GENEVA GRADUATE INSTITUTE

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

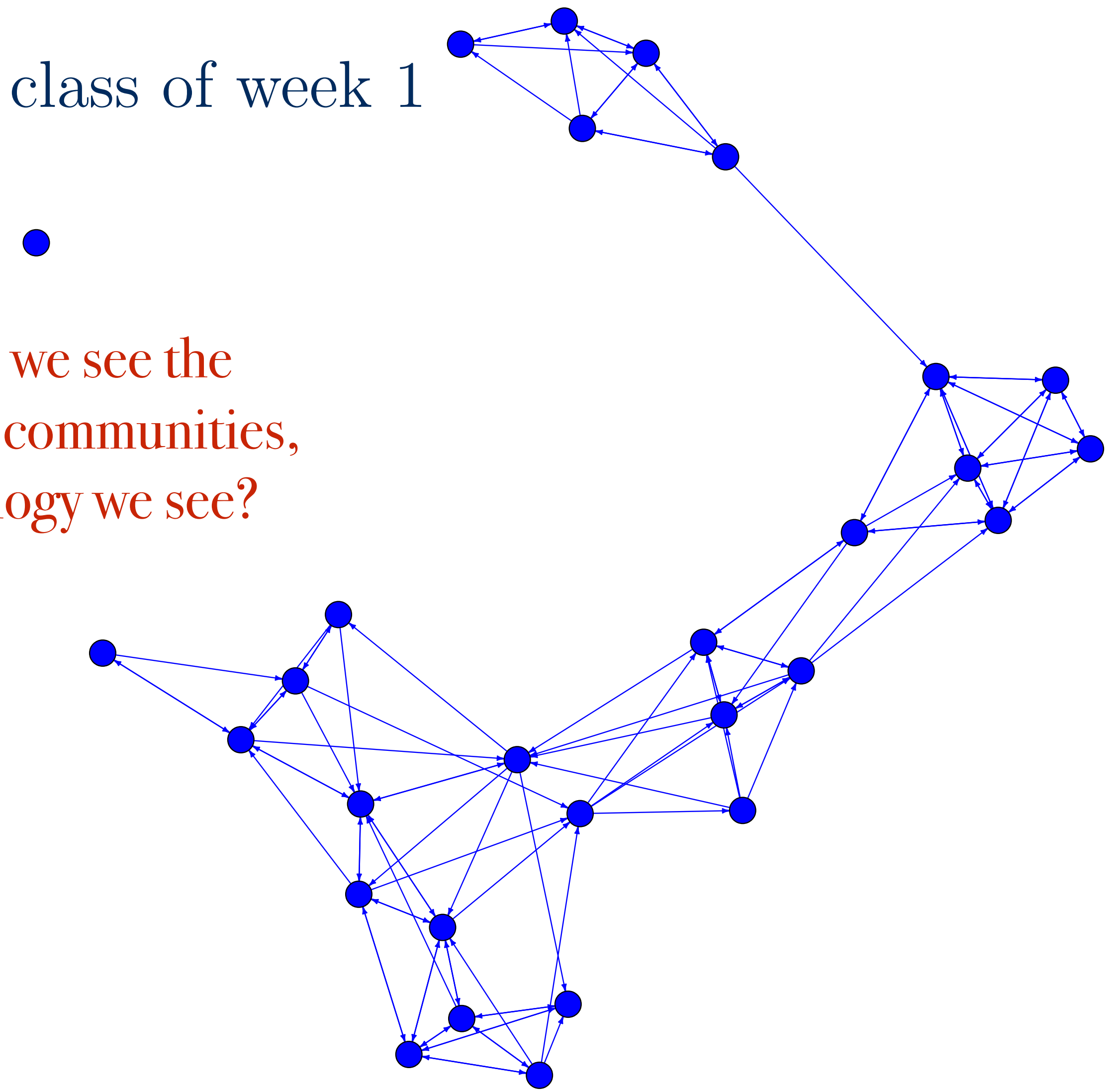
Testing

Introduction to Social Networks

James Hollway

The school class of week 1

Why do we see the
clustering, communities,
and topology we see?

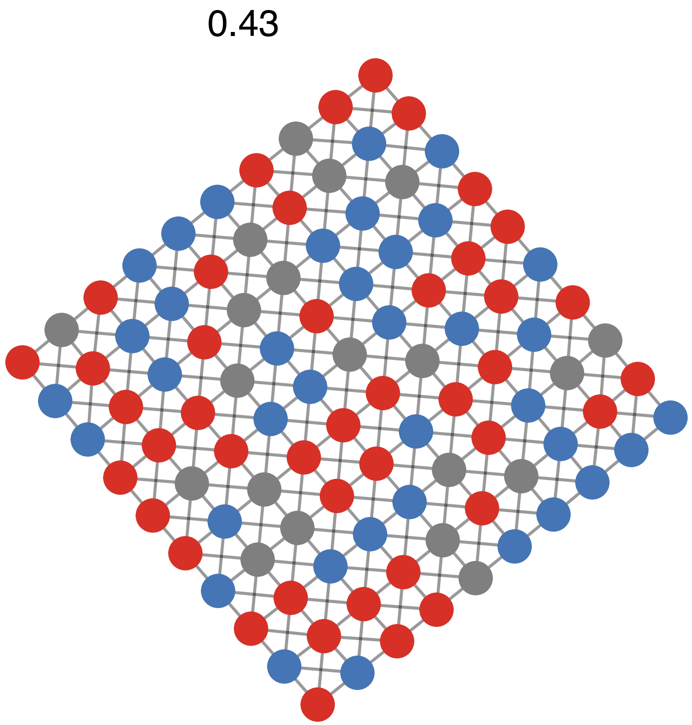


Shelling's simple model of segregation

- Thomas Schelling worked on multiple topics, including strategic interactions (game-theory)
  - Won Nobel Prize in 2005 and died in 2016 (aged 95!)
- One of his best books is called "Micromotives and Macrobehaviour" and treats the unintended consequences of individual action - you should read it!
- Here we will discuss one particular example from that book and link it to social networks

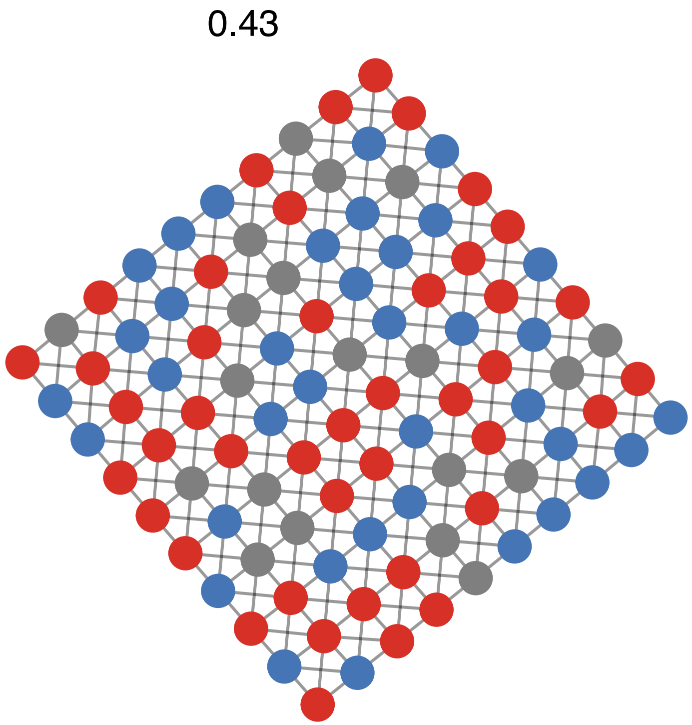
Schelling 1978, 2006

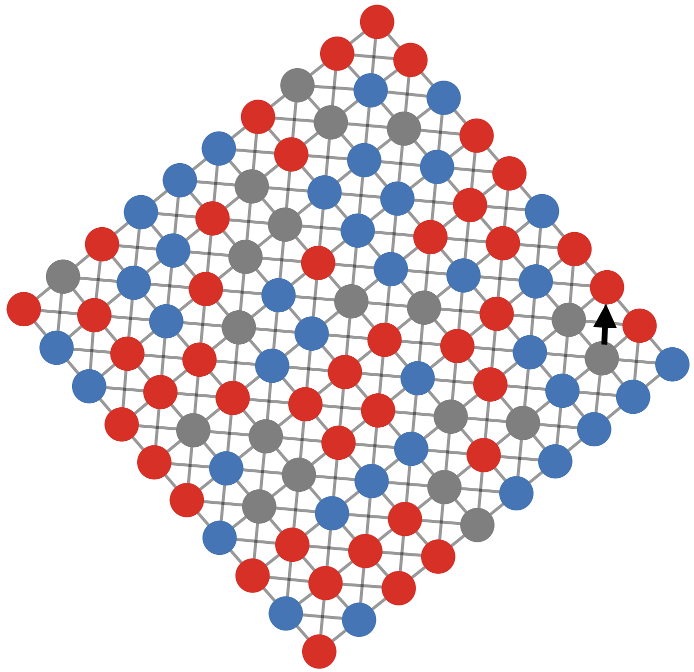
A chessboard model of residential segregation



- Each checker occupies a node on a lattice network, such that they have 3 (in the corners) up to 8 (in the middle) neighbours
- Each checker can be coloured red or blue (grey are unoccupied nodes), distributed at random with equal probability
- Most checkers are located near a mix of checkers of their own and the other color, with the result that they are more among others (0.43)

A chessboard model of residential segregation



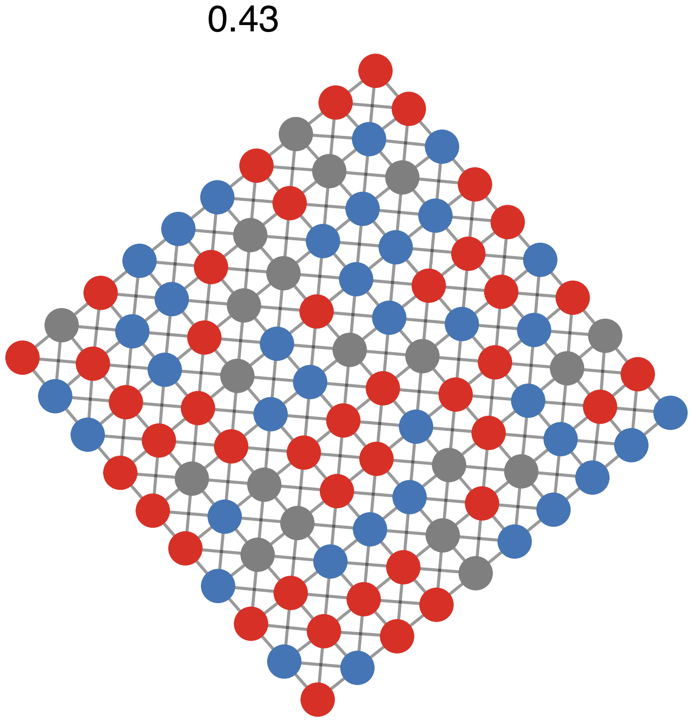


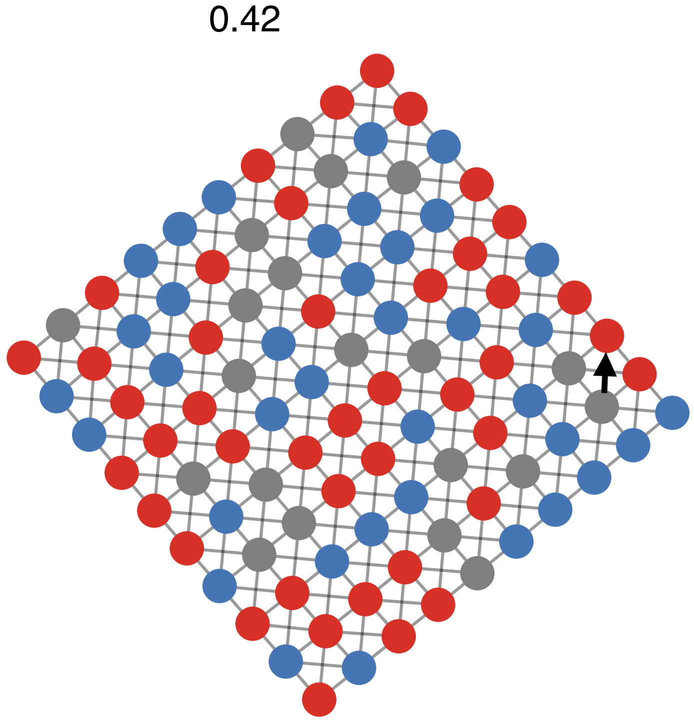
1 step in

- Let's say both red and blue checkers fairly happy with heterogeneity, just don't want to be *only* checker their colour around…
    - So happy to have 0.5 heterogeneity in their local environment (i.e. ¾ other)
- Simulation identifies a checker dissatisfied in local environment and offers them to move to next available space that satisfies their preference
- E.g. here red checker on node 19 moves to the vacant node 8, moving it from having 5 blue/1 red neighbours, to having 2 red / 1 blue neighbours…

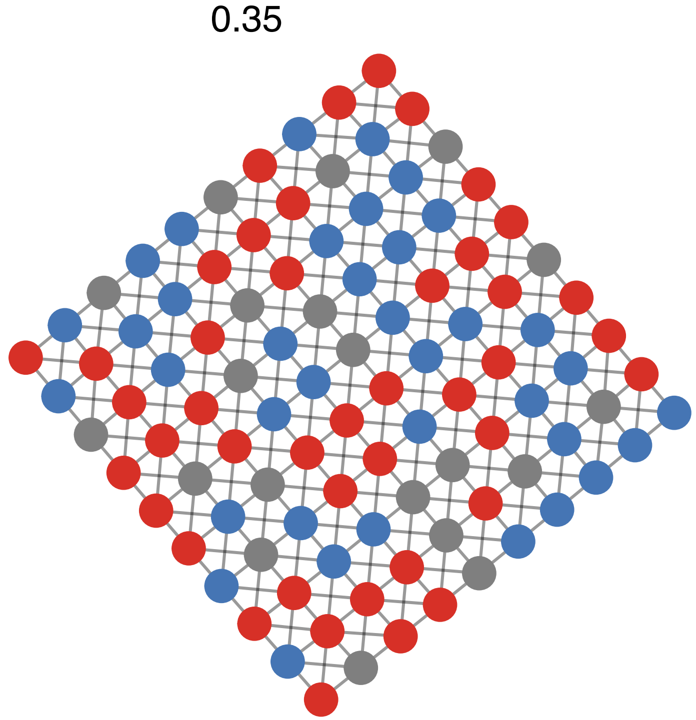
…what happens?

A chessboard model of residential segregation





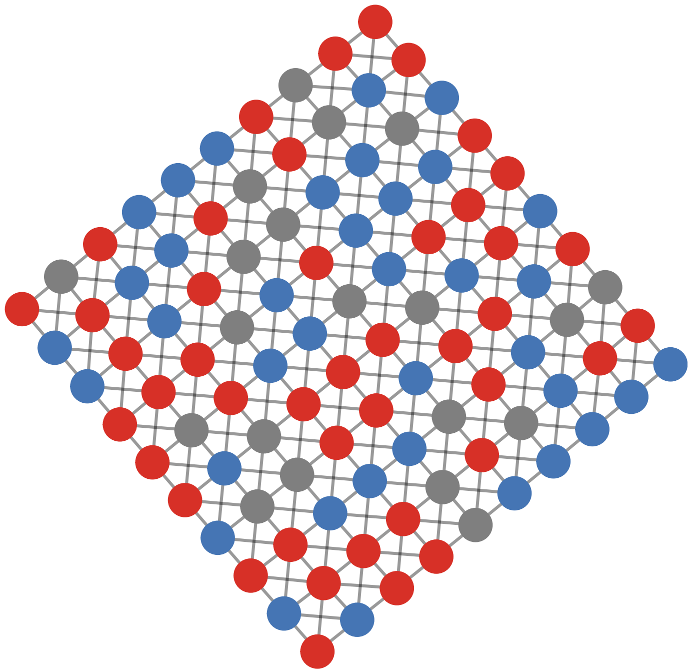
1 step in



50 steps in

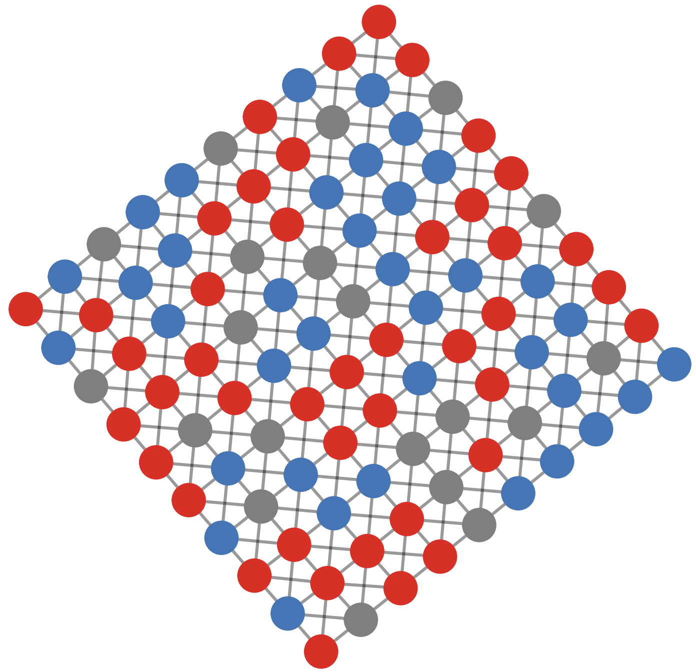
A chessboard model of residential segregation

0.43



Although checkers hold no prejudice *against* living around checkers of the other colour, micro-motives nonetheless result in segregation, an unintended macro-consequence that none of the actors desired

0.35



50 steps in

Lesson #1

Macro outcomes are a
result of micro motives

Coleman's boat

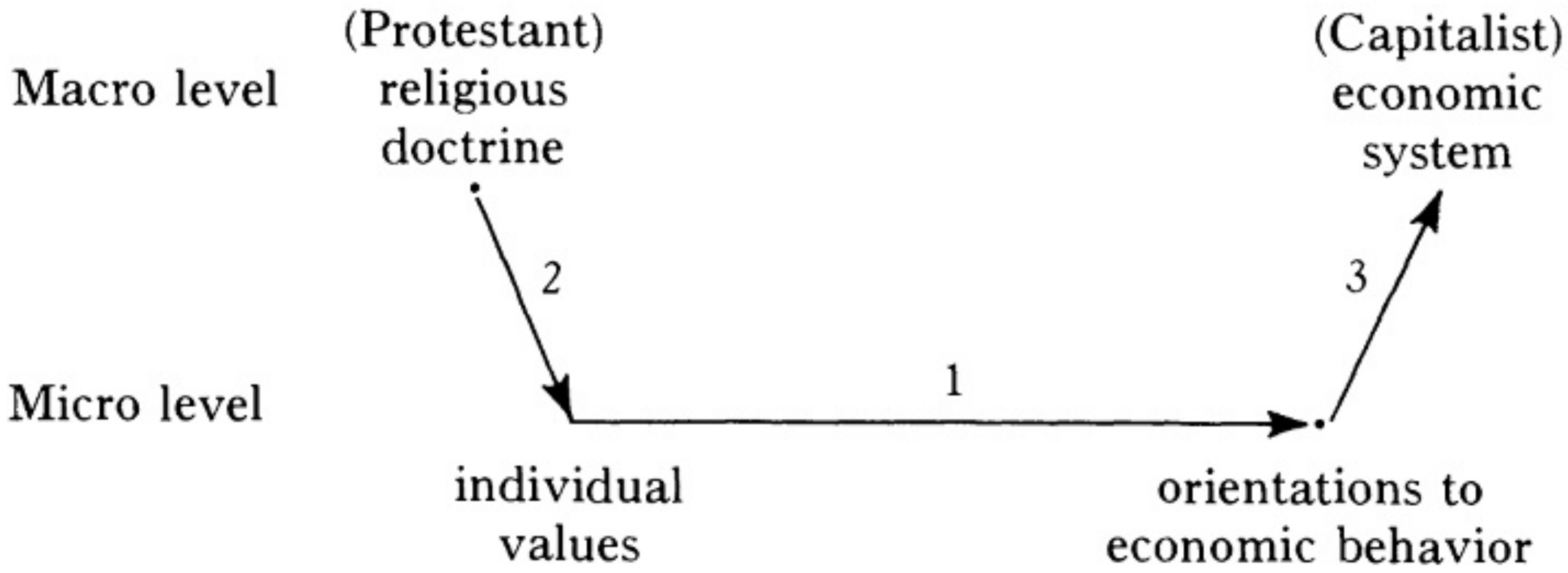
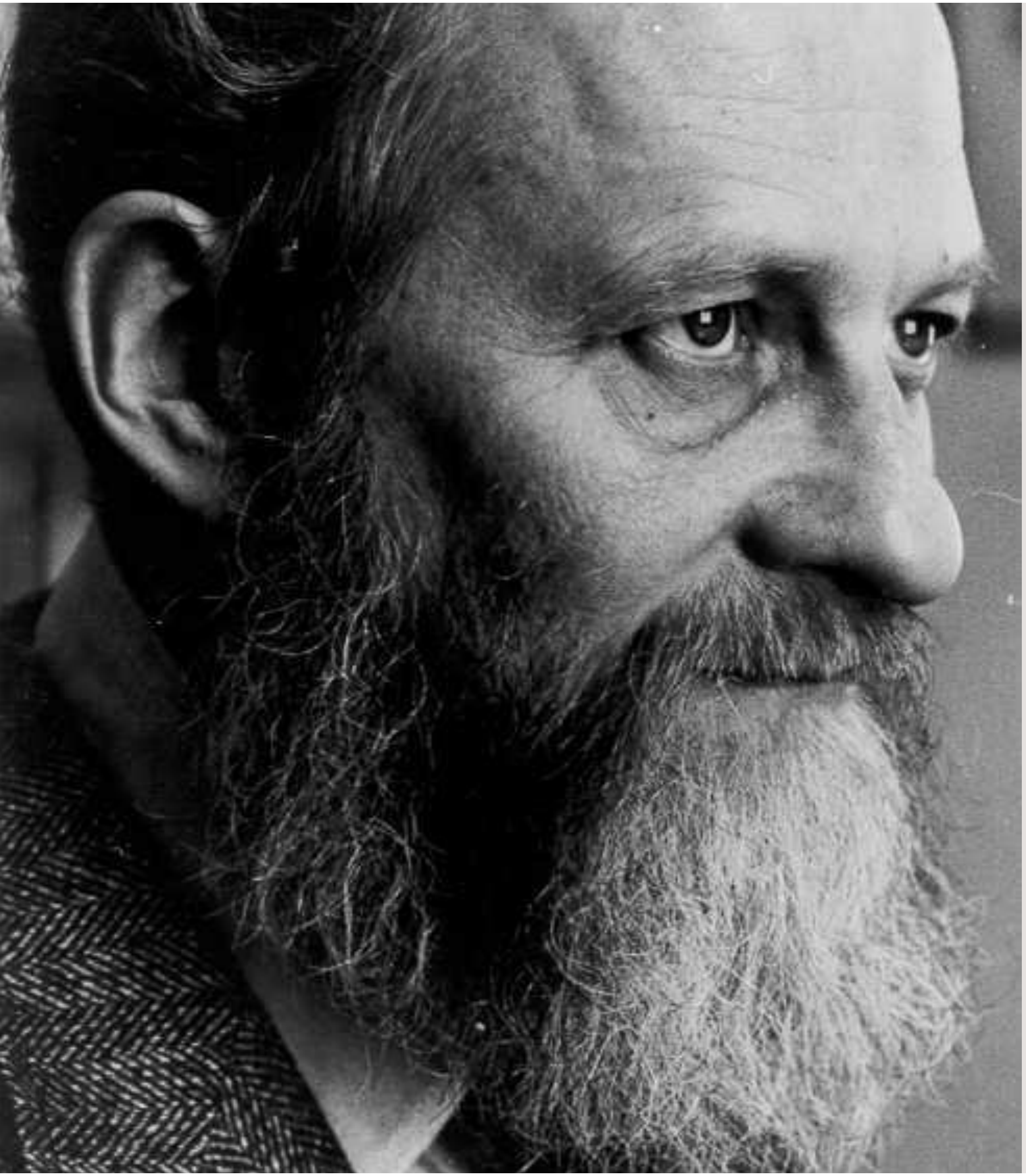


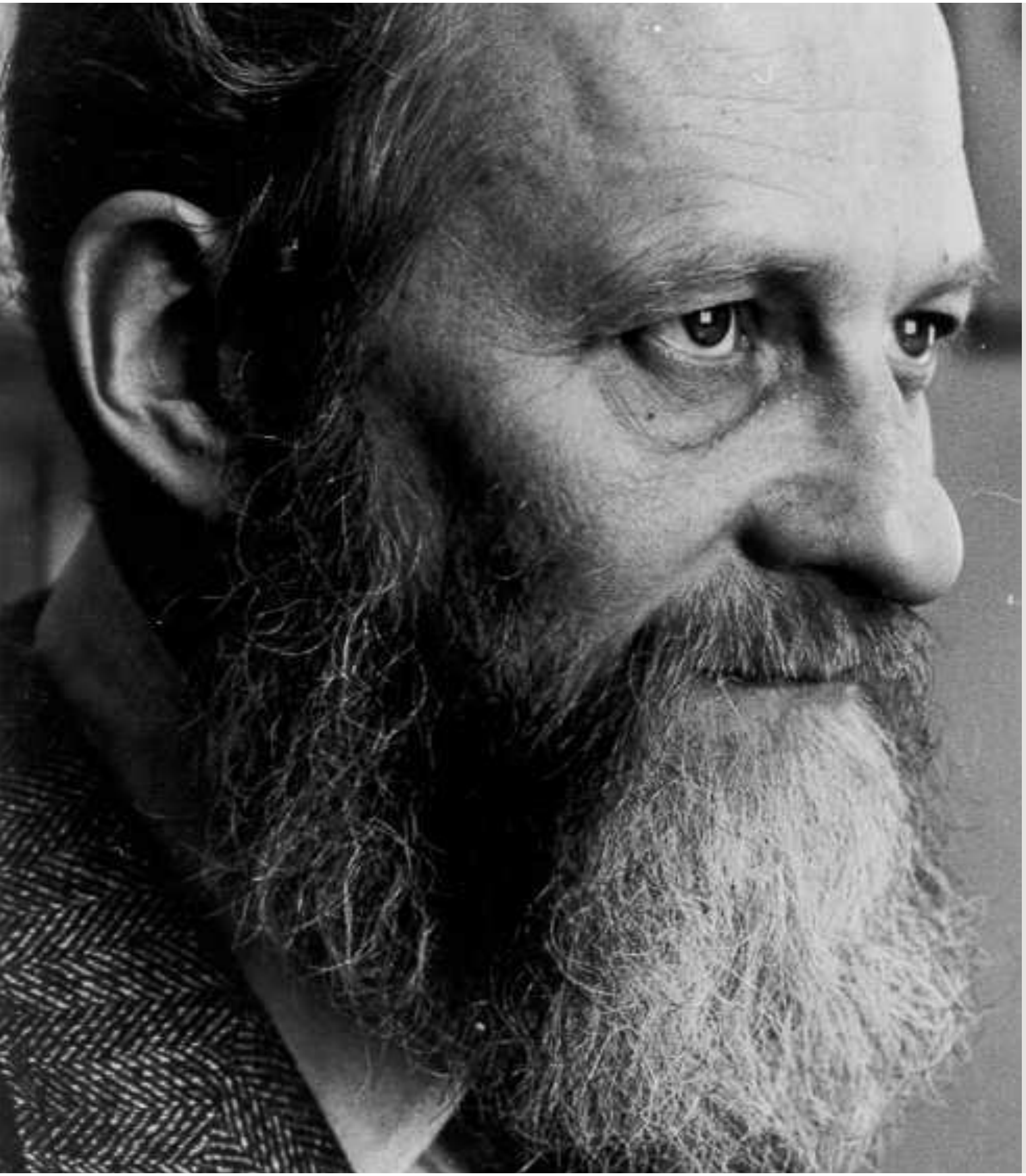
FIG. 2.—Macro-micro-macro relations: methodological individualism



Blau Index

$$1 - \sum p_i^2$$

- An index of variety or diversity of an attribute within a network or groups
- Based on the probability that two entities taken at random are the same (different)
- With or without 1-, known as the:
  - *(Gini-)Simpson Index* (statistics)
  - *Probability of Interspecific Encounter* (ecology)
  - *Hunter-Gaston Index* (microbiology)
  - *Herfindahl(-Hirschman) Index* (economics)
  - *Gibbs-Martin Index* (sociology)



Blau Index

$$1 - \sum p_i^2$$

- An index of variety or diversity of an attribute within a network or groups
- Based on the probability that two entities taken at random are the same (different)
- With or without 1-, known as the:
  - *(Gini-)Simpson Index* (statistics)
  - *Probability of Interspecific Encounter* (ecology)
  - *Hunter-Gaston Index* (microbiology)
  - *Herfindahl(-Hirschman) Index* (economics)
  - *Gibbs-Martin Index* (sociology)
- Distributional, but no structure

Homophily vs Heterophily

- **Homophily** means actors tie to those who are the same on some socially salient attribute more often than by chance
  - Status homophily: similar characteristics
    - Ascribed: ethnicity, age, gender?
    - Acquired: religion, education, occupation
  - Value homophily: similar preferences
    - e.g. political, sexual, musical preferences

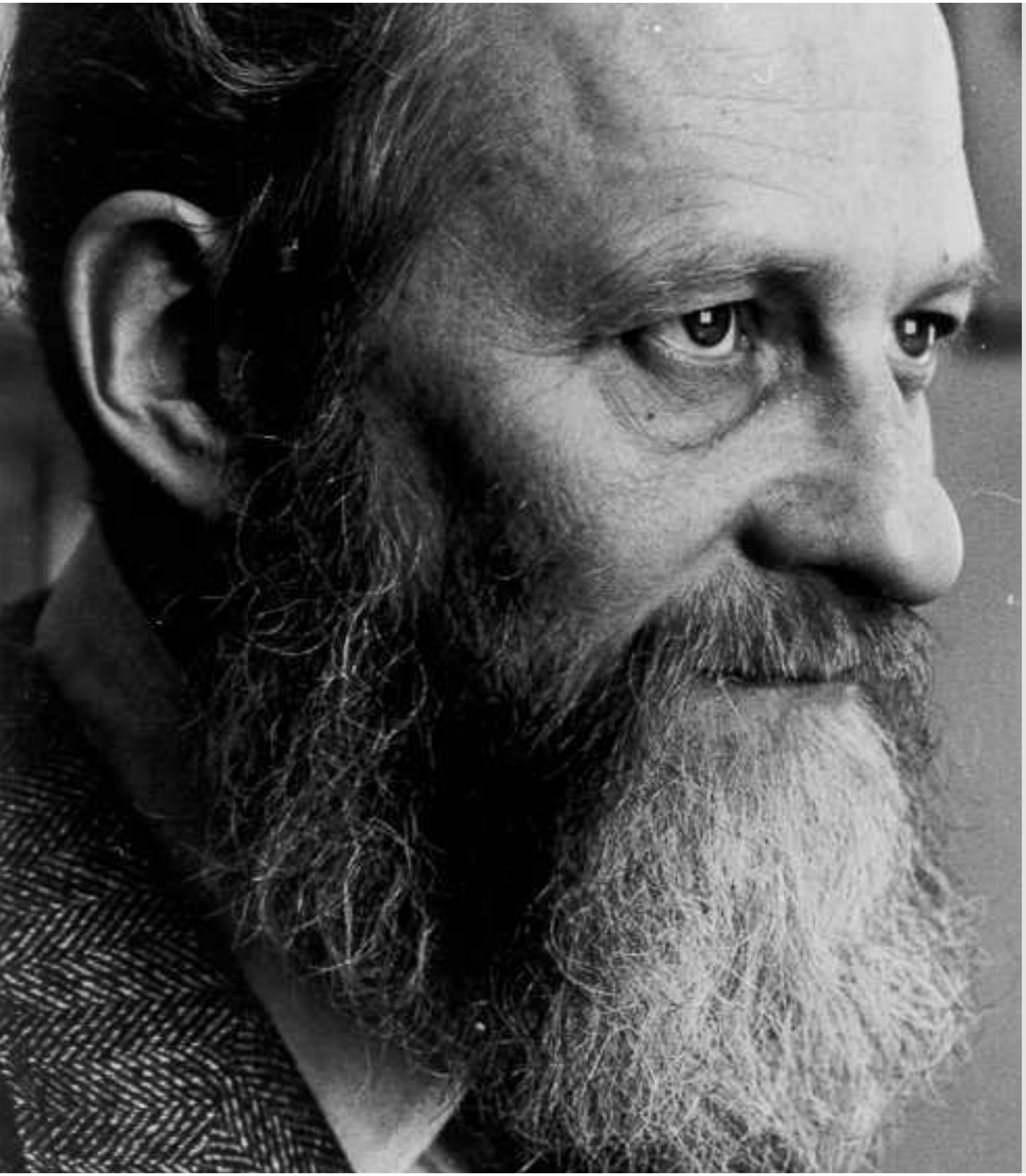
- **Heterophily** is when actors prefer those who are different on such attributes

Lazarsfeld and Merton 1954, Block and Grund 2014…
see also Plato [1968], Aristotle [1934]

Shared homophily

- English: *"Birds of a feather flock together"*
- French: *"Those who resemble each other assemble with each other" (Qui se ressemble s'assemble)*
- Italian: *"God makes them then couples them" (Dio li fa e poi li accoppia)*
- Japanese: *"Racoon dogs from the same den" (Onazi ana no mujina)*

PS: I'm collecting such translations, so if there are more that you are aware of in other languages, please let me know!



“One cannot marry an eskimo,
if no eskimo is around.”

–Peter Blau

“One cannot marry an eskimo,
if no eskimo is around.”

–Peter Blau

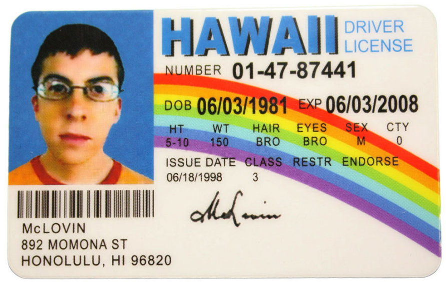
Cognition: Choice vs Induced

- Choice homophily: people associate disproportionately with similar others because human beings prefer (for rational or irrational reasons) similar others
  - I.e. if choice homophily, then if people enter a room with similar and dissimilar strangers, they will seek those who are similar and avoid those who are dissimilar

- Induced homophily: people form social ties with the people they encounter, and those whom they encounter are those that are similar (not because of any particular psychological preference)
  - I.e. if induced homophily, then if people enter a room with only similar strangers, then they will make relationships with similar people even if they do not have a particular preference

McPherson & Smith-Lovin 1987

Baseline vs Inbreeding



- The qualifier "more often than by chance" is crucial, because chance alone may explain the emergence of a great deal of similarity, especially when structured
- Baseline homophily created by the demography of the potential tie pool; level of homophily expected from random mixing in the population
- Inbreeding homophily explicitly over and above the opportunity set; level of homophily in excess of that baseline

McPherson, Smith-Lovin, and Cook 2001

E-I Index

- Krackhardt and Stern's (1988) E-I index is a simple descriptive of homogeneity in a network

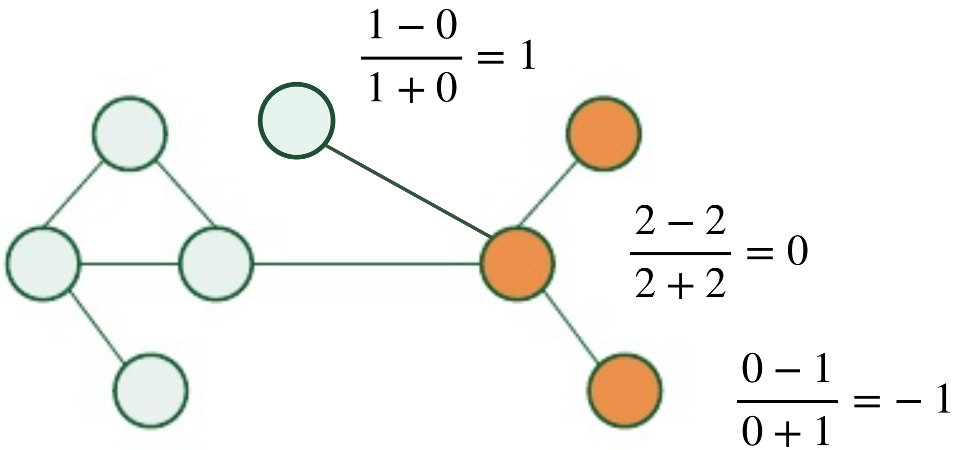
$$EI(x) = \frac{E - I}{E + I}$$

  - Measures the difference between the number of external (E, between-group) and internal (I, within-group) ties, normalised by the total number of ties
  - Can be on the network, group, or individual level

- What is the range?

- What is the expected value without choice homophily?

E-I example

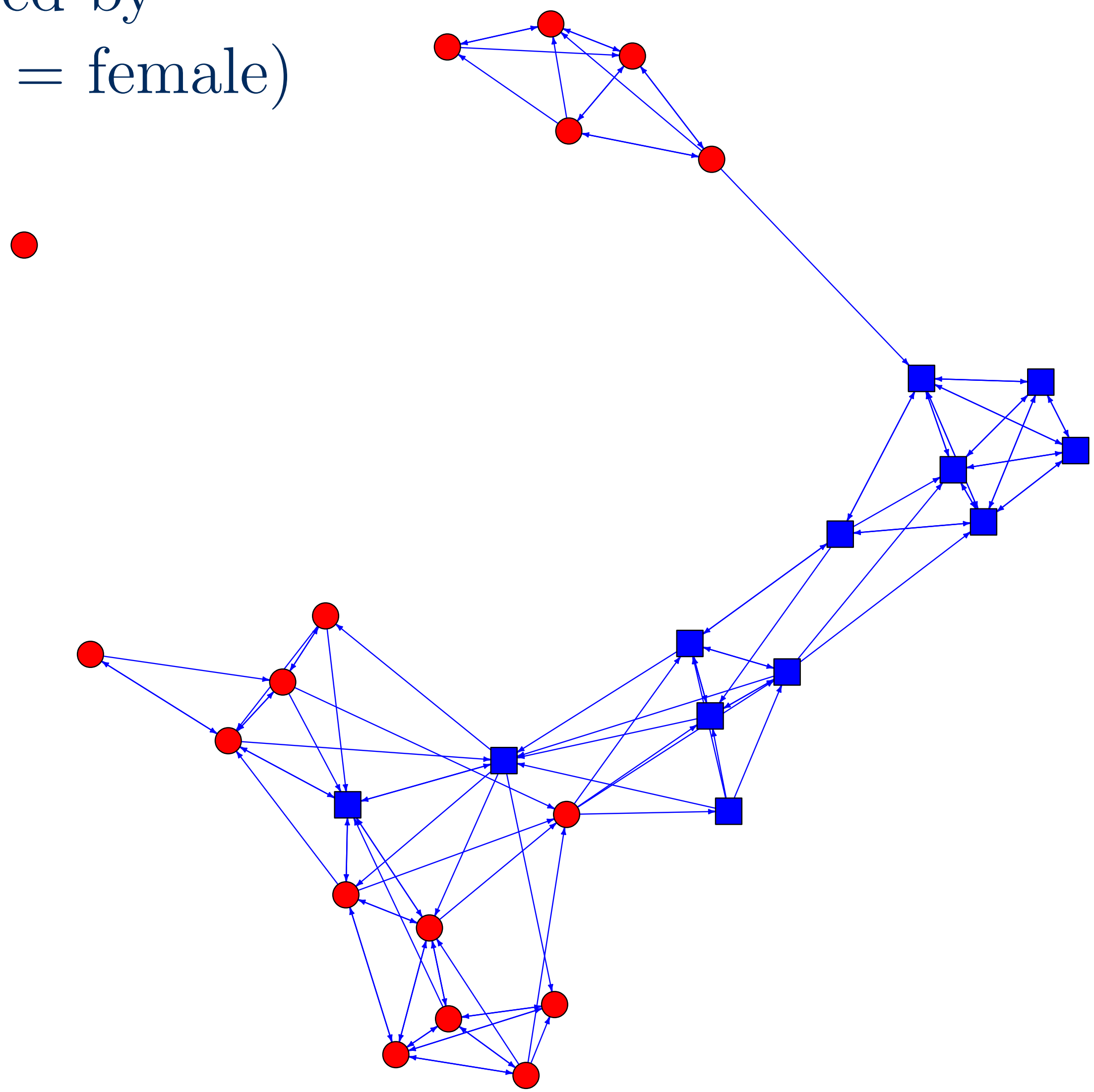


$$\frac{2-6}{2+6} = -0.5$$

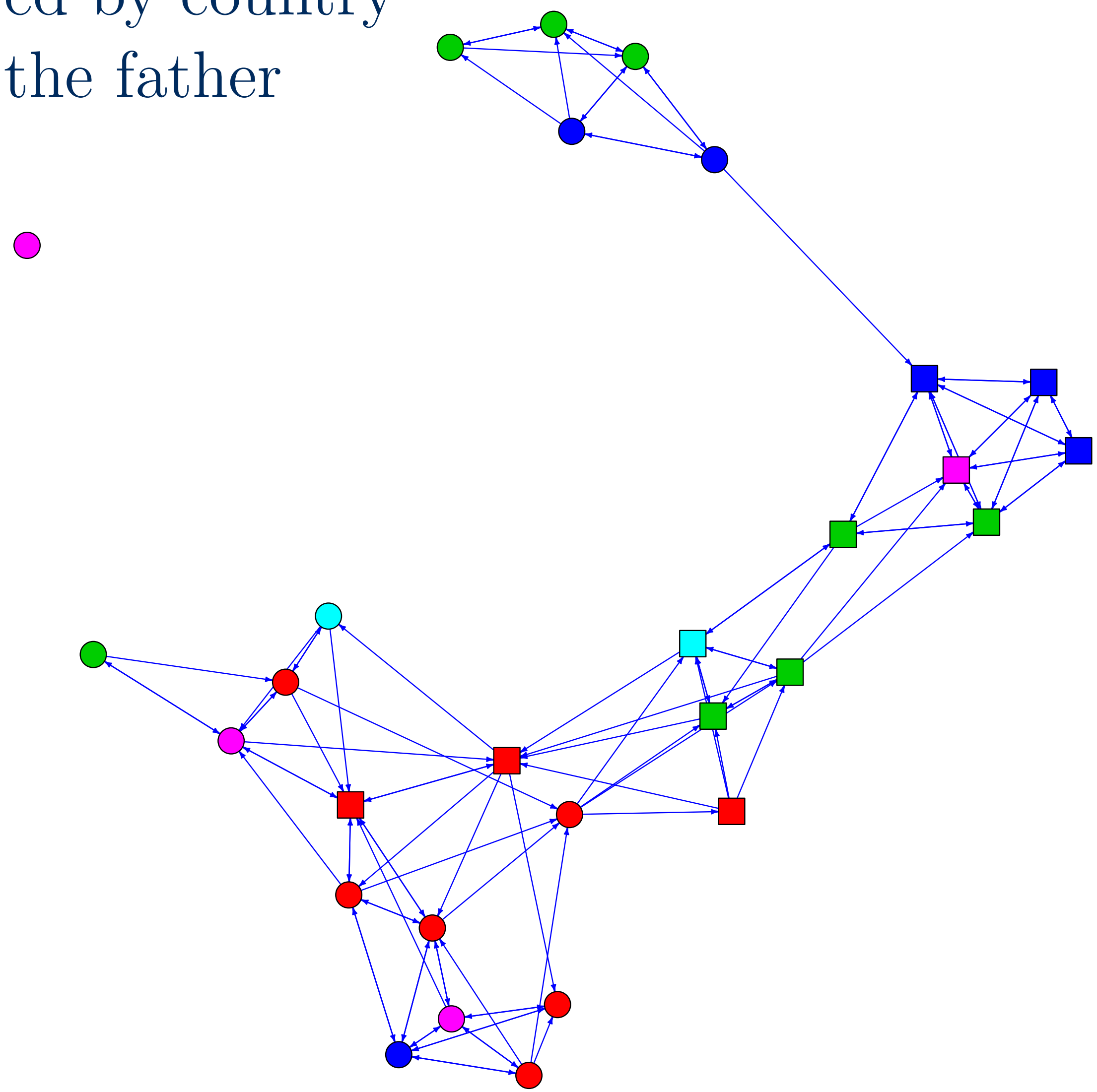
Lesson #2

Isolating mechanisms such as choice/inbreeding homophily is difficult

Nodes colored by
gender (red = female)

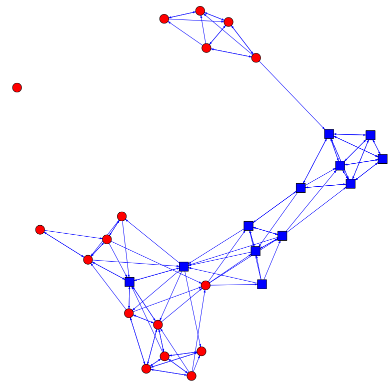


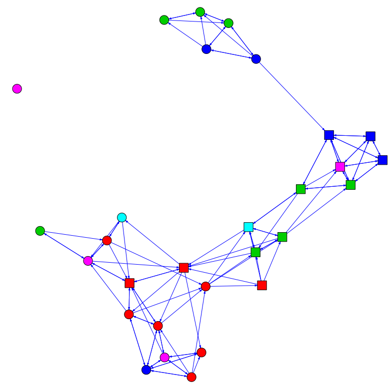
Nodes colored by country
of origin of the father



E-I Index of the classroom network

- Gender
  - External ties: 19
  - Internal ties: 92
  - $EI_{gender}(x) = -0.66$
- Ethnicity
  - External ties: 71
  - Internal ties: 40
  - $EI_{ethnicity}(x) = 0.28$
- What do we learn from these numbers?



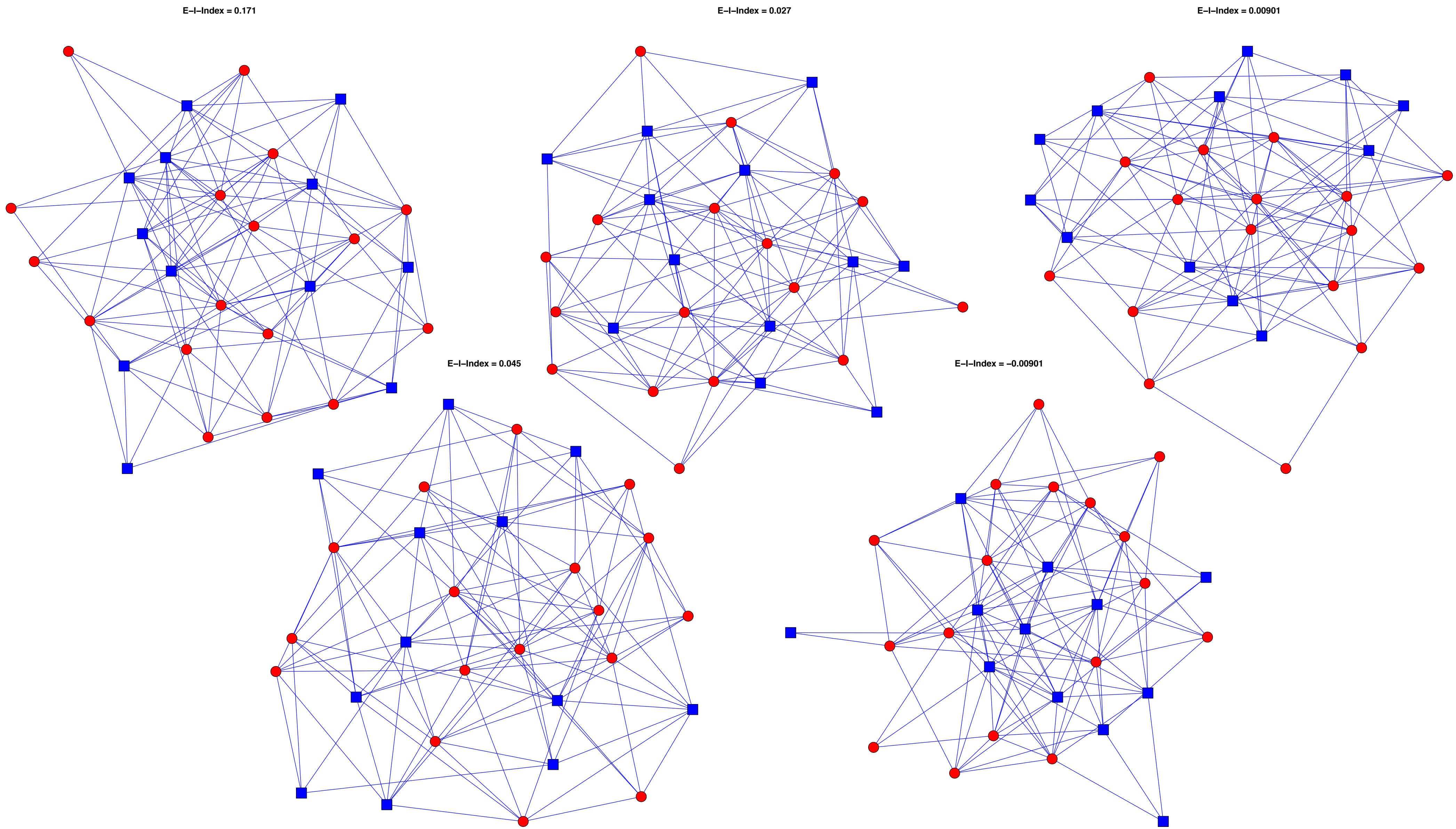


But this is just a score…
Is this more or less than we
should expect?

Option 1: obtain a baseline from random graphs

Null Hypothesis I: Random Graphs

1. Calculate E-I index of network x
2. Generate 1k random networks with same density and distribution of attributes as x
3. Calculate E-I index of random networks with given attribute
4. Identify where E-I index of observed network lies in the distribution of networks
5. If E-I index unlikely to be observed randomly, suggests some homophily/heterophily

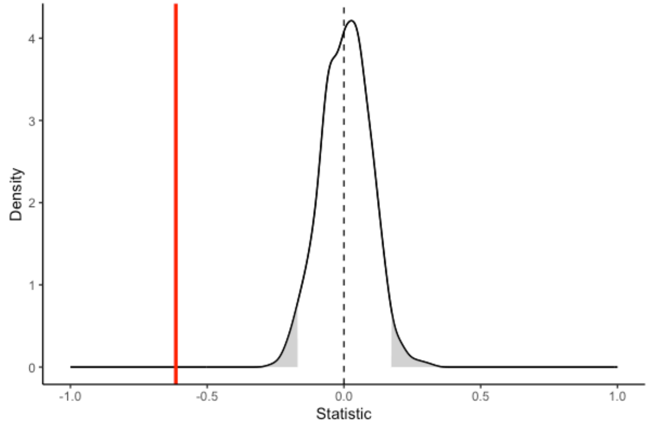

E–I–Index = 0.171
E–I–Index = 0.027
E–I–Index = 0.00901
E–I–Index = 0.045
E–I–Index = −0.00901

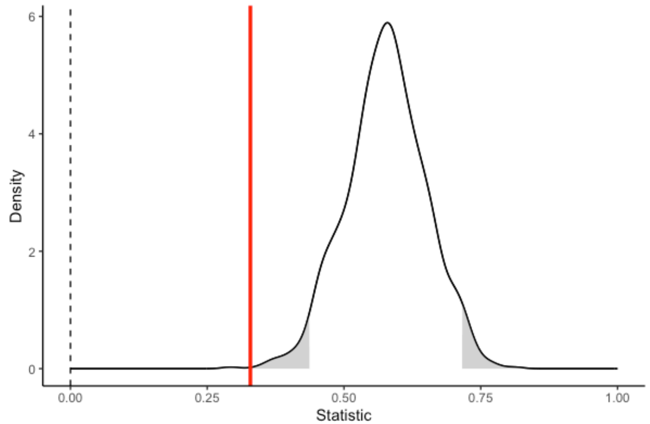
So is there homophily?

- Both E-I indices significantly differ from the expected value of a random base line network (with fixed density)
  - Significant gender homophily
  - Significantly less ethnicity-based heterophily

- Can you think of metrics for doing the same with reciprocity, transitivity, centralisation?

- Would you expect these to be statistically significant or not? Why?

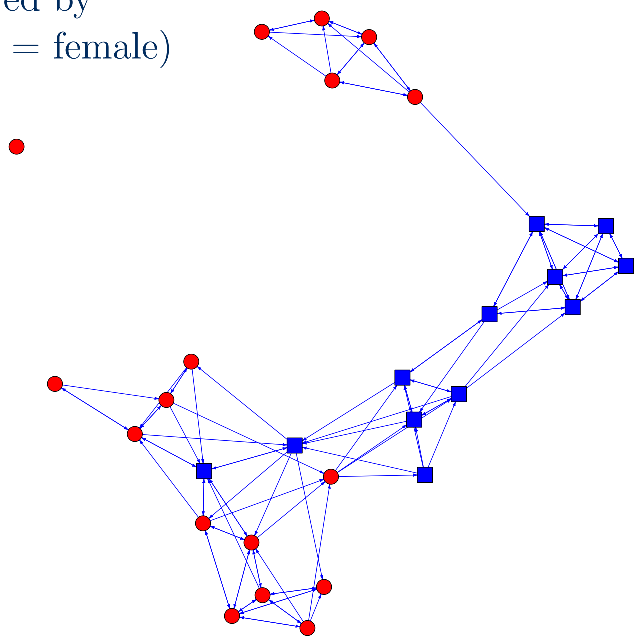


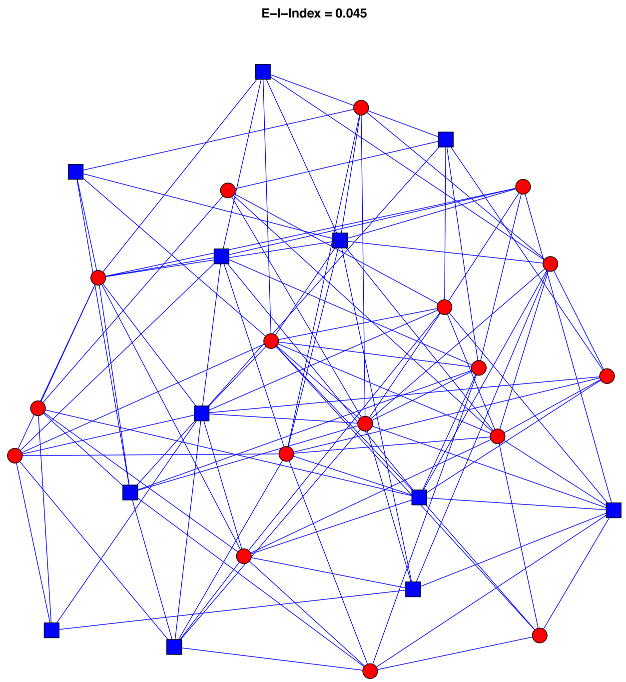


Problems with this test

- *Is a random network a good baseline?*
  - It fixes density/dimensions, but loses all structure…

Nodes colored by gender (red = female)

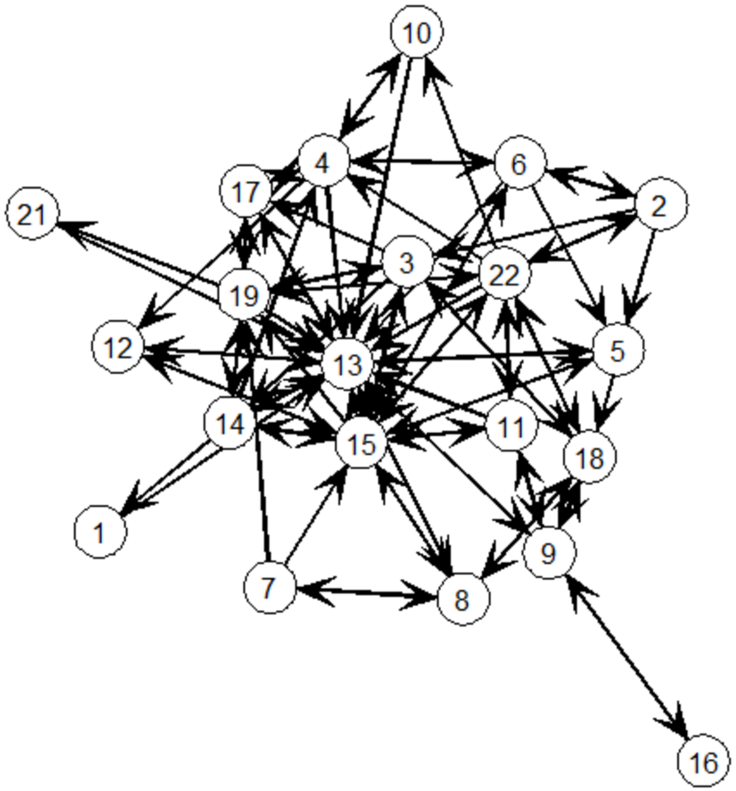




Consequences of ignoring network dependencies

- Yes, we *can* correlate networks (explain one "dependent" network as a function of other "independent" networks): regression/correlation coefficients are estimated correctly as association between networks
  - But… standard errors not reliable, as they rely on the assumption of *independence of observations* – which is not a justifiable assumption here
- Let's take a look at another example to illustrate what we mean
- We need a different (non-parametric) method of assessing whether the correlations are "significant"

Case study: organisational emails



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0	0	0	0	0	1	1	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0
2	0	0	3	1	1	4	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	2
3	0	1	0	4	1	0	2	0	0	0	0	0	2	1	1	0	7	2	6	0	0	1
4	0	1	1	0	0	5	0	0	0	1	0	1	3	0	0	0	3	0	0	0	0	1
5	0	0	0	0	0	1	0	2	0	0	0	0	3	0	2	0	0	0	0	0	0	0
6	0	2	0	2	2	0	0	2	1	0	0	0	1	0	5	0	0	1	0	0	0	0
7	1	0	1	0	0	0	0	1	0	0	0	0	0	0	8	0	0	0	1	0	0	0
8	0	0	0	0	2	1	1	0	2	0	0	0	4	0	6	0	0	6	0	0	0	0
9	1	0	0	0	1	2	0	3	0	0	2	7	6	0	1	1	0	30	0	0	0	1
10	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	2
11	0	0	0	0	0	1	0	0	1	0	0	1	4	0	3	0	1	2	0	0	0	6
12	0	0	0	0	0	1	0	0	5	0	1	0	6	0	3	0	0	0	0	0	0	0
13	0	0	2	0	1	0	0	1	5	0	0	6	0	1	0	0	5	0	0	0	0	3
14	2	0	0	1	0	0	0	0	0	0	0	0	4	0	6	0	1	0	3	0	0	0
15	0	3	1	0	3	6	1	5	1	0	5	3	7	8	0	0	1	0	1	0	0	3
16	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
17	0	0	1	3	0	0	0	0	0	1	2	0	6	0	0	0	0	0	5	0	0	0
18	0	0	2	0	0	3	0	9	34	0	3	0	0	0	0	0	0	0	0	0	0	8
19	1	0	7	1	0	1	0	0	0	0	0	0	2	1	0	0	15	0	0	0	1	2
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
22	0	2	1	2	0	0	0	0	0	3	7	0	5	0	1	1	0	5	1	0	0	0

Zooming in…

	9	10	11	12	13	14	15	16	17	18
9	0	0	2	7	6	0	1	1	0	30
10	0	0	0	0	1	0	0	0	1	0
11	1	0	0	1	4	0	3	0	1	2
12	5	0	1	0	6	0	3	0	0	0
13	5	0	0	6	0	1	0	0	5	0
14	0	0	0	0	4	0	6	0	1	0
15	1	0	5	3	7	8	0	0	1	0
16	2	0	0	0	0	0	0	0	0	0
17	0	1	2	0	6	0	0	0	0	0
18	34	0	3	0	0	0	0	0	0	0

Row dependencies

	9	10	11	12	13	14	15	16	17	18
9	0	0	2	7	6	0	1	1	0	30
10	0	0	0	0	1	0	0	0	1	0
11	1	0	0	1	4	0	3	0	1	2
12	5	0	1	0	6	0	3	0	0	0
13	5	0	0	6	0	1	0	0	5	0
14	0	0	0	0	4	0	6	0	1	0
15	1	0	5	3	7	8	0	0	1	0
16	2	0	0	0	0	0	0	0	0	0
17	0	1	2	0	6	0	0	0	0	0
18	34	0	3	0	0	0	0	0	0	0

Column dependencies

	9	10	11	12	13	14	15	16	17	18
9	0	0	2	7	6	0	1	1	0	30
10	0	0	0	0	1	0	0	0	1	0
11	1	0	0	1	4	0	3	0	1	2
12	5	0	1	0	6	0	3	0	0	0
13	5	0	0	6	0	1	0	0	5	0
14	0	0	0	0	4	0	6	0	1	0
15	1	0	5	3	7	8	0	0	1	0
16	2	0	0	0	0	0	0	0	0	0
17	0	1	2	0	6	0	0	0	0	0
18	34	0	3	0	0	0	0	0	0	0

Reciprocal dependencies…

	9	10	11	12	13	14	15	16	17	18
9	0	0	2	7	6	0	1	1	0	30
10	0	0	0	0	1	0	0	0	1	0
11	1	0	0	1	4	0	3	0	1	2
12	5	0	1	0	6	0	3	0	0	0
13	5	0	0	6	0	1	0	0	5	0
14	0	0	0	0	4	0	6	0	1	0
15	1	0	5	3	7	8	0	0	1	0
16	2	0	0	0	0	0	0	0	0	0
17	0	1	2	0	6	0	0	0	0	0
18	34	0	3	0	0	0	0	0	0	0

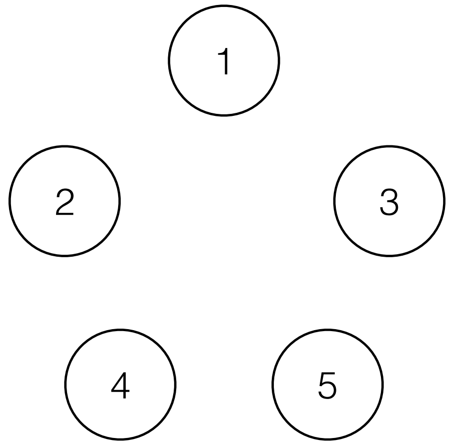
Some hypotheses (monadic and dyadic)

- Longer acquaintances exchange more emails (dyadic)
- Same department exchange more emails (dyadic)
- Same sex individuals exchange more emails (dyadic)
- More senior individuals send more emails (monadic)
- More senior individuals receive more emails (monadic)

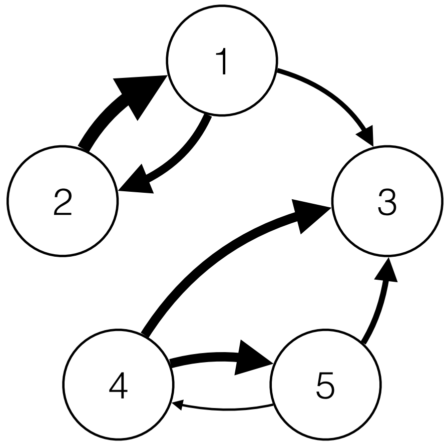
Null Hypothesis II: Permutations

- Permuting a network means changing the network in a way that keeps the structure of a network intact, but changes the positions of nodes in the network
  - Usually done by repeatedly swapping the rows and columns of an adjacency matrix
- Repeated permutations creates distributions of structurally similar networks that might otherwise be confounded with homophily, giving us a more realistic null hypothesis

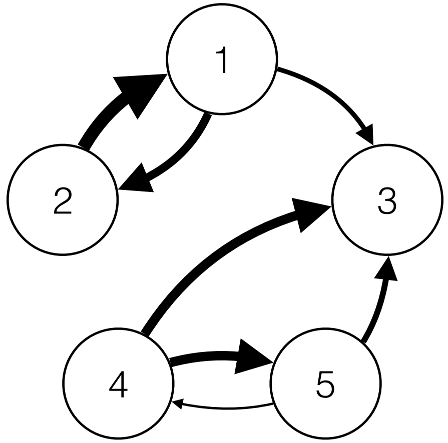
$$\begin{matrix} 0 & 5 & 3 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 6 \\ 0 & 0 & 4 & 2 & 0 \end{matrix} \quad == $$



$$
\begin{matrix}
0 & 5 & 3 & 0 & 0 \\
8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 6 \\
0 & 0 & 4 & 2 & 0
\end{matrix}
\quad ==
$$



$$
\begin{matrix}
0 & 5 & 3 & 0 & 0 \\
8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 6 \\
0 & 0 & 4 & 2 & 0
\end{matrix}
\quad ==
$$

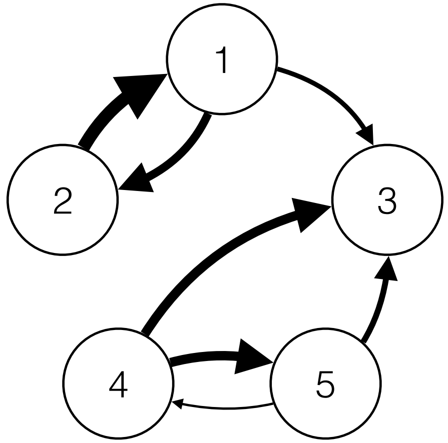


5, 3, 1, 2, 4

$$
\begin{matrix}
0 & 0 & 4 & 2 & 0 \\
0 & 5 & 3 & 0 & 0 \\
8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 6
\end{matrix}
\rightarrow
\begin{matrix}
0 & 0 & 0 & 4 & 2 \\
0 & 0 & 5 & 3 & 0 \\
0 & 8 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 6 & 0
\end{matrix}
\rightarrow
\begin{matrix}
0 & 0 & 0 & 4 & 2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 5 & 3 & 0 \\
0 & 8 & 0 & 0 & 0 \\
6 & 0 & 0 & 6 & 0
\end{matrix}
\rightarrow
\begin{matrix}
0 & 4 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 3 & 0 & 5 & 0 \\
0 & 0 & 8 & 0 & 0 \\
6 & 6 & 0 & 0 & 0
\end{matrix}
$$

0	5	3	0	0
8	0	0	0	0
0	0	0	0	0
0	0	6	0	6
0	0	4	2	0

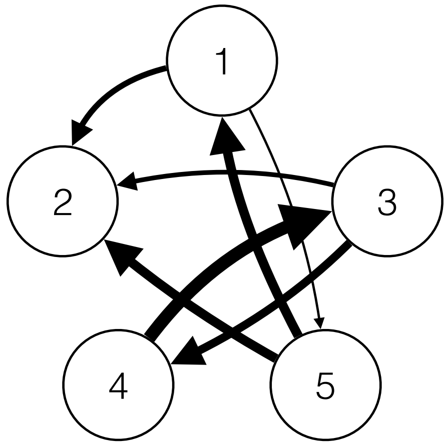
==



5, 3, 1, 2, 4

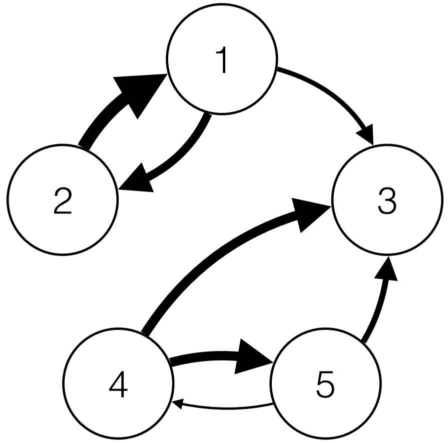
0	4	0	0	2
0	0	0	0	0
0	3	0	5	0
0	0	8	0	0
6	6	0	0	0

==



0	5	3	0	0
8	0	0	0	0
0	0	0	0	0
0	0	6	0	6
0	0	4	2	0

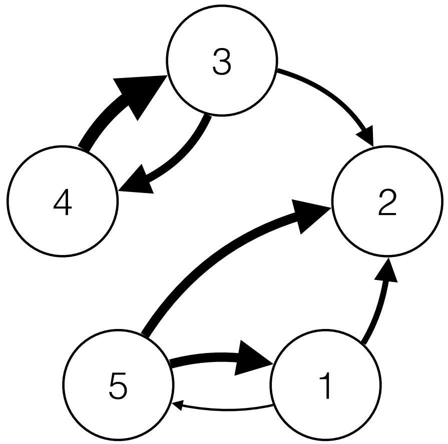
==

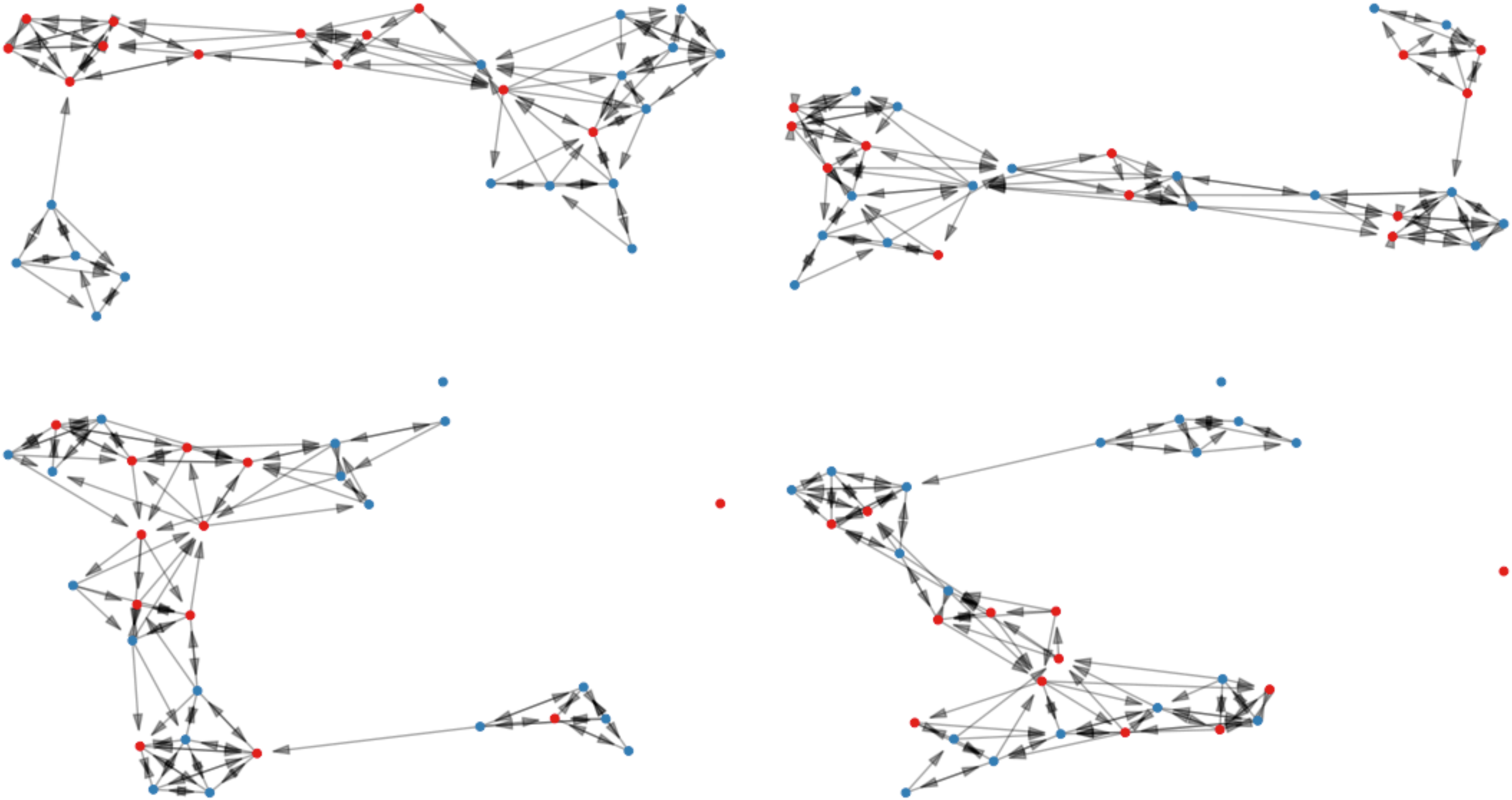


5, 3, 1, 2, 4

0	4	0	0	2
0	0	0	0	0
0	3	0	5	0
0	0	8	0	0
6	6	0	0	0

==



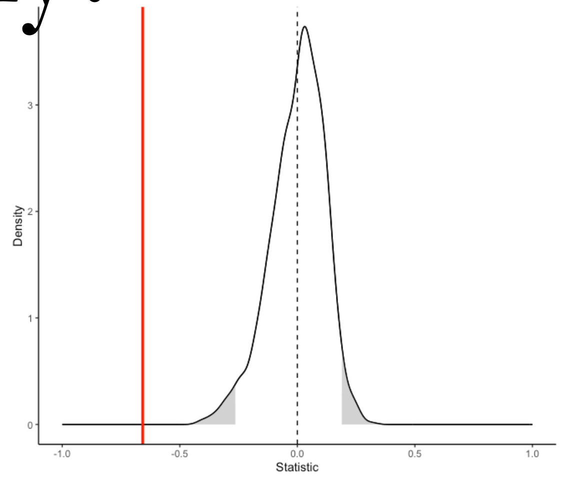


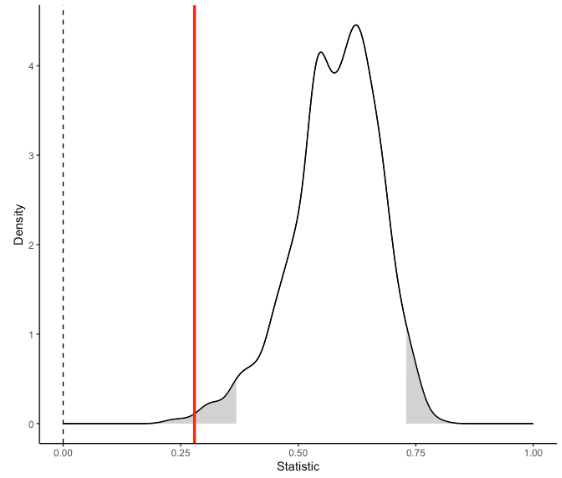
So is there homophily?

- Both E-I indices significantly differ from the expected value from permutations of the network
  - Significant gender homophily
  - Significantly less ethnicity-based heterophily

- Note that the distributions are broader than those using a random baseline

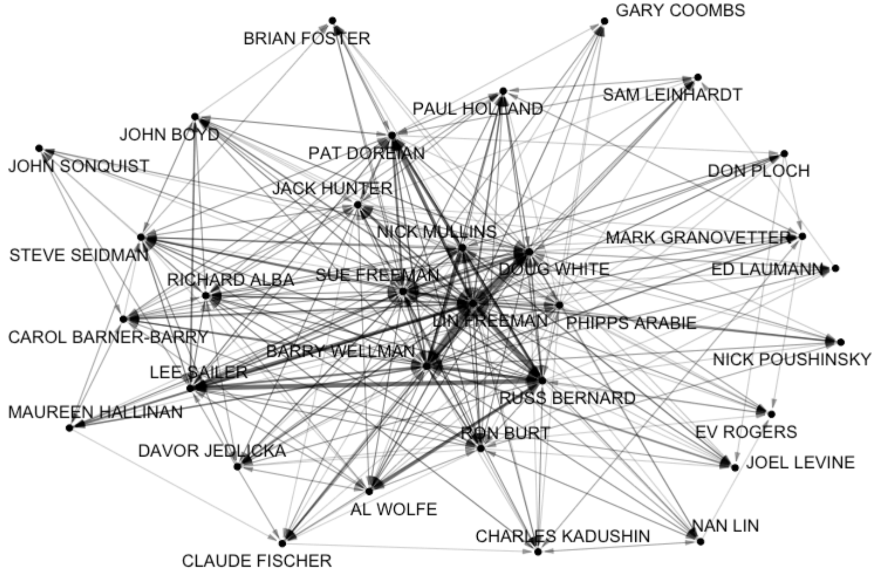
- How would this work with reciprocity, transitivity, centralisation? Why?





Larger networks

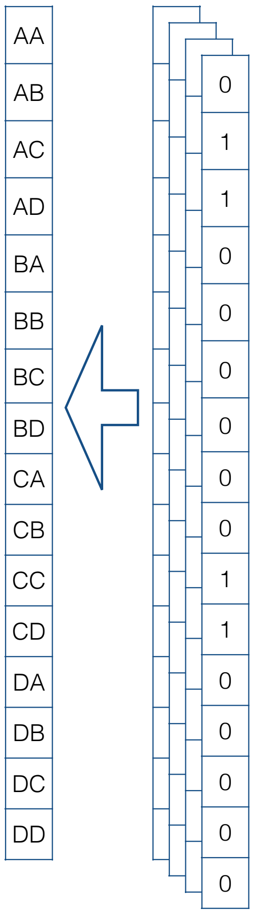
- For network of 5 nodes, number of permutations is tractable (5! = 120)
- However, in many cases we want to analyse larger networks
  - For a network of 10, 32, and 100 nodes there are 3.6e6, 2.6e35, and 9.3e157 possible permutations
  - For these cases, random draws of permutations are used to create the distribution
- Principle of *sampling* used repeatedly in the statistical analysis of social networks



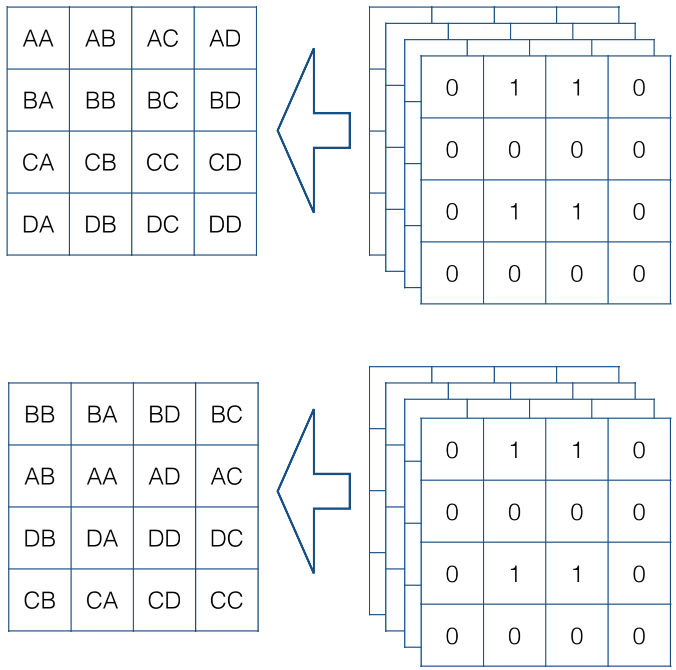
Possible confounds

- What are some possible confounds for homophily? (i.e. other explanations for homogeneous macro outcomes that are not about choice homophily?)
  - propinquity (geography)
  - kinship (family)
  - foci (organisational)
  - isomorphic roles (occupational, family, informal)
  - robustness (weak tie dissolution under crisis)

What should we do differently?

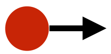


$$M_y = \beta_0 M_1 + \beta_1 M_{x1} + \beta_2 M_{x2} + \ldots + Z$$

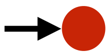


Available effects for network_regression

- ...: explains the network's ties (values) by (values of) another network



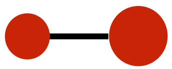
- `ego(...)`: explains the network's ties (values) by an attribute associated with the tie sender



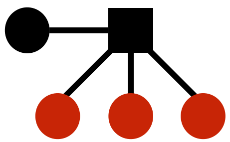
- `alter(...)`: explains the network's ties (values) by an attribute associated with the tie recipient

- `same(...)`: explains the network's ties (values) by the dyadic matching of attributes

- `dist(...)`: explains the network's ties (values) by an attribute associated with the tie sender



- `sim(...)`: explains the network's ties (values) by the proportional similarity between



- `tertius(...)`: explains the network's ties (values) by (sum/mean of) attributes associated other nodes sending to tie recipient

Haunss and Hollway 2023

Sometimes endogenous mechanisms in operation too though…

- In networks though, an observation of a tie may depend on other observations of a tie, e.g.:
  - One tie may depend on a tie in the other direction (reciprocity)
  - One tie may depend on other ties to that alter (popularity)
  - One tie may depend on ties to a third node (transitivity)
- This introduces *endogeneities*
  - Traditional statistics might find a nuisance and advise finding a way to exclude
  - But network statistics finds a crucial part of the story
- MRQAP helps us *take into account* these dependencies, but sometimes they are the *mechanism of interest…*
  - To explore them in more detail we might need models such as ERGMs and SAOMs…