Homophily and assortativity

Introduction to Network Science
Carlos Castillo
Topic 08

Universitat Pompeu Fabra Barcelona

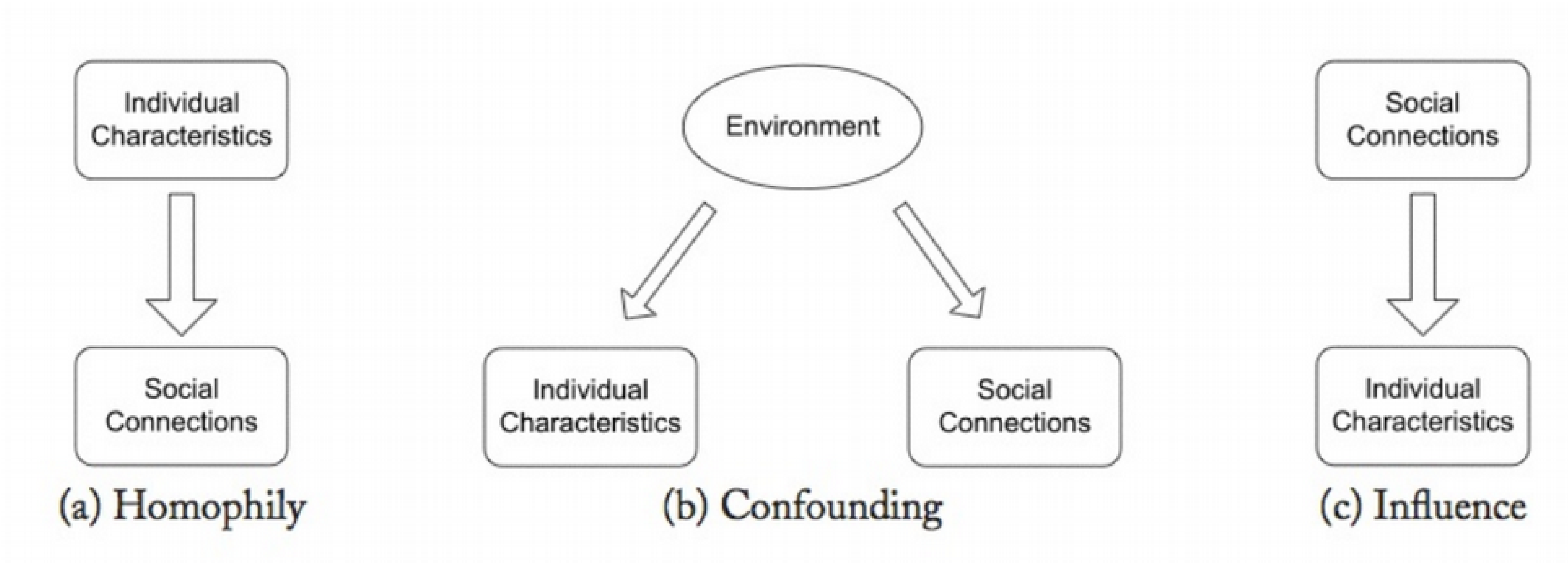
Contents

- Models of social correlation
- Homophily
- Social influence (more on this later)
- Assortativity

Sources

- Albert László Barabási: Network Science. Cambridge University Press, 2016. Ch 07
- Networks, Crowds, and Markets Ch 03 and 04
- Nicola Barbieri's tutorial on homophily and influence in social networks, 2016
- C. Castillo: Link prediction slides 2016

Three models of social correlation



Three models (cont.)

- **Homophily**: tendency of agents to be connected to similar agents
- **Confounding**: correlation between agent actions can be explained by external factors
- **Influence**: agent actions influence the actions of their connections

Homophily

"*Cada olleta té la seva tapadoreta*"

"Cada ovella amb sa parella"

"Cada qual amb son igual"

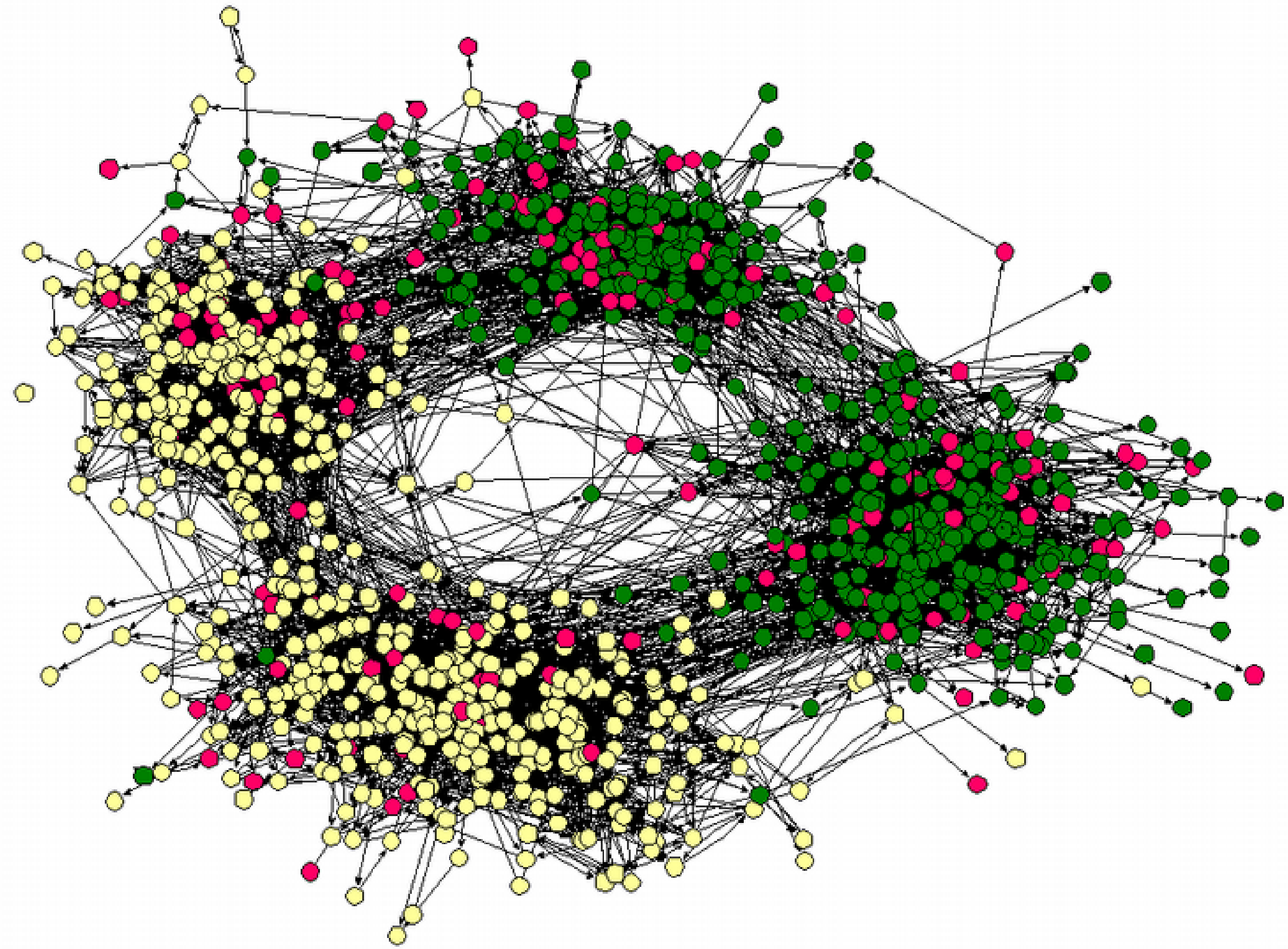
+ similar Catalan sayings

Think of your own friends

- Your friends are not a random sample
  - Similar age, affluence, interests, beliefs, … to you (most of them)
- Long-standing observation
  - Plato ("similarity begets friendship")
  - Aristotle (people "love those who are like themselves")

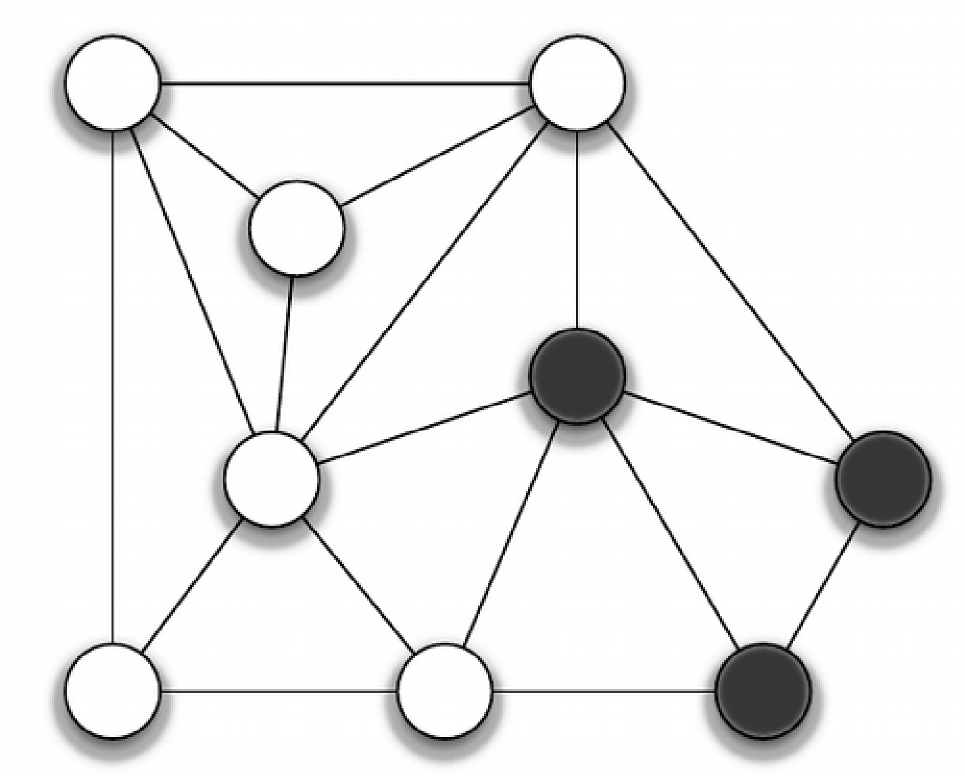
Friendships by race

(Middle school in the US, ca 2001)



Moody, J. (2001). Race, school integration, and friendship segregation in America. American journal of Sociology, 107(3), 679-716.

How to measure homophily?



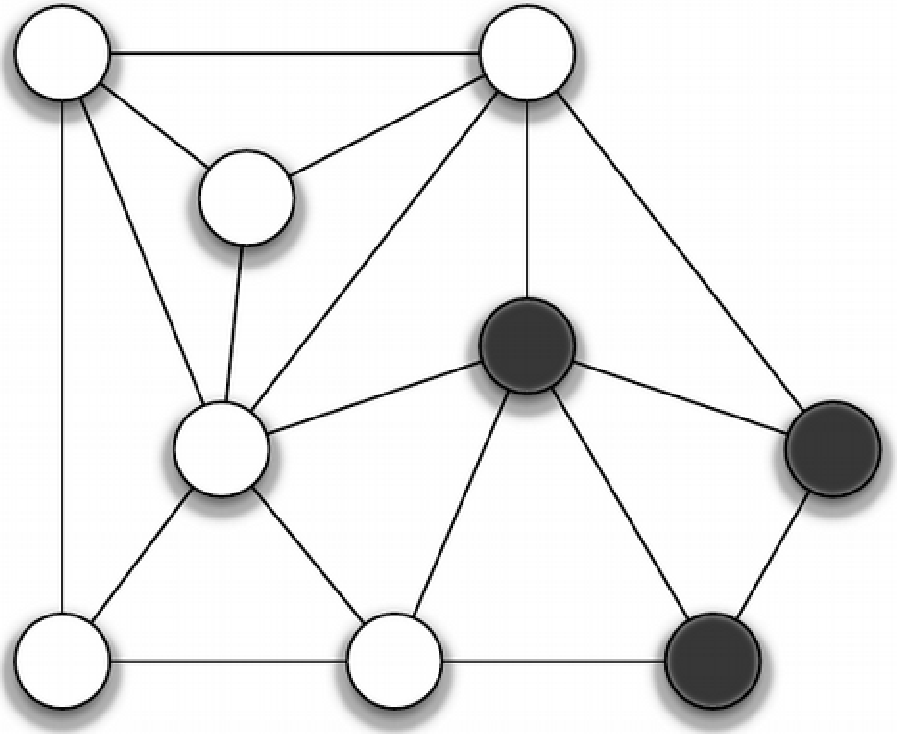
- Count homogeneous edges (white-white or black-black)
- Count heterogeneous edges (white-black)

Homophily on a graph

- Suppose the probability of being white is p
- The probability of being black is $q = 1 - p$
- What is the probability of:
  - A white-white edge
  - A black-black edge
  - A white-black edge

Homophily on this graph

- With 6 white nodes and 3 black nodes, how many homogeneous edges do we expect?
- Hence, does this graph exhibit homophily?



Homophily measurement: one-tailed binomial test

- Given G=(V,E) with colors assigned at random

  (p|V| white nodes and (1-p)|V| black nodes)

$$\forall (i,j) \in E : X_{ij} = Pr[(i,j) \text{ is an heterogeneous edge}]$$

$$X_{ij} \sim \text{Bernoulli}(2p(1-p))$$

- The number of heterogeneous edges follows

$$\sum_{(i,j) \in E} X_{ij} \sim \text{Binomial}(L, 2p(1-p))$$

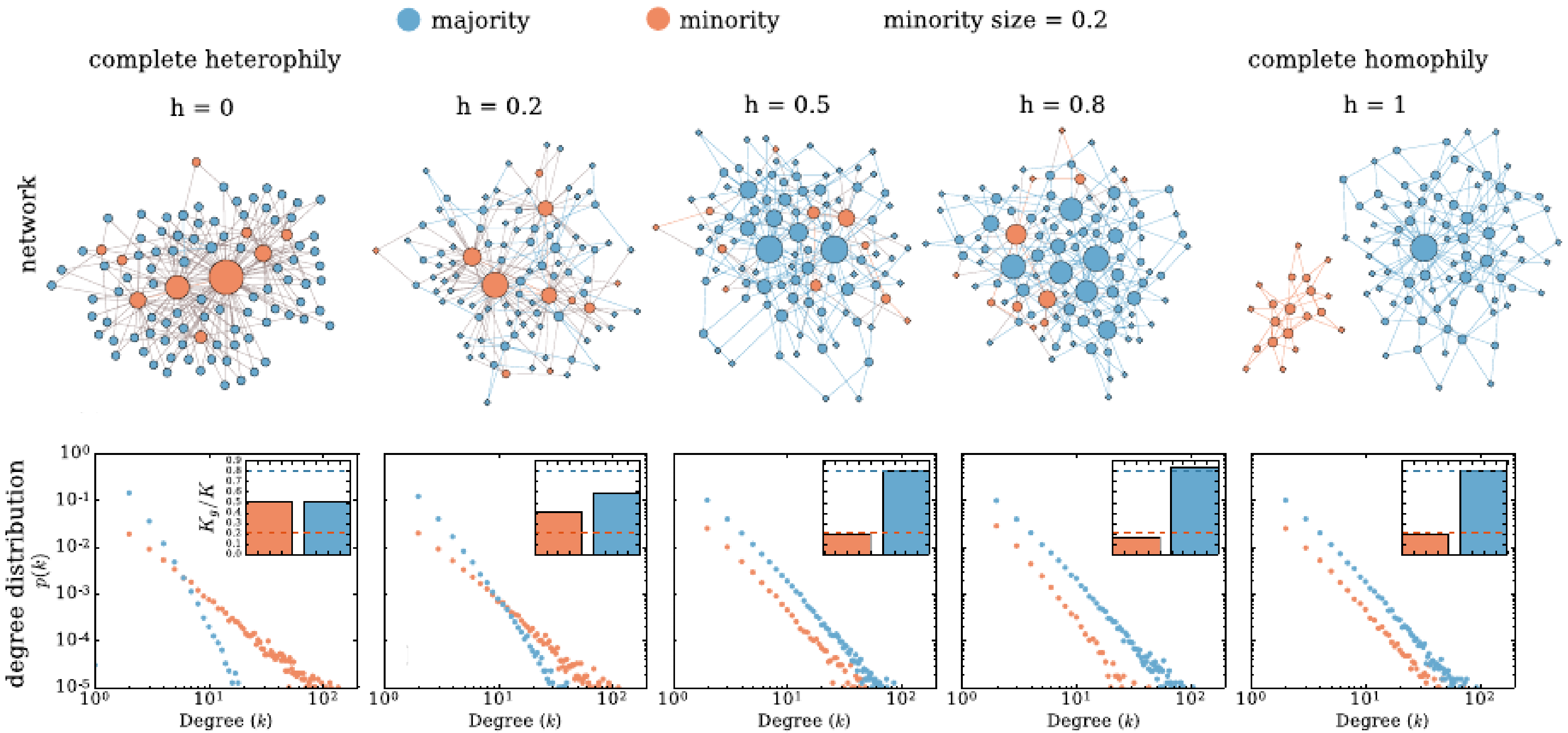
In our example

- We compute

$$
\begin{aligned}
Pr[\text{Binomial}(L, 2pq) \leq 5] &= Pr[\text{Binomial}(18, 4/9) \leq 5] \\
&= 0.1174
\end{aligned}
$$

- Hence observing this number of heterogenous edges or less has a probability of more than 10%
- Hence homophily in this case is not significant at 0.05

Preferential attachment with homophily



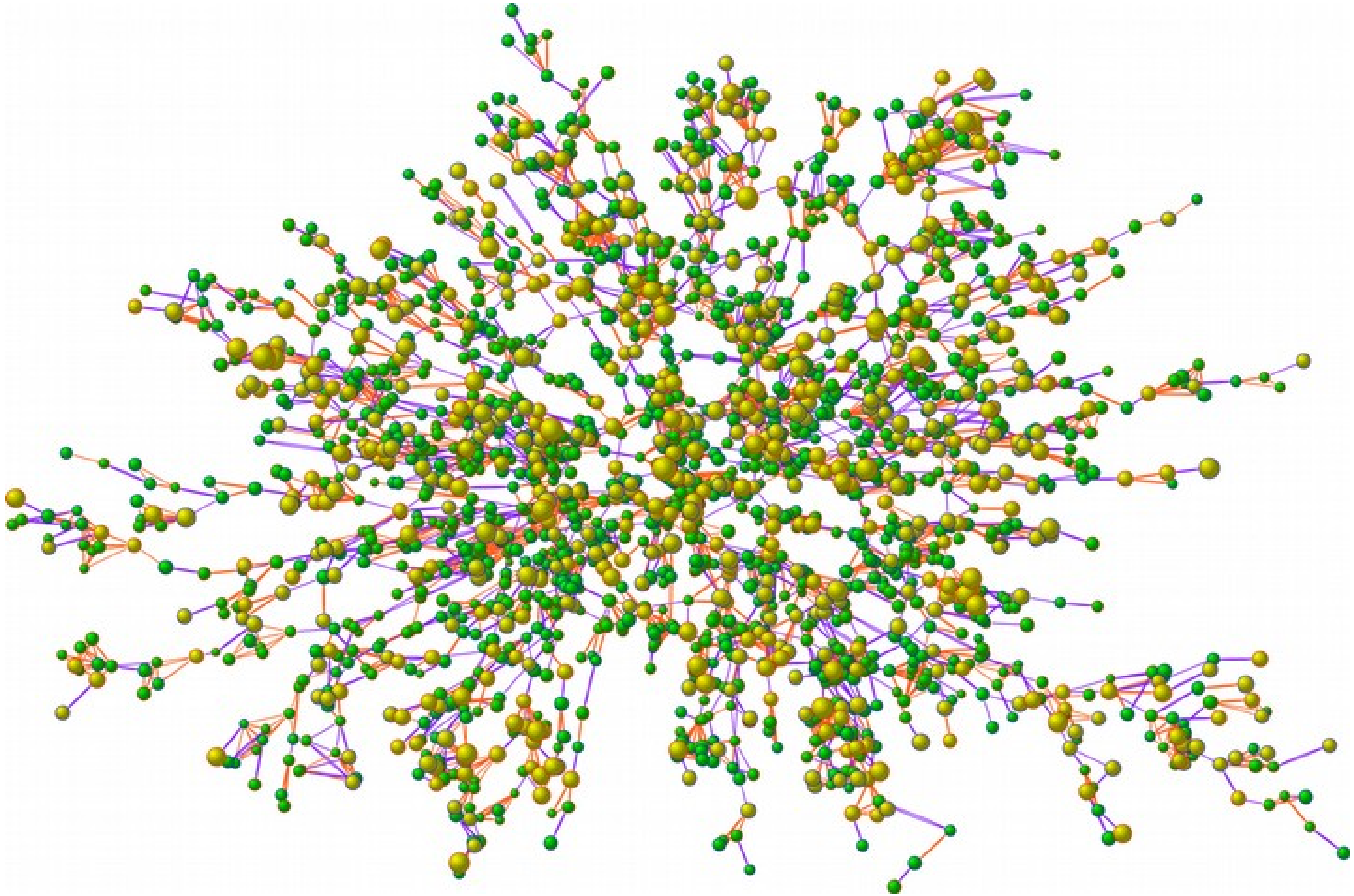
Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. Scientific reports, 8.

Shuffle tests

- Many network characterization questions can be addressed through **shuffle tests**
- For homophily, one can compare a characteristic (e.g., probability of having an heterogeneous edge) with the observation in a network in which node labels are shuffled
- This is a general, very powerful technique

Anagnostopoulos, A., Kumar, R., & Mahdian, M. (2008, August). Influence and correlation in social networks. In Proc. KDD (pp. 7-15). ACM.

Is obesity contagious?

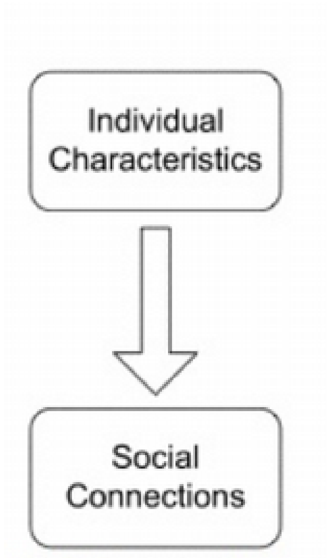


Size: BMI

Yellow: obese

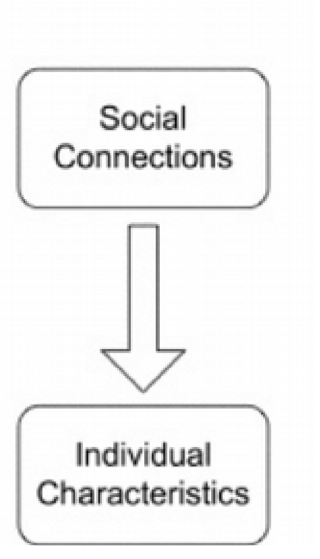
Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England journal of medicine, 357(4), 370-379.

Selection and social influence



Selection is the process by which we chose to connect to people based on our characteristics

Social influence is the process by which we transform and are transformed by our connections



What do we see in the obesity study?

Christakis and Fowler show evidence that in their sample:

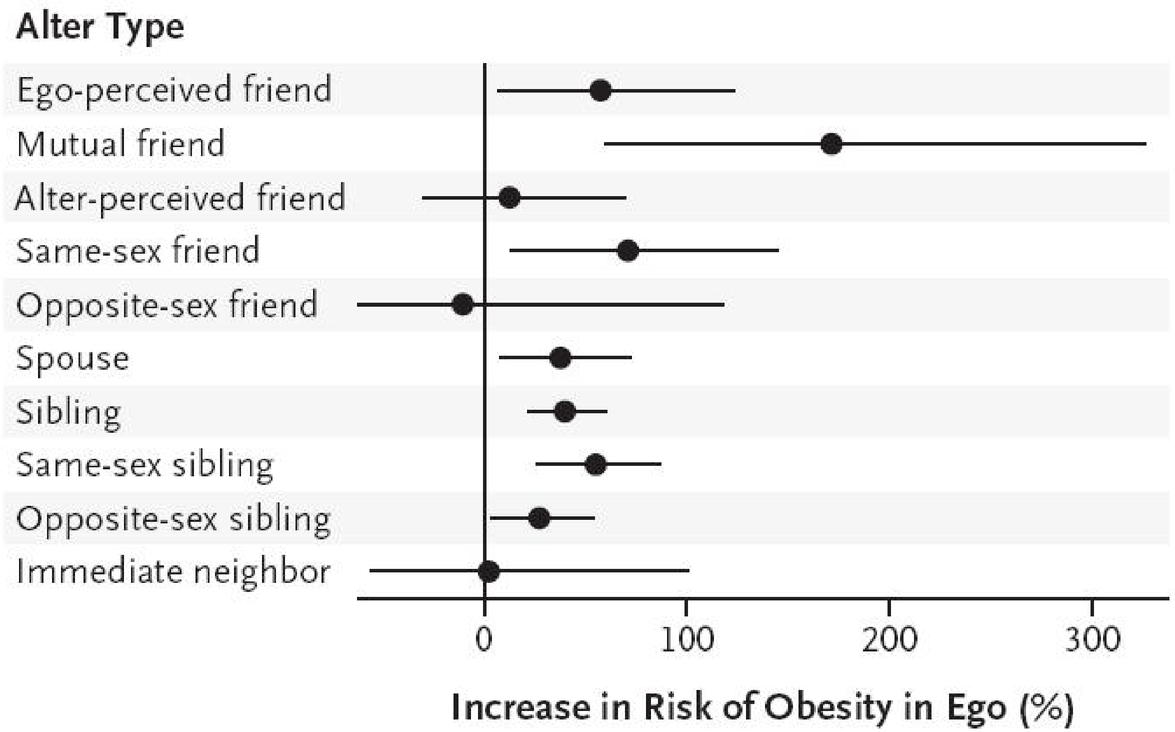
1) People indeed connect to other similarly obese or not obese (homophily)
2) People are affected by external influence to be more or less obese (confounding)
3) People change the behavior of others (social influence)

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England journal of medicine, 357(4), 370-379.

Obese Friend → 57% increase in chances of obesity

Obese Sibling → 40% increase in chances of obesity

Obese Spouse → 37% increase in chances of obesity



(more on social contagion later)

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England journal of medicine, 357(4), 370-379.

Thought experiment

Suppose you are asked to design a program to prevent drug addiction by focusing resources in ensuring well-connected people do not become addicted to drugs

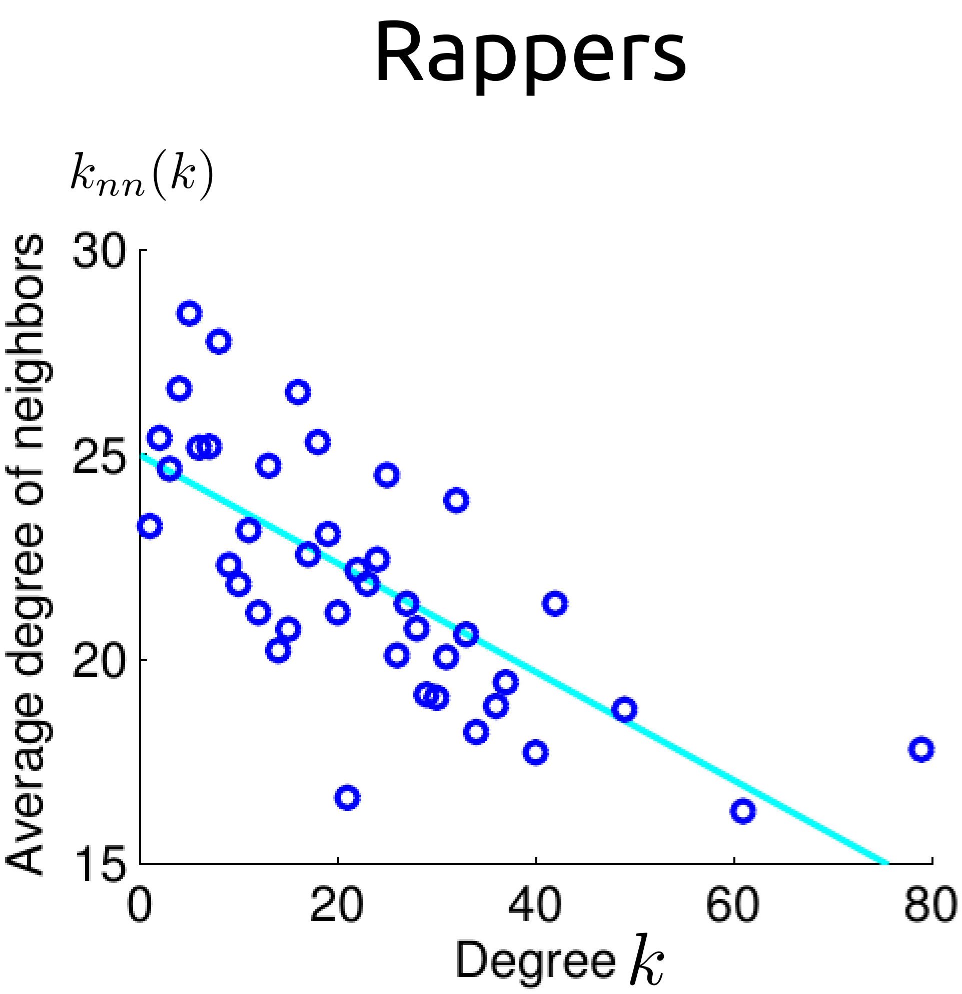
Under which circumstances this program would be more successful? Less successful?

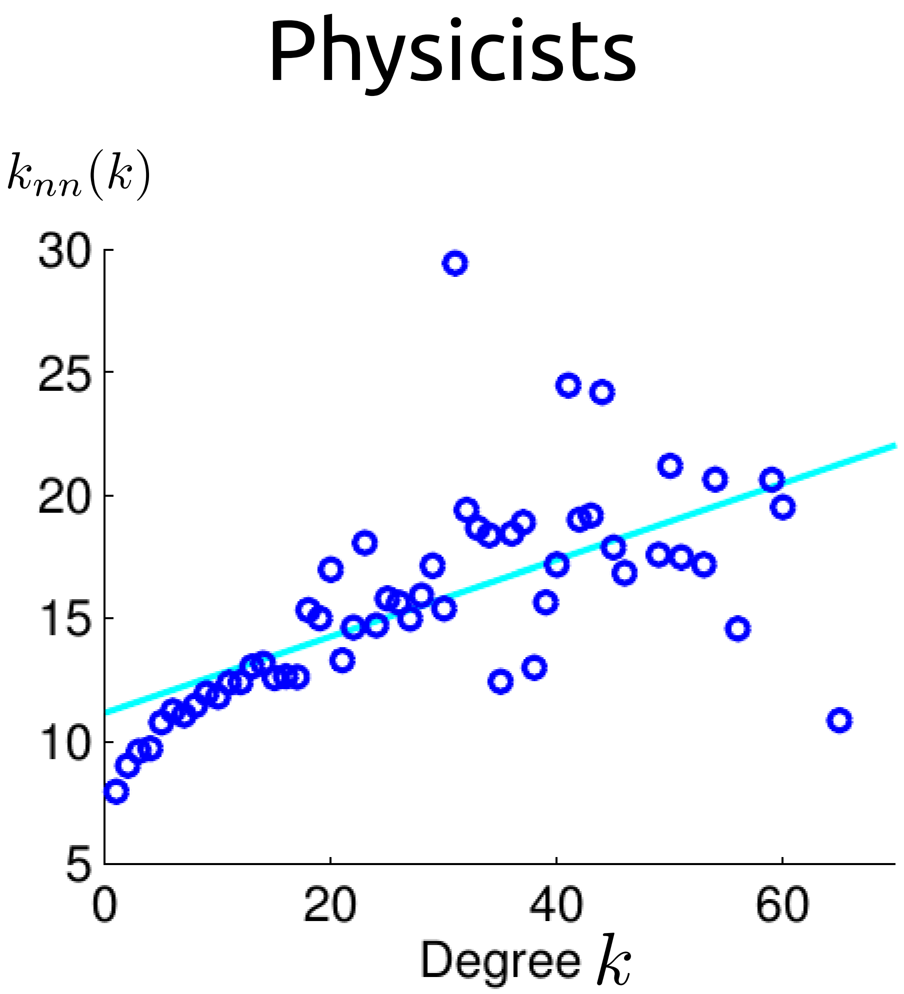
Assortativity

(≃Homophily by degree)

Evidence of assortative behaviors

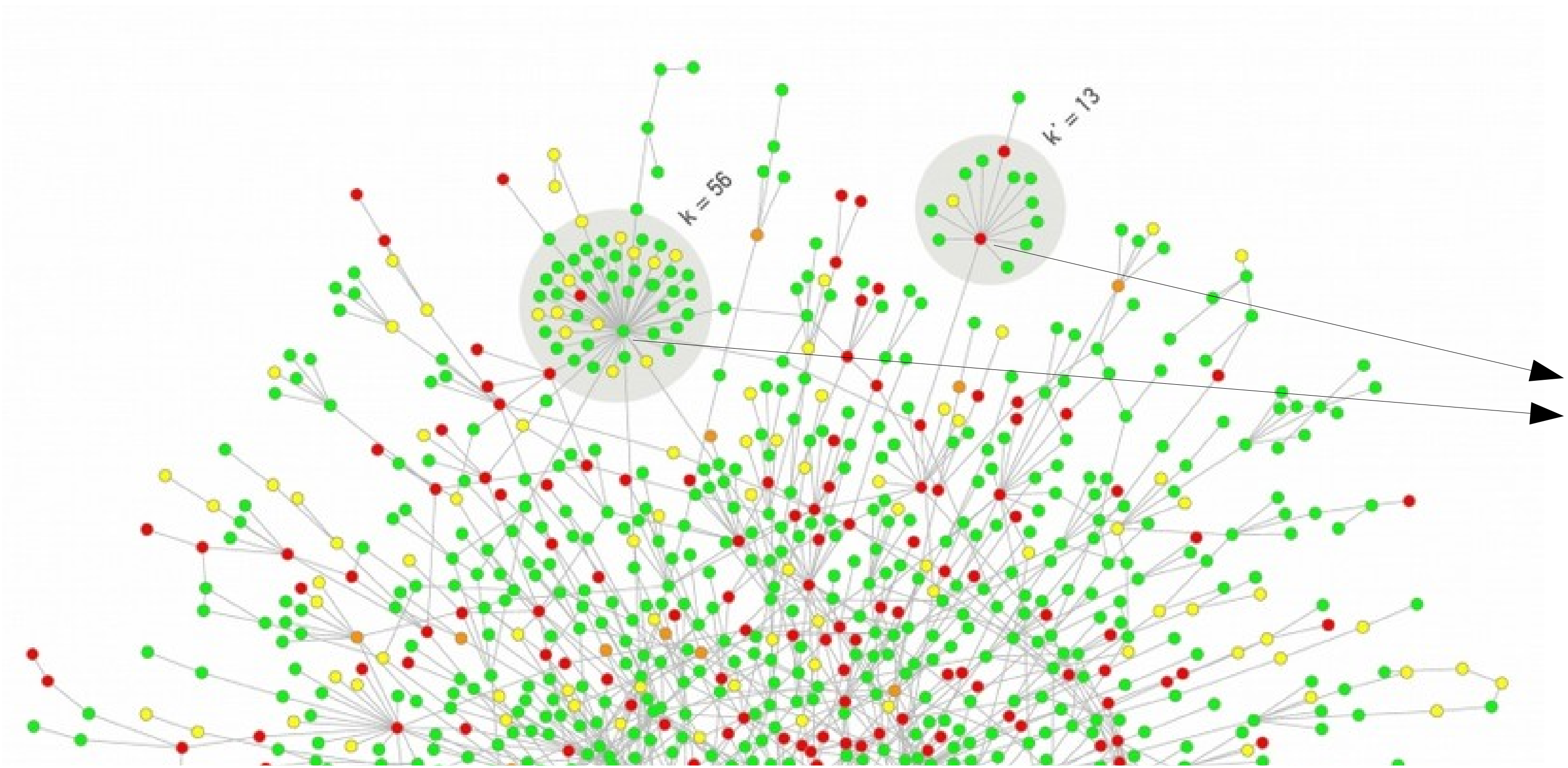
- Celebrities marrying celebrities
- Big companies having people in their boards who are in many boards of big companies
- Famous physicists work with each other
- However, famous rappers don’t …





https://blogs.aalto.fi/mining4meaning/2014/03/07/rappers-and-physicists/assortativity-2/

A disassortative network: protein interactions



The two largest hubs link to many small-degree nodes

Describing degree correlations

- Degree correlation matrix e_{ij}
  - Probability that in a randomly selected node, one end has degree i and the other end has degree j

Degree correlations

- Probability that a node at the end of a randomly chosen link has degree k

$$q_k = \frac{kp_k}{\langle k \rangle}$$

- Degree correlation matrix:

$$e_{ij} = q_i q_j$$

High-degree and high-probability nodes are more likely to be found

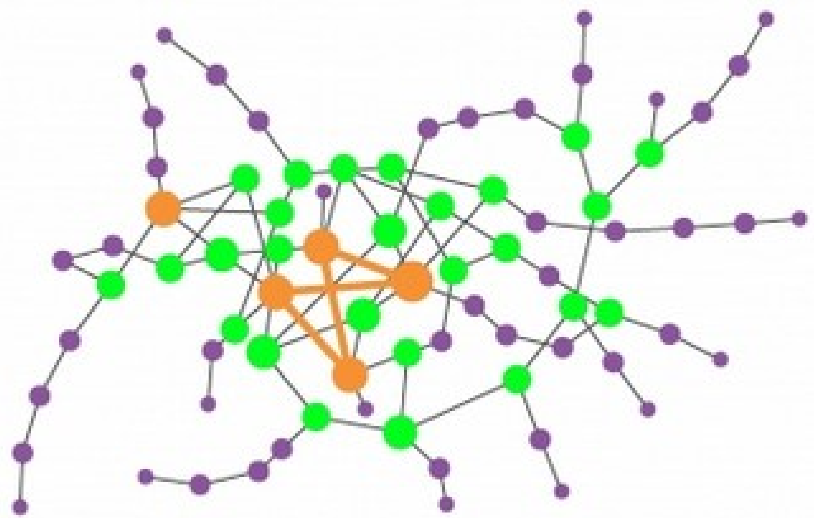
$$q_k = Ckp_k$$

$$\sum_k q_k = \sum_k Ckp_k = 1$$

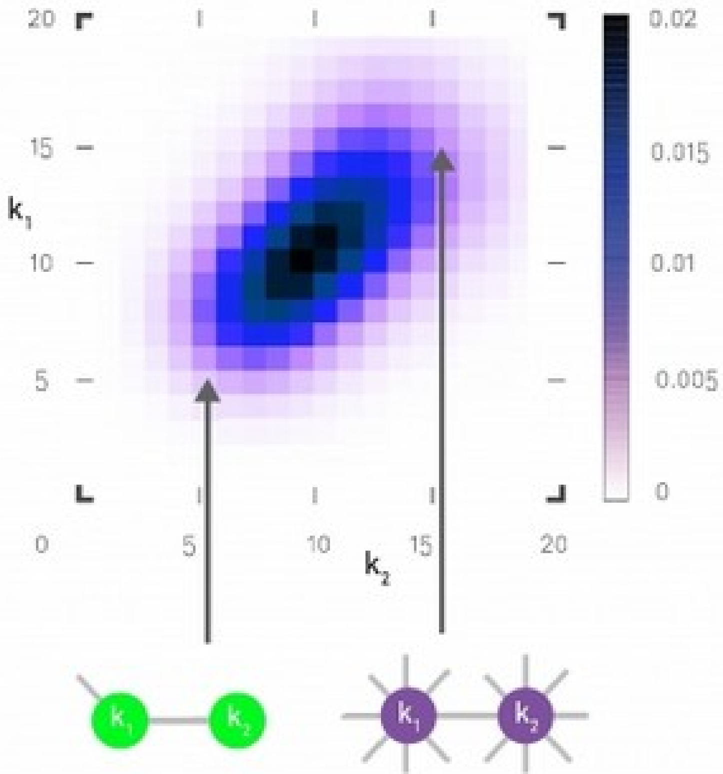
$$C = \frac{1}{\sum_k kp_k}$$

$$C = \frac{1}{\langle k \rangle}$$

Assortative behavior

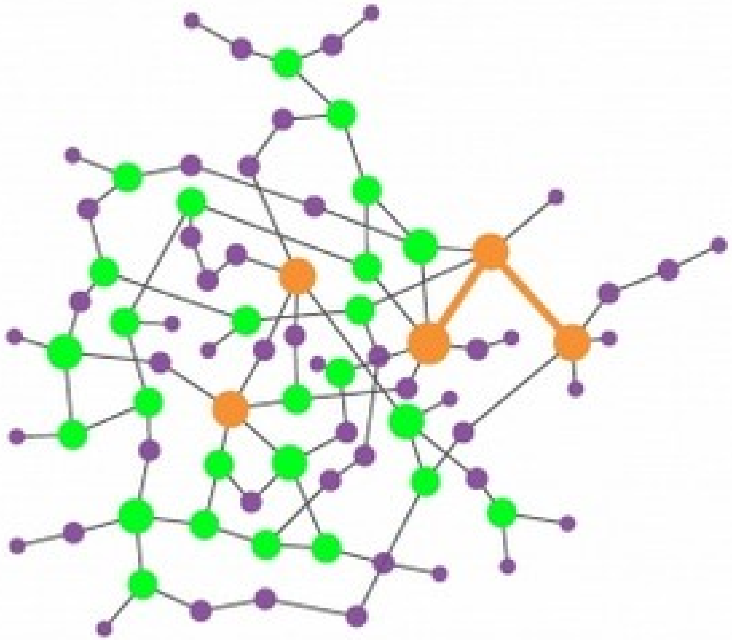


$e_{k_1 k_2}$

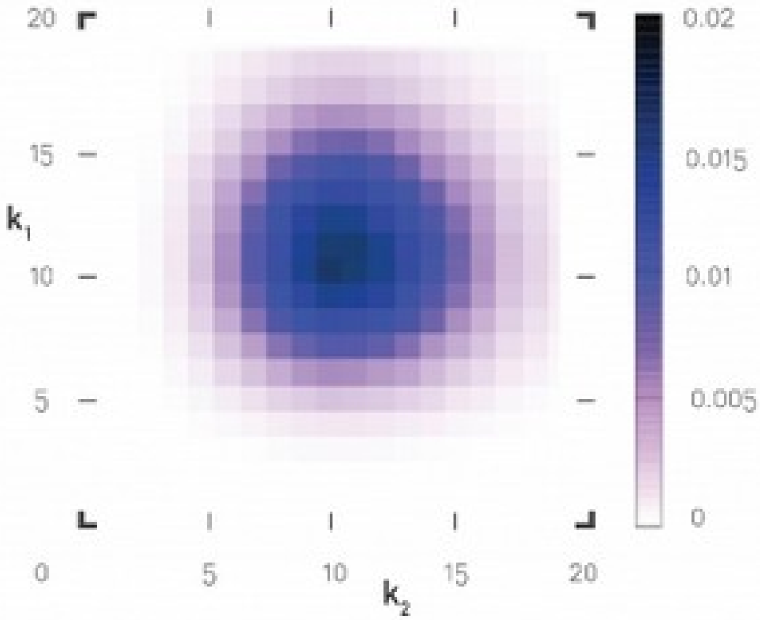


Colors by degree: low medium high

Neutral behavior

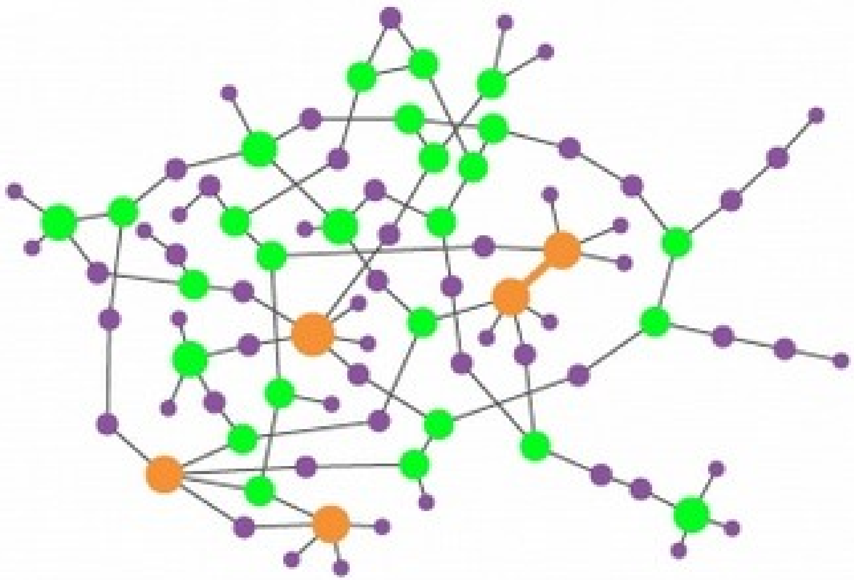


$e_{k_1 k_2}$

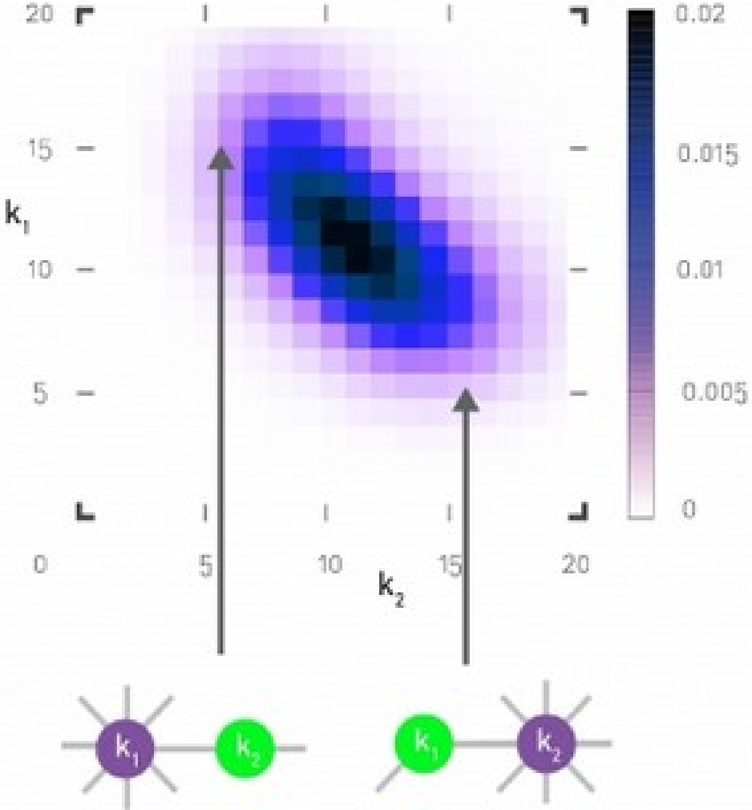


Colors by degree: low medium high

Disassortative behavior



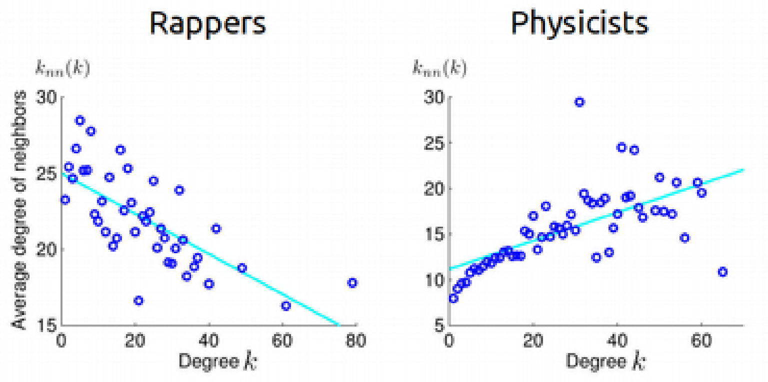
$e_{k_1 k_2}$



Colors by degree: low medium high

Degree correlation function

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



$P(k \to k')$ is the probability that by following a link of a node with degree k, we reach a node with degree k'

Compute the degree correlation function for a neutral network, in which $e_{ij} = q_i q_j$

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

$$q_k = \frac{k p_k}{\langle k \rangle}$$

Note that in a neutral network:

$$P(k \to k') = \frac{e_{kk'}}{\sum_{k'} e_{kk'}}$$

Neutral network

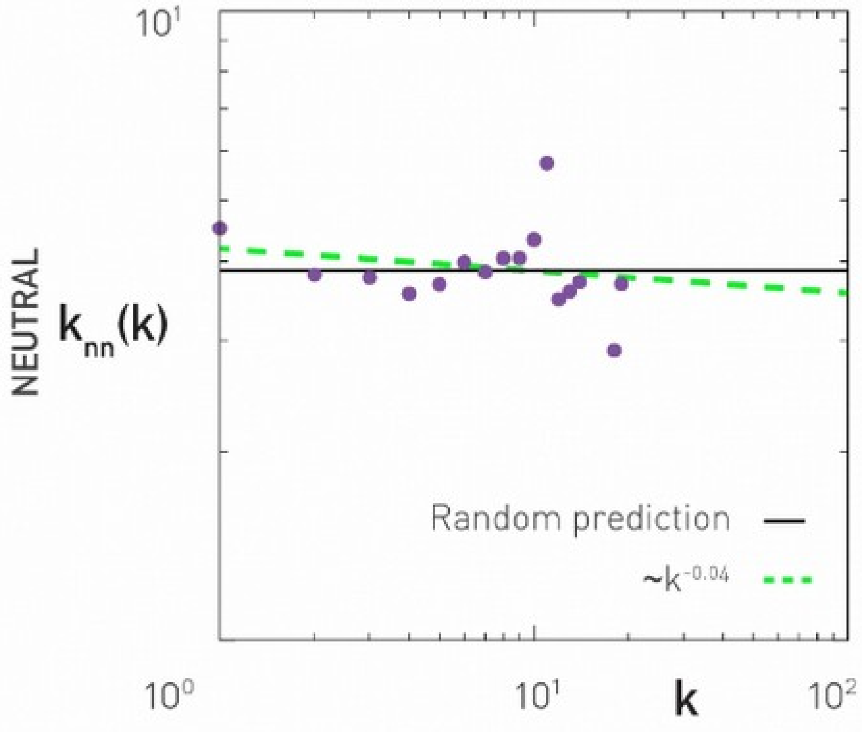
$$k_{nn}(k) = \sum_{k'} k' P(k \to k')$$

In a neutral network

$k_{nn}(k)$ is independent of k

Assortative: increases

Disassortative: decreases



Model for degree correlations

$$k_{nn}(k) = ak^{\mu}$$

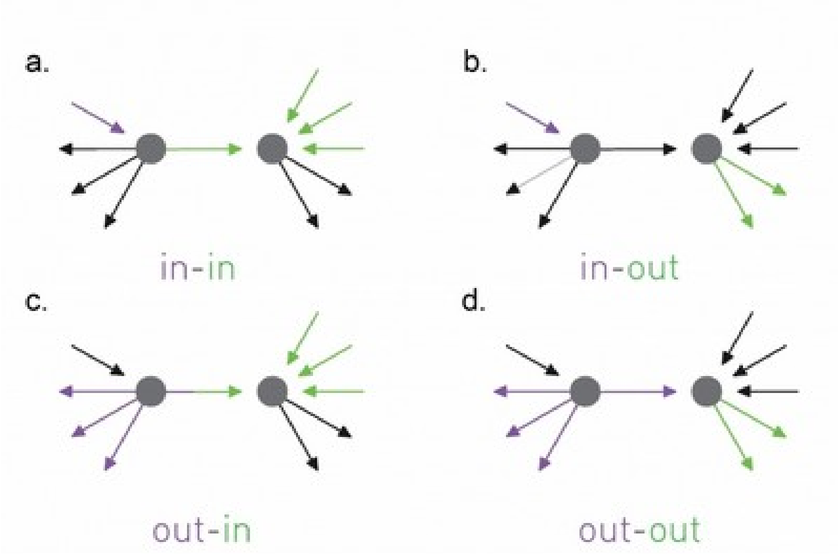
- If $\mu > 0$ the network is assortative
- If $\mu = 0$ the network is neutral
- If $\mu < 0$ the network is disassortative

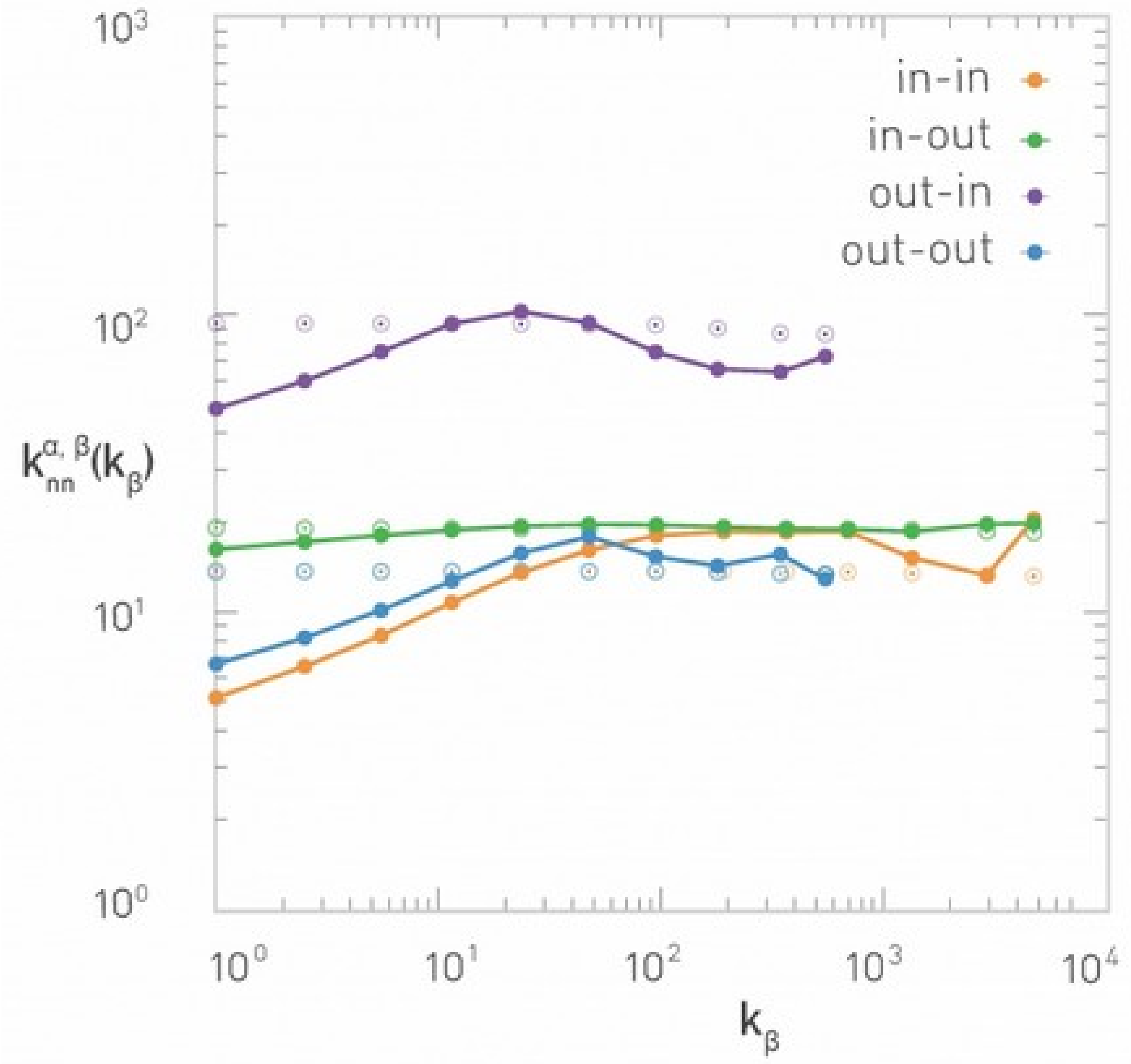
Alternative model

$$k_{nn}(k) = ak + b$$

- If $a > 0$ the network is assortative
- If $a = 0$ the network is neutral
- If $a < 0$ the network is disassortative

Degree correlations in directed networks





Note

- Assortative/disassortative observations can be explained in part simply by degree sequences in the network
- This can be addressed with a shuffle test