

Communities and assortativity

Max Pellert

University of Konstanz

Social Media Data Analysis

Outline

1. Structural holes

2. Assortativity

3. Permutation tests

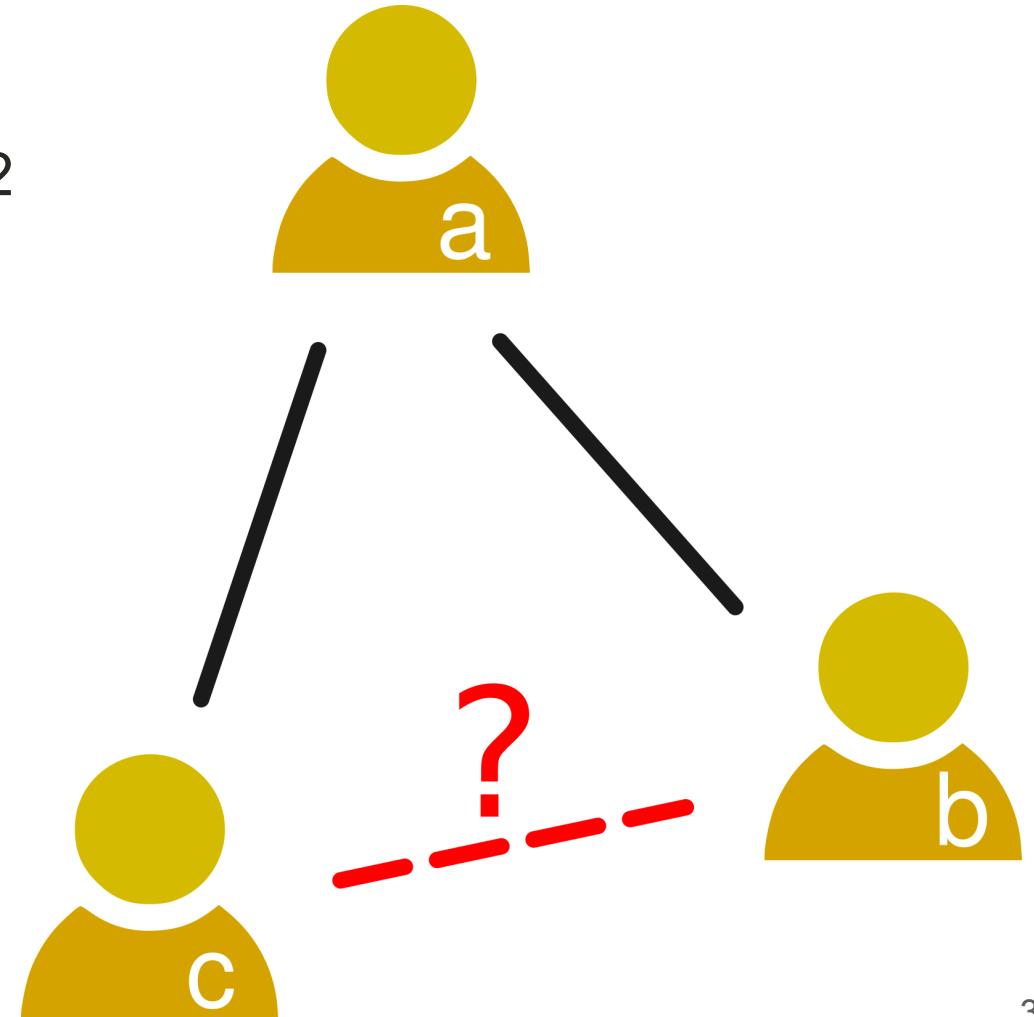
Triadic closure

Triadic closure: The tendency of individuals in social networks to create links to neighbors at distance 2

Local clustering coefficient:

$$C_i = \frac{2 * t(i)}{d_i * (d_i - 1)}$$

- $d_{out}(i)$ is the outdegree of $i (>1)$
- $t(i)$ is the number of pairs of neighbors of i that are connected
- can be averaged or computed globally



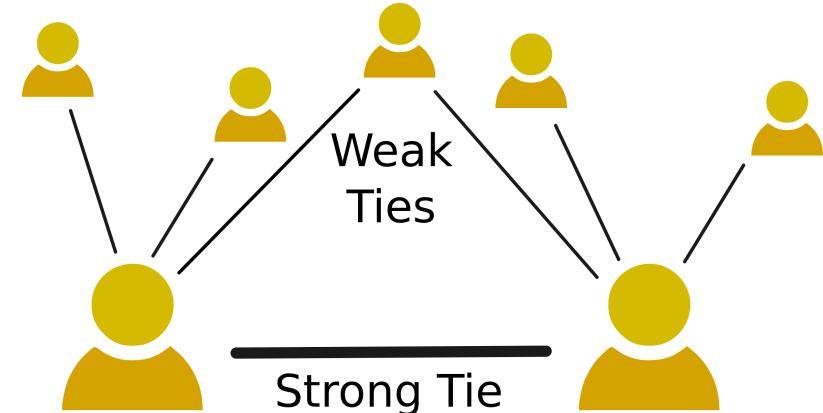
The strength of weak ties

Strong ties: People who are very close to you (family, friends, etc.)

- Provide reciprocal emotional support
- Long-lasting, require energy and time

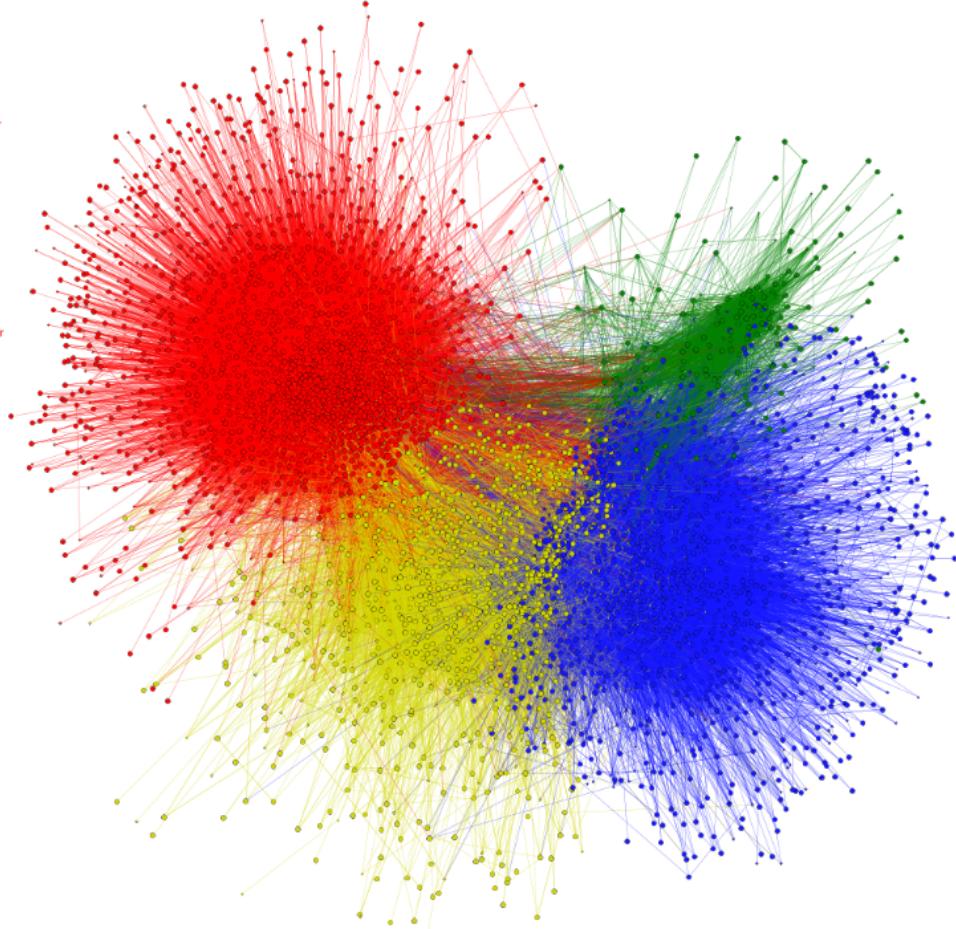
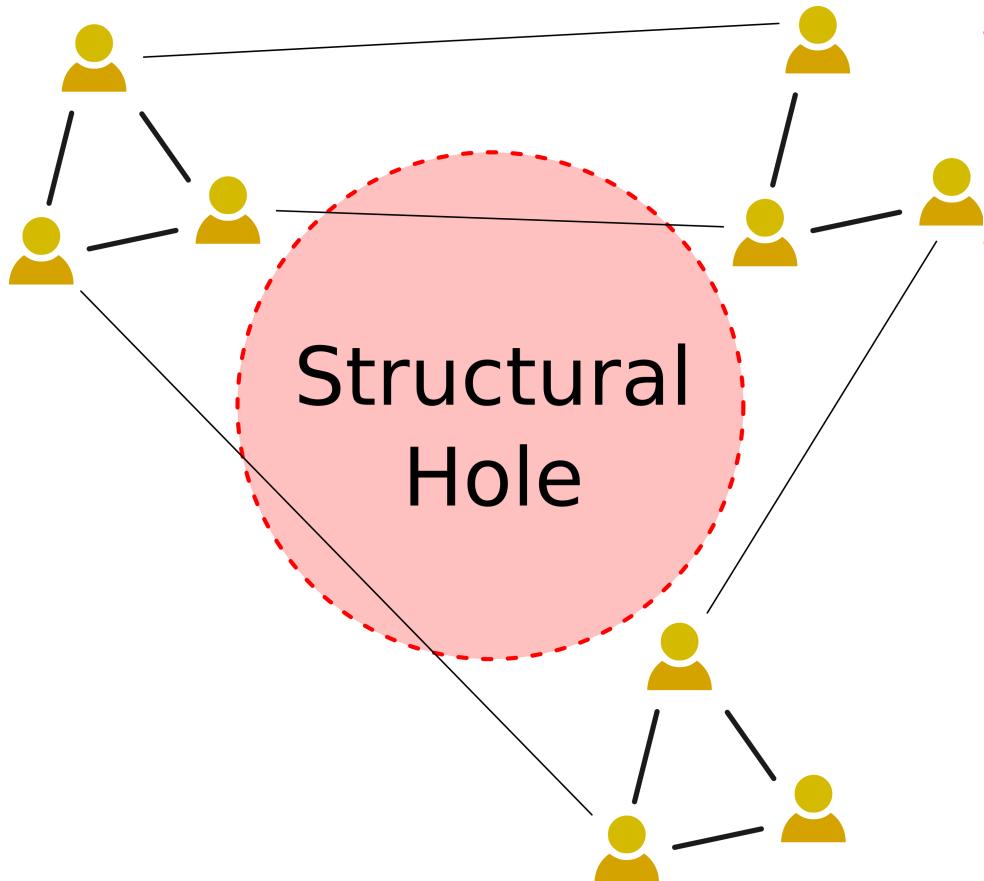
Weak ties: People you know but not very close to you (colleagues from work, school, etc.)

- Provide information, for example about employment and housing
- Are also faster to disappear



The strength of weak ties: Weak ties are weaker than strong ties, but are also many more than strong ties. (Mark Granovetter)

Structural holes and communities



When holes are bridged: The small world concept

Small world networks have the following three properties:

- Most nodes are not neighbors (sparse network)
- The mean distance between two nodes is small
- They have high global clustering coefficient

Stanley Milgram's experiment:

- 160 people in Omaha try to reach one person in Boston by mail to their acquaintances
- 44 letters reached the target with six steps on average
- Short path length as evidence of small-world
- Average path length on Facebook is 4.5



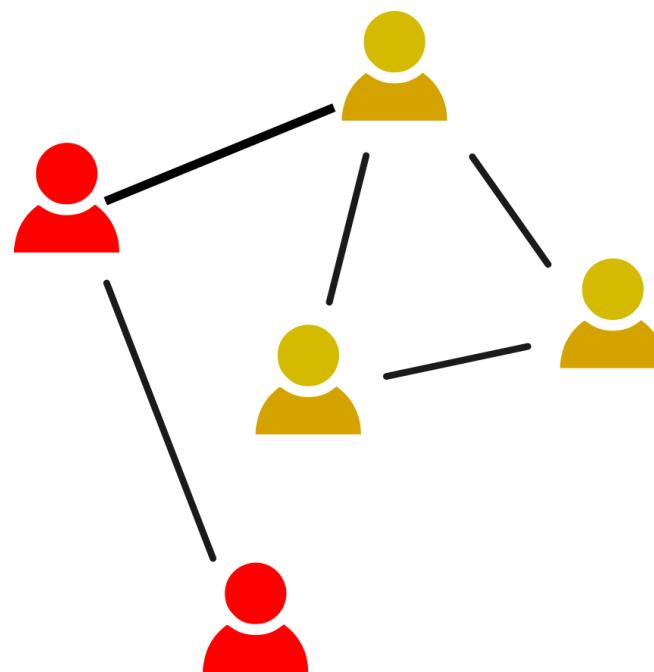
Assortativity

1. Structural holes

2. *Assortativity*

3. Permutation tests

Assortativity



Assortativity: A network property by which similar nodes are more likely to be connected than dissimilar nodes

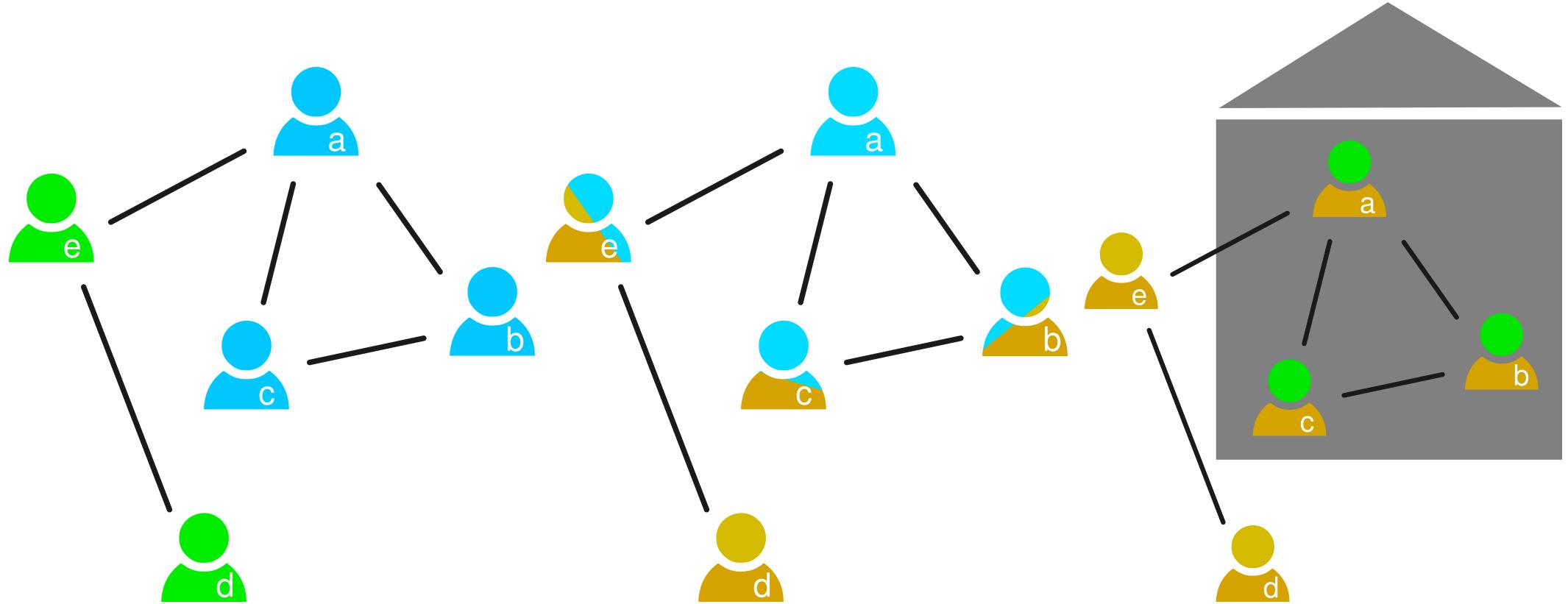
Measuring Assortativity

Similarity is based on some node property, e.g. degree, age, political orientation, gender... Can be measured with the **assortativity coefficient**:

$$r = \frac{\sum_c e_{cc} - \sum_c a_c b_c}{1 - \sum_c a_c b_c}$$

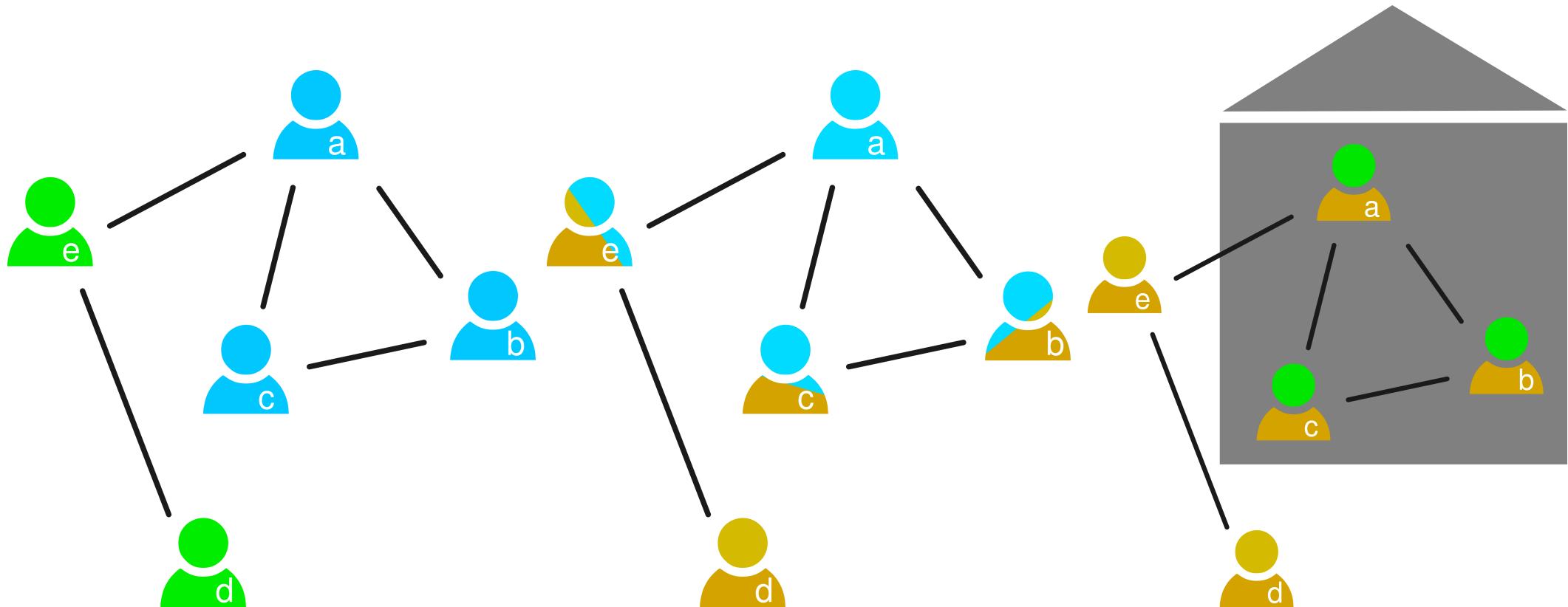
- c are the classes of nodes
- e_{cd} is the **fraction of all edges** connecting nodes of class c with nodes of class d
- e_{cc} is the fraction of all edges connecting two nodes of class c
- $a_c = \sum_d e_{cd}$, and $b_c = \sum_d e_{dc}$ (endpoint class fractions)

Three mechanisms that generate assortativity



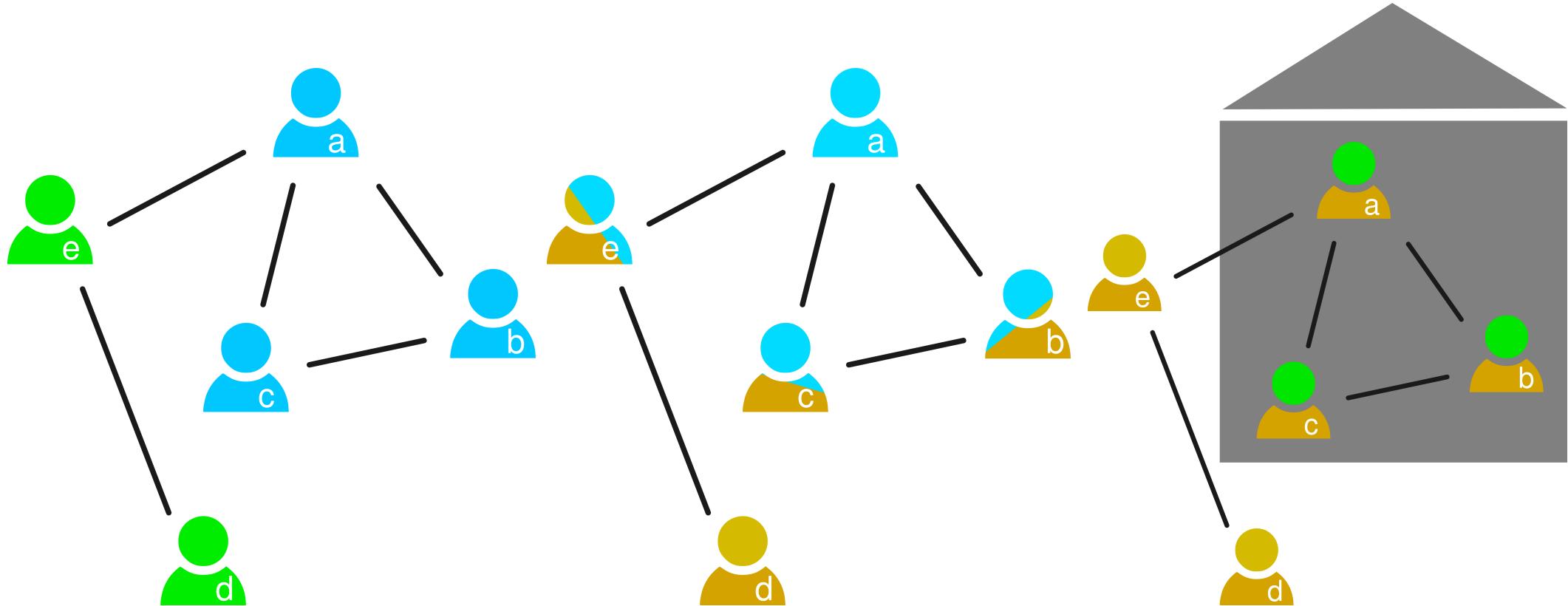
Homophily, social contagion, and confounding

Homophily



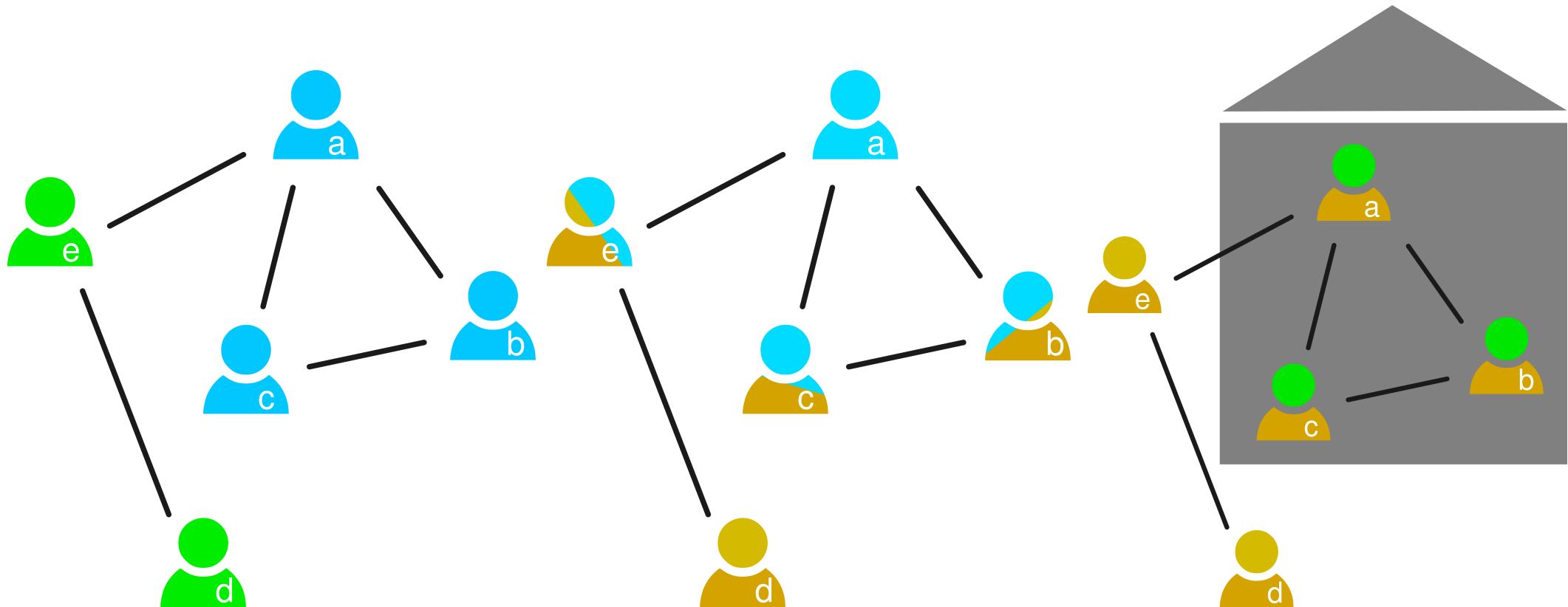
Homophily: Phenomenon in which people are more likely to create social links to similar than to dissimilar people

Social Contagion



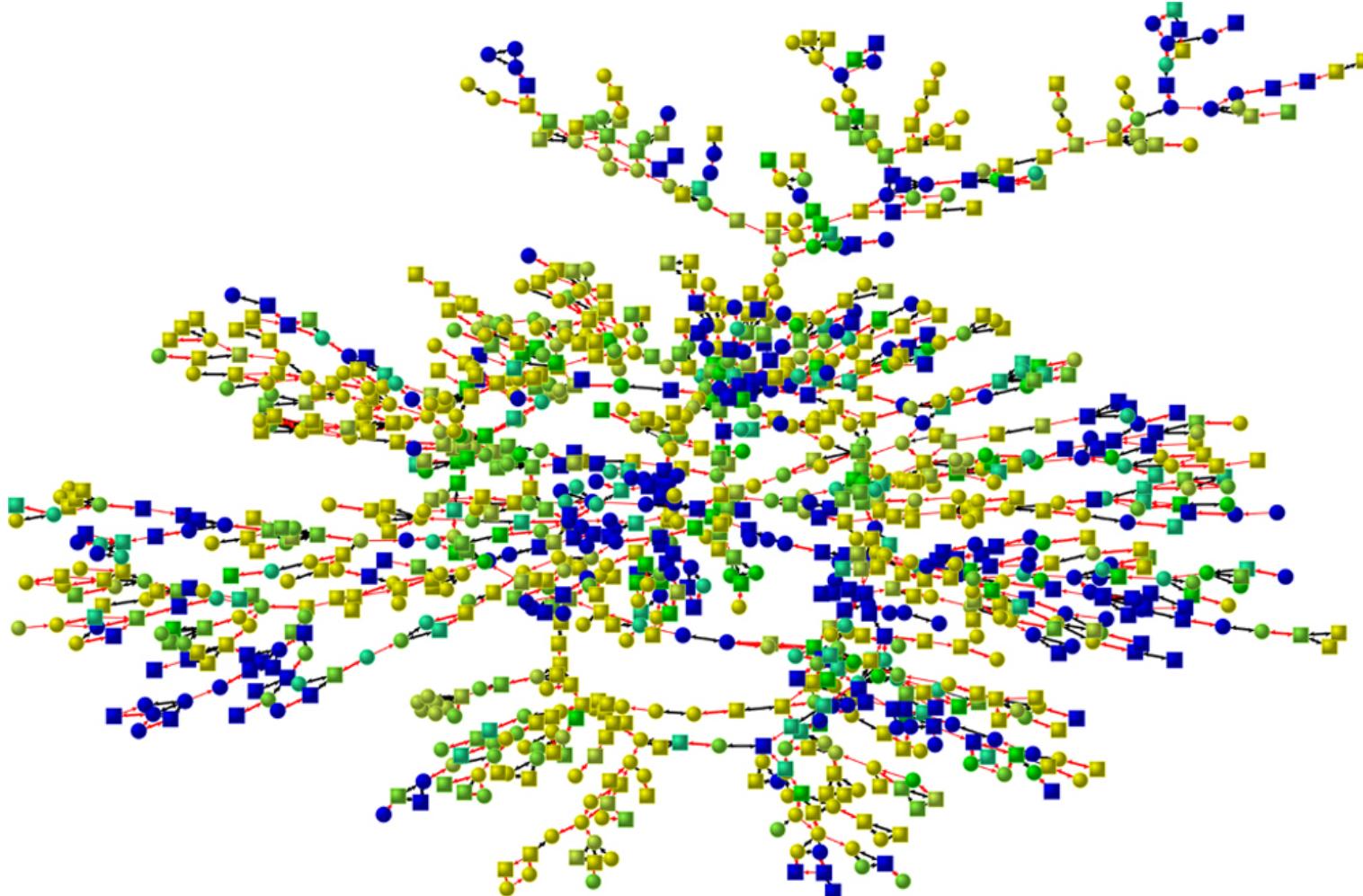
Social contagion: Phenomenon in which people tend to adopt behaviors of their social contacts

Confounding



Confounding: External effects can produce both link creation and node attributes (independent of the network topology)

Assortativity of happiness



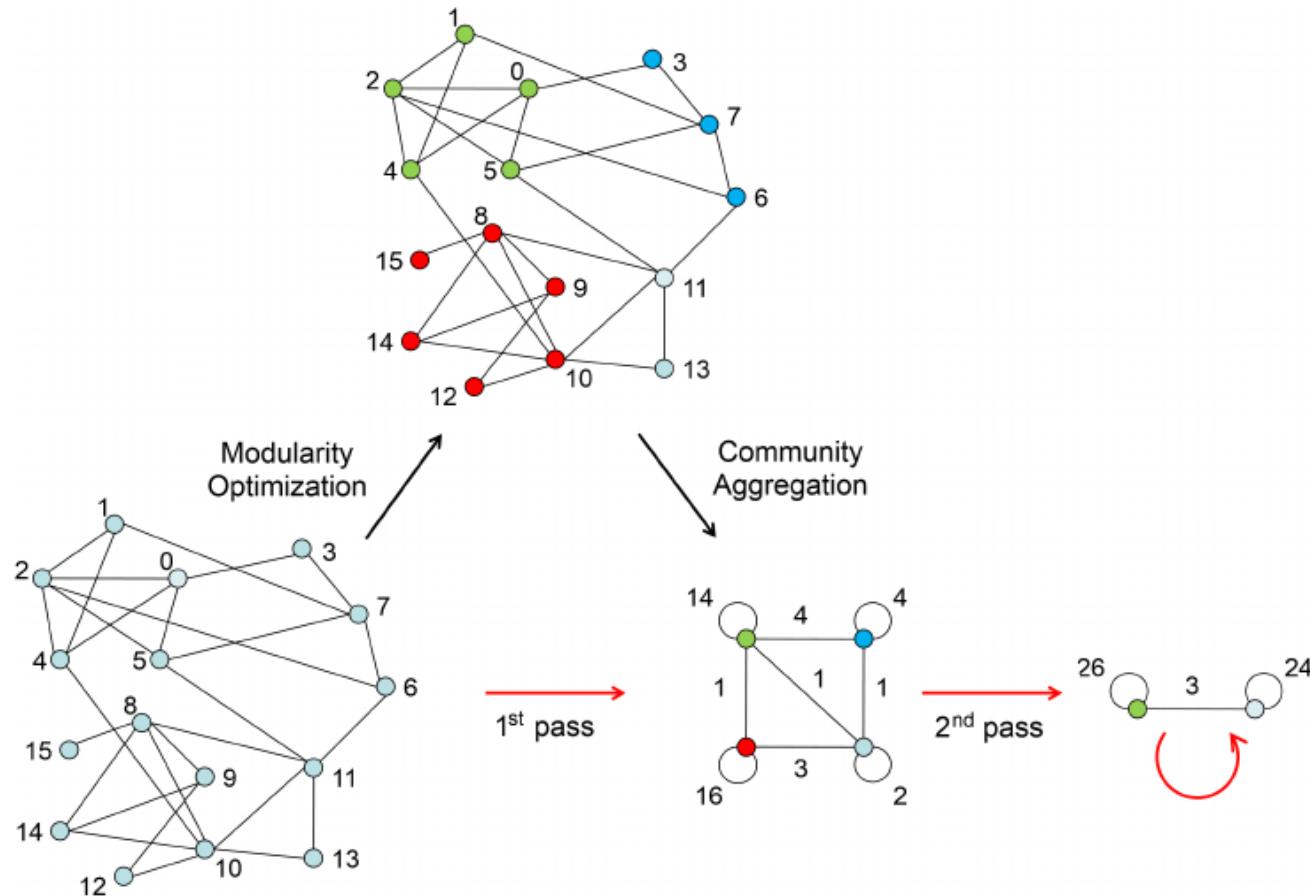
Another view on assortativity: Q-modularity

The measure of **modularity** quantifies the tendency of a network to be divided in densely connected subgroups that are sparsely connected across groups.
Q-modularity measures this tendency:

$$Q = \frac{1}{2m} \sum_{u,v} [A_{u,v} - \frac{k_u k_v}{2m}] \delta(c_u, c_v)$$

- u and v are two nodes of the network. The sum goes over all pairs of nodes.
- m is the number of edges
- A is the adjacency matrix ($A_{u,v} = 1$ if and only if u and v are connected)
- k_u and k_v are the degrees of u and v
- c_u is the community of u and c_v is the community of v
- $\delta(c_u, c_v)$ takes value 1 if and only if u and v are in the same community, otherwise it's 0

Finding community structures



Example: Louvain algorithm

Permutation tests

1. Structural holes

2. Assortativity

3. *Permutation tests*

Null models

Once we have measured something or got a statistical result from our analysis, we might ask ourselves: What is the probability that this result can appear due to chance alone?



- If the coin is fair: $P(H) = 1/2$ and $P(HHHHH) = 1/32$
- In the case of getting five heads, how plausible is that the coin is fair?

Permutation tests

- Example: correlation coefficient between two variables X and Y : $\rho(X, Y)$
- What is the probability that we could measure a similar or even stronger correlation if X and Y were completely independent?
- We can answer this questions with a permutation test
- A permutation test has four components:
 1. A **test statistic**, e.g. $\rho(X, Y)$
 2. A **null hypothesis**, which is a statement about the value of our test statistic when "nothing interesting happens".
 - It is often denoted as H_0 and in this case it would be $H_0 : \rho(X, Y) = 0$.

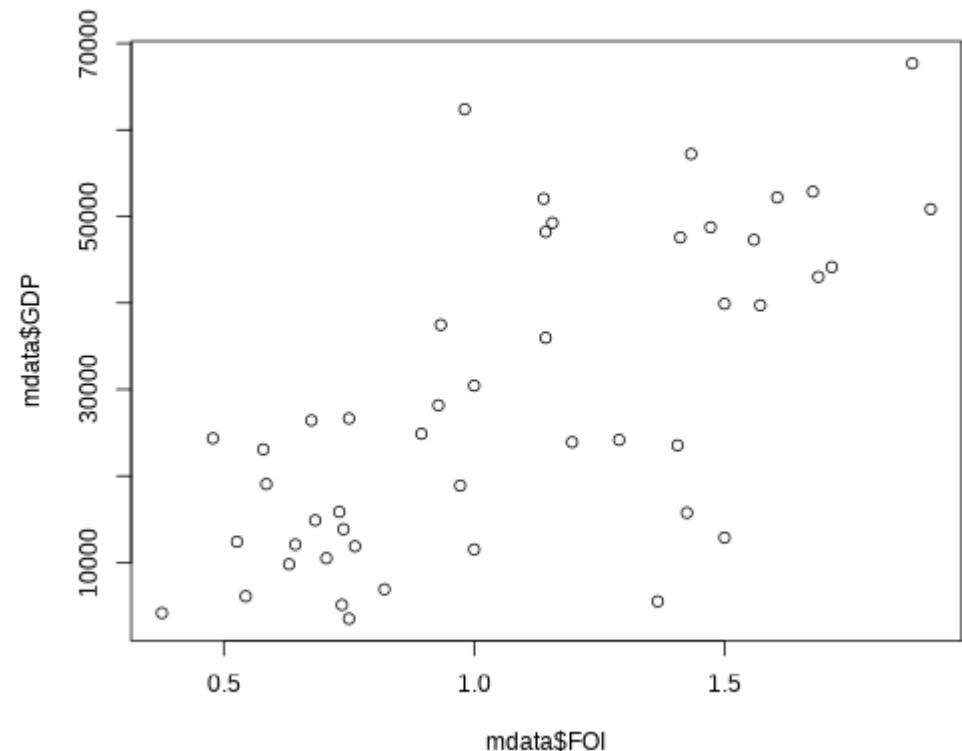
Permutation tests

1. A **test statistic**, e.g. $\rho(X, Y)$
2. A **null hypothesis**, which is a statement about the value of our test statistic when "nothing interesting happens".
 - It is often denoted as H_0 and in this case it would be $H_0 : \rho(X, Y) = 0$.
3. An **alternative hypothesis**, which is another statement about the value of the test statistic that stems from our theory or research question. It is often denoted as H_1 and, for example, it could be $H_1 : \rho(X, Y) > 0$.
 - The alternative hypothesis and the null hypothesis have to be contradictory. Both can't be true at the same time, but both could be false.
4. A **permutation set** with N random permutations or shuffles of the data.
Permutations in this set model a world in which the null hypothesis is true and any measurement different than the one specified in the null hypothesis is due to chance alone.

Permutation tests: FOI and GDP

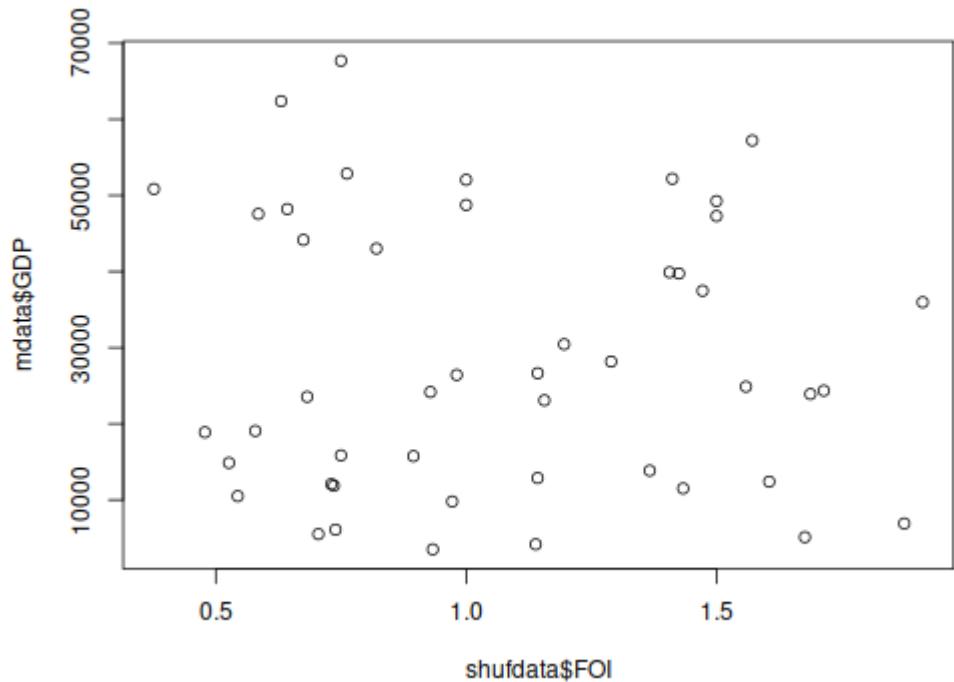
- Example: correlation between the Future Orientation Index and the GDP per capita of countries

Correlation in this example:
0.6841329

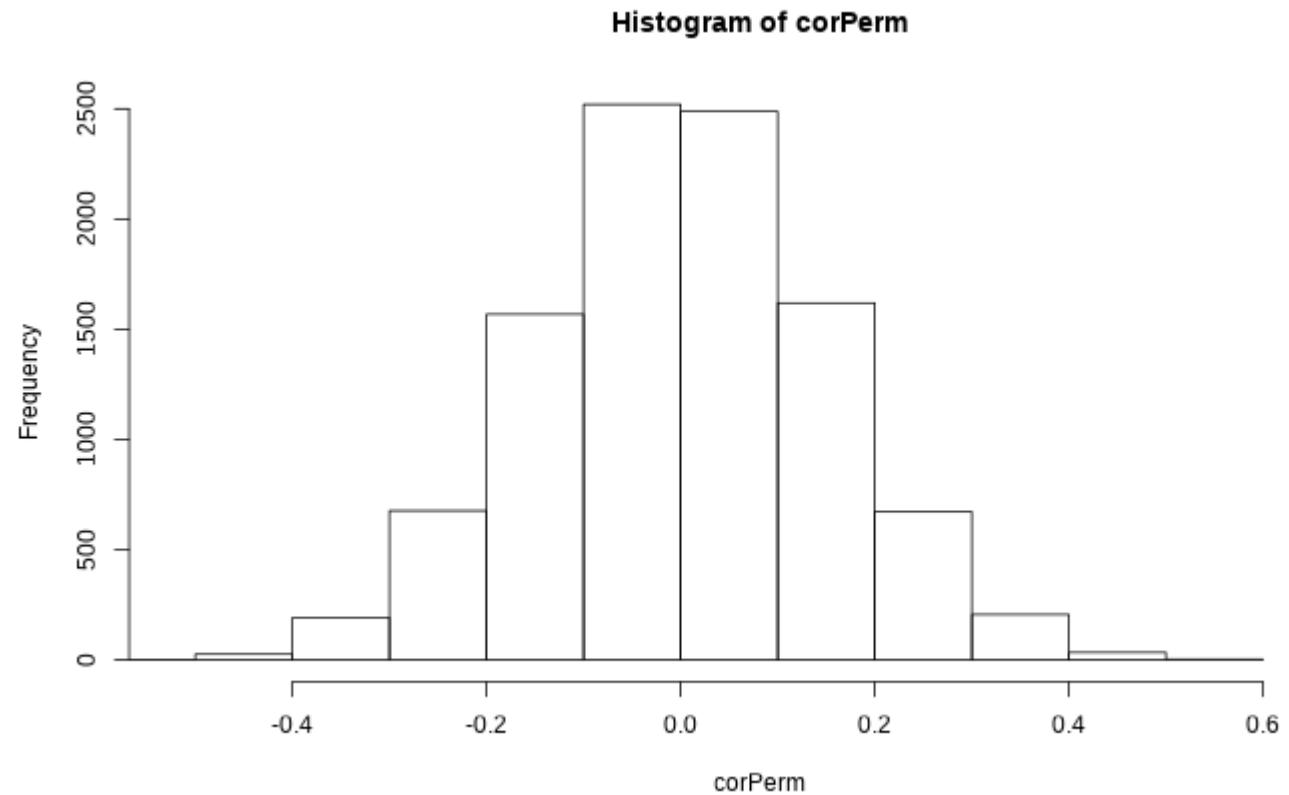


Permuted data

- If we shuffle one of the columns, the correlation will be close to zero, for example 0.08600673
- Idea: repeat this many times to estimate the chances of having a high correlation even though data was permuted



Histogram of permutation results



The red line shows the original correlation without reshuffling

The p-value of a permutation test

The p-value is a way to summarize the results of a permutation test.

p-value: Given that the null hypothesis is true, the p-value is the probability that we measure a statistic at least as extreme as the observed result

- A common misconception is that the p-value measures the probability that the null hypothesis is true.
- The p-value measures the plausibility of what we measure under the null hypothesis, which is very different.

Calculating a p-value

- *one-sided p-value*: proportion of permutations with a value of the statistic at least as large as the observed one.
- *two-sided p-value*: proportion of permutations with an absolute value of the statistic at least as large as the observed one.

We can calculate the one-sided p-value from the permutation test:

```
p_value_Cor = (sum(corPerm>=corObserved)+1)/length(corPerm)
```

- We add one to the numerator to ensure we don't say zero
- The result is "significant" ($p<0.05$), but careful interpreting that word!
- How many times to permute: 10000 times is safe for ($p<0.05$)

Summary

- **Social network structures**
 - Triadic closure
 - The strength of weak ties
 - Structural holes and communities
 - Small worlds
- **Assortativity: when nodes have attributes**
 - Assortativity coefficient
 - Three processes to generate assortative networks
- **Permutation tests**
 - How plausible is that I get a result as extreme as this purely by chance?
 - Permuting to simulate a null model
 - P-values as a way to summarize the result