

# Graph models and hypothesis testing

Leto Peel  
[l.peel@maastrichtuniversity.nl](mailto:l.peel@maastrichtuniversity.nl)  
@PiratePeel

Why Graph models?

# Why Graph models?

Graph models allow us to generate synthetic networks

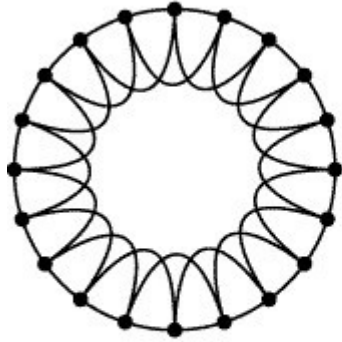
They allow us to capture and model properties observed in real networks

Graph models can serve as hypotheses for mechanisms of network formation

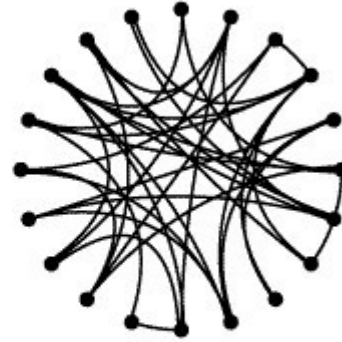
Allows us to explore how “similar” networks behave

# Regular and random graphs

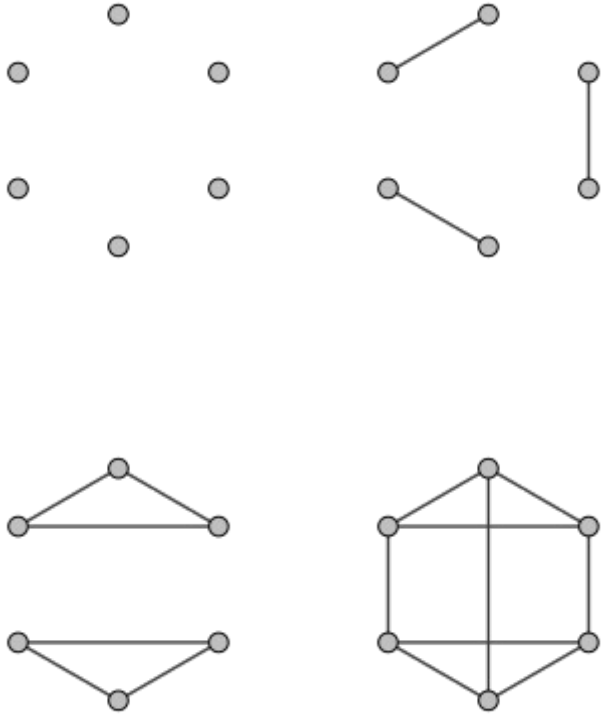
Regular



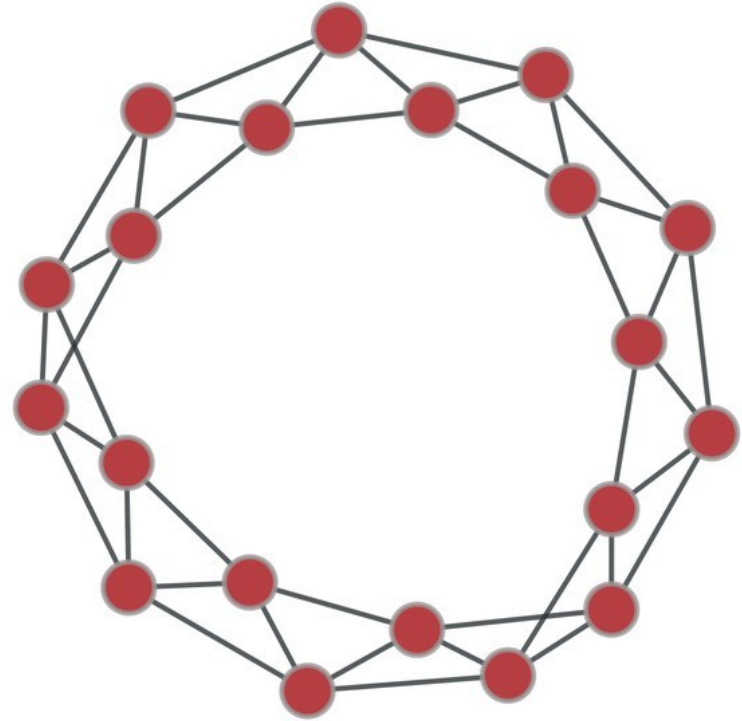
Random



# Regular graphs

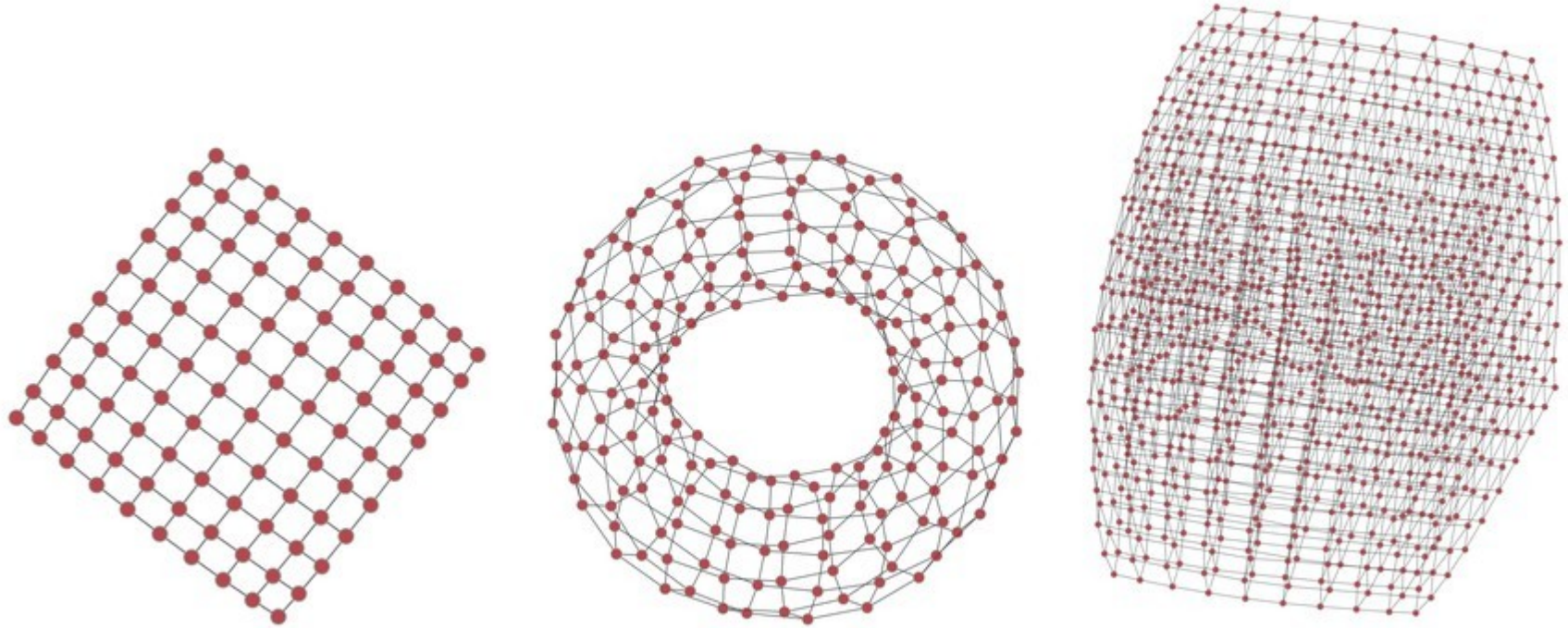


Regular graphs with 0 – 3 degree nodes



Regular Ring Lattice

# Lattice graphs



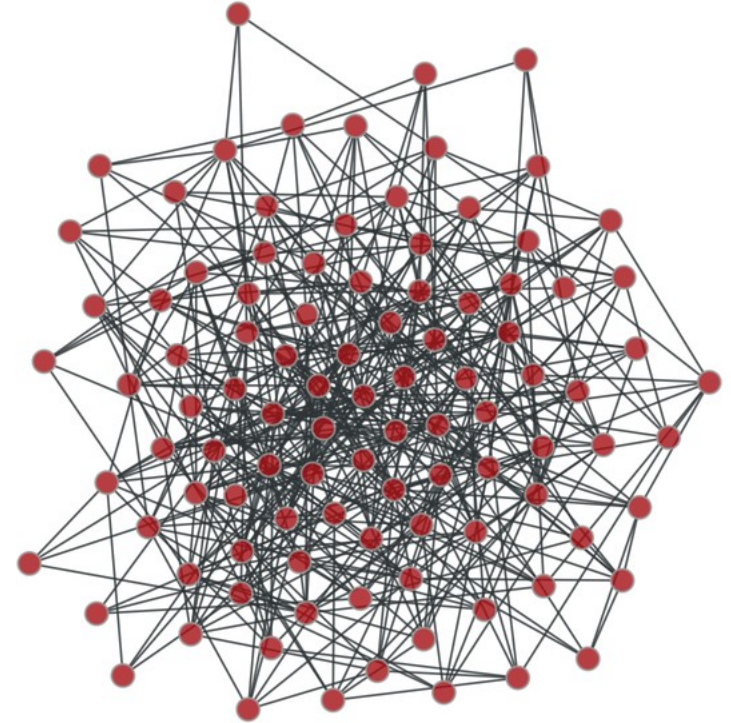
Every time we generate regular/lattice graphs we get the same output

# Random graphs

## The Erdos-Renyi model

Specified by a number of nodes,  $n$ ,  
and either:

- a number of edges,  $m$
- a probability of connection,  $p$



All edges are equally likely to exist

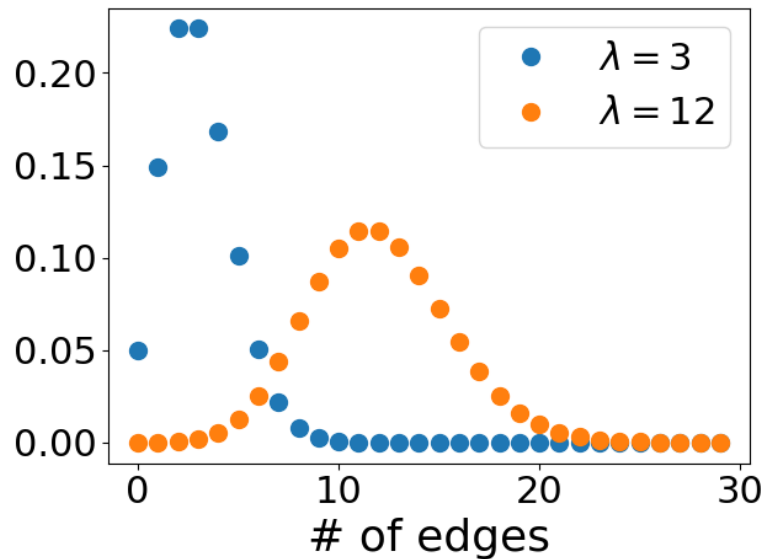
# Properties of random graphs

- Locally tree-like
- Large connected component for mean degree greater than 1
- Short path lengths



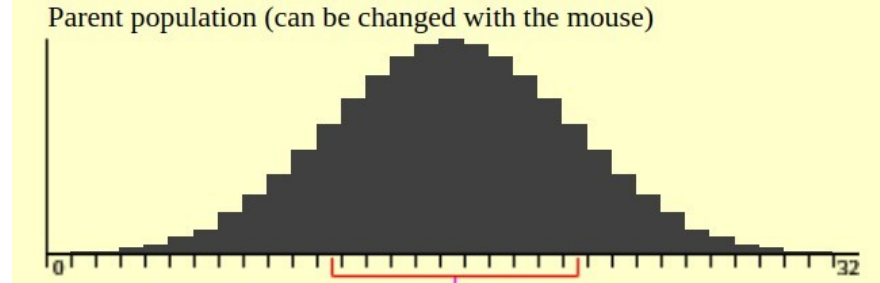
# When is the ER not so good?

- Degree distribution
- Clustering



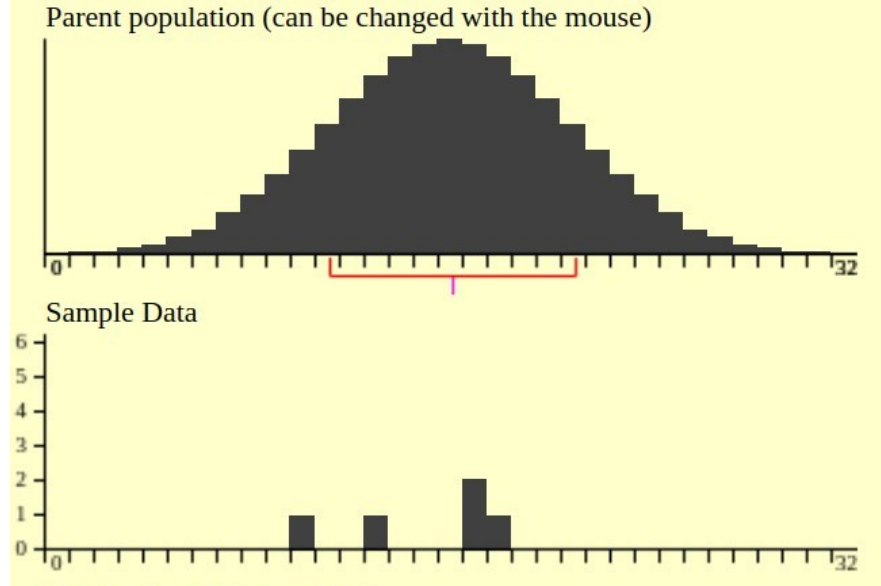
# Null hypothesis testing

Population distribution



Mean = 16.00  
Sd = 5.00

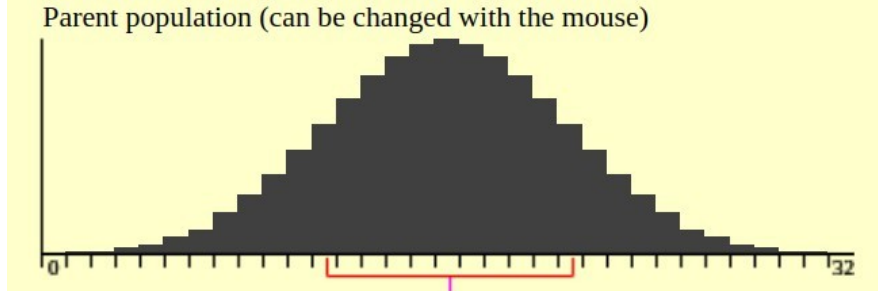
Population distribution



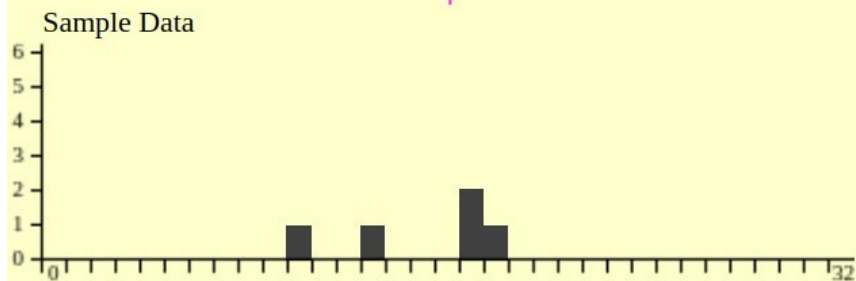
Mean = 16.00  
Sd = 5.00

One sample  
(size = 5)

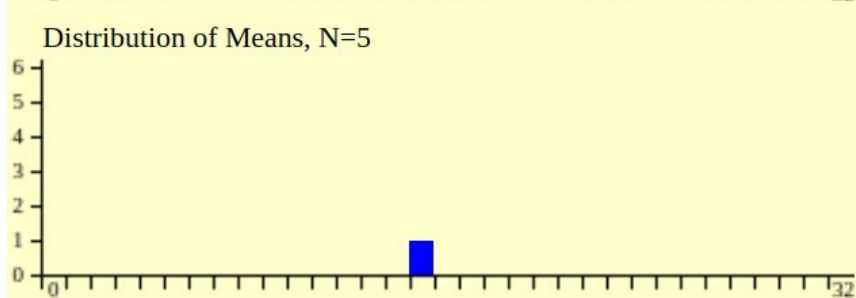
Population distribution



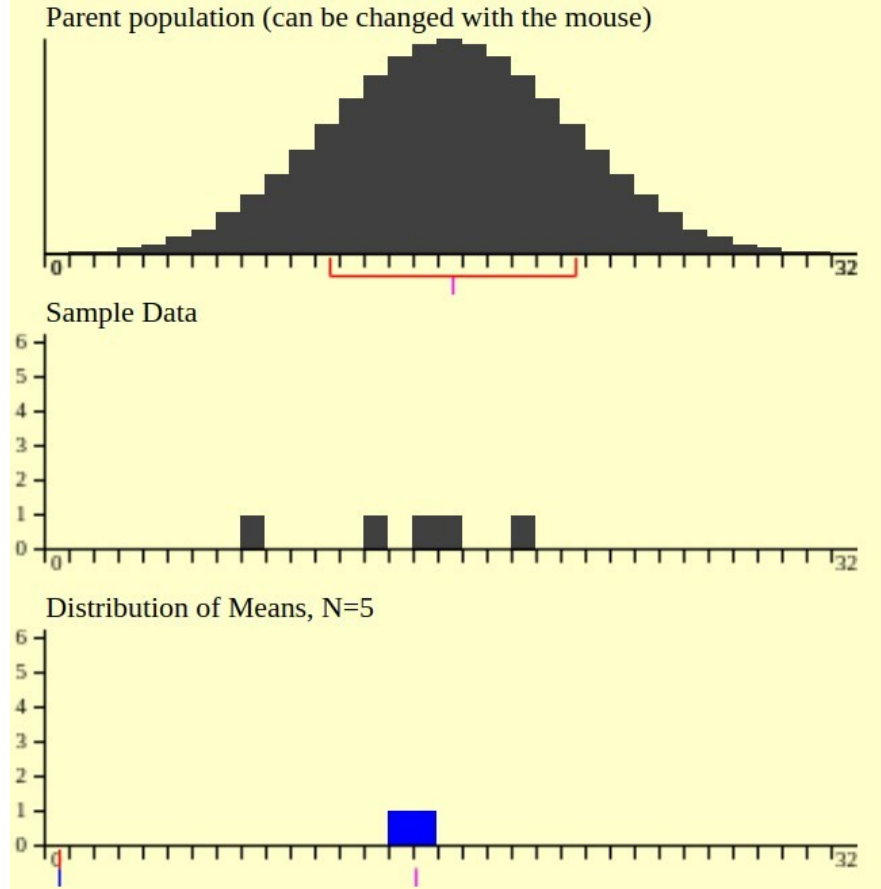
One sample  
(size = 5)



Distribution of means  
(1 sample)



Population distribution

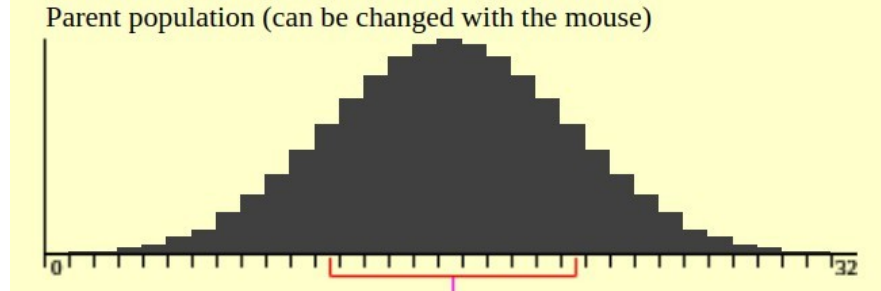


Mean = 16.00  
Sd = 5.00

One sample  
(size = 5)

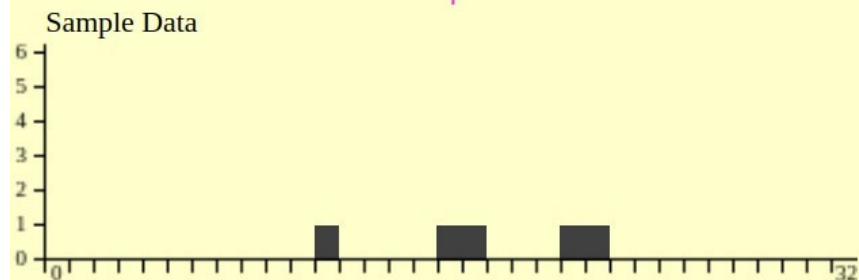
Distribution of means  
(2 samples)

Population distribution

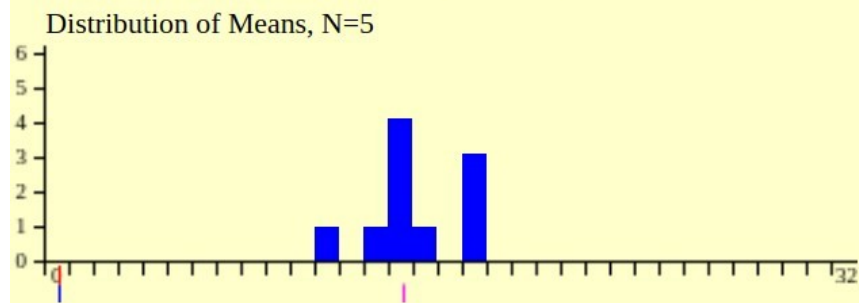


Mean = 16.00  
Sd = 5.00

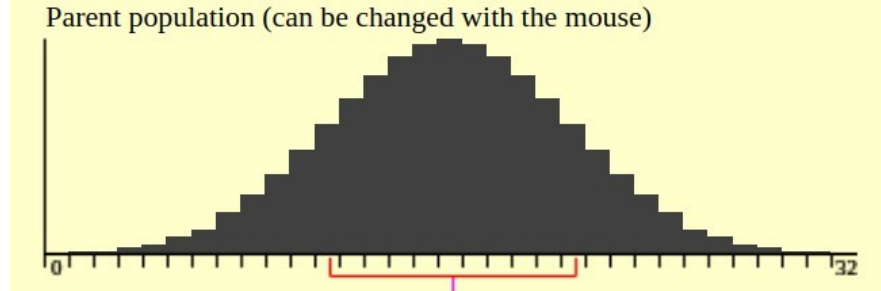
One sample  
(size = 5)



Distribution of means  
(10 samples)

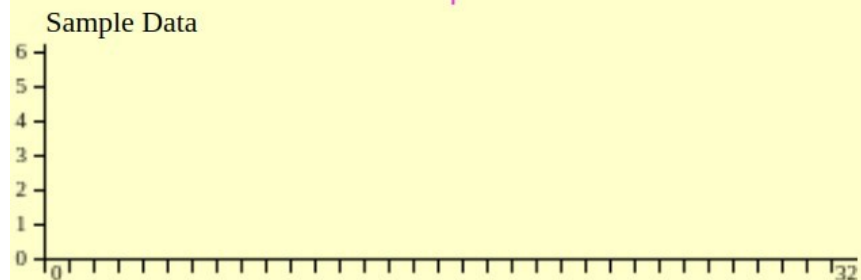


Population distribution

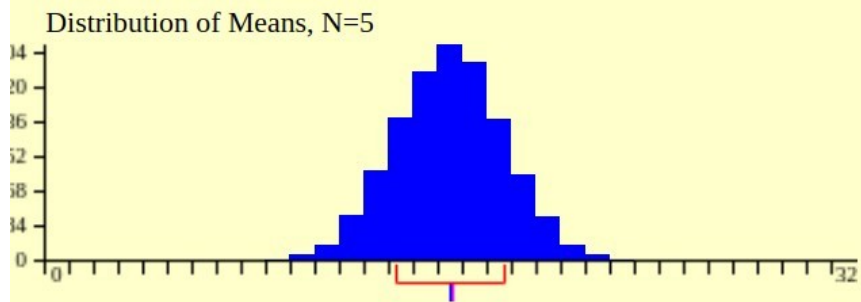


Mean = 16.00  
Sd = 5.00

One sample  
(size = 5)



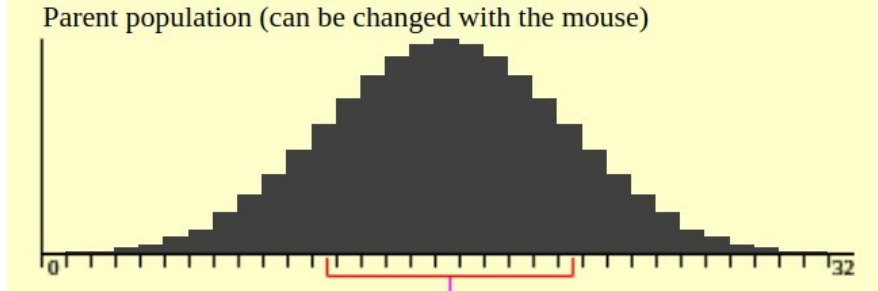
Distribution of means  
(10,000 samples)



Mean = 15.99  
Sd = 2.25

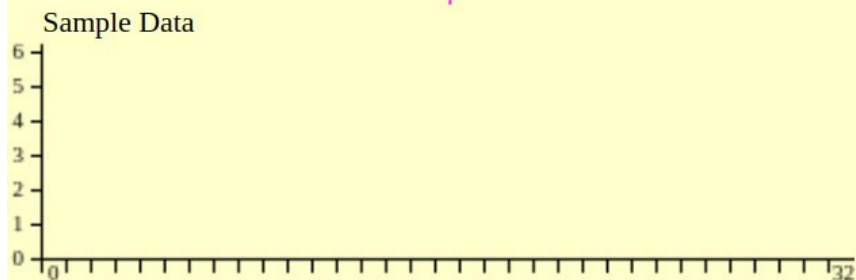


Population distribution

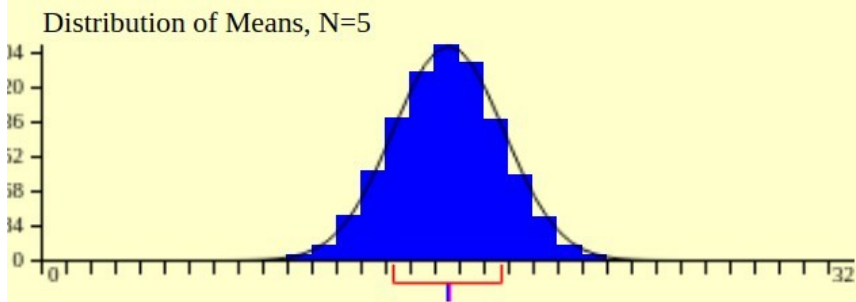


Mean = 16.00  
Sd = 5.00

One sample  
(size = 5)



Distribution of means  
(10,000 samples)



Mean = 15.99  
Sd = 2.25

Central limit theorem

# The infamous P-value

## *P*-VALUE

The probability, computed assuming that  $H_0$  is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. The smaller the *P*-value, the stronger the evidence against  $H_0$  provided by the data.

# The infamous P-value

## P-VALUE

The probability, computed assuming that  $H_0$  is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the **P-value** of the test. The smaller the P-value, the stronger the evidence against  $H_0$  provided by the data.

Definition, pg 405  
*Introduction to the Practice of Statistics, Fifth Edition*  
© 2005 W.H. Freeman and Company

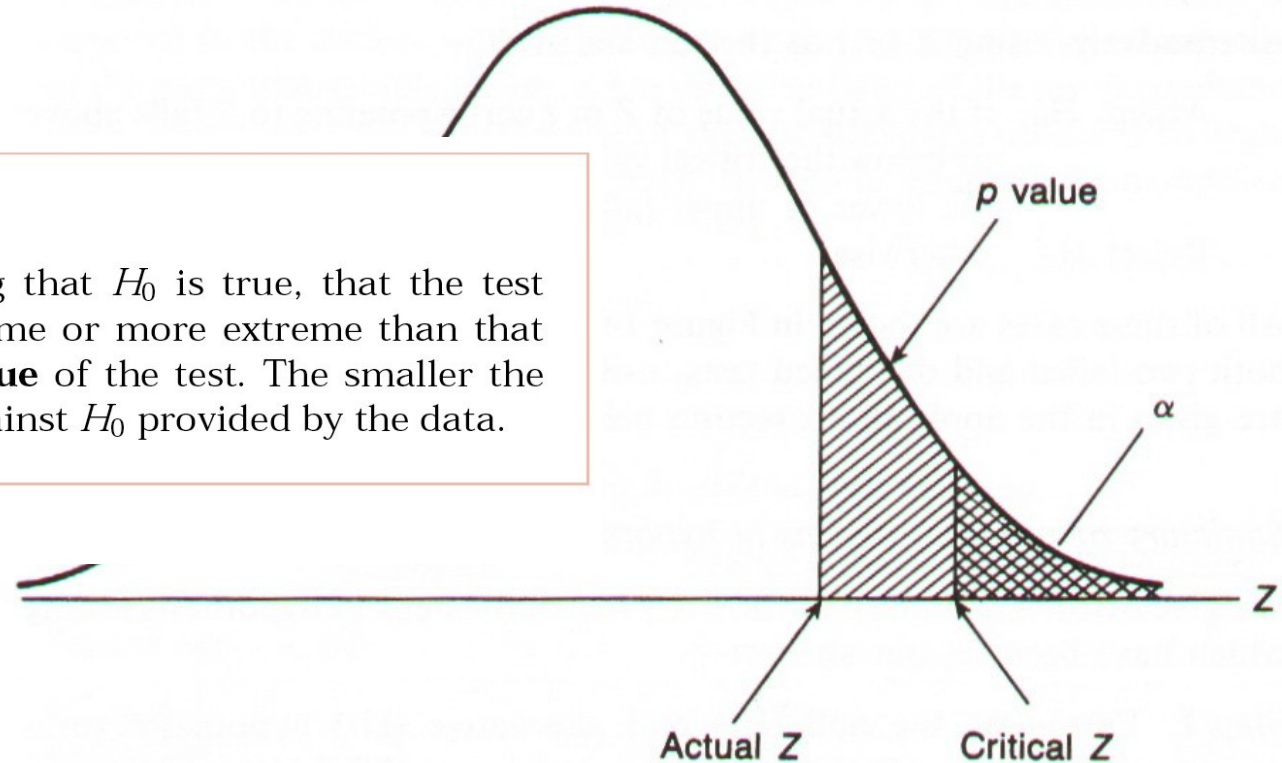


Figure 14.2 Comparison of  $p$  values and critical values of  $Z$  in a one-tailed test

# The infamous P-value

## P-VALUE

The probability, computed assuming that  $H_0$  is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the **P-value** of the test. The smaller the P-value, the stronger the evidence against  $H_0$  provided by the data.

Definition, pg 405  
*Introduction to the Practice of Statistics, Fifth Edition*  
© 2005 W.H. Freeman and Company

If  $p\text{-value} < \alpha$ ,  
then we reject the  
null hypothesis

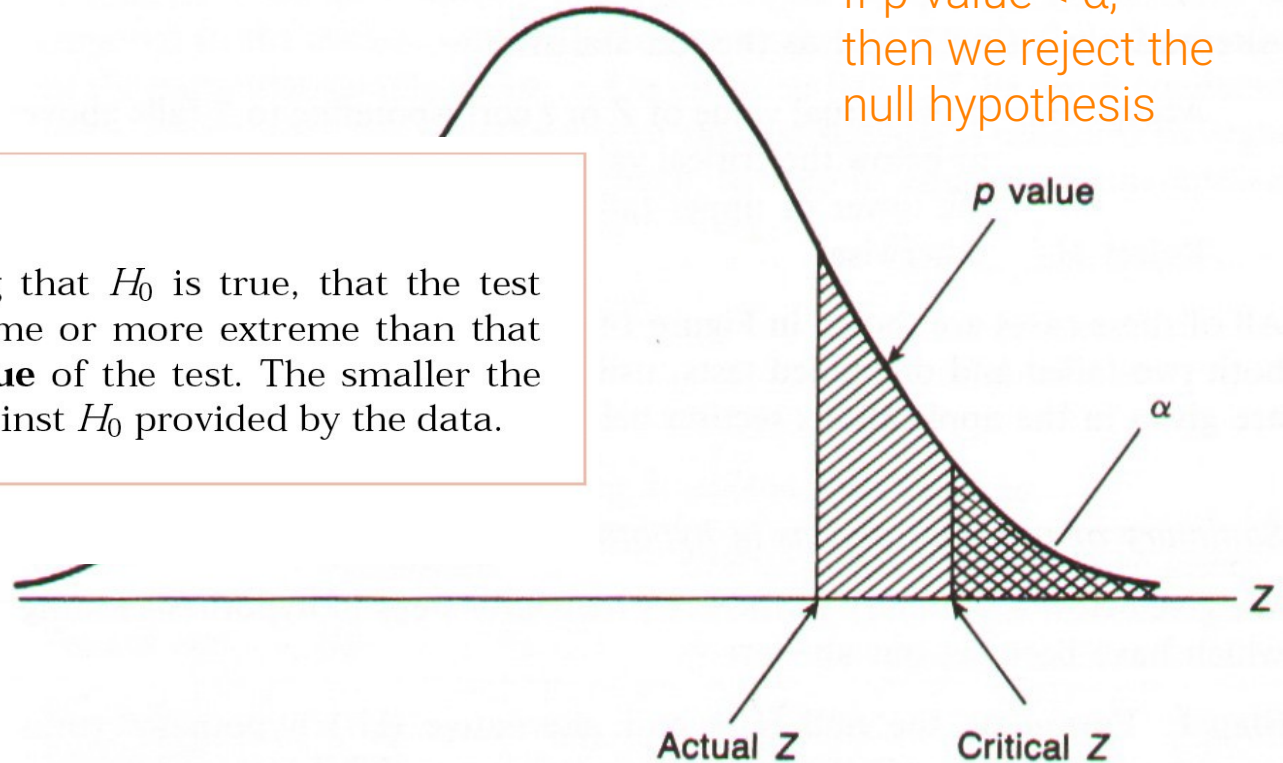


Figure 14.2 Comparison of  $p$  values and critical values of  $Z$  in a one-tailed test

# For networks we can use graph models as a null model

- Does the network appear to be significantly different from a random graph?
- Can specific properties of the observed network (e.g., clustering coefficient) be explained by a particular generative process?
- Two approaches to create samples:
  - permute edges/nodes in a way that is consistent with the graph model
  - “fit” a model to an observed network and generate networks from it

# Practical Part 1

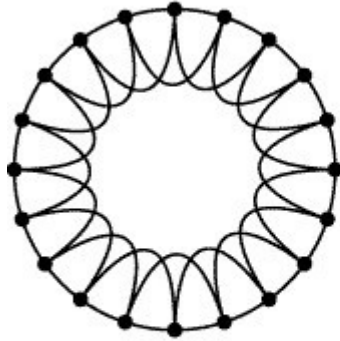
# Small-world networks

It's a network after all

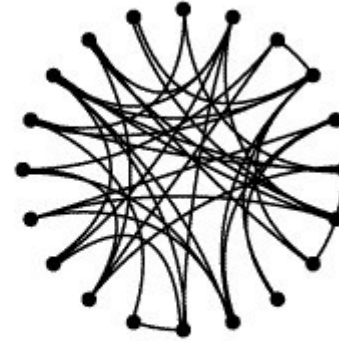
# Small-world networks

It's a network after all

Regular



Random

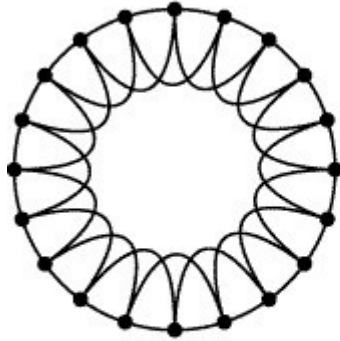




# Small-world networks

It's a network after all

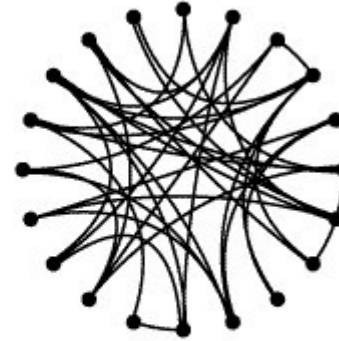
Regular



(triangles)

High **clustering coefficient**

Random

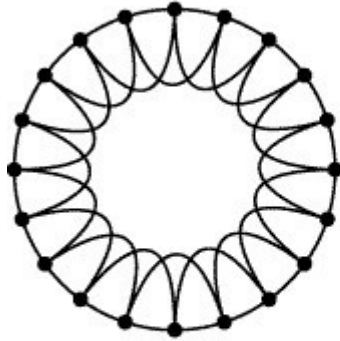


Low **clustering coefficient**

# Small-world networks

It's a network after all

Regular



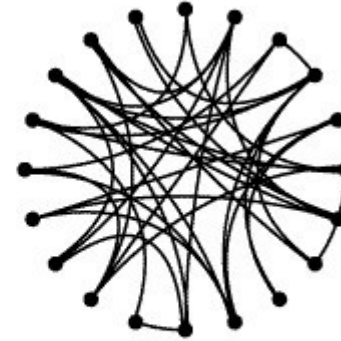
(triangles)

High **clustering coefficient**

(shortest  
paths)

High mean **geodesic path**

Random



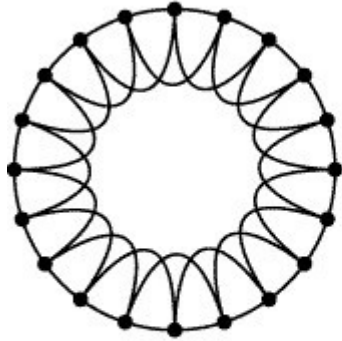
Low **clustering coefficient**

Low mean **geodesic path**

# Small-world networks

It's a network after all

Regular



(triangles)

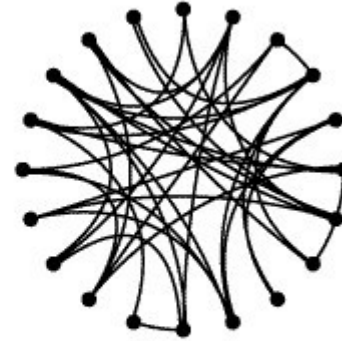
High clustering coefficient

(shortest paths)

High mean geodesic path

Real-world networks

Random

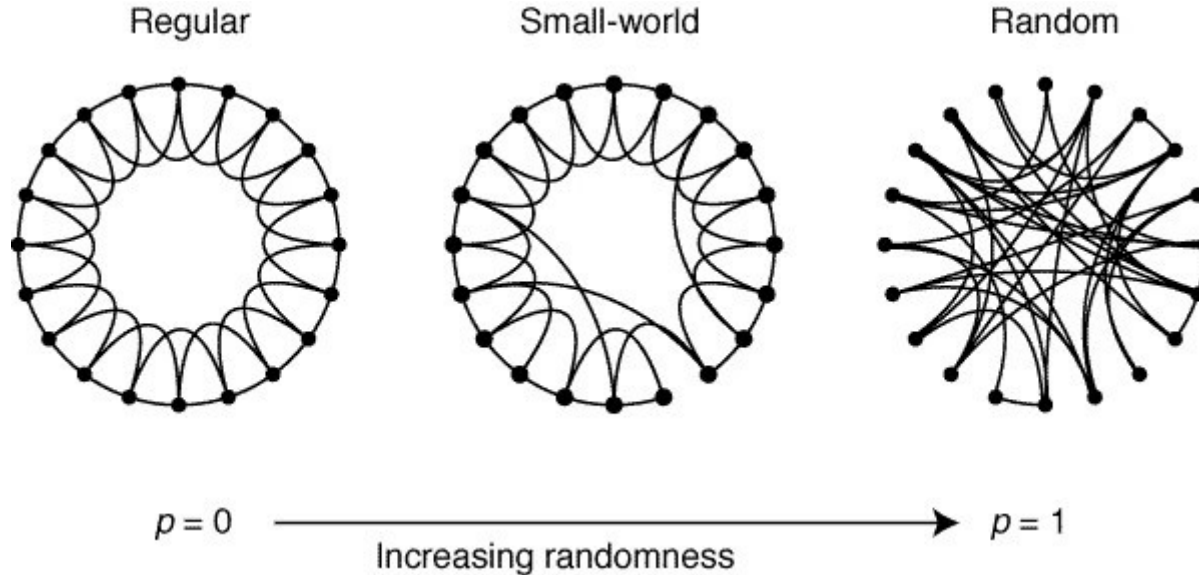


Low clustering coefficient

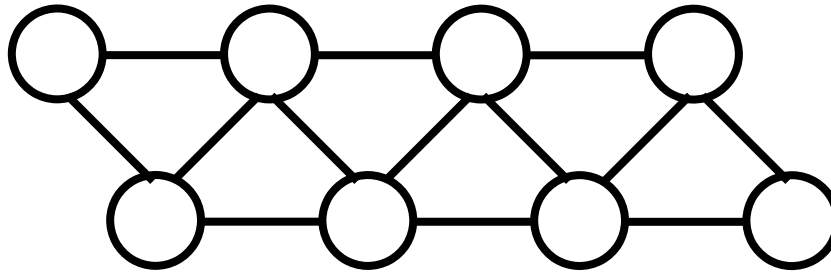
Low mean geodesic path

# Small-world networks

It's a network after all

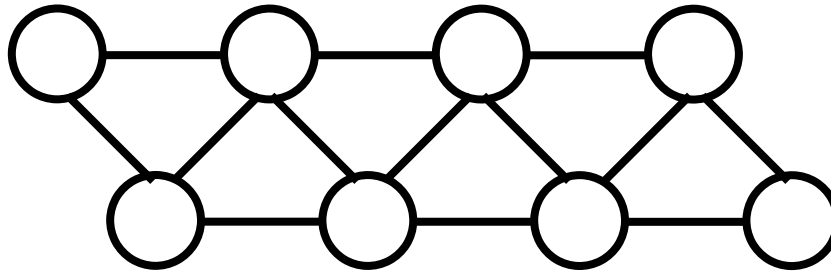


# Exercise



Calculate the **clustering coefficient**  
and mean **shortest path**

# Exercise

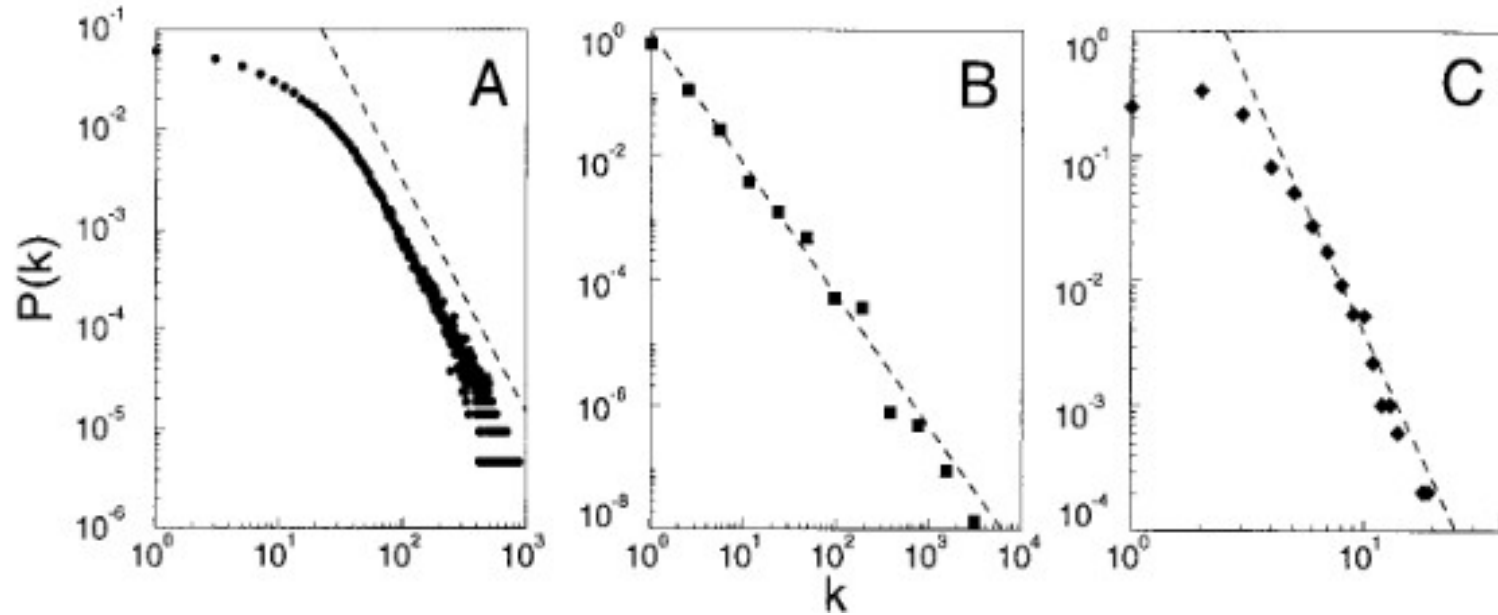


Calculate the clustering coefficient  
and mean shortest path

Now choose a pair of edges to  
randomly rewire and recalculate

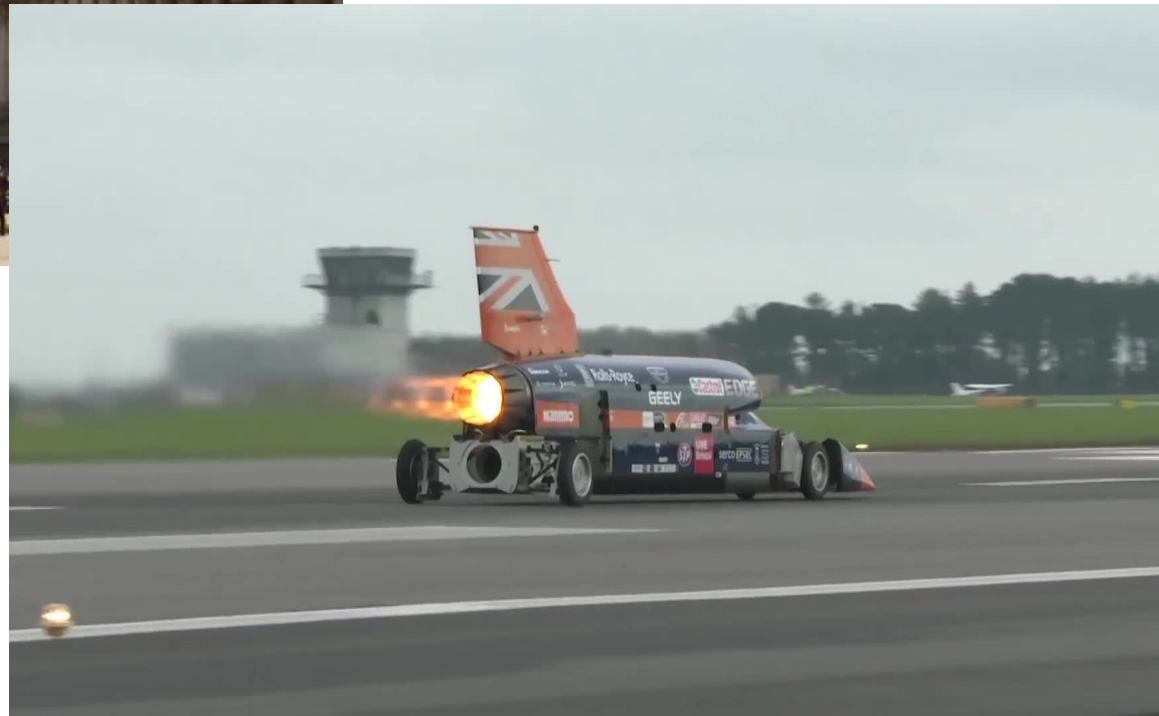
# "Scale-free" networks

# "Scale-free" networks









# The Price model

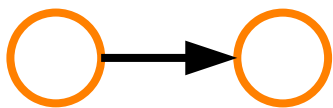
For undirected networks, this model is known as the Barabasi-Albert model

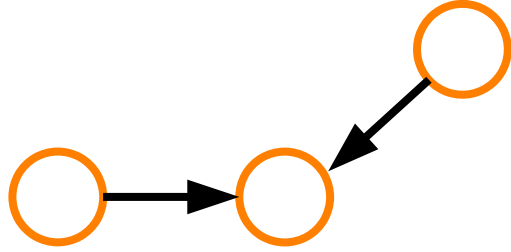
Add nodes to a network one at a time.

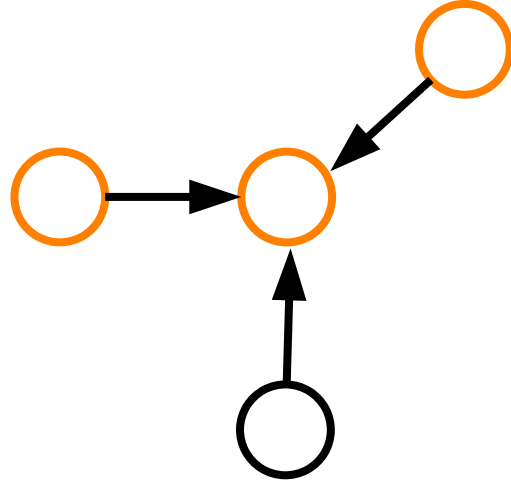
Connect to existing nodes with probability **proportional to their degree**

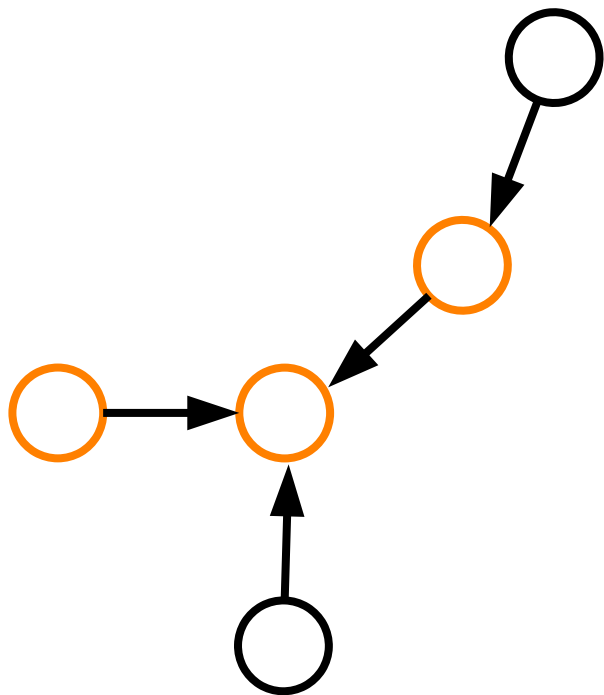
$$p_i = \frac{k_i}{\sum_j k_j},$$



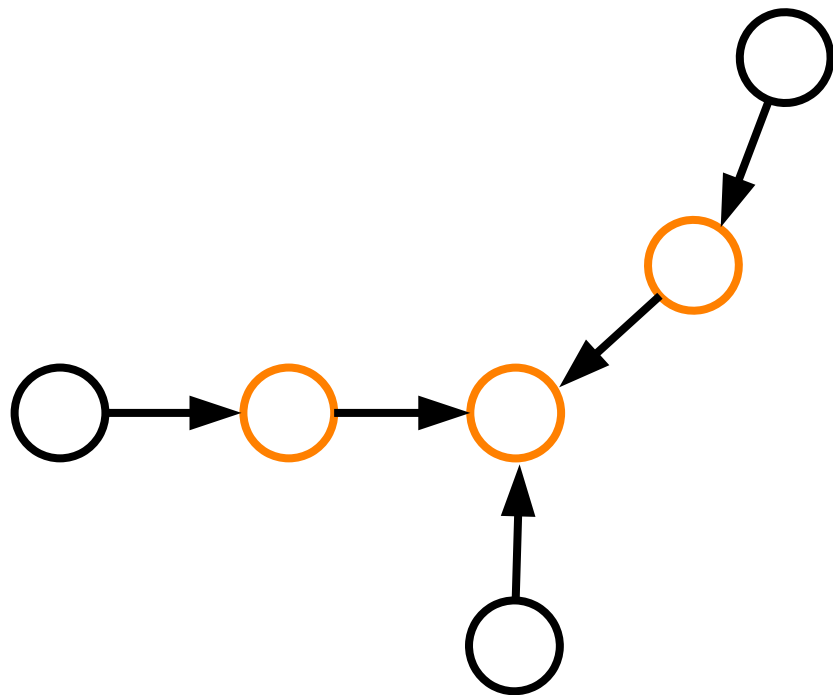


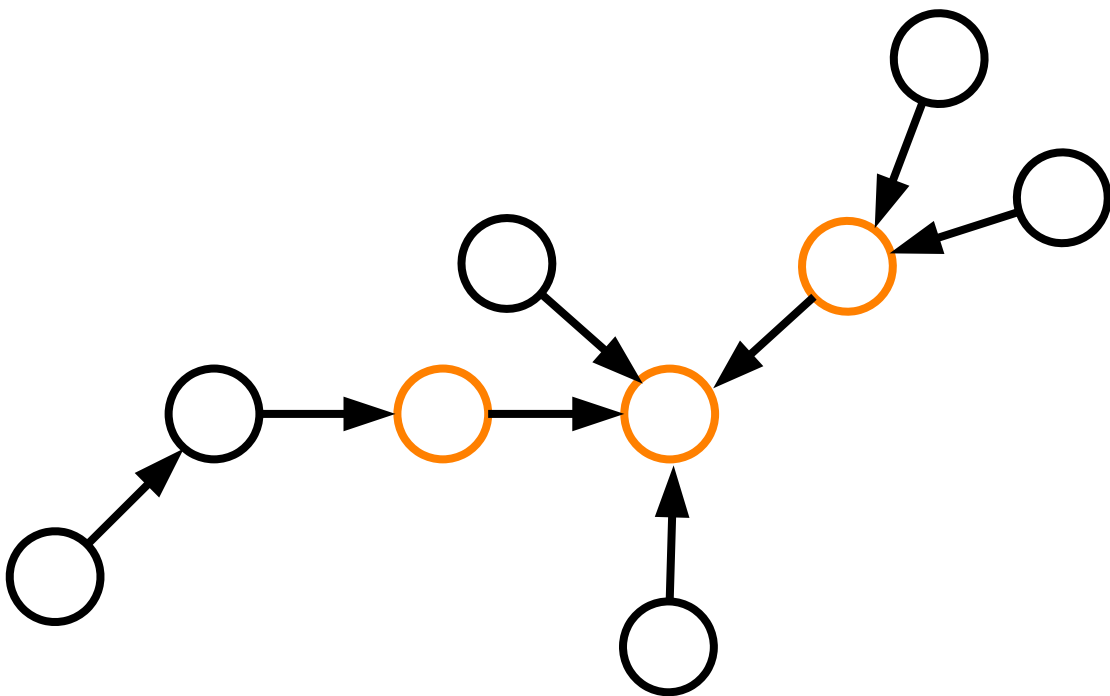


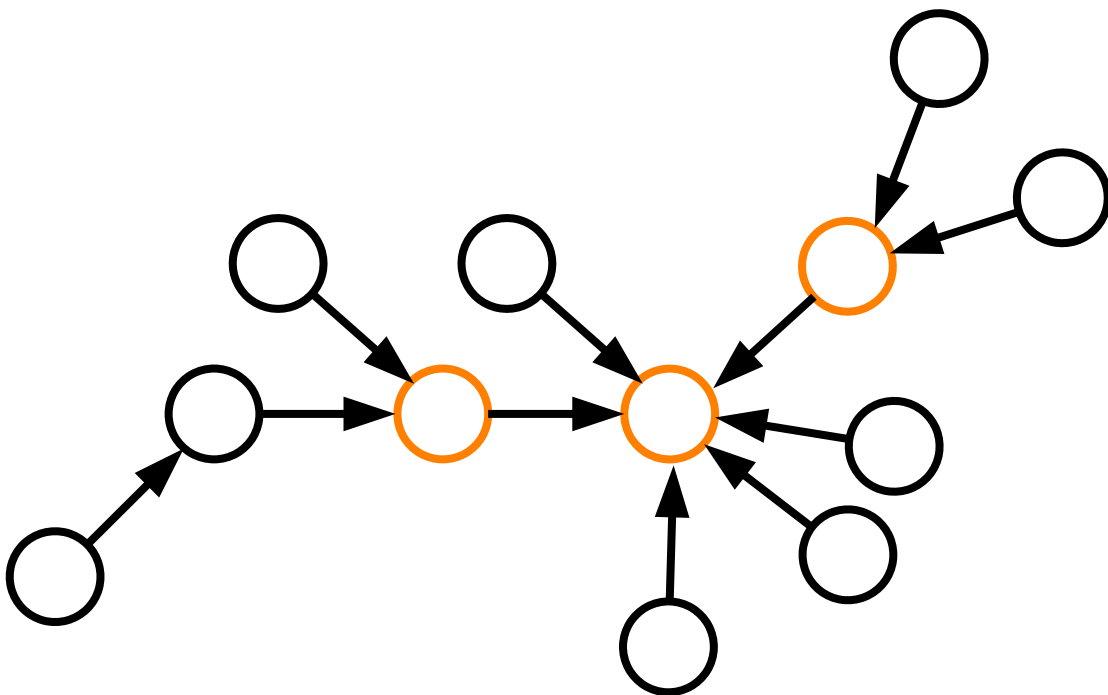


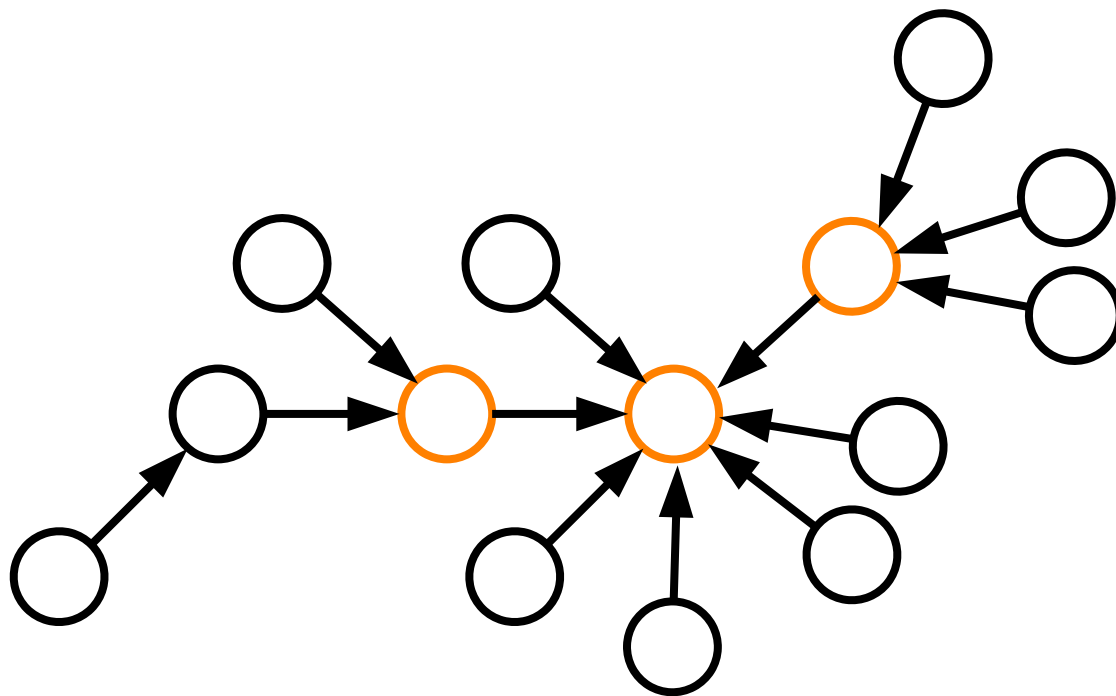




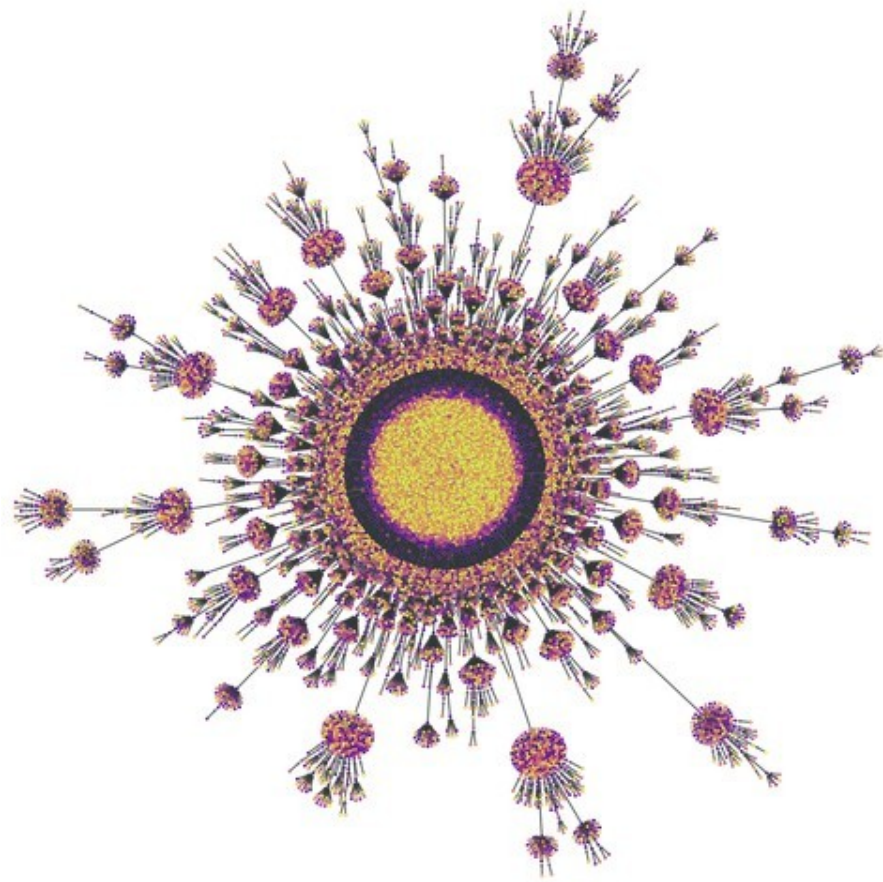
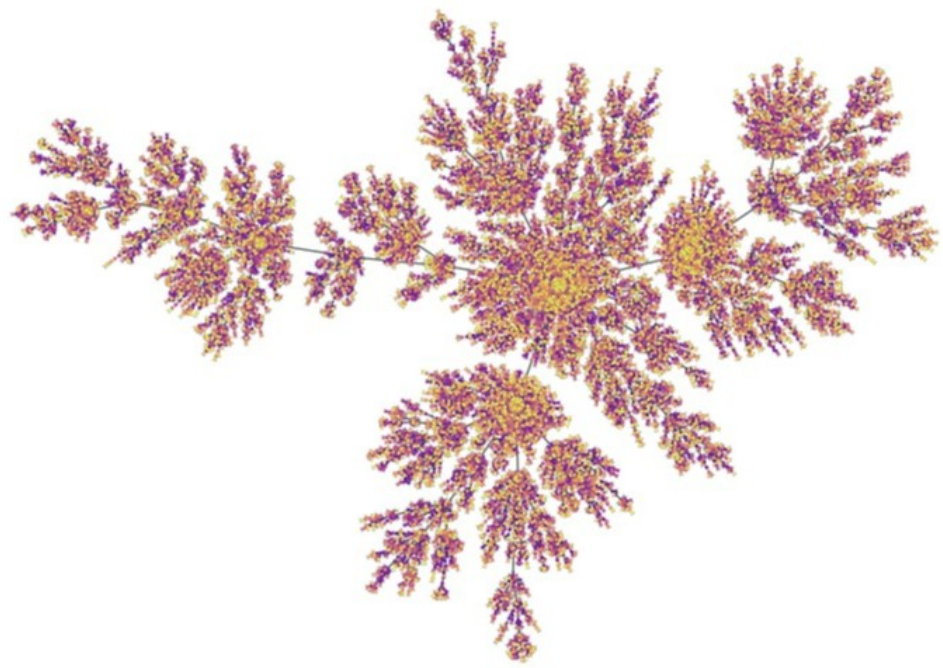








Nodes that join the network earlier have higher degree



# Configuration model

- More realistic than ER
- Edges are random conditional on a specified degree sequence
- Unlike the ER model, probability of the degree of a randomly selected node is not the same as the probability of the degree of its neighbours

# The friendship paradox

Your friends are more popular than you are

# The friendship paradox

Your friends are more popular than you are

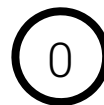
Some people have **no friends**.



# The friendship paradox

Your friends are more popular than you are

Some people have **no friends**. But because **they appear in nobody's friendship circles**, they're not making anyone else feel popular.





# The friendship paradox

Your friends are more popular than you are

average friend  
(count node proportional  
to their degree)

average person  
(count each node once)

So popular people are oversampled as friends. Hence the paradox.

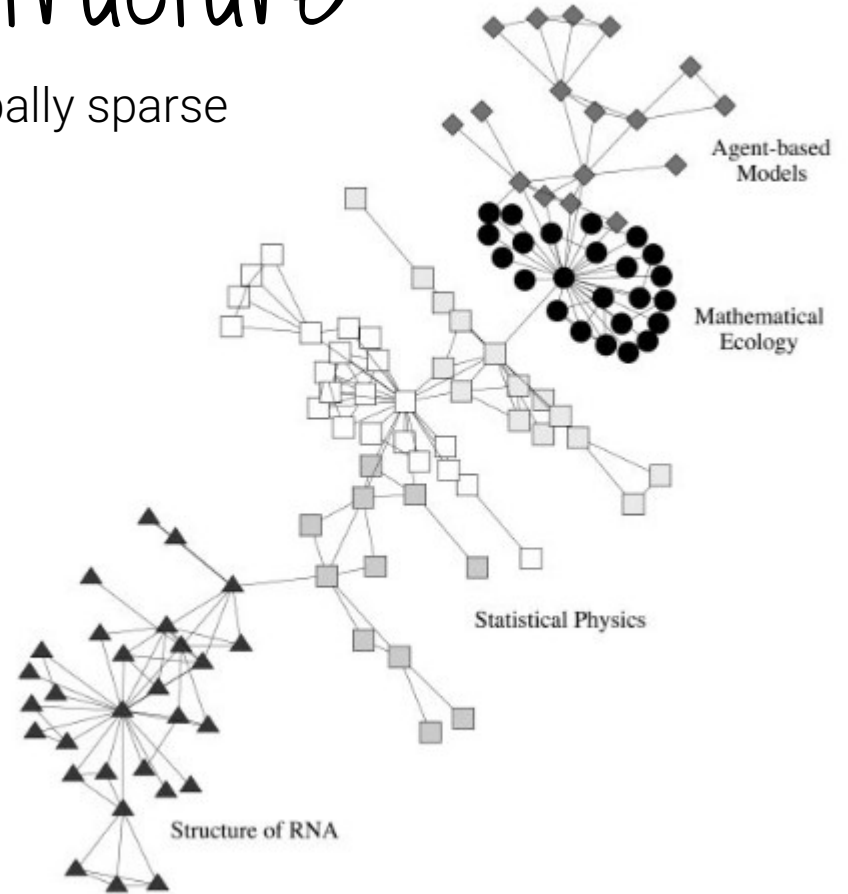
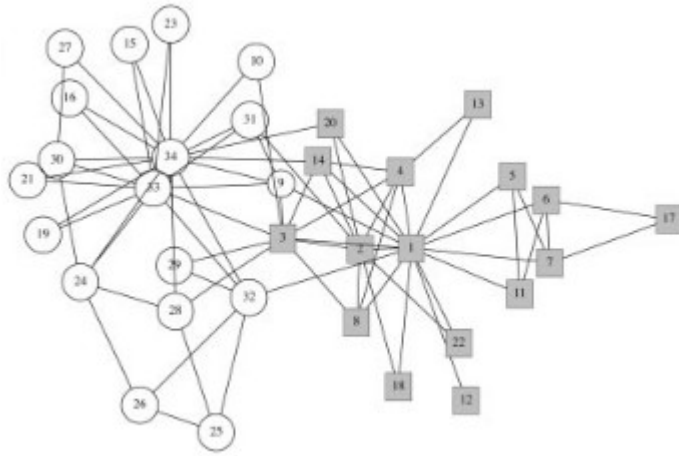
# Practical Part II

# Community structure

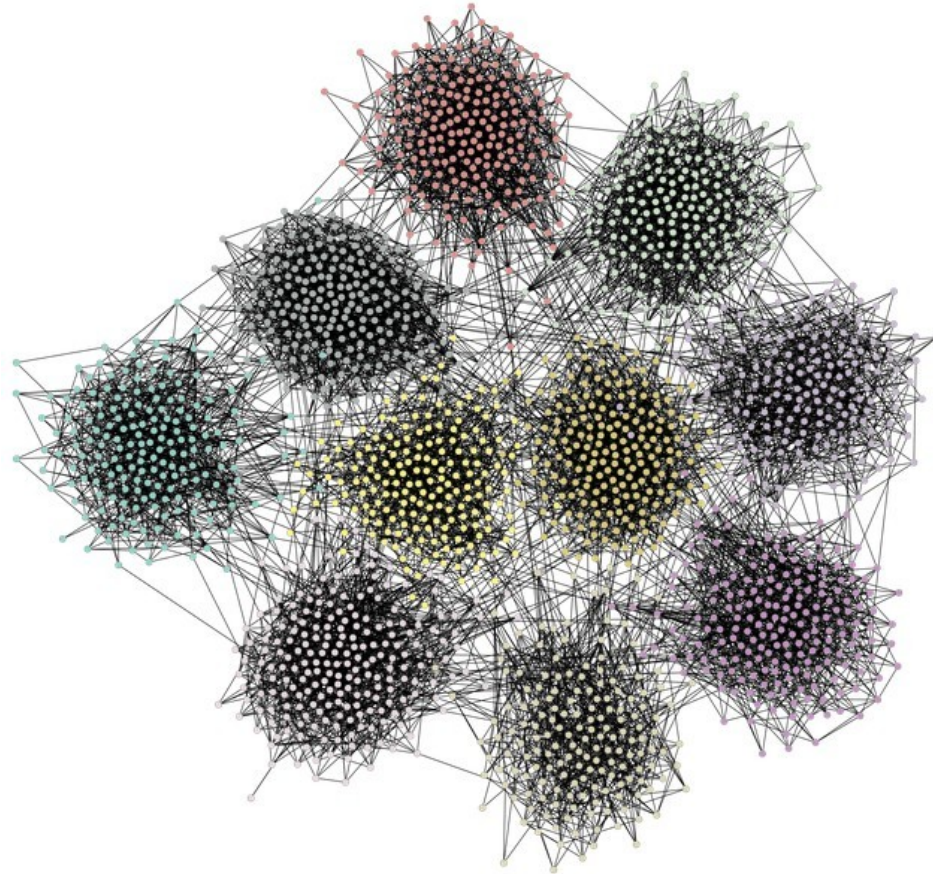
Locally dense, globally sparse

# Community structure

Locally dense, globally sparse



# Stochastic Block Models



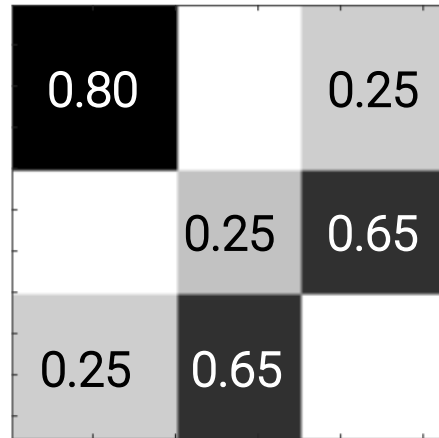
# Generating a network using the SBM

Step 1 : Assign each node to a group



# Generating a network using the SBM

Step 1 : Assign each node to a group



Mixing Matrix

- Step 2 : Select some connection probabilities (mixing matrix)

# Generating a network using the SBM

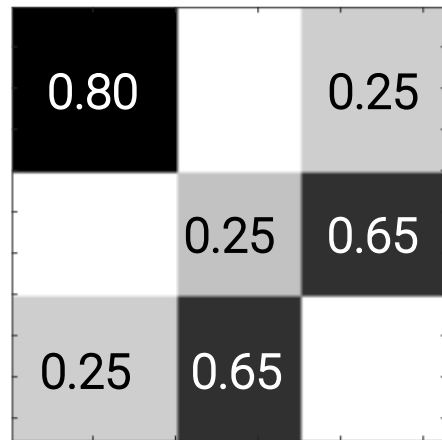
Step 1 : Assign each node to a group



Mixing Matrix

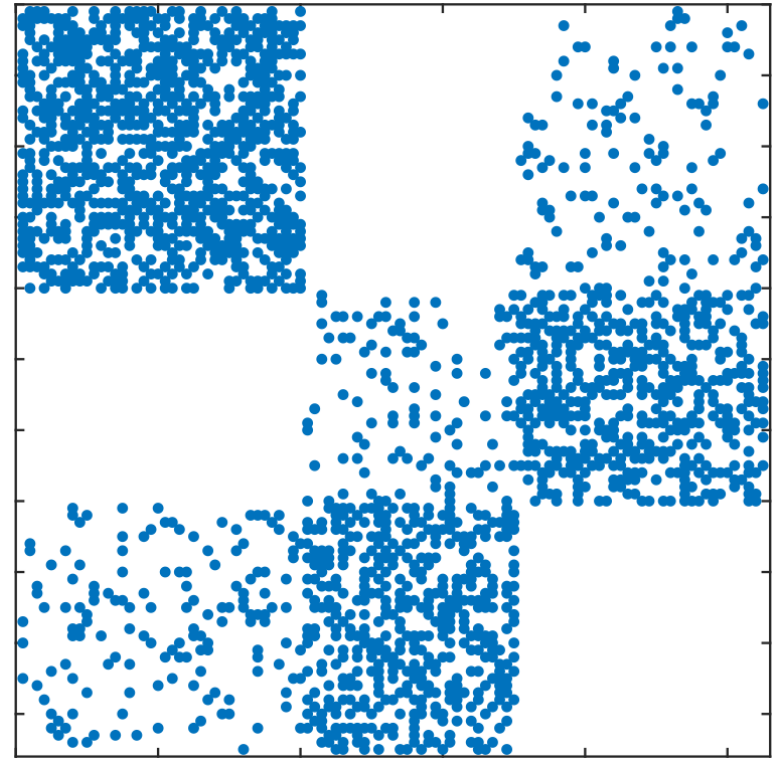

- Step 2 : Select some connection probabilities (mixing matrix)
- Step 3 : For each pair of nodes, add an edge with probability according to the group memberships

# Generating a network using the SBM



Mixing Matrix

generation



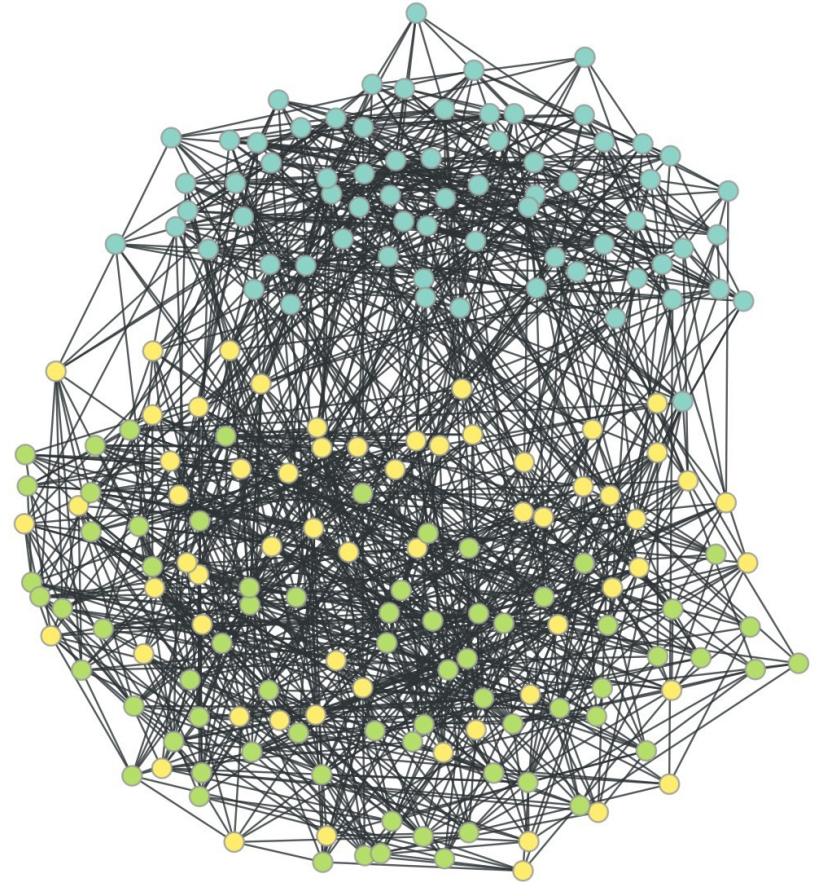
Adjacency Matrix

# Generating a network using the SBM

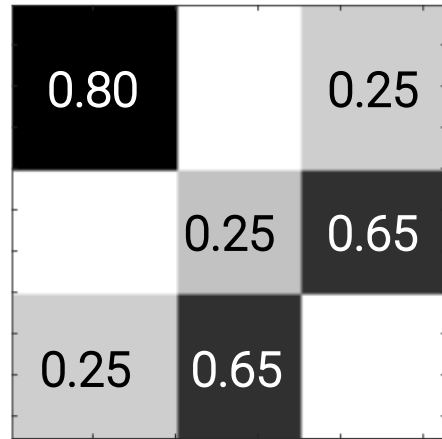


Mixing Matrix

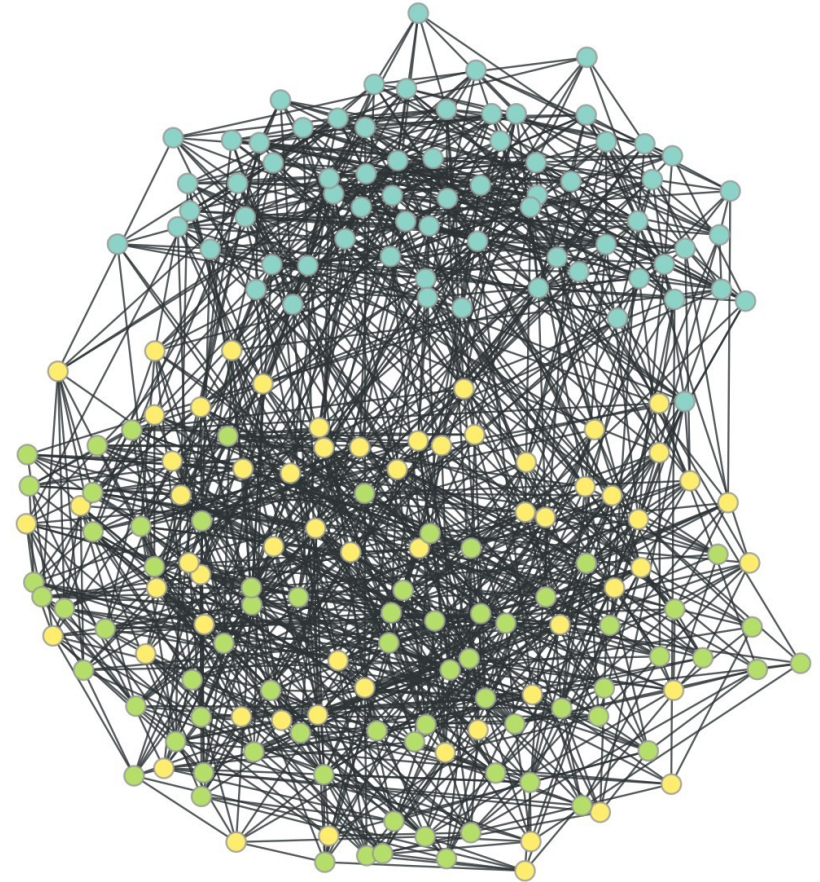
generation



# Generating a network using the SBM

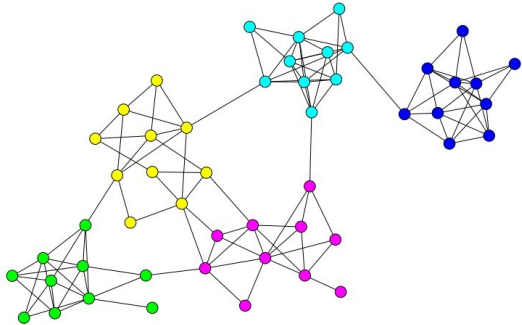
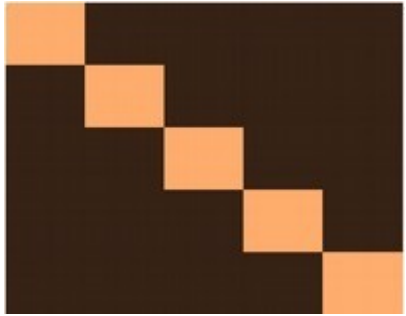


Mixing Matrix



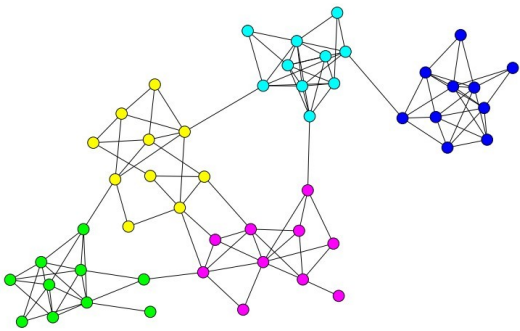
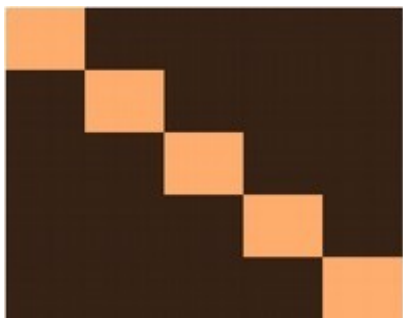
# Different types of structure

assortative

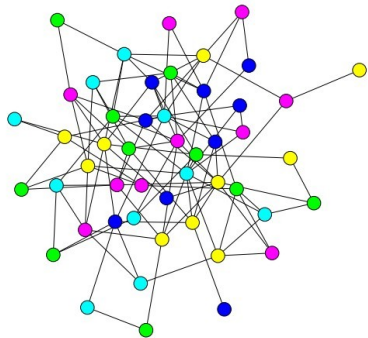
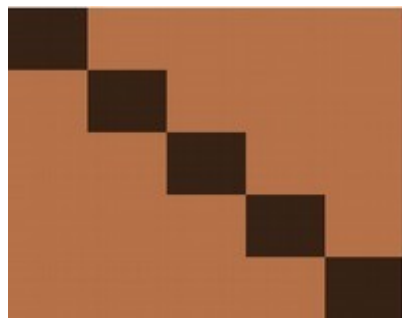


# Different types of structure

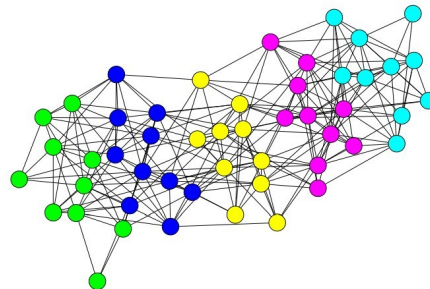
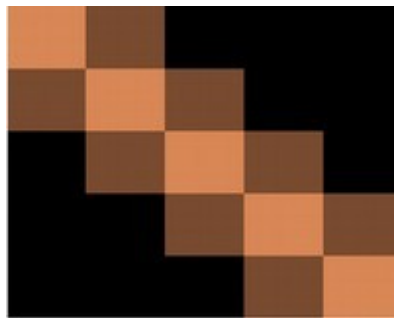
assortative



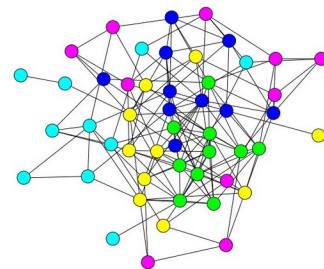
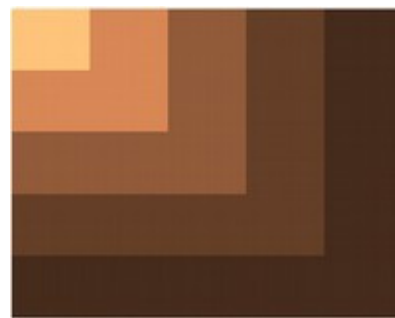
disassortative



ordered



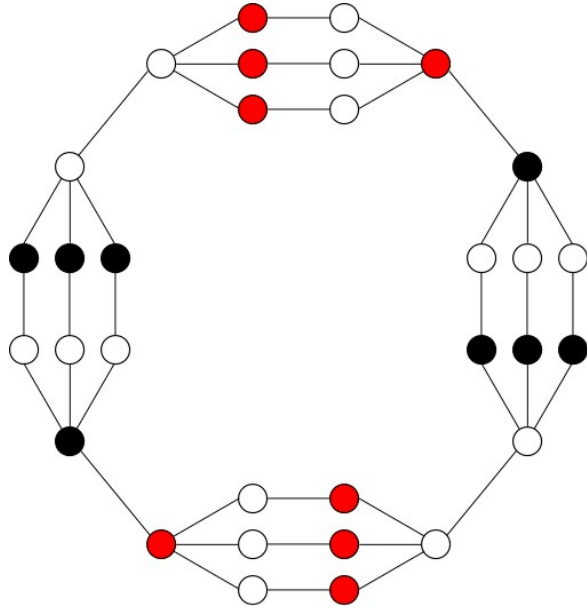
core-periphery



# Practical Part III



Network nodes can have properties or attributes (metadata)



- Metadata (M) values
- Metadata (M) unknown

social networks *age, sex, ethnicity, race, etc.*

food webs *feeding mode, species body mass, etc.*

internet *data capacity, physical location, etc.*

protein interactions *molecular weight, association with cancer, etc.*

# Blockmodel Entropy Significance Test

*How well do the metadata explain the network?*

# Blockmodel Entropy Significance Test

*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .

# Blockmodel Entropy Significance Test

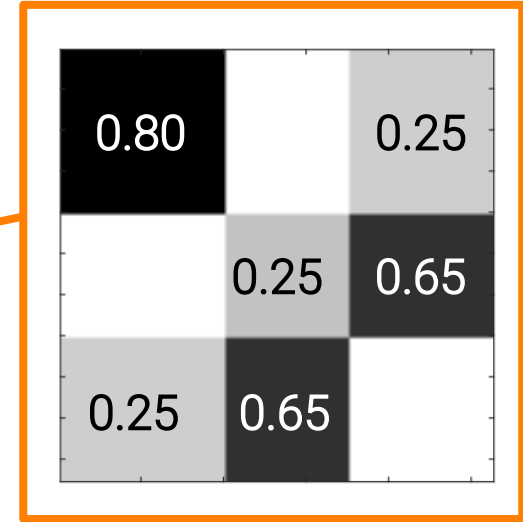
*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .
2. Fit the parameters of an SBM and compute the entropy  $H(G,M)$

# Blockmodel Entropy Significance Test

*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .
2. Fit the parameters of an SBM and compute the entropy  $H(G,M)$



# Blockmodel Entropy Significance Test

*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .
2. Fit the parameters of an SBM and compute the entropy  $H(G,M)$

Test statistic



# Blockmodel Entropy Significance Test

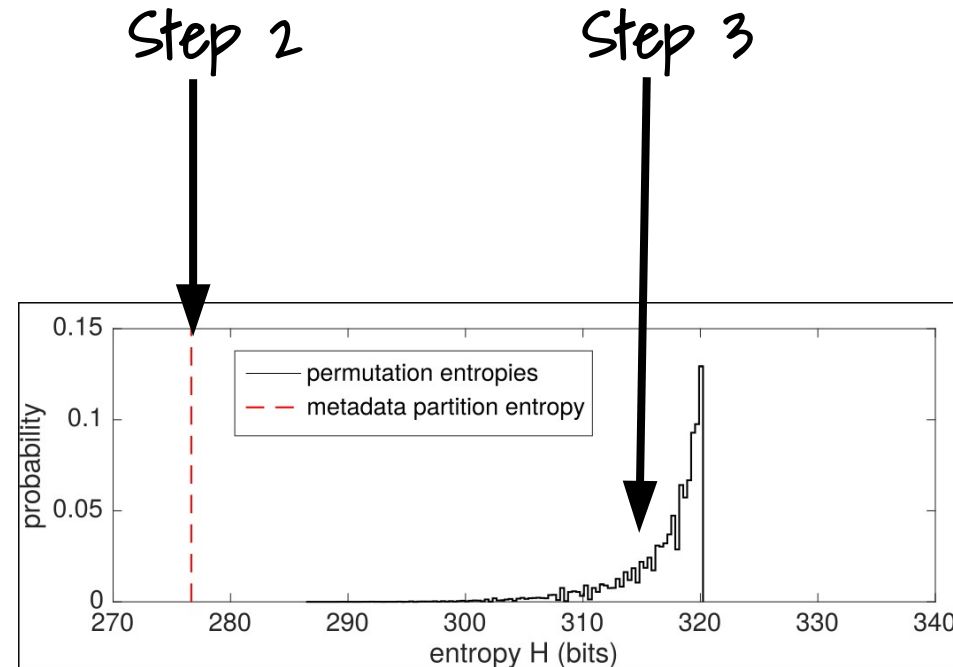
*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .
2. Fit the parameters of an SBM and compute the entropy  $H(G, M)$
3. Compare this entropy to a distribution of entropies of networks partitioned using random permutations of the metadata labels.

# Blockmodel Entropy Significance Test

*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .
2. Fit the parameters of an SBM and compute the entropy  $H(G, M)$
3. Compare this entropy to a distribution of entropies of networks partitioned using random permutations of the metadata labels.





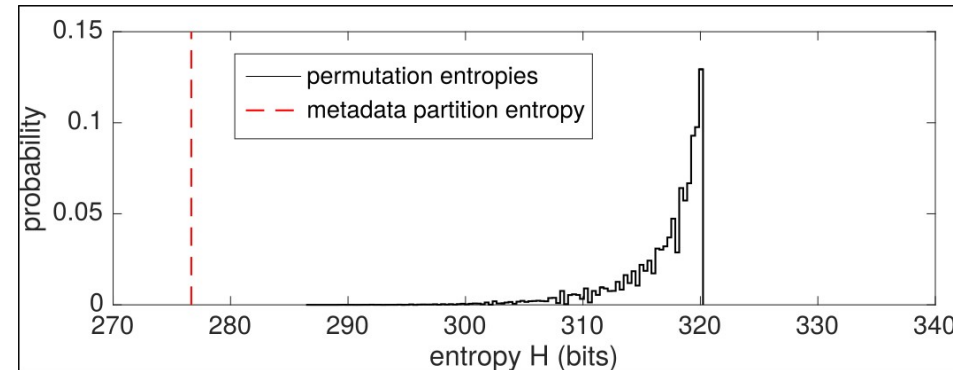
# Blockmodel Entropy Significance Test

*How well do the metadata explain the network?*

1. Divide the network  $G$  into groups according to metadata labels  $M$ .
2. Fit the parameters of an SBM and compute the entropy  $H(G,M)$
3. Compare this entropy to a distribution of entropies of networks partitioned using random permutations of the metadata labels.

metadata is randomly assigned  
→ model gives no explanation, high  $H$

metadata correlates with structure  
→ model gives good explanation, low  $H$



# Multiple networks; multiple metadata attributes

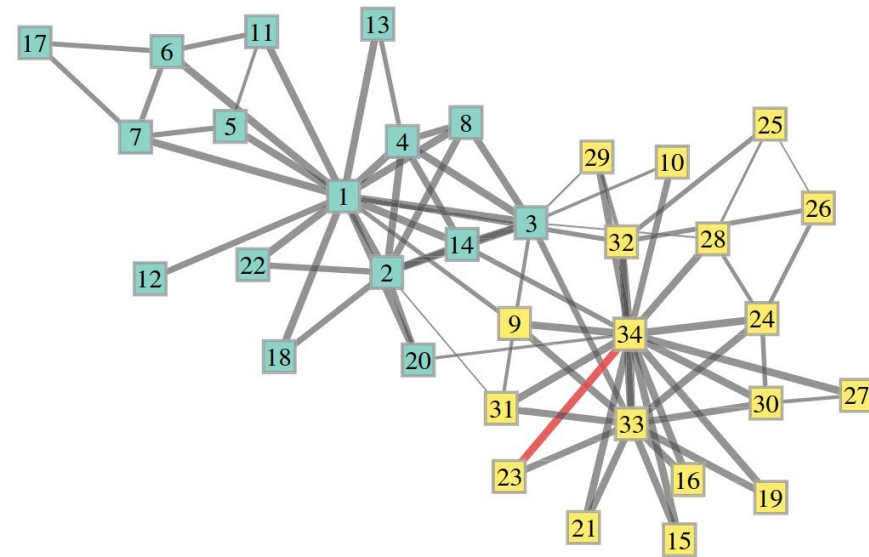
Network	Status	Gender	Office	Practice	Law School
Friendship	$< 10^{-6}$	0.034	$< 10^{-6}$	0.033	0.134
Cowork	$< 10^{-3}$	0.094	$< 10^{-6}$	$< 10^{-6}$	0.922
Advice	$< 10^{-6}$	0.010	$< 10^{-6}$	$< 10^{-6}$	0.205

Multiple sets of metadata provide a significant explanation for multiple networks.

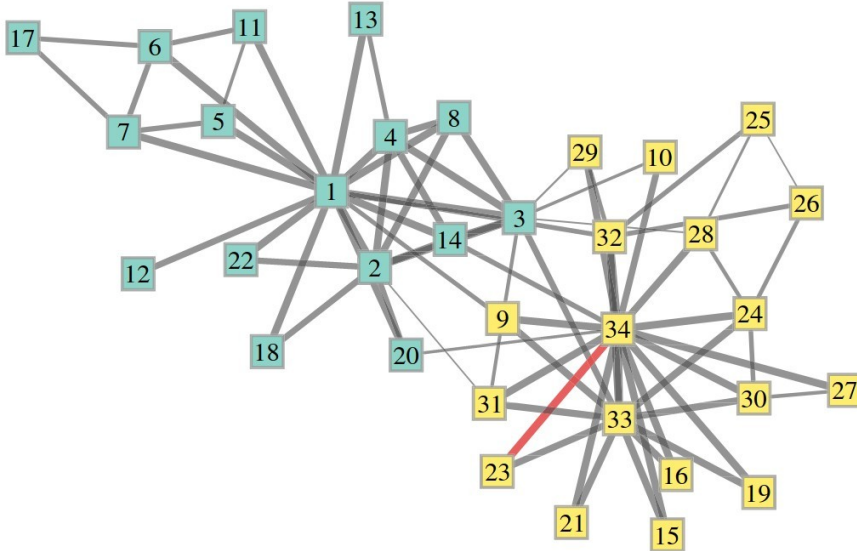
# Practical Part III

Errors in networks and reconstruction

# Zachary's Karate Club

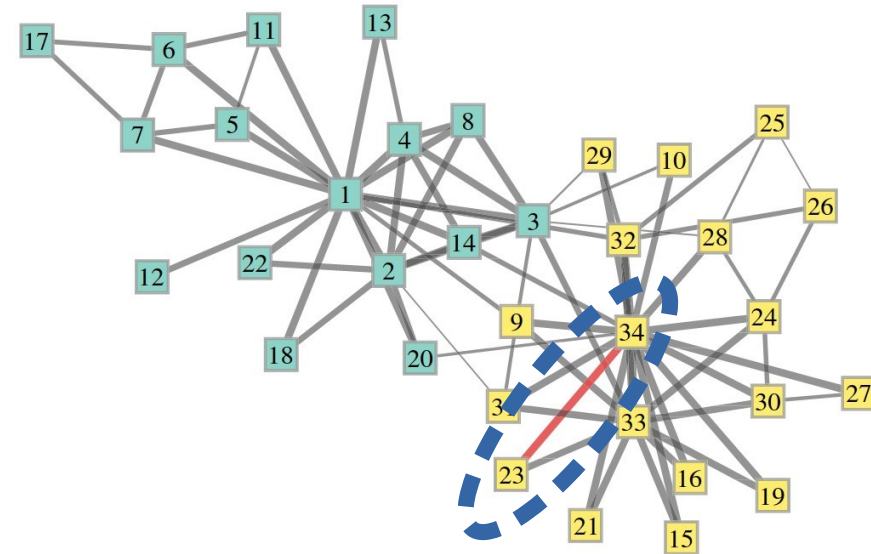


# Karate Club

[illegible]

# Zachary's Karate Club

Individual Number																																				
	1	2	3	4	5	6	7	8	9	0	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3		
1	0	4	5	3	3	3	3	2	2	0	2	3	2	3	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	
2	4	0	6	3	0	0	0	4	0	0	0	0	0	5	0	0	0	1	0	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0		
3	5	6	0	3	0	0	0	4	5	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	3	0		
4	3	3	3	0	0	0	0	3	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
5	3	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
6	3	0	0	0	0	0	5	0	0	0	3	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
7	3	0	0	0	2	5	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
8	2	4	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
9	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	3		
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2		
11	2	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
12	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
13	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
14	3	5	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3		
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2		
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4		
17	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
18	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2		
20	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1		
22	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	4	0	2	0	0	0		
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	0	2	0		
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	0	0	0	0	0	0	7	0		
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	2		
28	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	0	4		
29	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2		
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	4	0	0	0	0	3	2		
31	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3		
32	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	7	0	0	2	0	0	4		
33	0	0	2	0	0	0	0	0	3	0	0	0	0	0	3	3	0	0	1	0	3	0	5	0	0	0	0	0	4	3	4	0	5	0	0	
34	0	0	0	0	0	0	0	4	2	0	0	0	3	2	4	0	0	2	1	1	0	3	0	0	2	4	2	2	3	4	5	0	0	0	0	



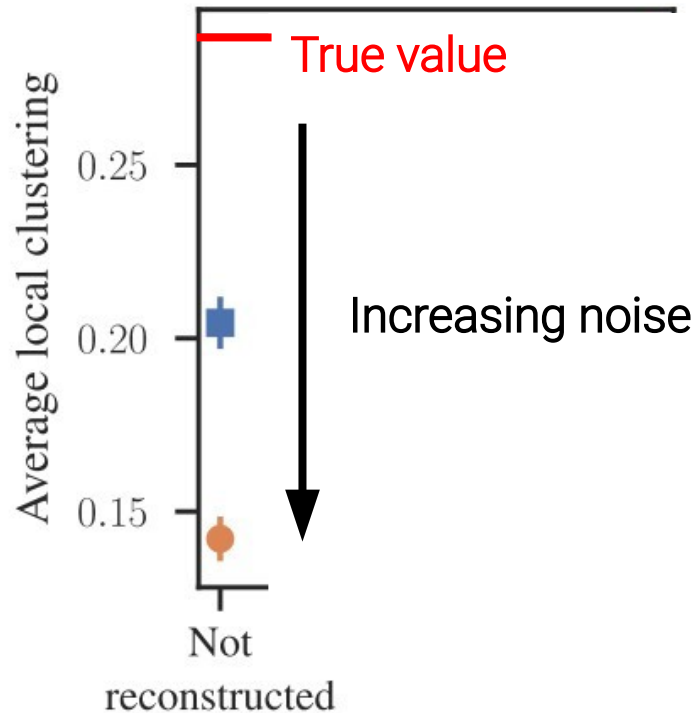
Does this edge exist?

Errors in network data create systematic biases...

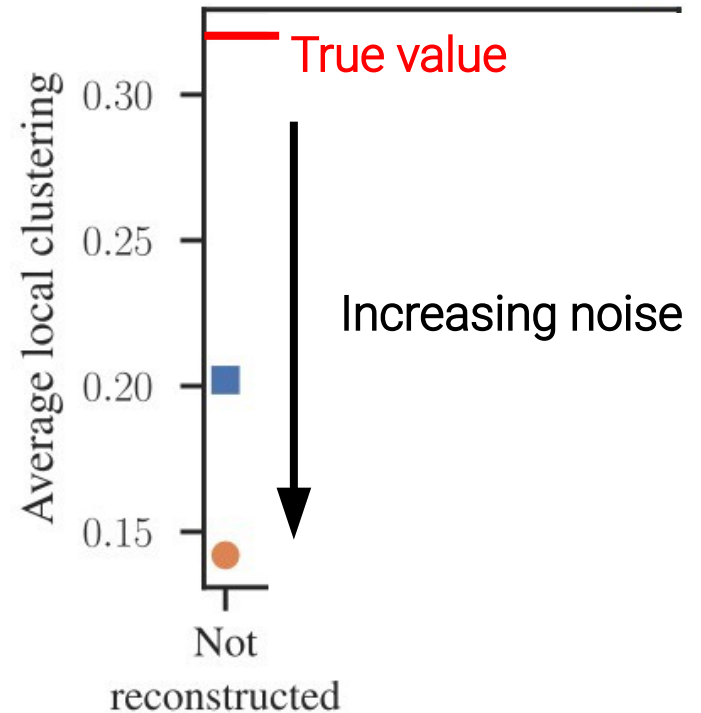


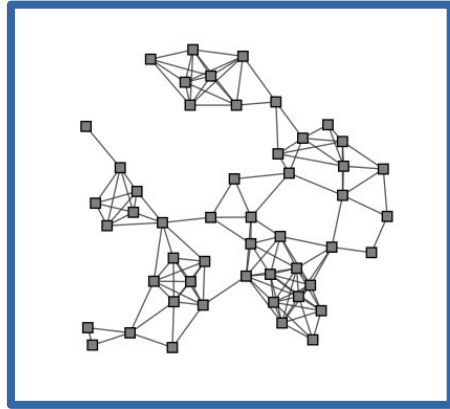
# Errors in network data create systematic biases...

(a) High-school friendships

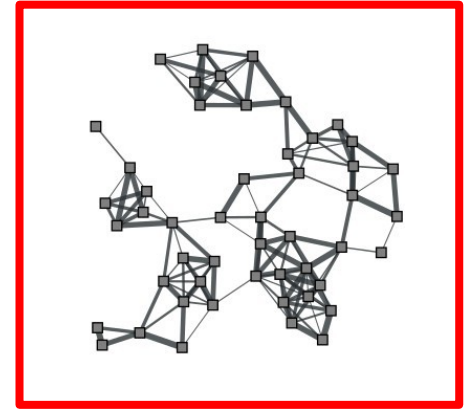


(b) Political blogs





True Network



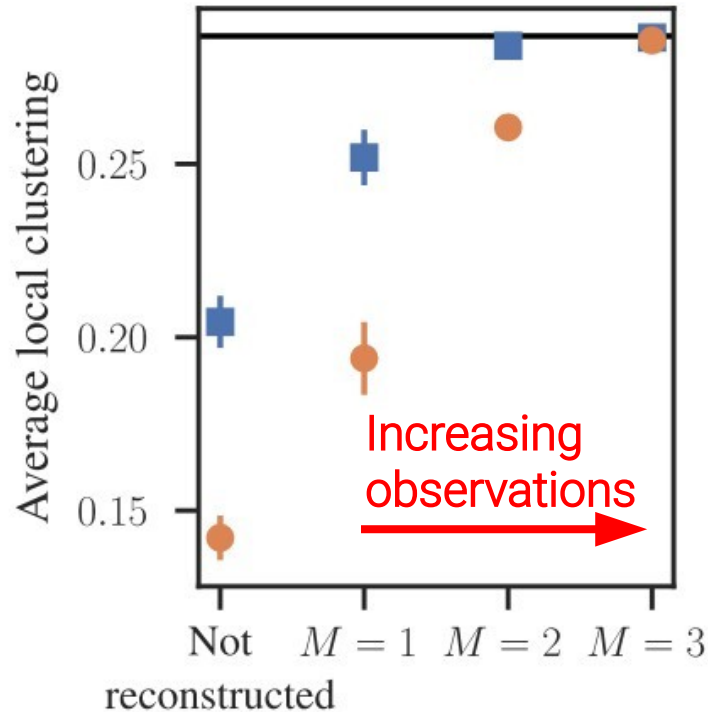
Reconstructed Network

We don't know if the network represents the system

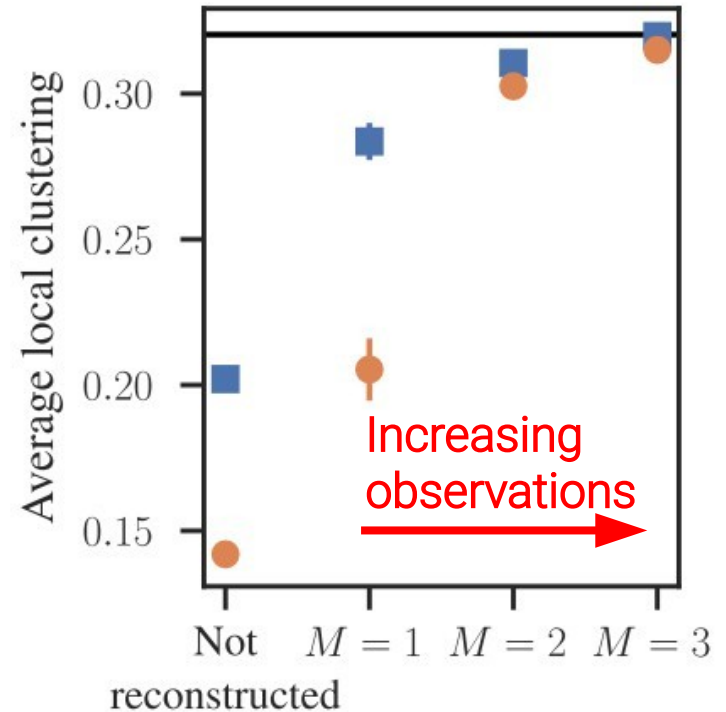
$$P(\mathbf{A}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{D})} .$$

## Bayesian inference

(a) High-school friendships



(b) Political blogs





# Practical Part IV

# Summary...

We can use graph models to simulate networks and better understand the effects of network structure

We can use graph models to:

- test hypotheses
- reconstruct uncertain networks