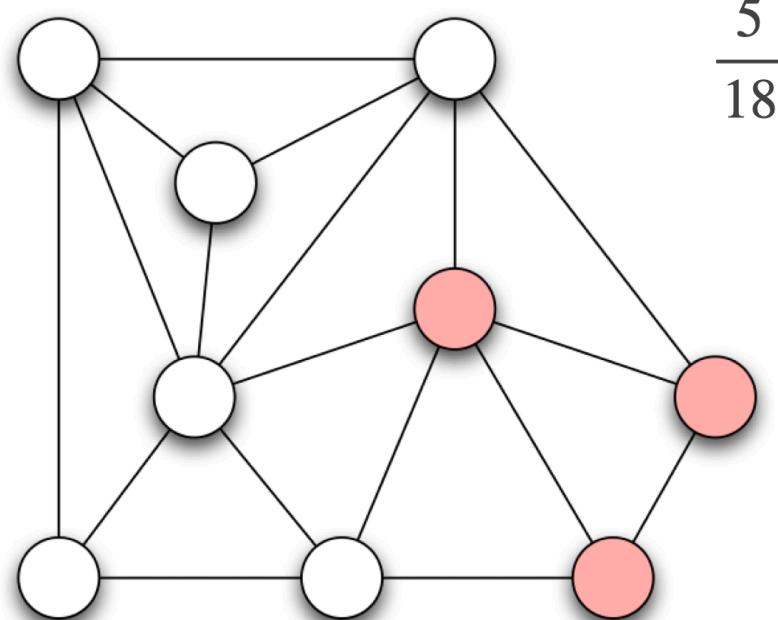


$$q \cdot q = q^2 = \frac{1}{9}$$



$$2 \cdot p \cdot q = 2 \frac{2}{3} \frac{1}{3} = \frac{4}{9}$$

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

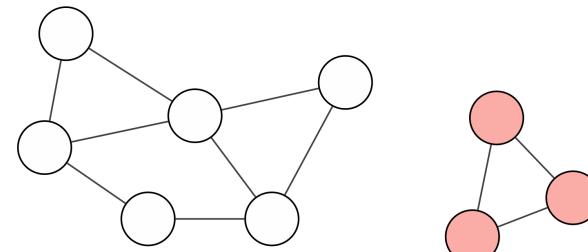


$$\frac{5}{18}$$

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



$$\frac{5}{18} < \frac{4}{9} \Rightarrow \text{homophily}$$

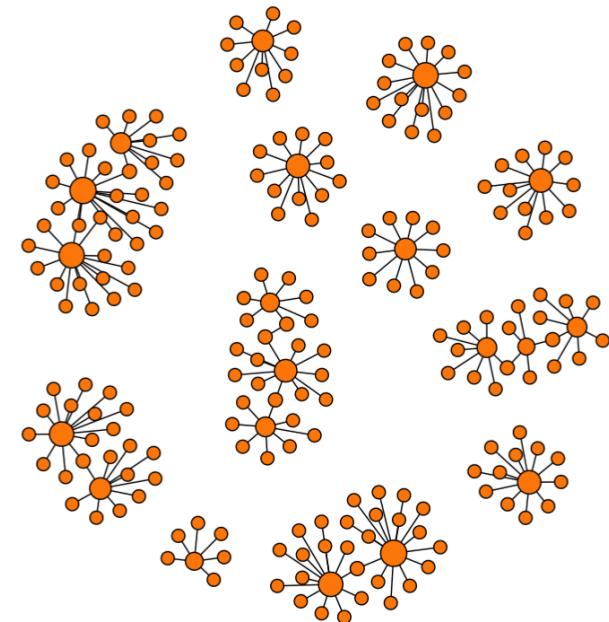
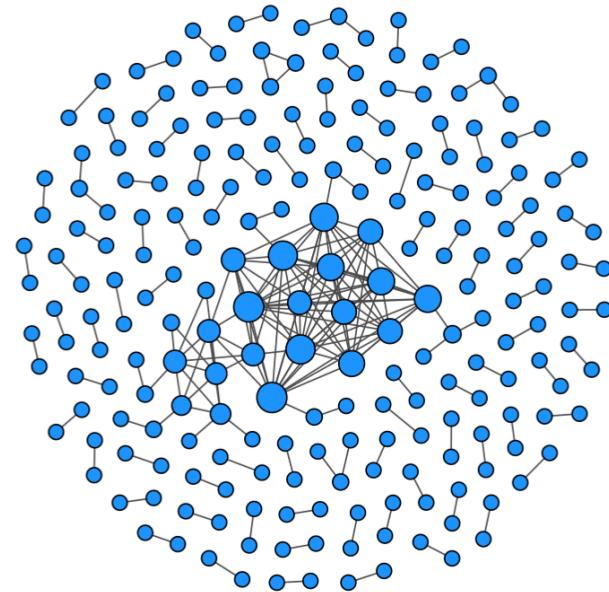


$$0 < \frac{4}{9} \Rightarrow \text{perfect homophily}$$

- Standard measures of statistical significance (quantifying the significance of a deviation below a mean) can be used to define "significantly less than".
- **Inverse homophily:** cross-edges  $\gg 2pq$

## 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



## Measuring assortativity by neighbors

A way to compute the degree assortativity is by measuring the **degree correlation function**:

- **the correlation between the degree and the average degree of the neighbors of nodes with that degree.**

Let's calculate the average degree of  $i$ 's neighbors, first:

$$k_{nn}(i) = \frac{1}{k_i} \sum_j A_{ij} k_j$$

With a little abuse of notation, let's define the degree correlation function as the average of all the  $k_{nn}(i)$  for all the nodes whose degree is  $k$ .

$$k_{nn}(k) = \langle k_{nn}(i) \rangle_{i:k_i=k}$$

To visually check the network assortativity, we plot:

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

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 results in a horizontal line.

- Reference:
  - See Chapter 7 of [\[ns3\]](#) for more mathematical insights

## Example

node  $i$  with  $k_i$  neighbors

$$k_{nn}(i) = \frac{1}{k_i} \sum_j a_{ij} k_j$$

that is the average of  $i$ 's neighbors' degrees

Let's compute  $k_{nn}(i)$  for each node:

$$k_{nn}(0) = \frac{1}{3}(2 + 2 + 1) = \frac{5}{3}$$

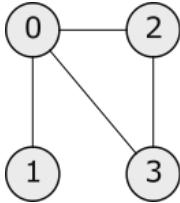
$$k_{nn}(1) = 3$$

$$k_{nn}(2) = \frac{1}{2}(3 + 2) = \frac{5}{2}$$

$$k_{nn}(3) = \frac{1}{2}(3 + 2) = \frac{5}{2}$$

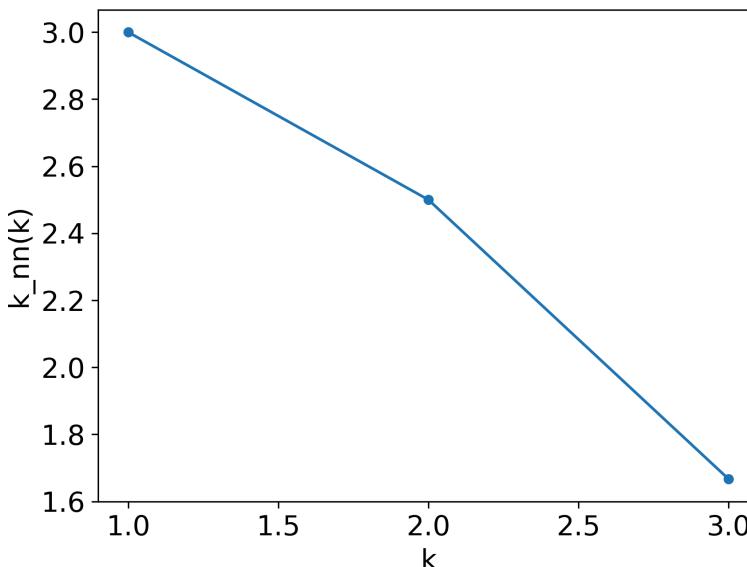


## Example



$$k_{nn}(0) = \frac{5}{3}, \quad k_{nn}(1) = 3, \quad k_{nn}(2) = \frac{5}{2}, \quad k_{nn}(3) = \frac{5}{2}$$

Let us plot  $(k, k_{nn}(k))$

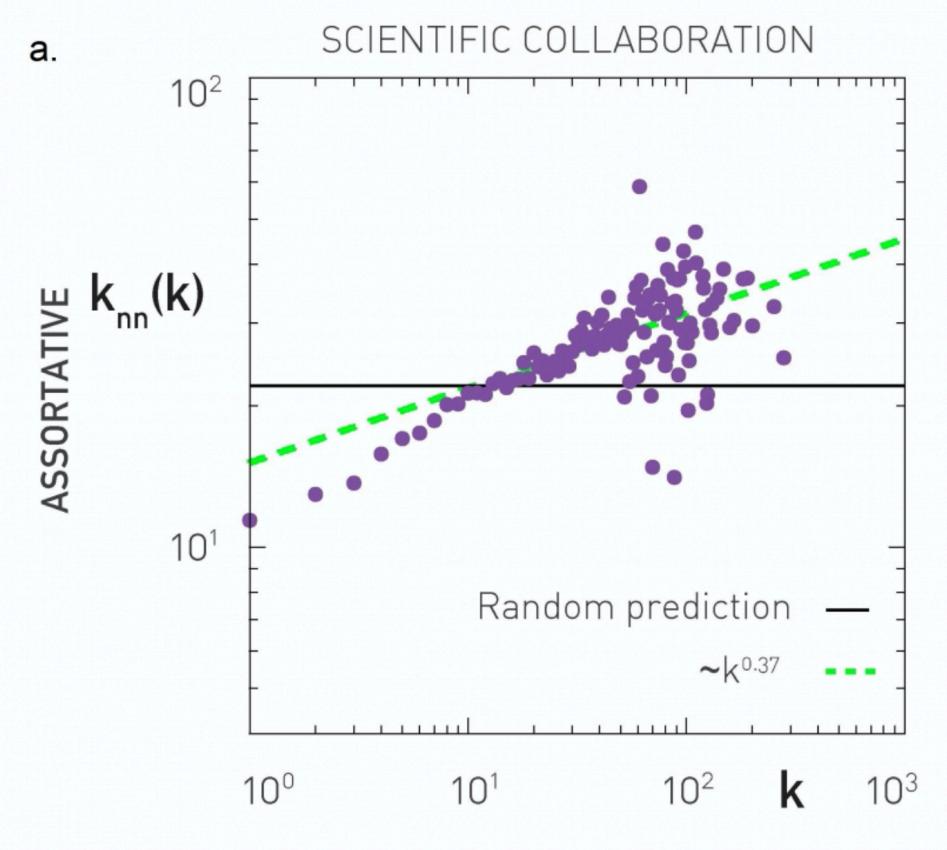
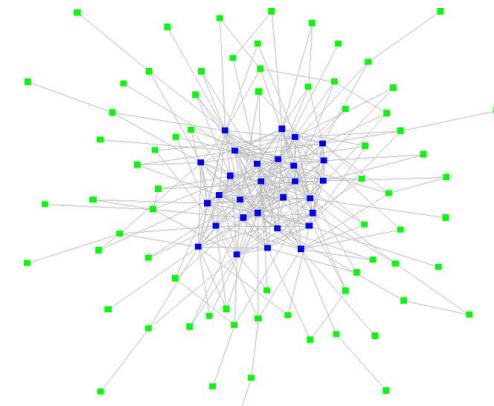


## Positive assortativity

- Scientific collaboration network

The increase with  $k$  indicates that the network is assortative.

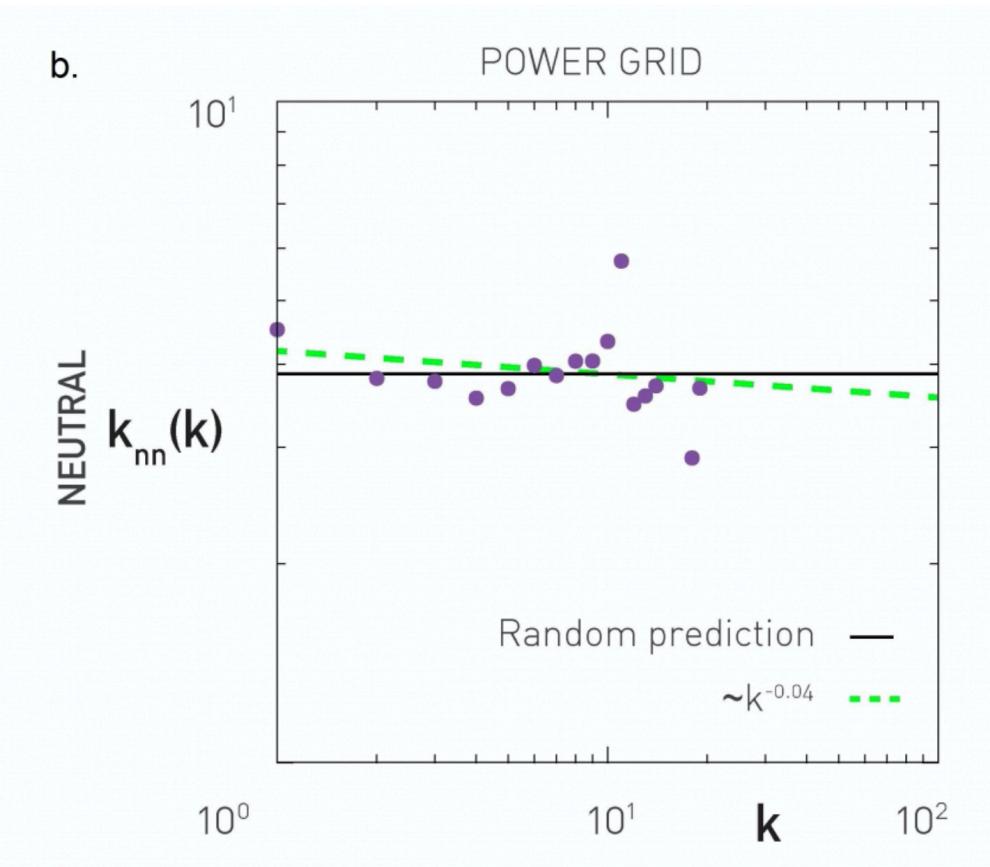
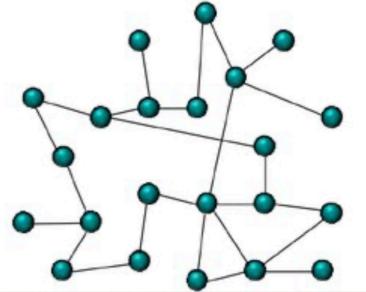
- ex. core-periphery



## Neutral assortativity

- Power grid

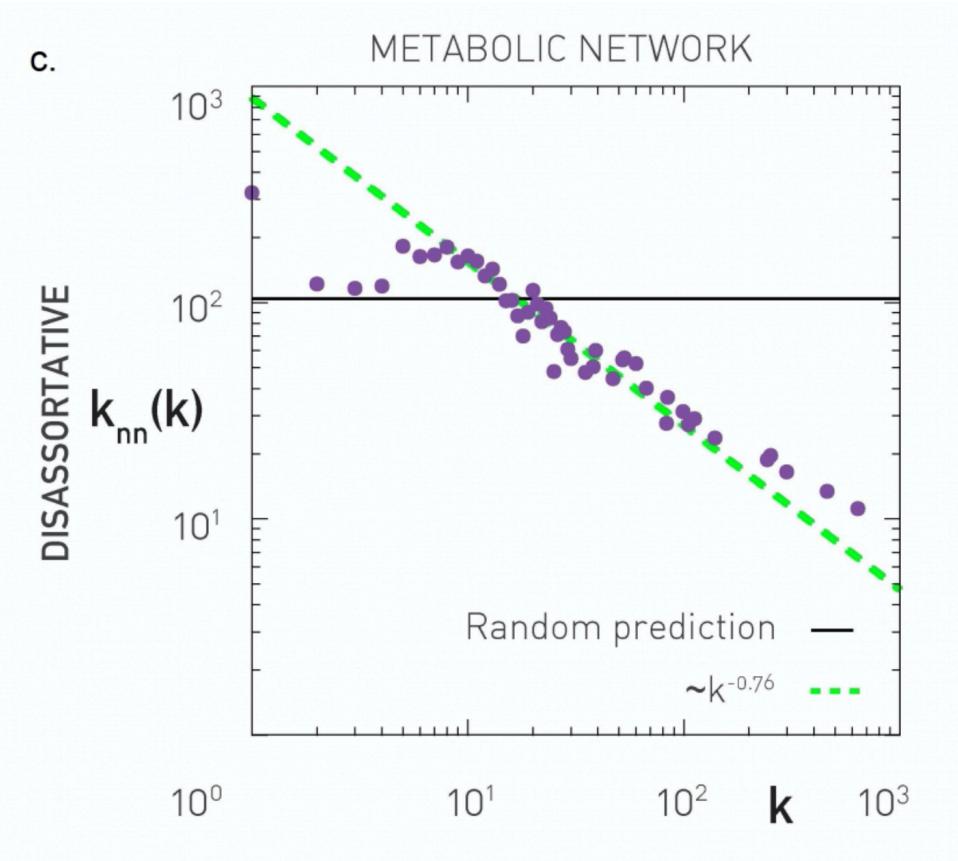
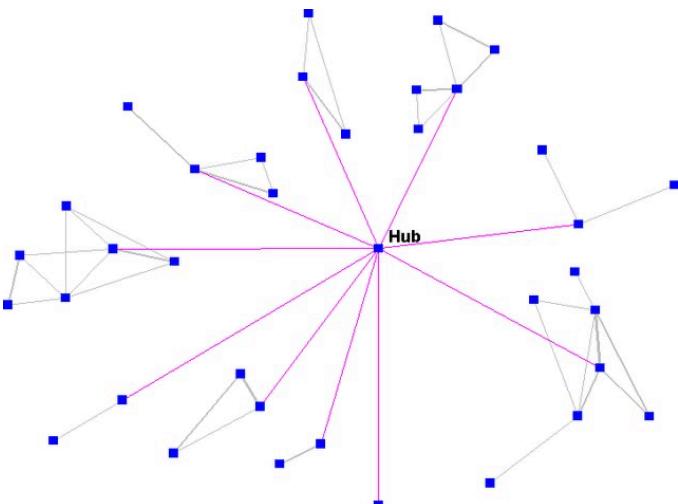
The horizontal indicates the lack of degree correlations, in line with our expectations for neutral networks.



## Negative assortativity

- Metabolic Network

The decreasing documents the network's disassortative nature.





# Mechanisms Underlying Homophily: Selection and Social Influence

## Underlying mechanisms of homophily

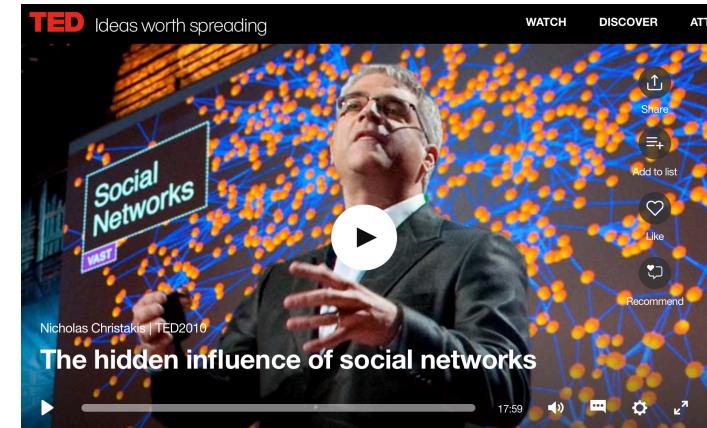
- Two possible mechanisms by which homophily (also assortativity) emerges naturally:
  1. **Selection:** similar nodes become connected
    - especially, but not limited to, in case of immutable attributes
  2. **(Social) influence:** connected nodes become more similar over time
    - it applies to mutable attributes
- 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 (connections and individual behaviors) for a long period
  - observe **both factors in action**
  - this makes it possible to see the **behavioral changes that occur after changes in an individual's network connections**, as opposed to the **changes to the network that occur after an individual changes his or her behavior**.
  - how do we quantify the impact?
- **Example 1: drug dependency:**
  - Students show a greater likelihood to use drugs when their friends do
  - Important for how to design interventions
  - Is targeting students who use drugs enough? influence vs. selection

## Example 2: obesity "contagion"

- dataset: 12,000 people
- they tracked obesity status and social network structure over 32 years
- obese vs. non-obese: there is a tendency toward clustering
- homophily test: passed
- The problem is to understand **why** is this the case
  - i. Selection?
  - ii. Confounding effects: homophily that correlates with something else?
  - iii. Social influence?
- Authors observe (i) and (ii) but **mainly (iii)** hinting at a form of "contagion" effect
  - [\[The hidden influence of social networks, Nicholas Christakis, TED2010\]](#)





# Affiliation

# Affiliation

- Can we represent the surrounding context through networks?
- An **affiliation network** is a network in which actors are connected via their **membership in groups** of some kind.
- **foci** (or groups): **focal points** of social interaction
- a **focus**: social, psychological, legal, or physical entities around which joint activities are organized (e.g. workplaces, voluntary organizations, hangouts, etc.)
- We can use **bipartite graphs**

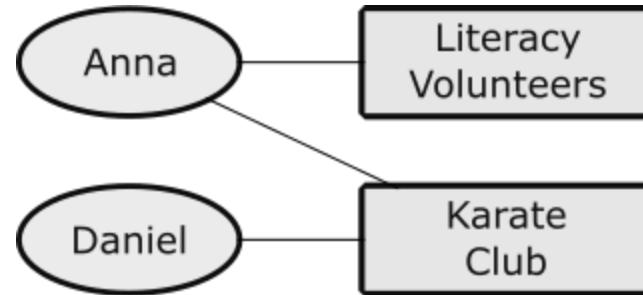
$G = (U, V, E)$  where  $U$  are persons and  $V$  are foci

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



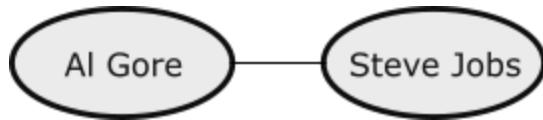
## Affiliation networks

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

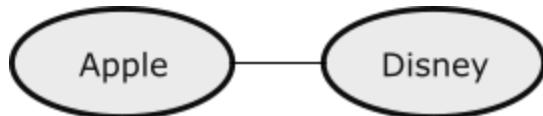


## Using projections

- Example: memberships to corporate boards of directors
- **Projection 1:** Network of interaction among board members



- **Projection 2:** Network of interaction among companies



- Can you do the same with the movie networks?

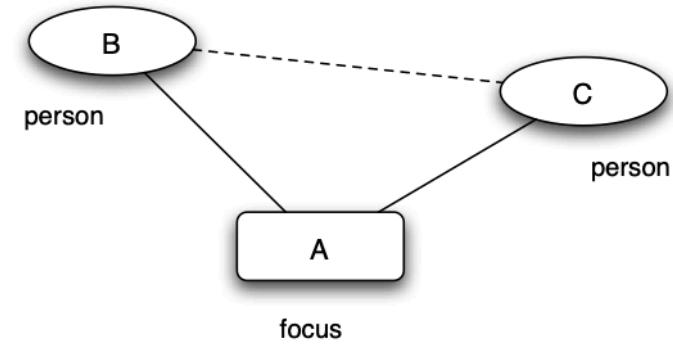


## Co-evolution of social and affiliation nets

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

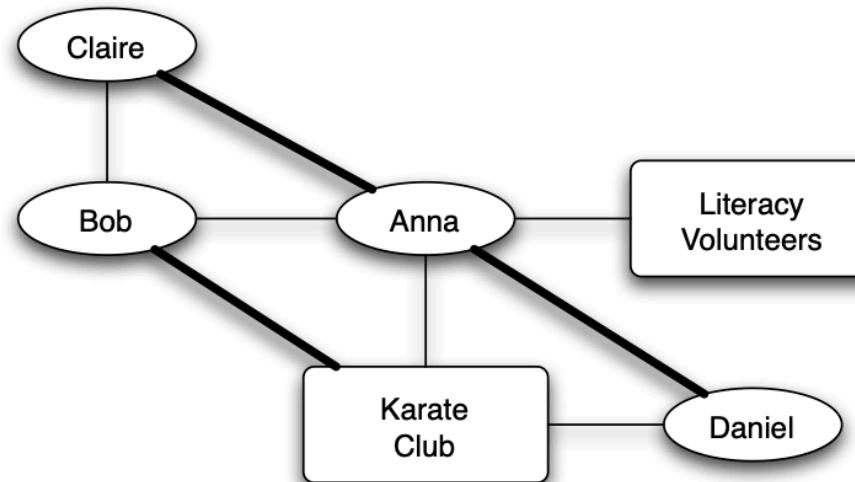


## Closures

(a) *Triadic closure*(b) *Focal closure*(c) *Membership closure*

## More on Closures

- In which links is more likely that the following factors played a role?
  - Friendship transitivity?
  - Focal closure
  - Membership closure



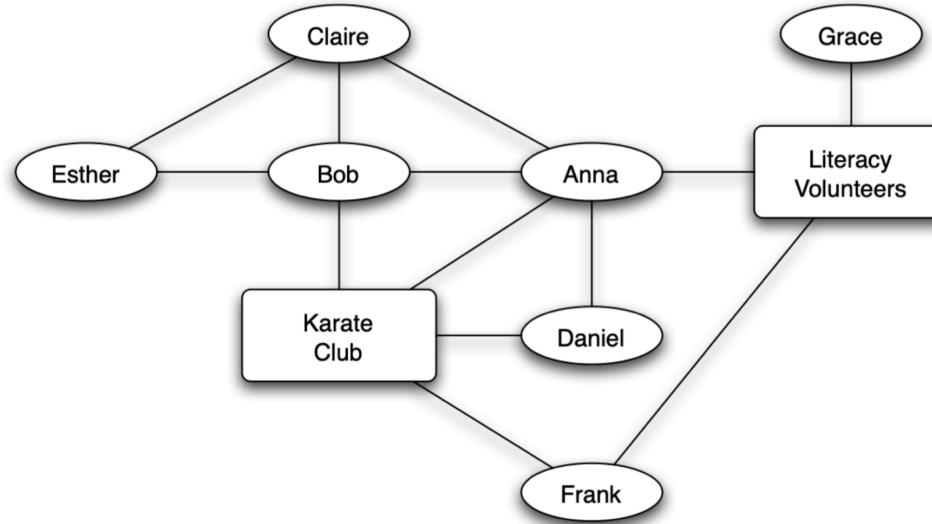
# Tracking Link Formation in Online Data

## Data from online platforms

- **caveat:** it is never a priori clear how much one can extrapolate from digital interactions to interactions that are not computer-mediated, or even from one computer-mediated setting to another.

## 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?



# Experimental setting

- 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
- 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:
  - $T_{baseline}(k) = 1 - (1 - p)^k$
  - $T_{baseline'}(k) = 1 - (1 - p)^{k-1}$

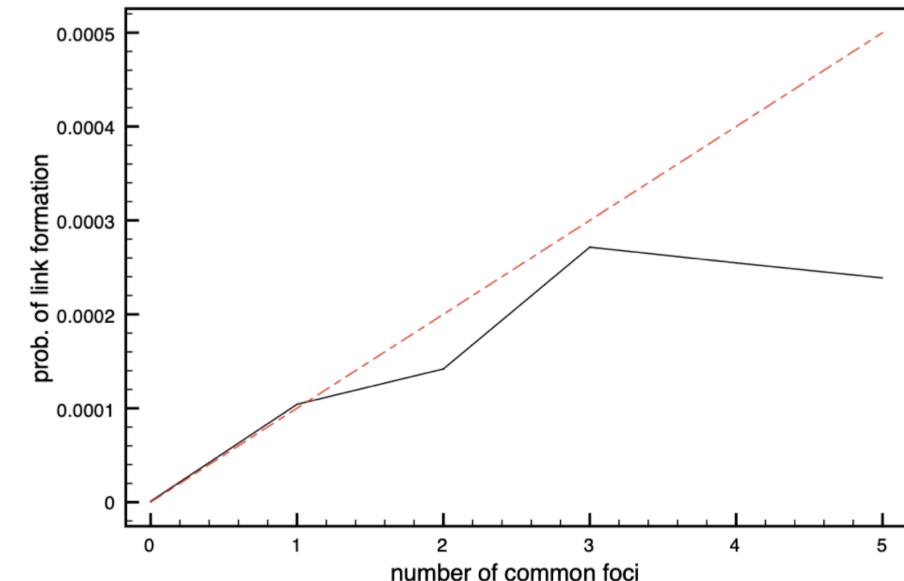




- Reference:
  - Kossinets and Watts, e-mail communication among roughly 22,000 undergraduate and graduate students over a one-year period at a large U.S. university
- **The common friends assumption is too simple!**
- There is "something" more than this

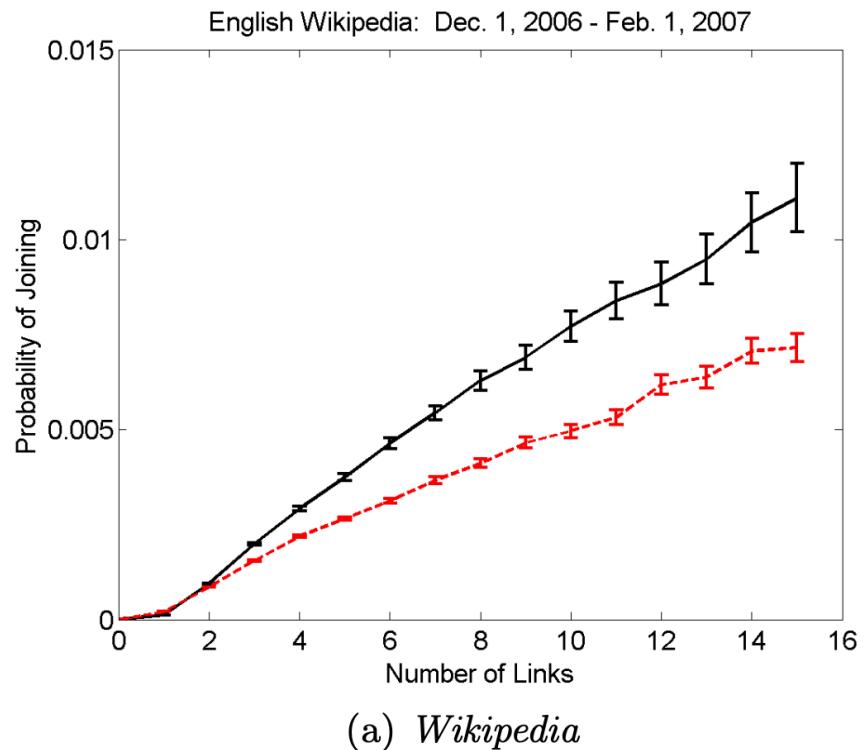
## Focal closure: looking for evidence

- **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: it turns **downward** and appears to approximately **level off**
  - subsequent shared classes after the first produce a **diminishing returns** effect



# Membership closure: Wikipedia case study

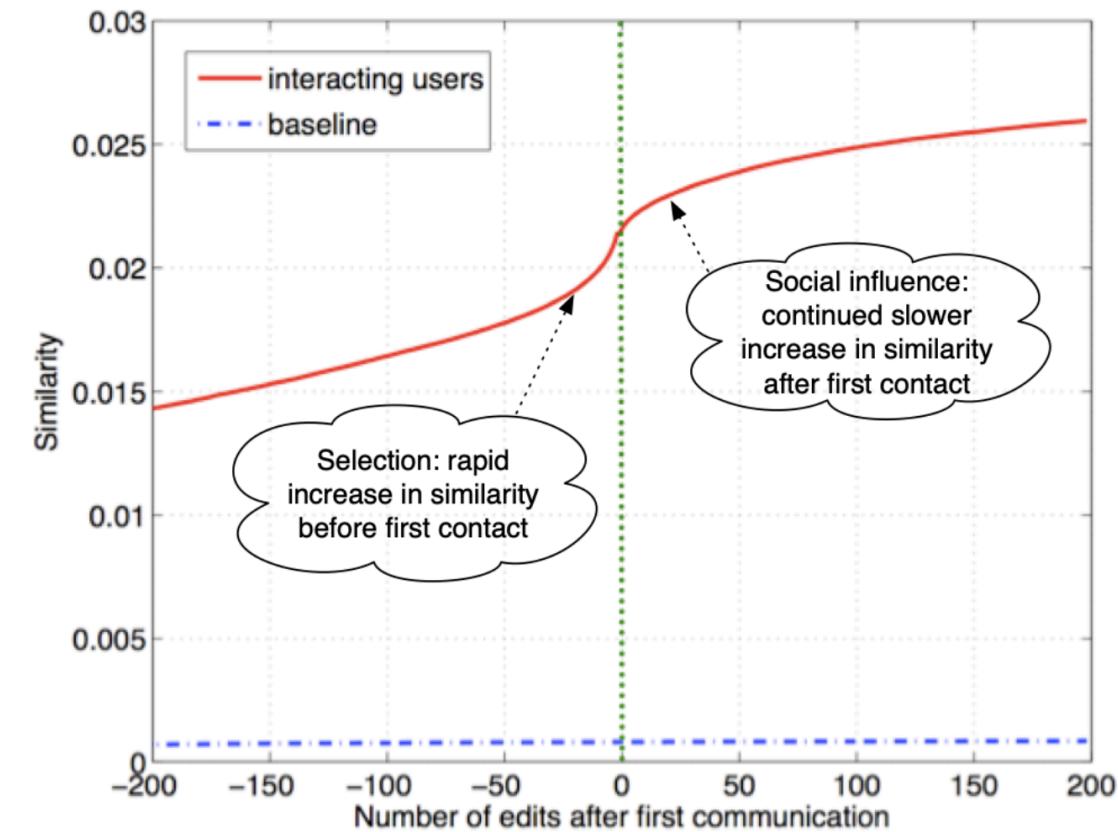
- **membership closure:** what is the probability that a person becomes involved with a particular focus as a function of the number of friends already involved?
- Probability of joining a community based on  $k$  exposure via social ties (black) versus similarity ties (red)
- social connections via direct communication on the user talk page
- Reference:
  - Feedback Effects between Similarity and Social Influence in Online Communities



- $A$  and  $B$  editors;  $NA(X)$  is the number of articles edited by editor  $X$

$$sim(A, B) = \frac{NA(A) \cap NA(B)}{NA(A) \cup NA(B)}$$

- Average similarity of two editors on Wikipedia, relative to time (0) at which they first communicated
- Time (x-axis) measured in discrete units, where each unit is a single Wikipedia action taken by either of the two editors
- **Both selection and influence effects at work!**
- dashed blue line: baseline for non-interacting editors



## 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



# A Spatial Model of Segregation

# Natural spatial signature in cities

Formation of homogeneous (according to some type or class) neighbors in cities



(a) Chicago, 1940



(b) Chicago, 1960

Blocks colored yellow and orange the percentage of African-Americans is below 25, while in blocks colored brown and black the percentage is above 75

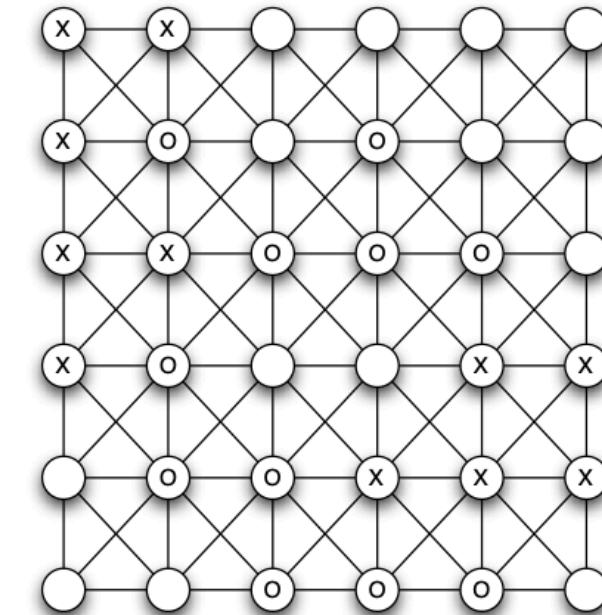
## The Schelling Model

- Can global spatial segregation arise from the effect of homophily operating at a local level?
- Assumption: no individual want segregation explicitly
- Agents:
  - two types
  - immutable characteristics
- Agents reside in a cell of a grid
  - some cells contain agents
  - some other cells are unpopulated
- Neighbors: 8 other cells "touching" an agent

- Each agent wants to have at least  $t$  neighbors of their type
- If an agent finds  $< t$  neighbors of the same type, then they are **unsatisfied**
- If unsatisfied, they want to move

x	x				
x	o		o		
x	x	o	o	o	
x	o			x	x
	o	o	x	x	x
		o	o	o	

(a) Agents occupying cells on a grid.



(b) Neighbor relations as a graph.

## The dynamics of movements

- Agents move in a sequence of rounds
- If there is no empty cells around an unsatisfied agent:
  - move the agent randomly or
  - leave the agent there

X1*	X2*				
X3	O1*			O2	
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

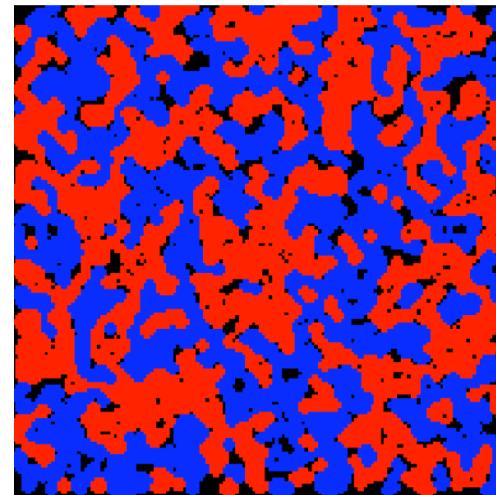
(a) An initial configuration.

X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

(b) After one round of movement.

## Larger examples

- Computer simulations to look for patterns at a larger scale
- We want to run different simulations and make some comparisons  $\Rightarrow$  integrated pattern?
- on the right: two runs of a simulations of the Schelling model with a threshold  $t$  of 3
  - 150x150 grid
  - 10,000 agents
- Segregation emerges even when agents accept to be a minority!



(a) A simulation with threshold 3.



(b) Another simulation with threshold 3.

# NetLogo

## Agent-based simulations



Try it out!

## Interpretation of the model

- Optimistic hypothesis: an agent would be perfectly happy to be in minority with  $t = 3$
- Realizations, as the one in the example, are possible
  - **segregated realizations are more likely**
- The basic (random) reasons behind segregation are at work even in an idealized setting where agents are perfectly happy of being a minority

x	x	o	o	x	x
x	x	o	o	x	x
o	o	x	x	o	o
o	o	x	x	o	o
x	x	o	o	x	x
x	x	o	o	x	x

If  $t$  is larger, segregation is amplified





# Reading material

## References

[ns1] Chapter 2 (2.1)

[ns2] **Chapter 4 (4.1-4.5)**



# Q & A

