Introduction to Network Modeling

what is a statistical model and why do we need them?

theory driven models for networks

statistical (network) models

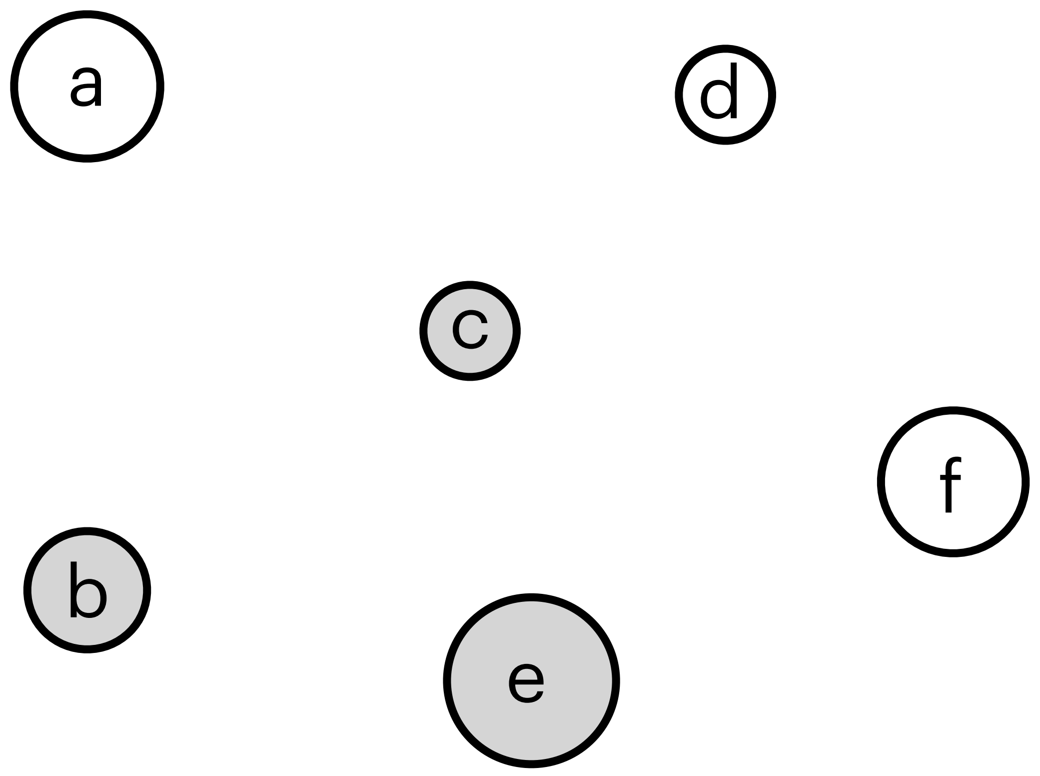
- start with assumptions on observed data but extend beyond the data
- encapsulate understanding (theory, hypothesis, conjecture) about mechanisms underlying the data
- an mathematical expression of rules governing the (null) world from which we think our data is from
- used to make inference: test hypotheses on processes assumed to have generated the network

no model can capture all of the niceties of the real world
models are idealizations and simplifications.

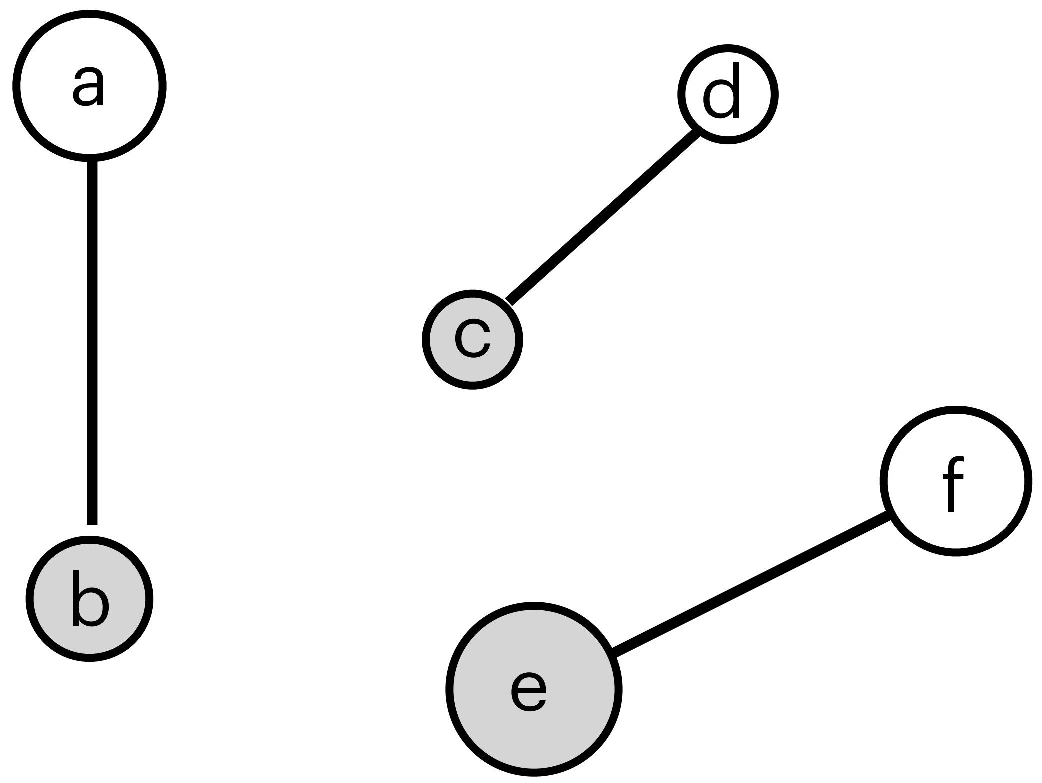
what is so special about network data?

comparing conventional monadic data to relational data

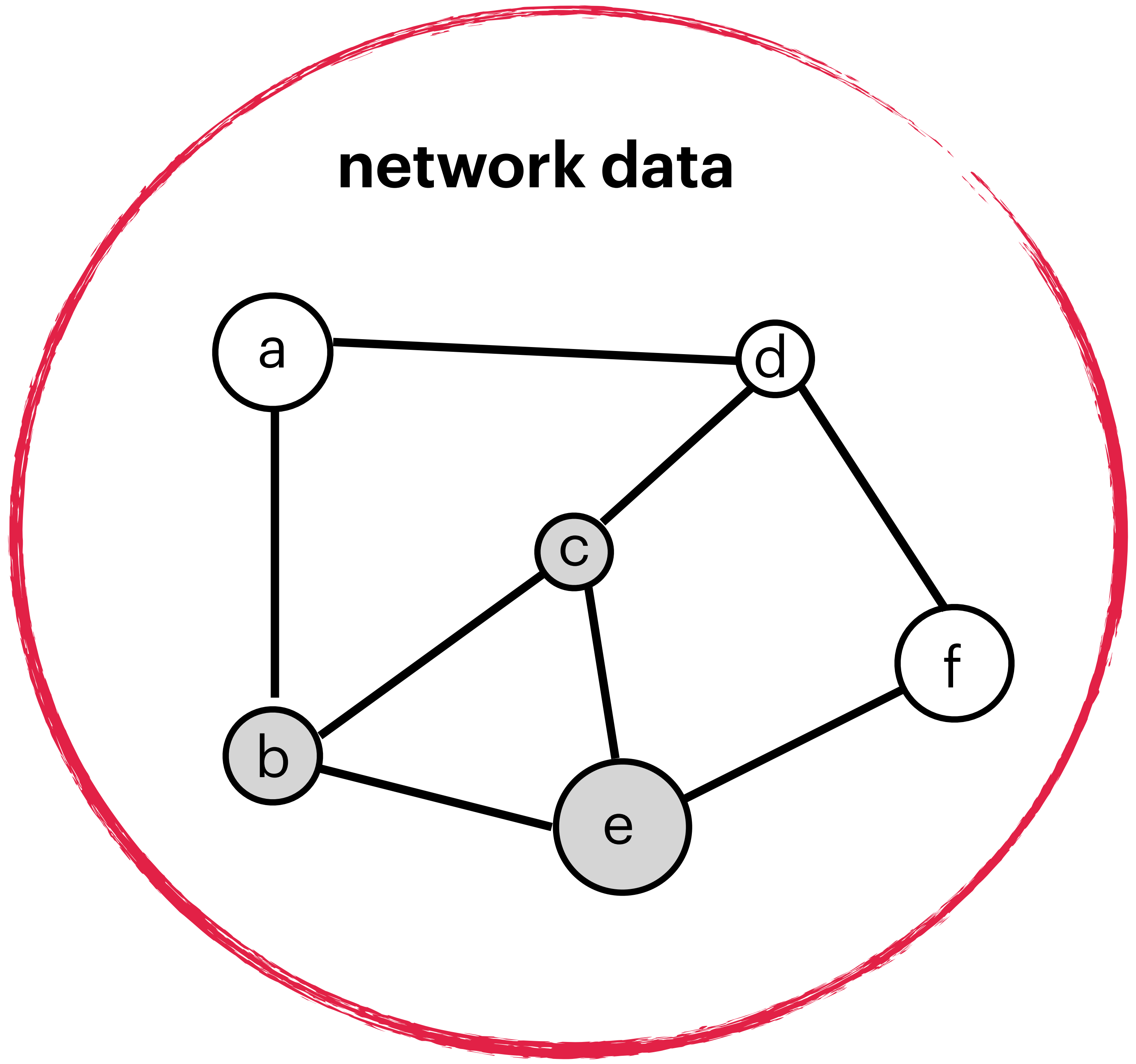
monadic data



dyadic data

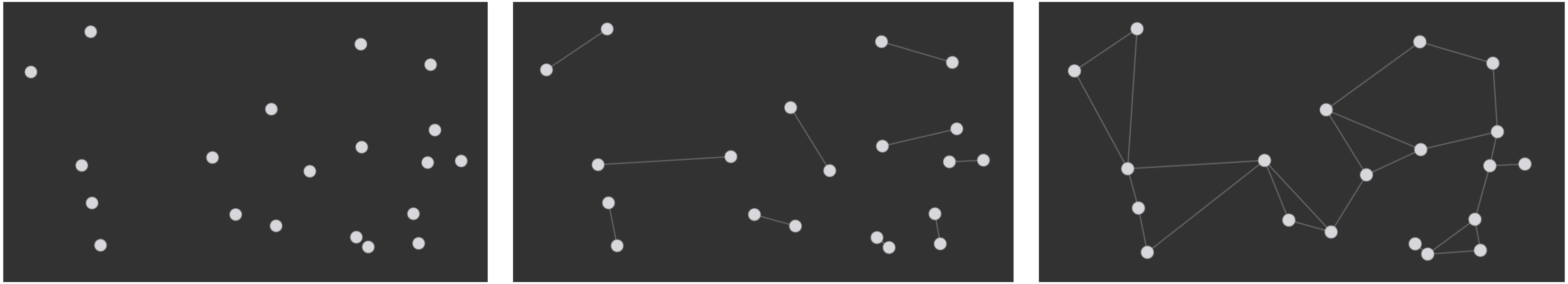


network data



what is so special about network data?

comparing conventional monadic data to relational data



- the unit of observation is ties (or edges or dyads)
- dyads are overlapping
- observations are interdependent
- the existence of a tie often changes the probability of other ties

} ~~*i.i.d.*~~

the emergence of social structure

the elements for social network theory

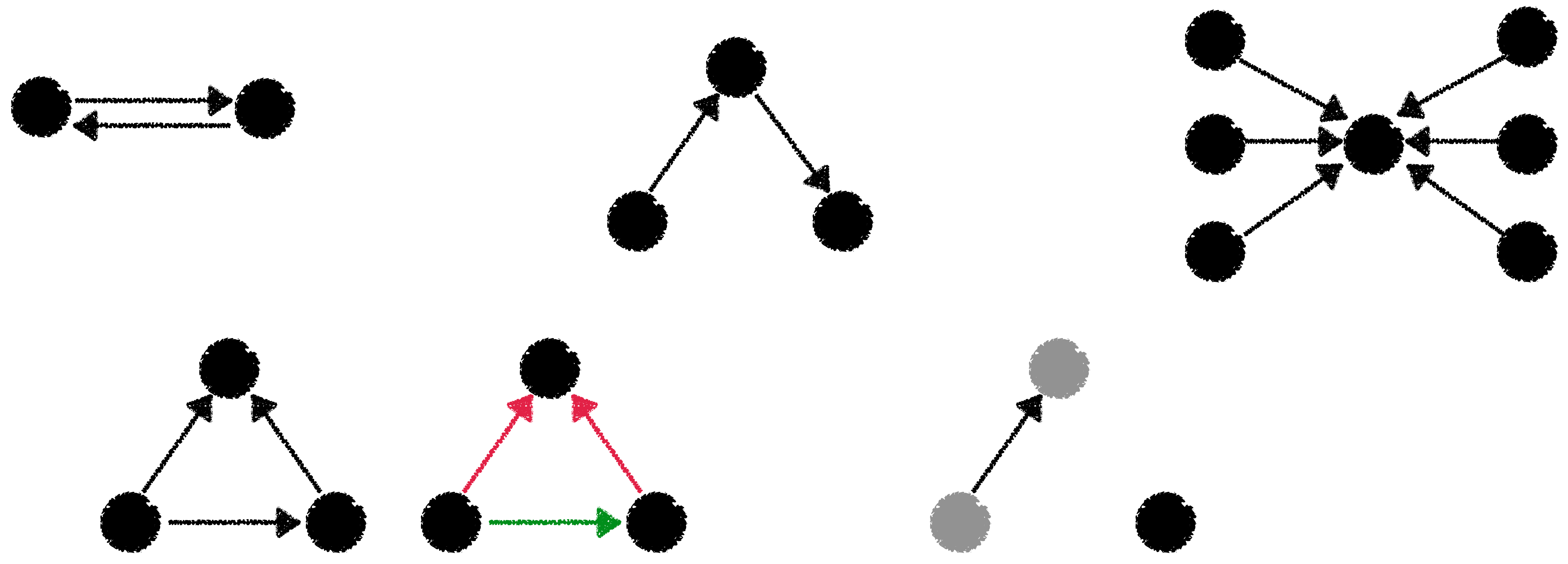
- structural patterns are locally emergent $\Longrightarrow$ local patterns form global structure
- network ties self organize through dependency between them:

  the presence of one tie may lead to another

- network patterns are evidence of several ongoing social processes operating simultaneously

the social rules we consciously and unconsciously adhere to

- *you scratch my back, I scratch yours*
- *a friend of a friend is a friend*
- *the enemy of my friend is my enemy*
- *brokerage*
- *bird of a feather flock together*
- *follow the crowd*



the emergence of social structure

the elements for social network theory

- structural patterns are locally emergent $\Longrightarrow$ local patterns form global structure
- network ties self organize through dependency between them:

  the presence of one tie may lead to another

- network patterns are evidence of several ongoing social processes operating simultaneously

the social rules we consciously and unconsciously adhere to

you scratch my back, I scratch yours	social exchange
a friend of a friend is a friend	structural balance
the enemy of my friend is my enemy	structural balance
brokerage	structural holes
bird of a feather flock together	homophily: social selection and social influence
follow the crowd	the Matthew effect

parametric vs. non-parametric methods

parametric

- tests based on theoretical distribution of summary statistics
- data follows some sort of theoretical probability distribution
- models that more or less incorporate dependencies among ties

non-parametric

- distribution free methods
- no assumption on the data is needed
- evaluate null against working hypothesis without assuming any parametric model
- *p*-values have same interpretation: probability of seeing such extreme data given the null hypothesis is true
- tests: shuffling ties while fixing an observed summary measure (i.e. null model)

Conditional Uniform Graph Distributions

non-parametric tests: conditional uniform graph distributions

null hypothesis

H_0 observed network is created from specified model that does *X*

alternative hypothesis

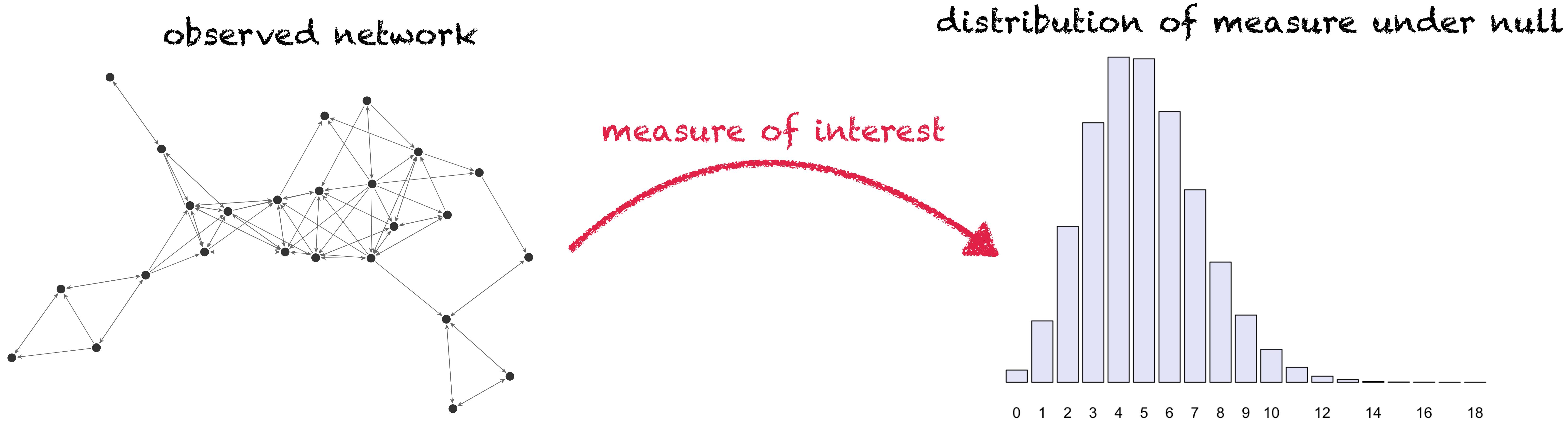
H_1 observed network is not created from specified model that does *X*

decision rule

if simulated networks from null model look like the observed in $\alpha(100\%)$ of cases

$\implies$ reject H_0 on the $\alpha(100\%)$ significance level

otherwise $\implies$ H_0 cannot be rejected



non-parametric tests: conditional uniform graph distributions

(1) create a null model to which we can test our observed network against

the null model corresponds to a world of hypothetical networks

(2) distribution of chosen statistic under null is generated by simulations from the null model

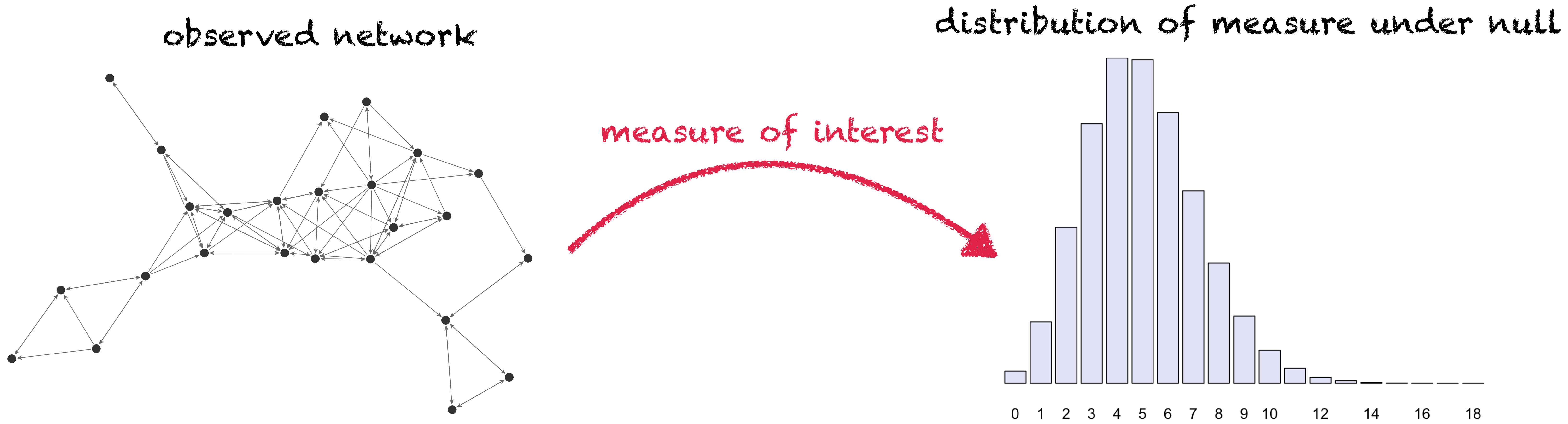
(3) check where the observed value of the statistic falls in this null distribution

does the observed value differ significantly from the expected?

(4) if yes ⟹ reject null hypothesis

is there a social phenomenon at play?

(5) if no ⟹ re-specify null model



non-parametric tests: conditional uniform graph distributions

statistical inference relies on the assumption of randomness in the data
we need to **model that randomness**

creating the null distribution is done by shuffling ties randomly while fixing

- the number of edges or density of graph: $\mathscr{U} \mid L$ or $\mathscr{U} \mid E(L)$
- the degree distribution: $\mathscr{U} \mid \mathbf{d}$ where $\mathbf{d} = (d_1, d_2, \ldots, d_n)$
- dyad census (mutual, asymmetric, null): $\mathscr{U} \mid$ MAN
- ...or some other summary measure

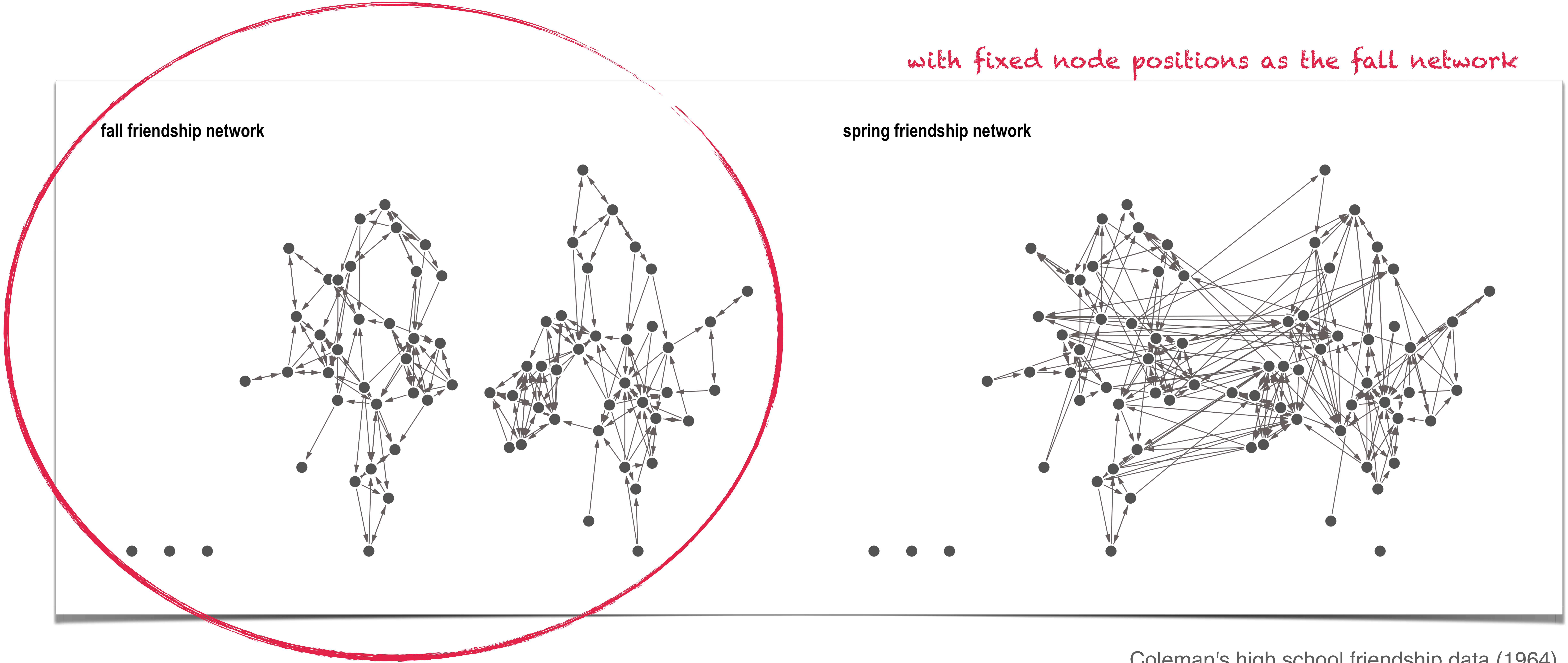
example.

a uniform distribution conditional on observed network's number of ties:

- graphs with specified number of ties are equally probable to appear
- graphs without specified number of ties have a probability of zero to appear

example. high school friendships

with fixed node positions as the fall network

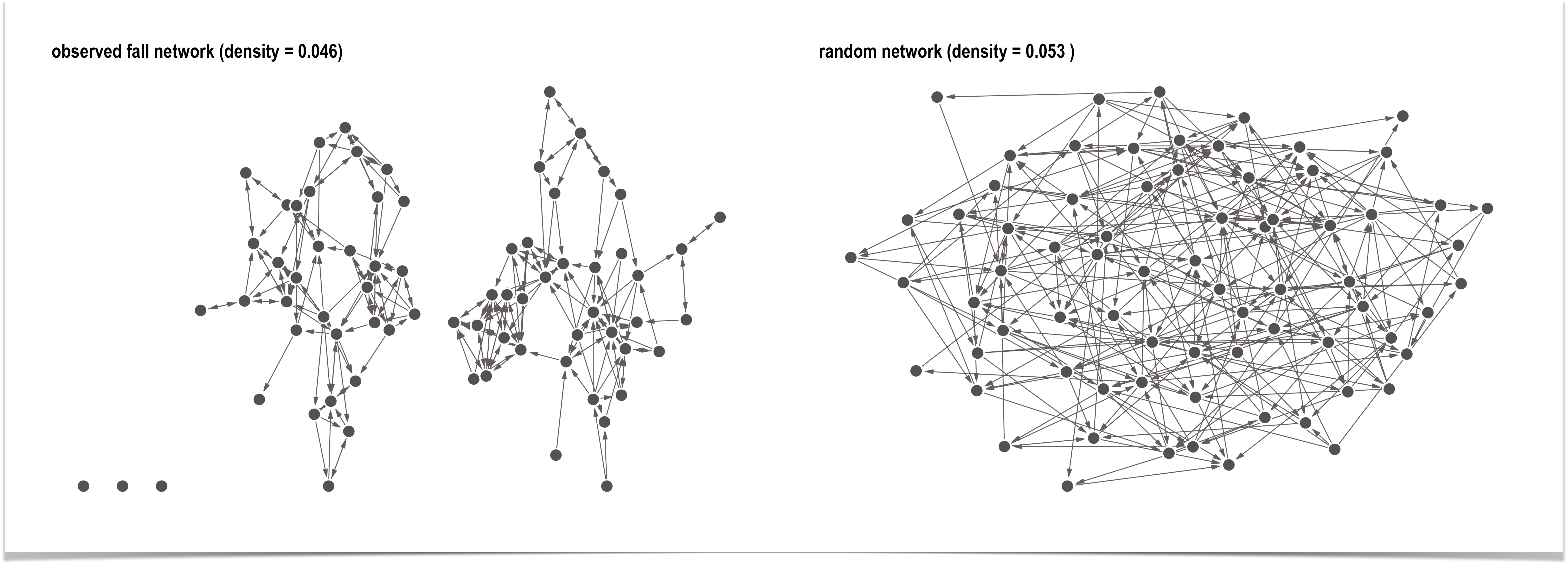


Coleman's high school friendship data (1964)

example. high school friendships

uniform graph distribution given expected density $\mathscr{U} \mid E(L)$

- calculate the density of the Coleman fall network ≈ 0.046
- generate **one** random graph with the same density on average as the observed network

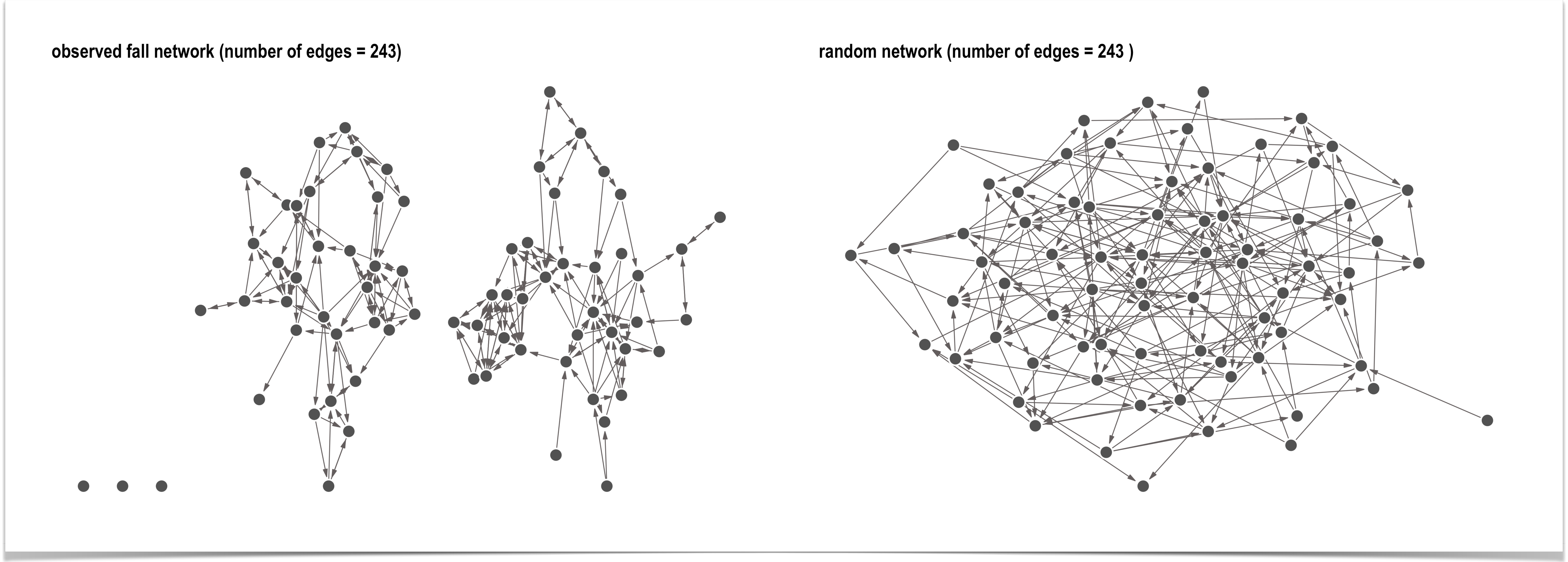


random network may not have the exact same number of ties as the observed one but stochastically it has the same density

example. high school friendships

uniform graph distribution given number of edges $\mathscr{U} \mid L$

- calculate the number of ties in the Coleman fall network = 243
- generate **one** random graph with the exact same number of ties as the observed network



example. high school friendships

uniform graph distribution given number of edges $\mathscr{U} \mid L$

- calculate the number of ties in the Coleman fall network $= 243$
- generate **one** random graph with the exact same number of ties as the observed network

compare dyad census for observed to random network

observed fall network

mutual	asymmetric	null
62	119	2447

random network

mutual	asymmetric	null
8	227	2393

even though the random network has the same density as the observed,
we have a completely different number of reciprocated ties

example. high school friendships

uniform graph distribution given number of edges $\mathcal{U} \mid L$

one random network had a very different count of mutual ties than the observed (62)

- was this a coincidence?
- do most random networks generated from this null model behave this way?

to answer how unusual mutual ties are in the alternative world
we need to generate more random networks from the null model

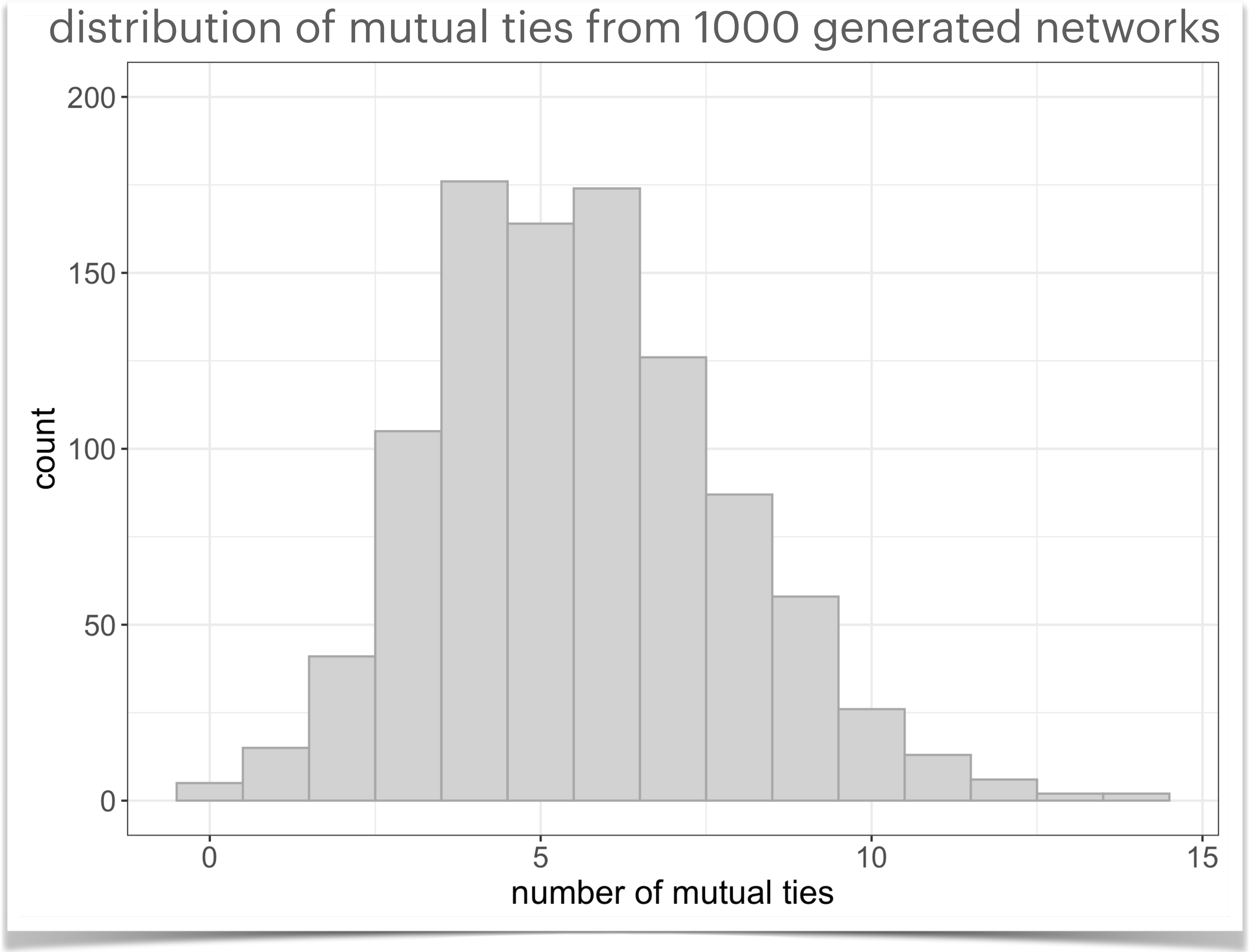
H_0 : observed reciprocity effect is created from $\mathcal{U} \mid L$

H_1 : observed reciprocity effect is **not** created from $\mathcal{U} \mid L$

do any of the 1000 random networks have
as large a number of mutual dyads as the observed?

reject or not reject the null hypothesis?

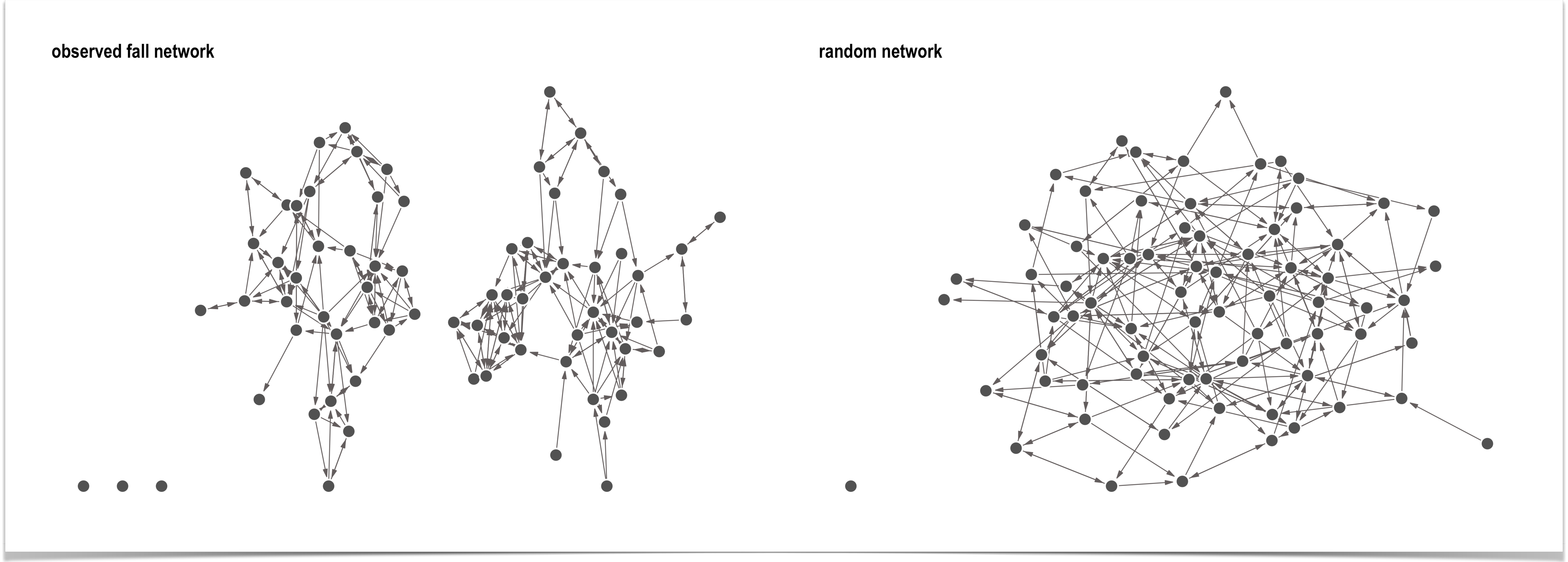
conclusion?



example. high school friendships

uniform graph distribution given dyad census $\mathcal{U} \mid \text{MAN}$

- dyad census of observed network: mutual $= 62$, asymmetric $= 119$, null $= 2447$
- generate **one** random graph with the same number of dyad counts as observed network



example. high school friendships

uniform graph distribution given dyad census $\mathcal{U}$ | MAN

- dyad census of observed network: mutual = 62, asymmetric = 119, null = 2447
- generate **one** random graph with the same number of dyad counts as observed network

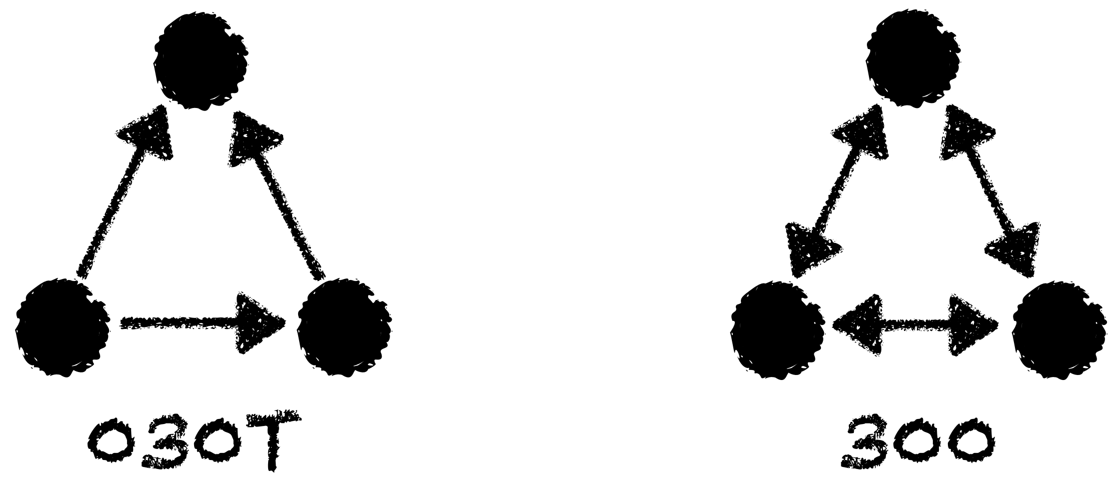
compare triad census for observed to random network

observed fall network

003	012	102	021D	021U	021C	111D	111U	030T	030C	201	120D	120U	120C	210	300
50171	7384	3957	64	121	128	139	70	23	1	20	43	10	9	34	22

random network

003	012	102	021D	021U	021C	111D	111U	030T	030C	201	120D	120U	120C	210	300
50223	7312	3808	88	107	157	185	205	5	1	86	4	0	4	9	2



interpretation:
had allocation of ties in the network been completely random given 'dyadic processes' it would be unlikely to observe any complete triangles

but we have only looked at one random graph...

example. high school friendships

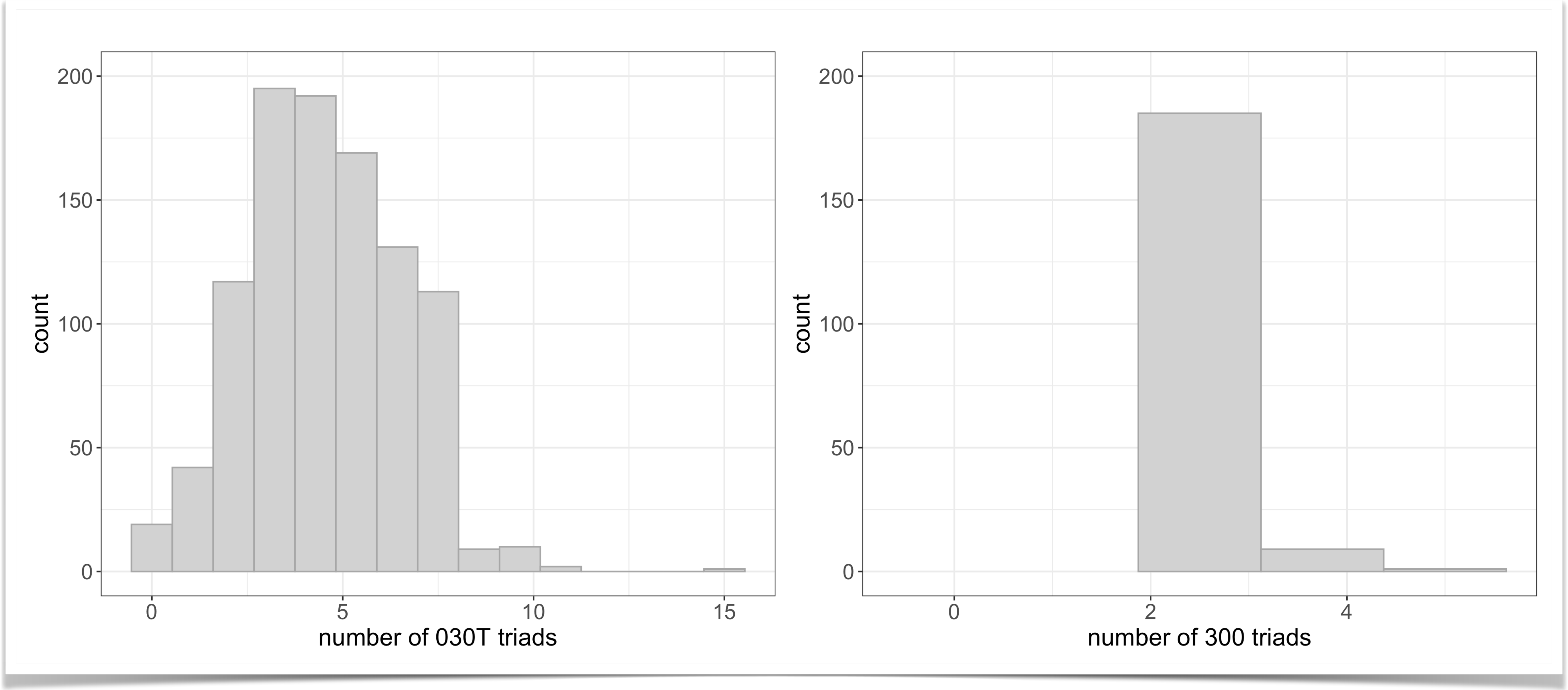
uniform graph distribution given dyad census $\mathscr{U} \mid$ MAN

- compare transitivity for observed network to 1000 random networks

H_0 : observed transitivity effect is created from $\mathscr{U} \mid$ MAN

H_1 : observed transitivity effect is **not** created from $\mathscr{U} \mid$ MAN

distribution of number of transitive triads under null is generated
by simulating 1000 random networks with the same dyad counts as the observed one



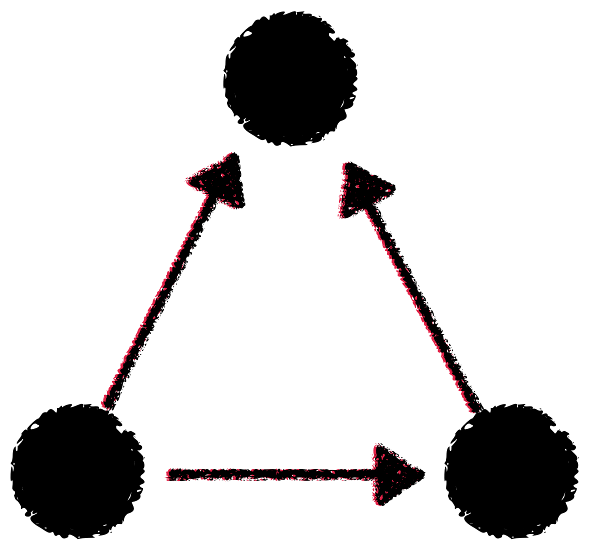
reject or not reject the null?

conclusion?

non-parametric tests:
conditional uniform graph distributions

limitation of using conditional uniform graph distributions:

subgraphs such as dyads and triads are nested in each other
and results may be confounded if we do not control for one when we counting the other



we need a model that can control for several configurations

parametric vs. non-parametric methods

parametric

- tests based on theoretical distribution of summary statistics
- data follows some sort of theoretical probability distribution
- models that more or less incorporate dependencies among ties

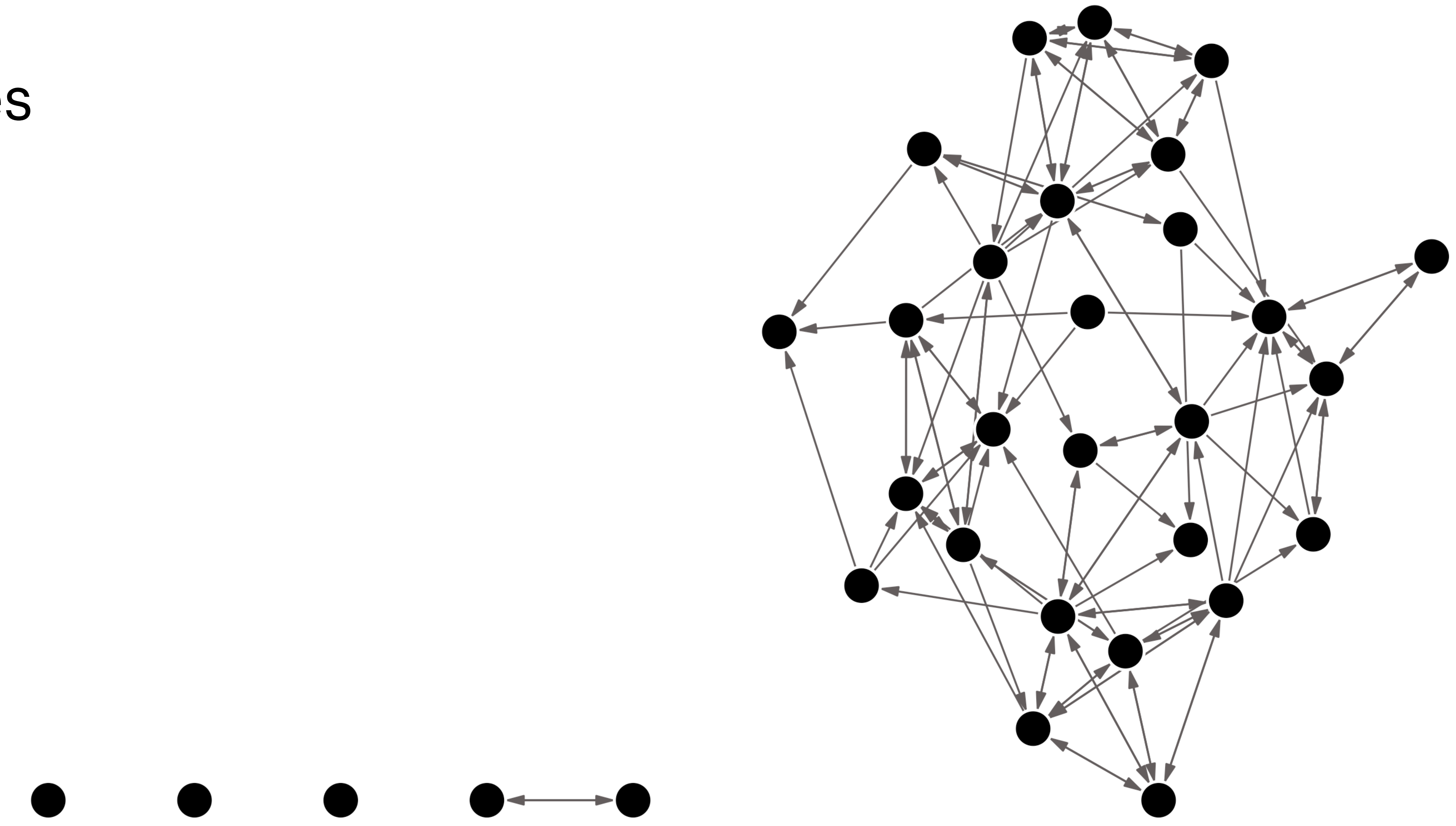
non-parametric

- distribution free methods
- no assumption on the data is needed
- evaluate null against working hypothesis without assuming any parametric model
- *p*-values have same interpretation: probability of seeing such extreme data given the null hypothesis is true
- tests: shuffling ties while fixing an observed summary measure (i.e. null model)

example: friendship among university freshmen

Van de Bunt (1999), data set available to download here

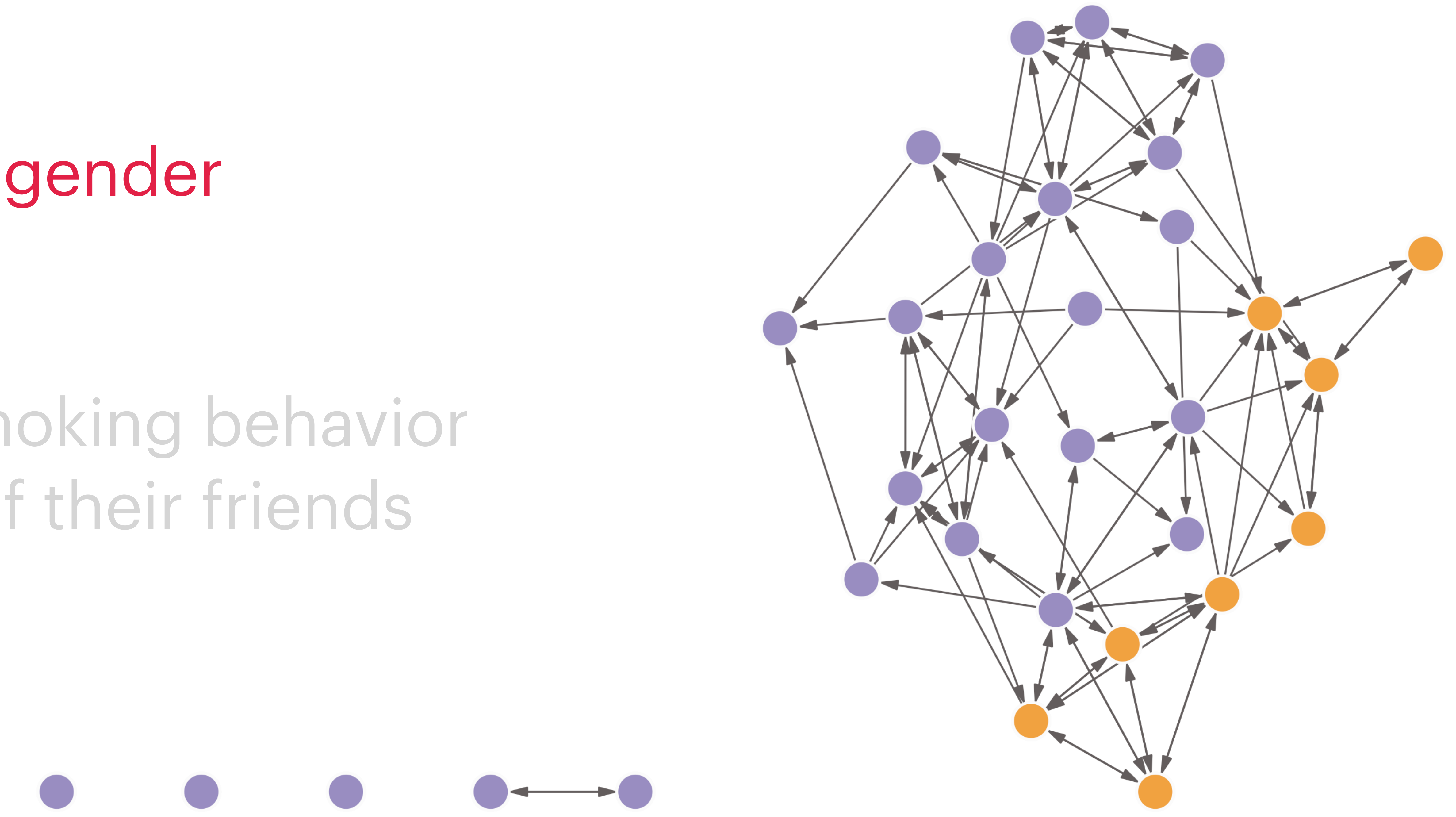
- directed network: 32 students, 110 ties
- constant actor attributes
  - gender (f/m)
  - program (2/3/4 year)
- changing actor attributes
  - smoke (y/n)



example: friendship among university freshmen

running hypotheses:

- pupils choose friends with the same gender
- pupils reciprocate friendship
- the friend of a friend is a friend
- pupils choose friends with similar smoking behavior
- pupils adopt the smoking behavior of their friends



is the probability of friendship between students of the same gender higher?

let's start with a non-parametric approach to study social selection by gender

example: friendship among university freshmen

observed values:

divide all possible pairs of students (*dyads*) in two groups

group 1 (G1): all dyads with same gender
group 2 (G2): all dyads with different gender

then compare observed proportion of ties in each group:

$$\frac{\#\text{ ties in G1}}{\#\text{ dyads in G1}} = \frac{91}{608} = 0.15 \qquad \frac{\#\text{ ties in G2}}{\#\text{ dyads in G2}} = \frac{19}{384} = 0.05$$

probability of friendship between student of same gender is 0.15
probability of friendship between students of different gender is 0.05

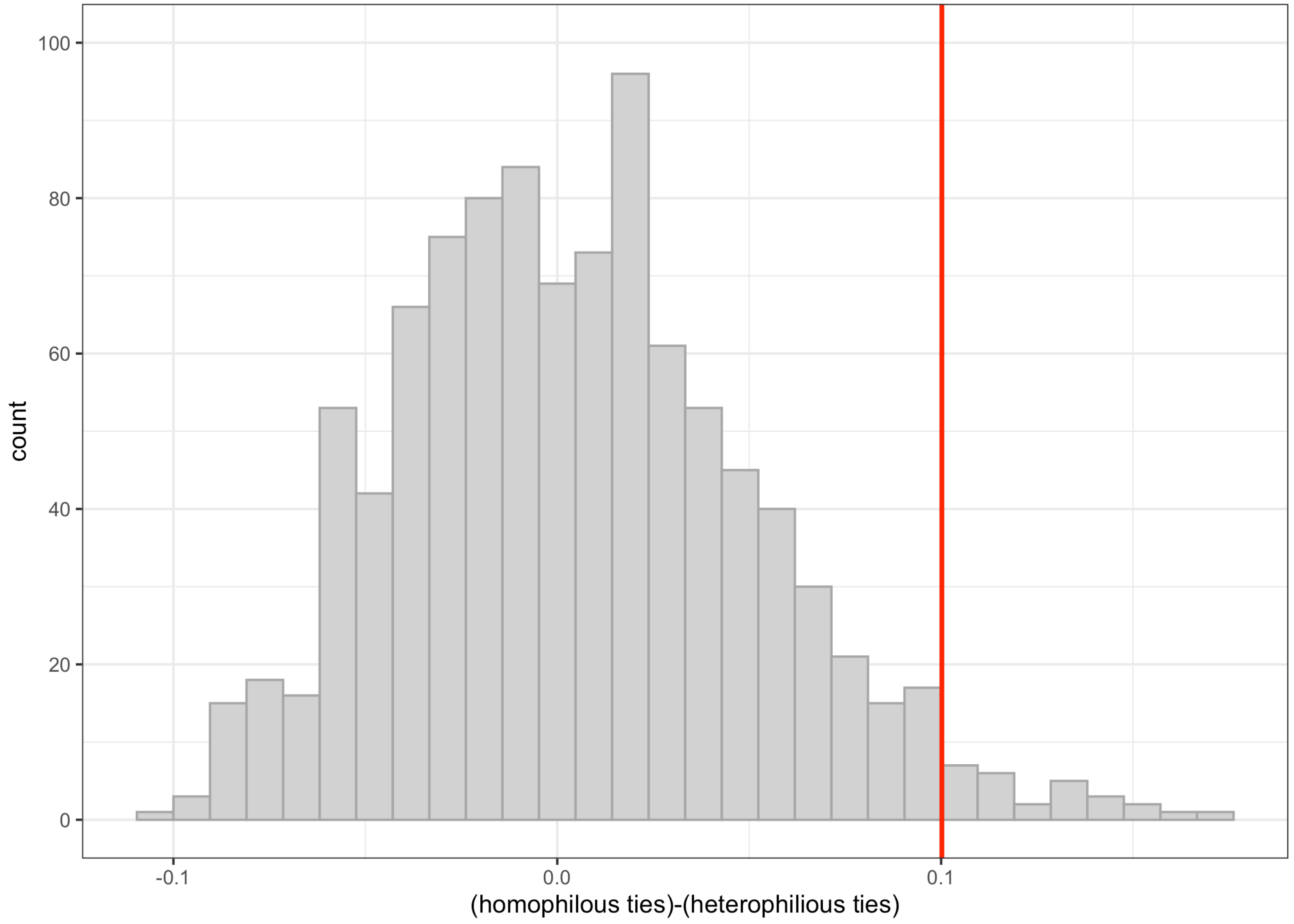
is this result accidental or significant?

example: friendship among university freshmen

compare observed values to those from simulated networks:

repeat the analysis 1000 times with random gender assignment

difference between same and different gender ties



➡ average difference is 0.005

➡ maximum difference is 0.17

observed difference: 0.15 — 0.05 = 0.10

we need a model that can control for the influence of other variables!

(for example behaviour, other ties in networks, etc.)