Random Graph Models

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

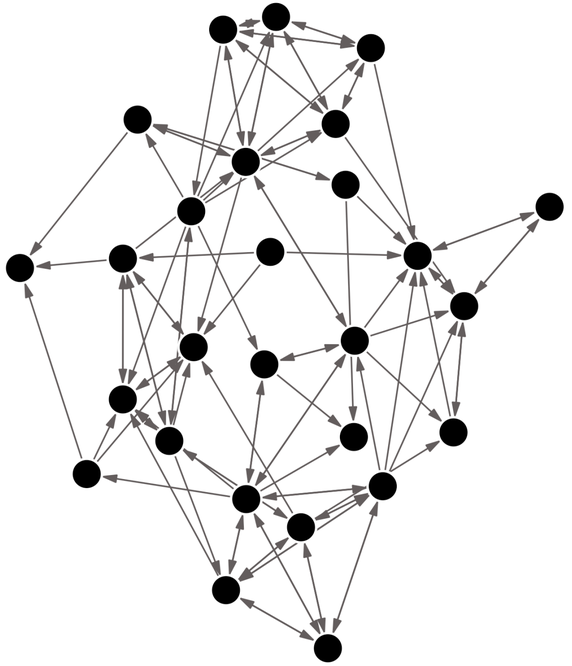
non-parametric

- 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

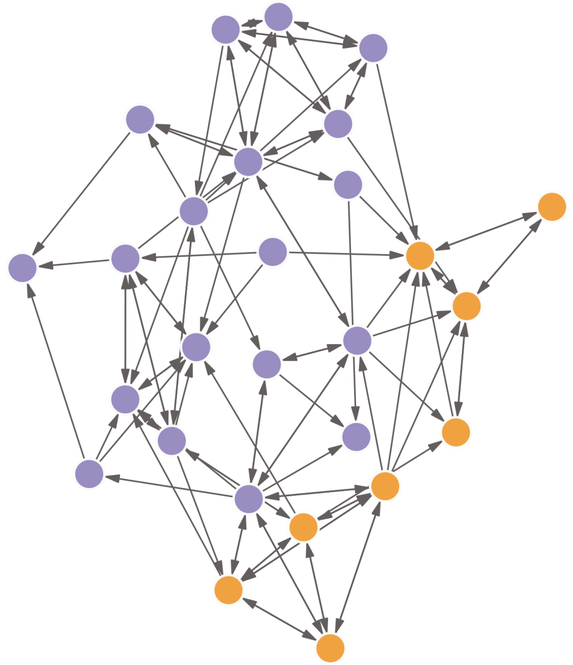
- directed network: 32 students, 110 ties
- constant actor attributes
  - gender (f/m)
  - program (2/3/4 year)
- changing actor attributes
  - smoke (y/n)



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:

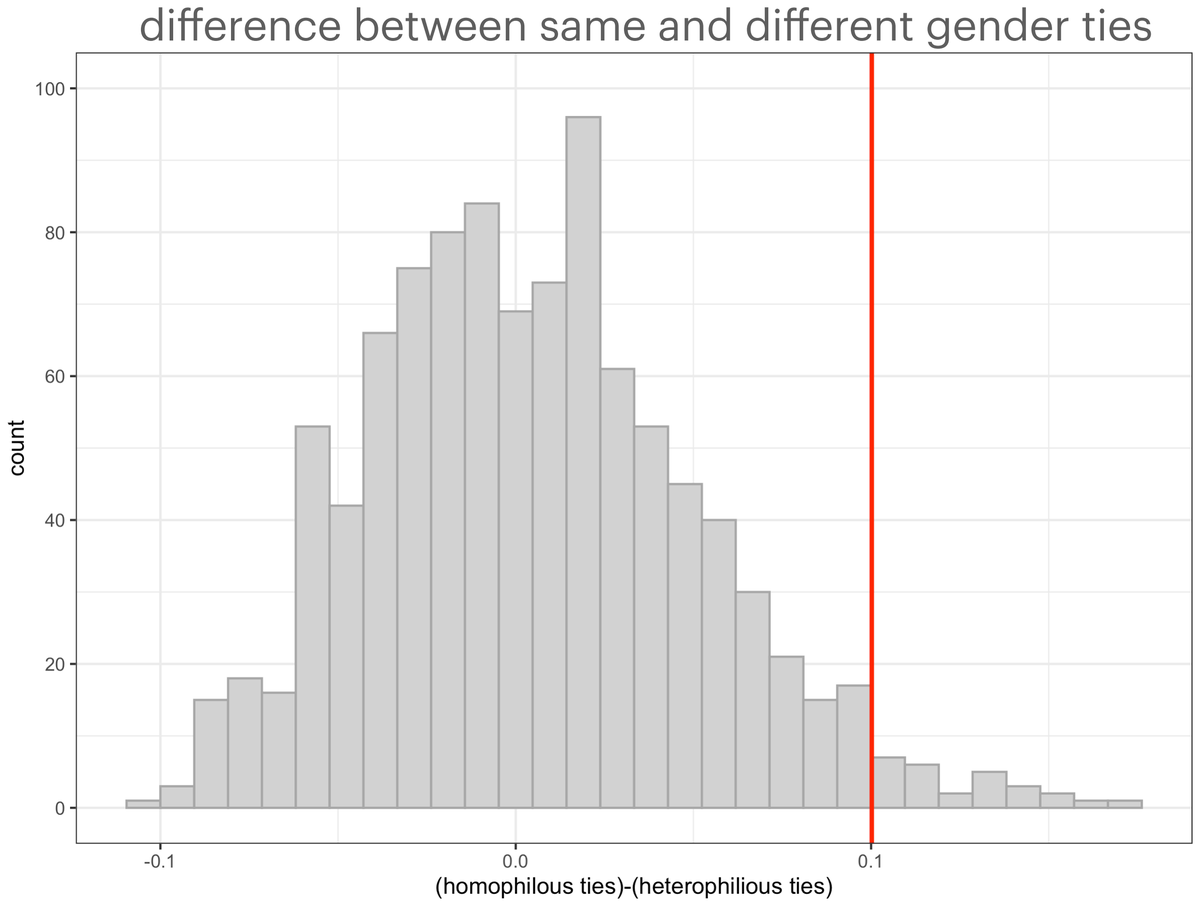
$$\frac{\text{\# ties in G1}}{\text{\# dyads in G1}} = \frac{91}{608} = 0.15 \qquad \frac{\text{\# ties in G2}}{\text{\# dyads in G2}} = \frac{19}{384} = 0.05$$

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



➡ 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.)

example: friendship among university freshmen

let's instead move to a more 'conventional' parametric model to study social selection by gender

logistic regression to model the presence/absence of ties:

random variable Y_{uv} denotes the ties between dyad (u, v):

$$Y_{uv} = \begin{cases} 1 & \text{with probability } P_{uv} \\ 0 & \text{with probability } 1 - P_{uv} \end{cases}$$

where $$p_{uv} = \text{logit}^{-1}(\theta \cdot s) = \frac{\exp(\theta \cdot s)}{1 + \exp(\theta \cdot s)}$$

$s = (s_1, \ldots, s_k) \subseteq \mathbb{R}^k$ statistics

$\theta = (\theta_1, \ldots, \theta_k) \subseteq \mathbb{R}^k$ parameters

$$\theta \cdot s = \sum_{i=1}^{k} \theta_i \cdot s_i$$

parameters are estimated using maximum likelihood estimation (MLE)

example: friendship among university freshmen

let's instead move to a more 'conventional' parametric model to study social selection by gender

logistic regression to model the presence/absence of ties:

in layman's terms: tie probability is determined using a function of statistics and parameters

- the statistics (explanatory variable) quantify properties of dyad (u, v) in observed network
- the parameters quantify the influence of those variables on the tie probability
  - ➡ a **positive** (**negative**) parameter means the higher the statistic the **higher** (**lower**) the probability
  - ➡ a zero parameter means the statistic has no influence on the tie probability

the parameters are estimated to maximize the probability of the observed network

note: the intuition behind this model and ERGM which we will also be covering later on is very similar

example: friendship among university freshmen

let's instead move to a more 'conventional' parametric model to study social selection by gender

logistic regression to model the presence/absence of ties:

model 1: friendship ties explained by gender equality

$$p_{uv} = \text{logit}^{-1}(\theta_0 + \theta_1 \cdot \text{same\_gender}(u, v))$$

```
Call:
glm(formula = frnd ~ same_gender, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5694  -0.5694  -0.5694  -0.3186   2.4520

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9655     0.2353 -12.560  < 2e-16 ***
same_gender   1.2183     0.2613   4.662 3.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

gender equality increases tie probability

implied tie probabilities:

- ➡ 0.15 for friendship of same gender
- ➡ 0.05 for friendship of different gender

let's control for other possible explanations...

example: friendship among university freshmen

let's instead move to a more 'conventional' parametric model to study social selection by gender

logistic regression to model the presence/absence of ties:

model 2: controlling for more variables

$$p_{uv} = \text{logit}^{-1}\left(\theta_0 + \theta_1 s_1 + \ldots + \theta_k s_k\right) = \left(\text{logit}^{-1} \sum_{i=0}^{k} \theta_i \cdot s_i\right)$$

```
Call:
glm(formula = frnd ~ same_gender + same_smoke + reciprocity +
    transitivity, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5556  -0.2606  -0.2305  -0.1589   2.9591

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.3654     0.3569 -12.231  < 2e-16 ***
same_gender    0.7505     0.3287   2.283   0.0224 *
same_smoke     0.2493     0.2873   0.868   0.3856
reciprocity    2.9128     0.2989   9.745  < 2e-16 ***
transitivity   1.2266     0.1525   8.043  8.8e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

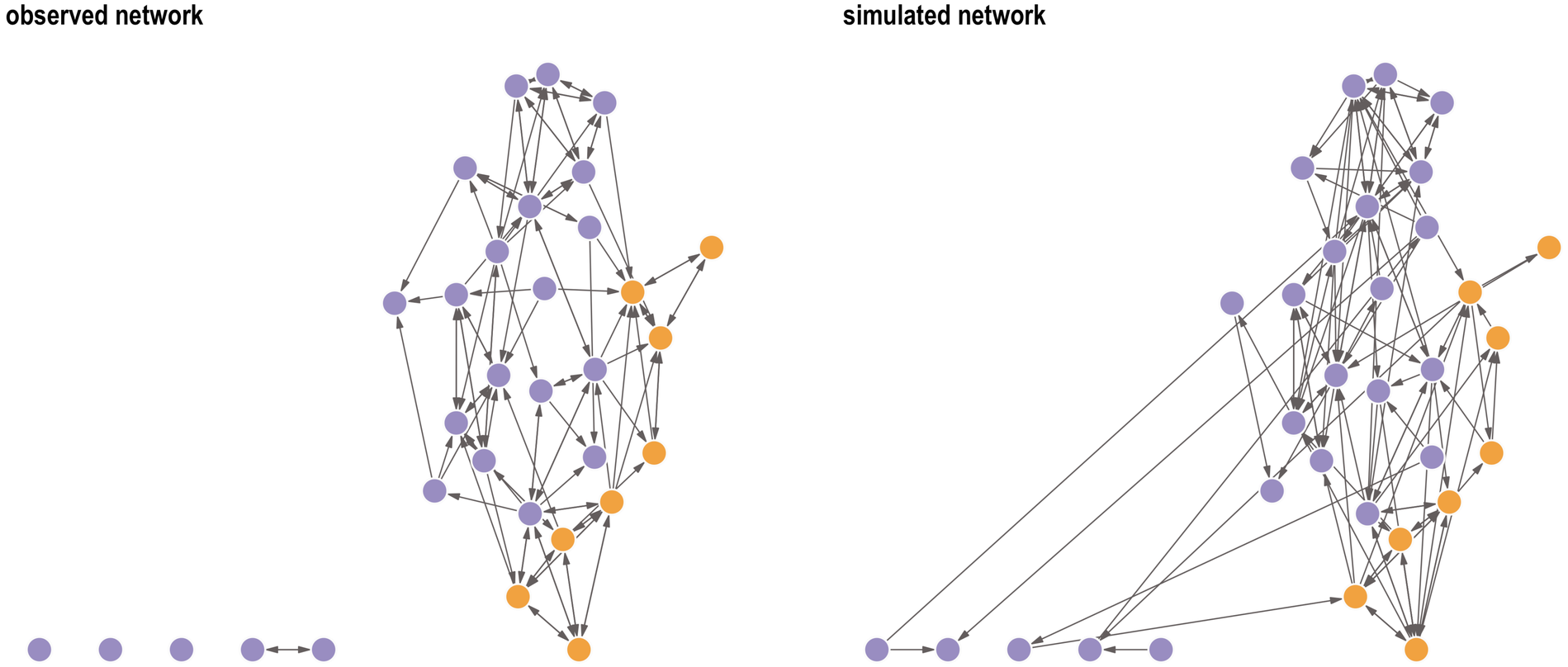
statistics (s_i)	**interpretation**
1	constant (intercept)
same_gender (u, v)	gender homophily
same_smoke (u, v)	behavior homophily
y_{vu}	reciprocity
$\sum_w y_{uw} y_{wv}$	transitivity

example: friendship among university freshmen

let's instead move to a more 'conventional' parametric model to study social selection by gender

logistic regression to model the presence/absence of ties:

model 2: observed versus simulated network

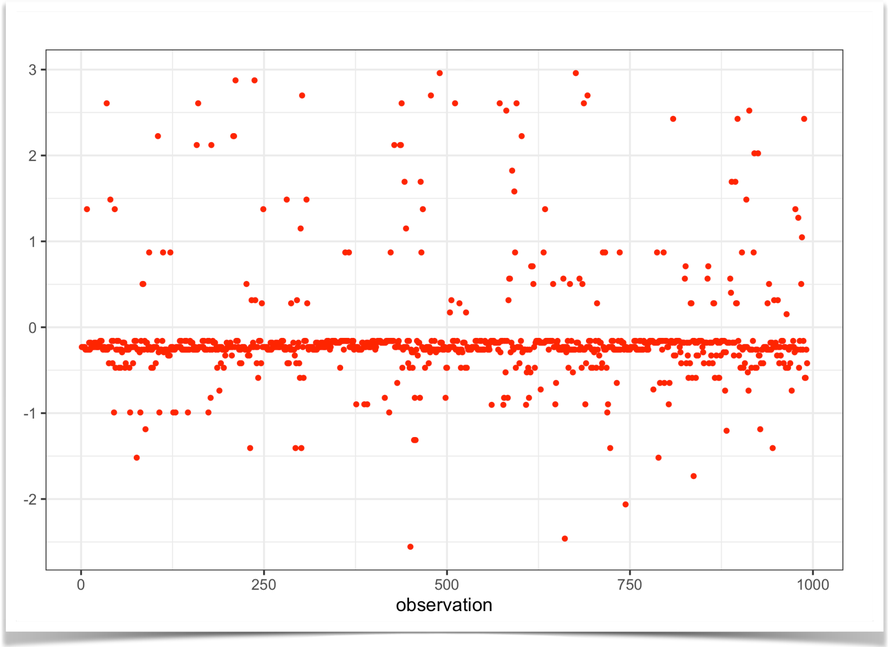


example: friendship among university freshmen

let's instead move to a more 'conventional' parametric model to study social selection by gender

logistic regression to model the presence/absence of ties:

but the analysis so far is **invalid**



- logistic model is only valid for independent observations
- for network data, different tie observations are **not** independent

let's turn to **random graph models**
where probability are assigned to graphs rather than individual ties

[note: we now have only **one** observation on the dependent variable which is the network]

random graph models

a random graph model is a probability space $(\mathcal{G}, P)$, where $\mathcal{G}$ is a (finite) set of graphs

a random graph model

- assigns probabilities to entire graphs (rather than to individual ties)
- implies tie probabilities (but is not determined by them)

let's look at three 'classic' random graph models:

(1) Erdös Rényi model (Bernoulli model)

(2) configuration model

(3) small world model

Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

a graph with n nodes where an tie exists with
independent random probability $0 < p < 1$ for each edge

short summary of model:

- the model is fully independent
- the tie probability of every dyad is equal to p
  - what is the most likely parameter value p?
  - MLE $\Longrightarrow$ the density of observed network $\Longrightarrow$ $\hat{p} = \dfrac{\text{number of edges}}{\text{number of dyads}}$
- degree distribution is binomial (and for large n approximately normal)
- expected degree of a node is approximately equal to $n \cdot p$
- same as the non-parametric $\mathcal{U} \mid E(L)$ model (can also be specified with ERGMs)

Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

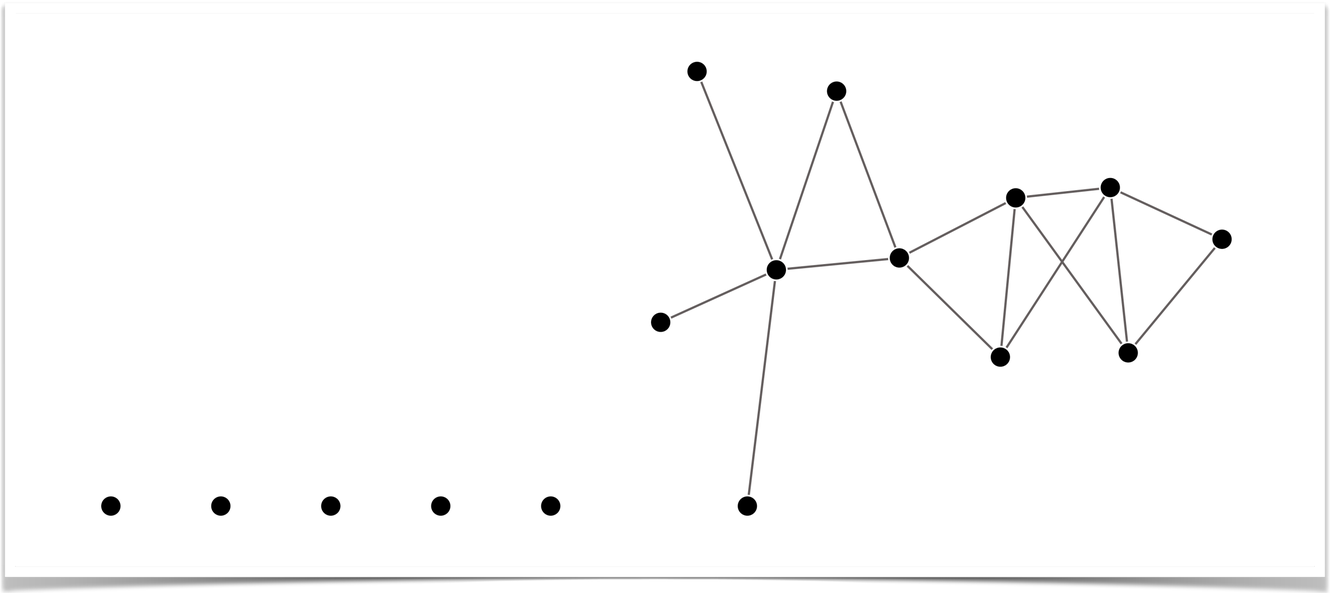
a graph with n nodes where an tie exists with
independent random probability $0 < p < 1$ for each edge

example. Florentine business network

number of ties = 15

number of dyads = 16(15)/2 = 120

density of network $\hat{p}$ = 15/120 = 0.125



to see how plausible the Bernoulli graph is as a representation of real world networks

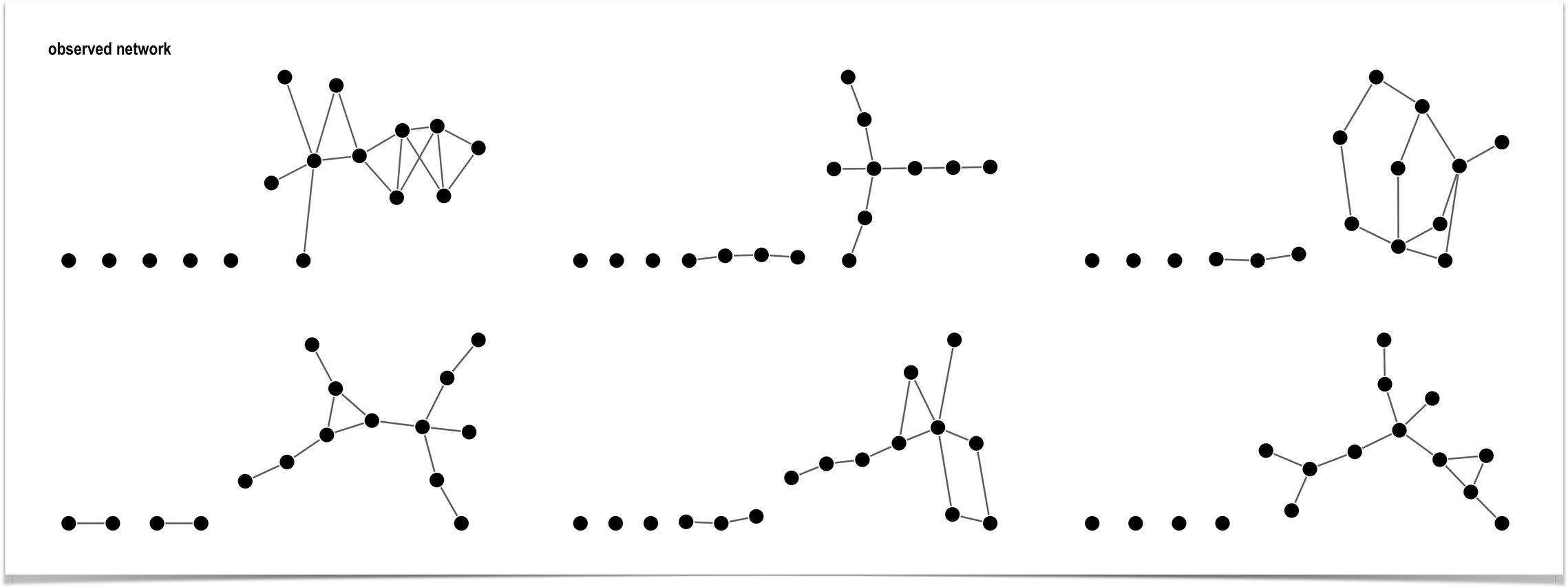
we simulate some networks from this model and look at some properties

Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

example. Florentine business network

5 simulated Bernoulli graphs with 15 ties: general structure



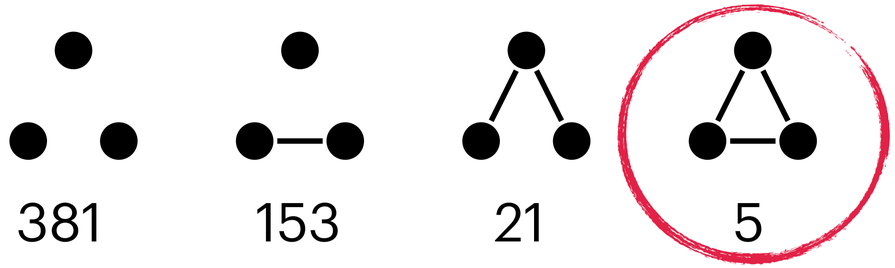
Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

example. Florentine business network

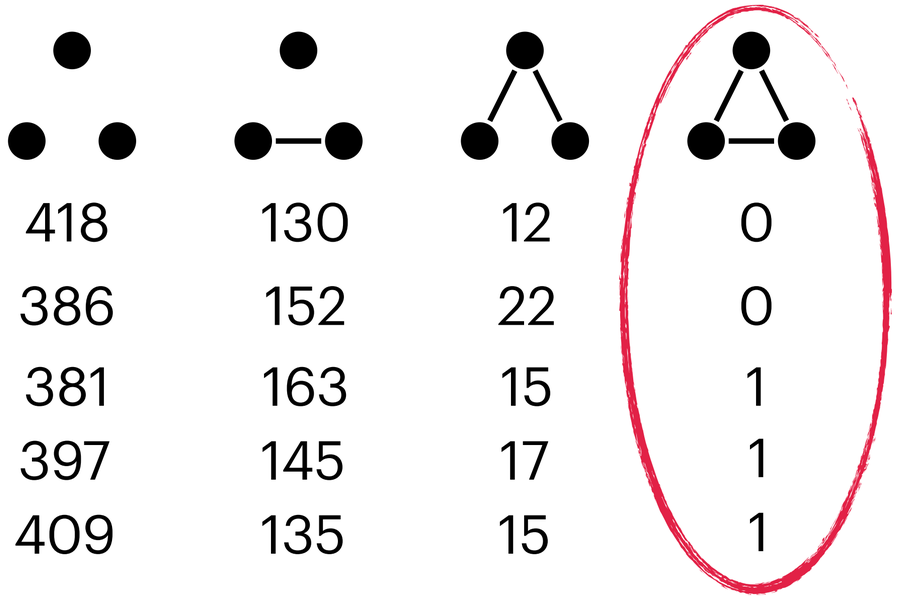
5 simulated Bernoulli graphs with 15 ties: triad census

observed network



381	153	21	5

simulated networks



418	130	12	0
386	152	22	0
381	163	15	1
397	145	17	1
409	135	15	1

transitivity effect

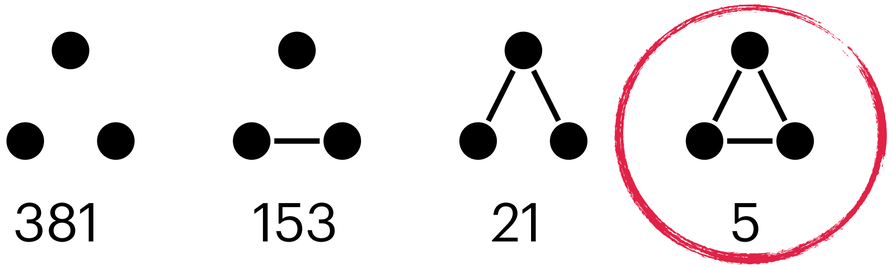
Erdös Rényi model (the Bernoulli graph)

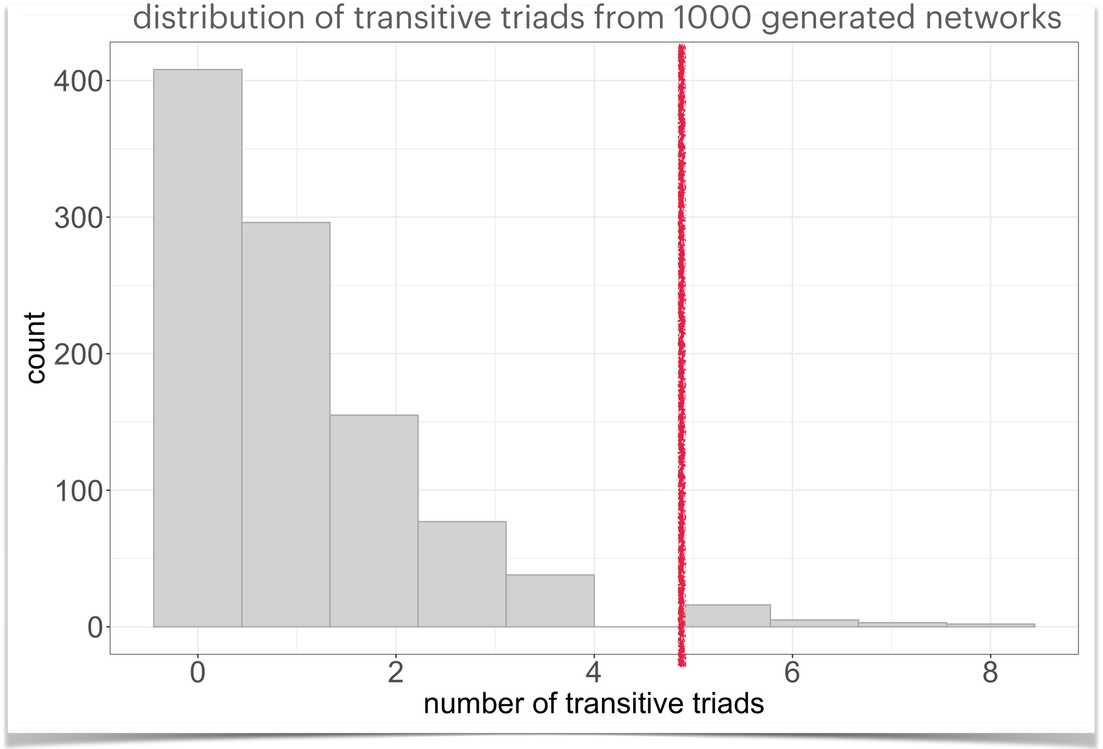
$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

example. Florentine business network

5 simulated Bernoulli graphs with 15 ties: triad census

observed network





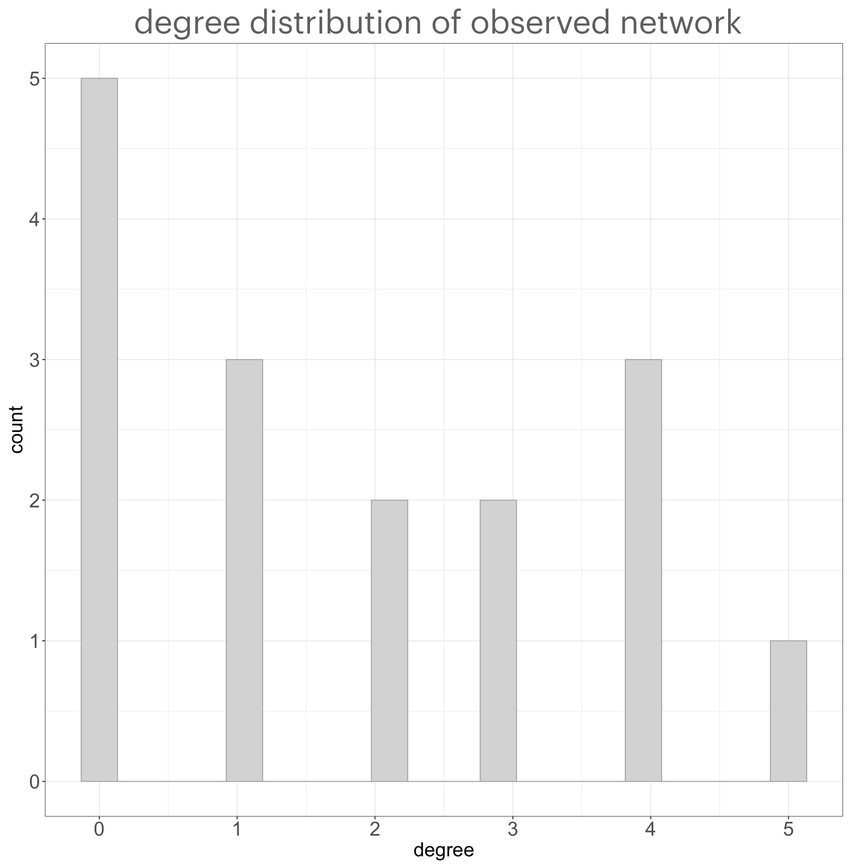
transitivity effect

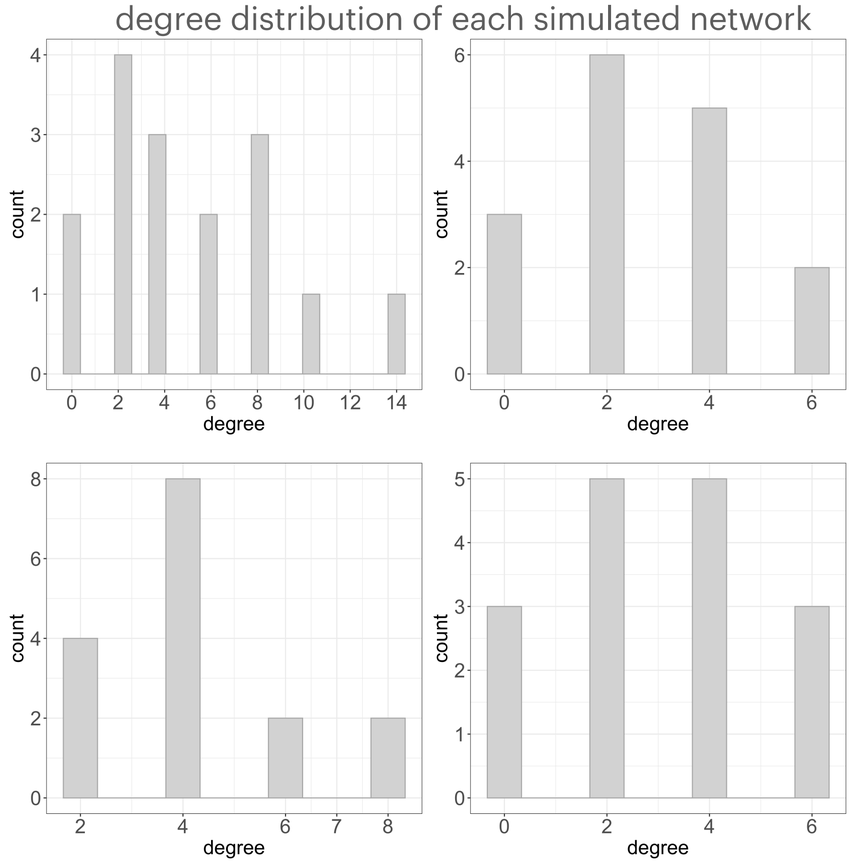
Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

example. Florentine business network

5 simulated Bernoulli graphs with 15 ties: degree distribution



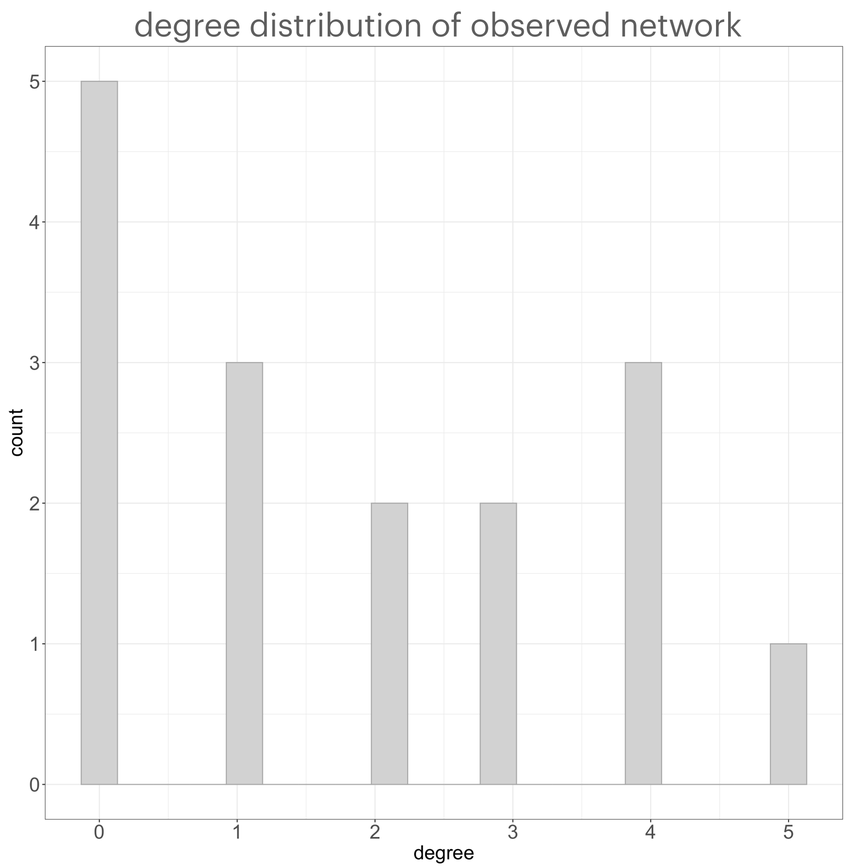


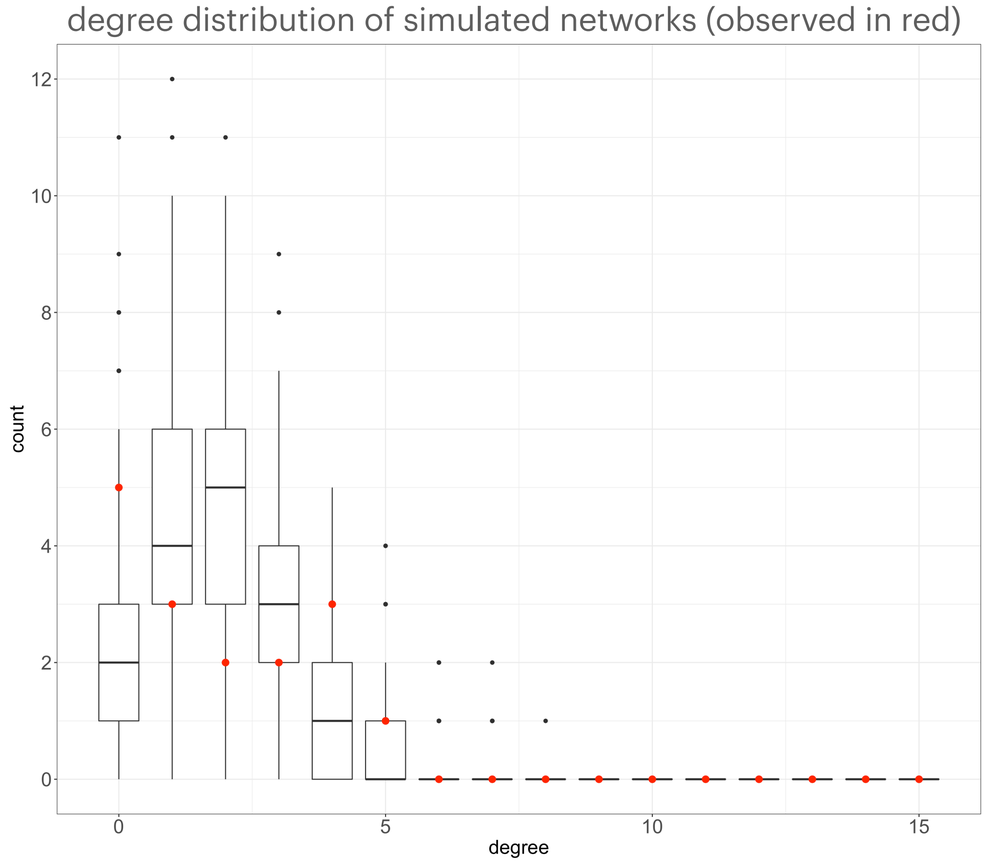
Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

example. Florentine business network

5 simulated Bernoulli graphs with 15 ties: degree distribution





Erdös Rényi model (the Bernoulli graph)

$\mathcal{G}(n, p)$ where n is number of nodes and p is edge probability

properties of Bernoulli graphs:

- most nodes are average linked (think normal distribution)
- the average distance between two nodes is small
- nodes do not tend to form clusters (no hubs)

because of these properties, Bernoulli graphs are not representative for real world networks

to account for the unrealistic degree distribution $\Longrightarrow$ **configuration model**

to account for small graph diameter and high clustering $\Longrightarrow$ **small world model**

configuration model

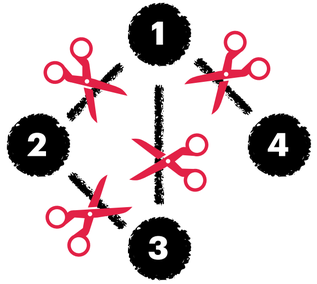
(Bender & Canfield, 1978)

random networks constrained by the observed degree sequence

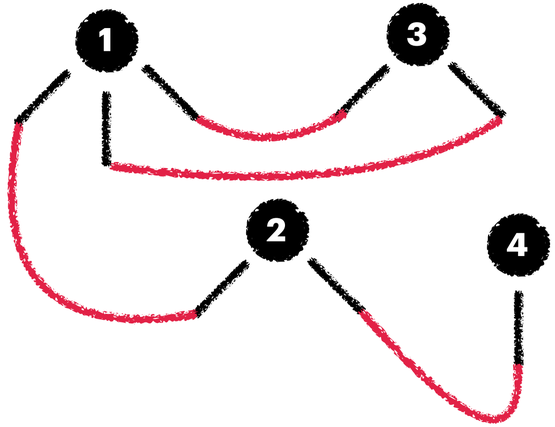
short summary of model:

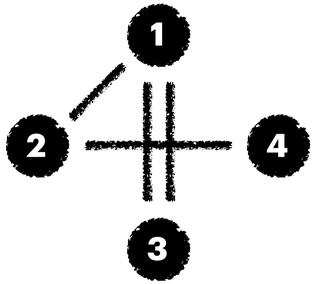
- the ties from an observed network are cut in half (stubs)
- these stubs are randomly coupled to form new ties in random networks

example.



$d_1 = 3,\ d_2 = 2,\ d_3 = 2,\ d_4 = 1$





$d_1 = 3,\ d_2 = 2,\ d_3 = 2,\ d_4 = 1$

limitations:

- loops and multi-edges
- sum of degree must be even

- same as the non-parametric $\mathscr{U} \mid \mathbf{d}$ model

small world model

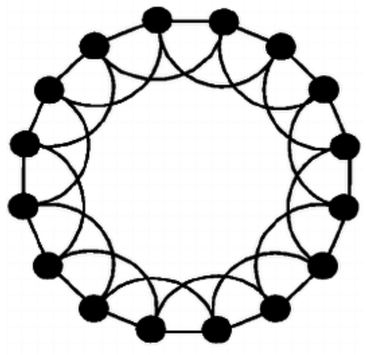
(Wattz & Strogatz, 1998)

random networks with short average path length (small world phenomena)

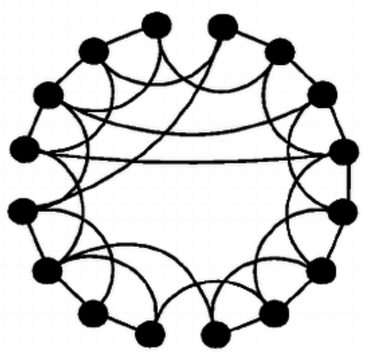
short summary of model:

- a circle network where each node is connected to a specified number of nearest neighbours
- rewire ties to new nodes given a probability p

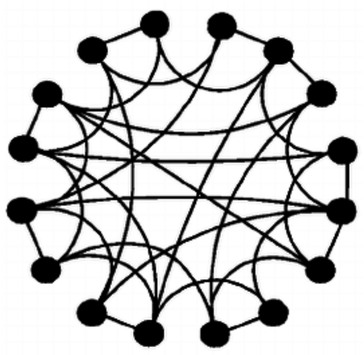
regular



small world



Bernoulli



$p = 0$ → $p = 1$

limitations:

- $0 < p < 0.5$

  all nodes have same degrees

- $p \approx 1$

  Bernoulli graph

beyond the basic random graph models

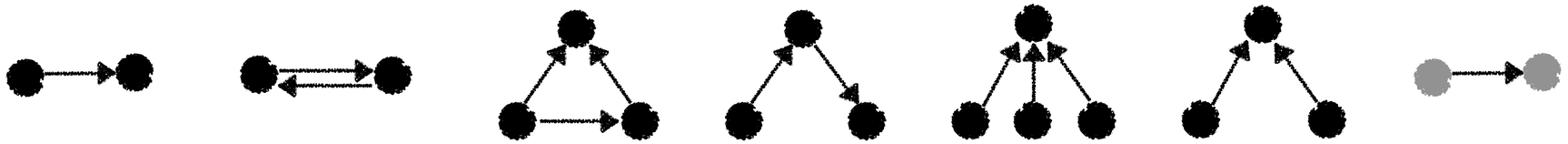
summary so far

- conditional uniform graph distributions
  - can only control for one network feature
- logistic regression
  - can control for several variables but invalid due to interdependent observations
- some random graph models
  - can only control for one network feature
  - found to be unrepresentative for real world networks

beyond the basic random graph models

we need a model that accounts for dependencies among the network ties
and can control for several network features at once

- these dependencies express various types of network self organization and
  - make network modelling interesting but difficult
  - are often the essence of social network theories
- dependence assumptions represent network pattern that are possible in the model
- these representations are **network configurations** which the network is built up of:



- network configurations
  - are nested in each other
  - represent competing explanatory mechanisms

four generations of dependence assumptions

- **Bernoulli dependence**

  network variables are independent of each other

- **dyadic dependence**

  dependence within dyads for directed networks

- **Markov dependence**

  network variables are conditionally dependent if they share at least one node

- **social circuit dependence**

  network variables are conditionally dependent if they create 4-cycles

[we will look at exponential random graph models (ERGMs) specified to include these dependence assumptions]