Graph models and hypothesis testing

Leto Peel
l.peel@maastrichtuniversity.nl
@PiratePeel

Why Graph models?

Why Graph models?

Graph models allow us to generate synthetic networks

Why Graph models?

Graph models allow us to generate synthetic networks

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

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

Why Graph models?

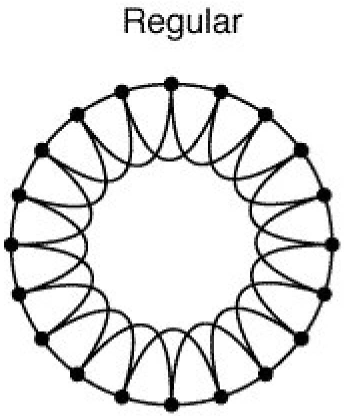
Graph models allow us to generate synthetic networks

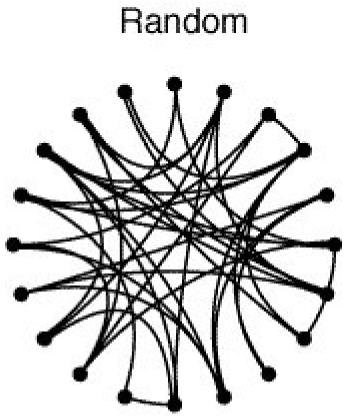
They allow us to capture and model properties observed in real networks

Graph models can serve as hypotheses for mechanisms of network formation

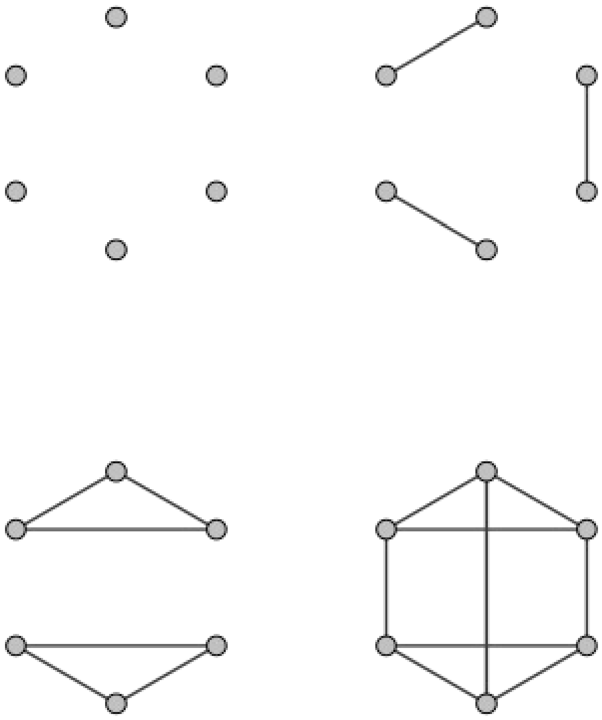
Allows us to explore how “similar” networks behave

Regular and random graphs

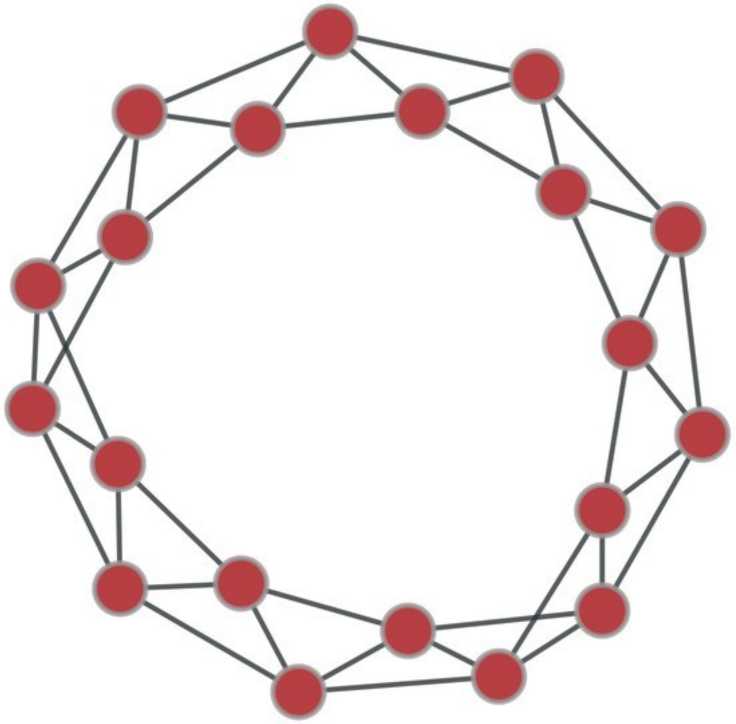




Regular graphs

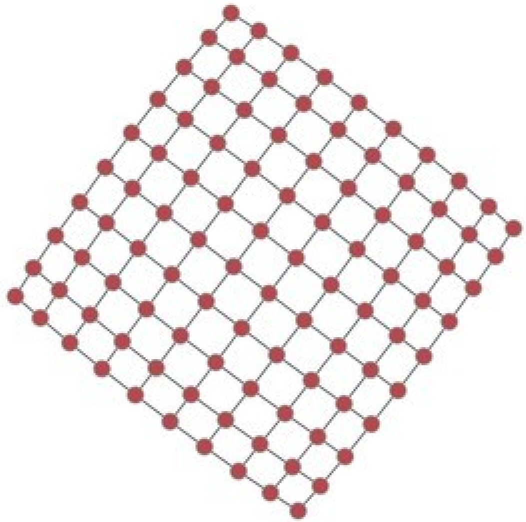


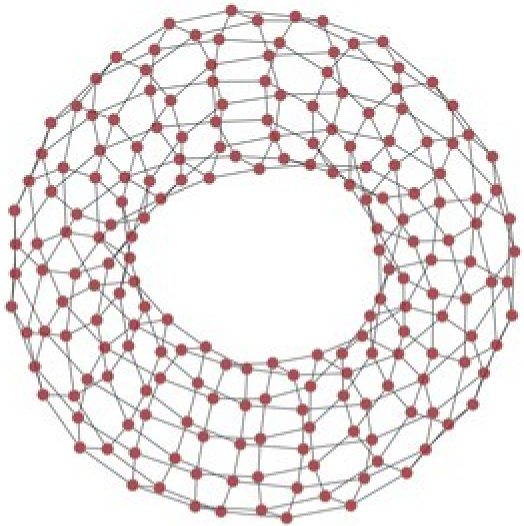
Regular graphs with 0 – 3 degree nodes

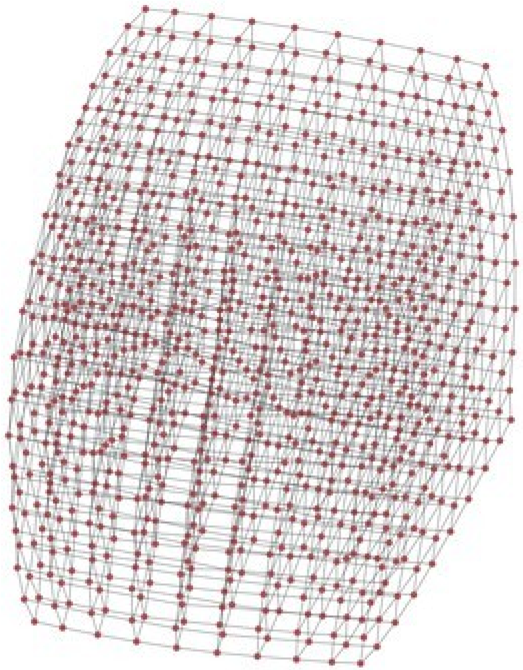


Regular Ring Lattice

Lattice graphs





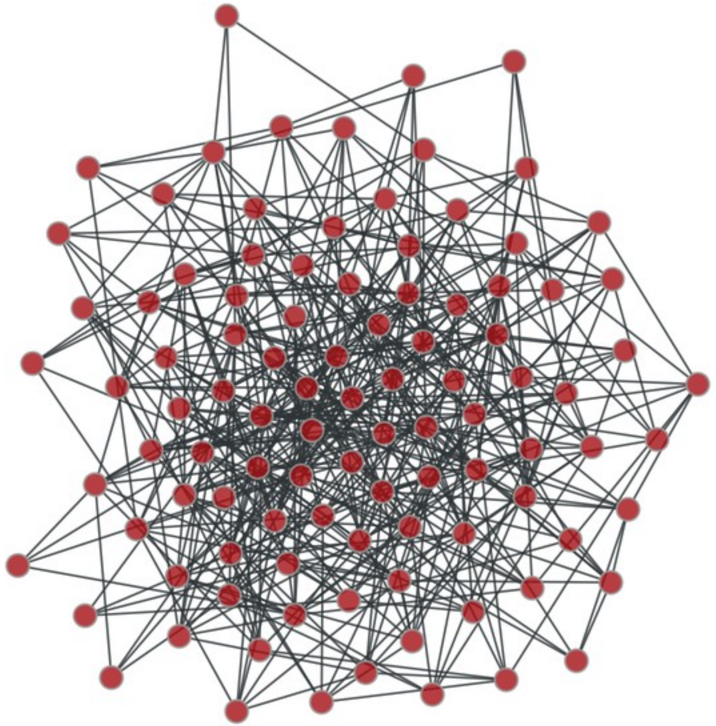


Random graphs

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

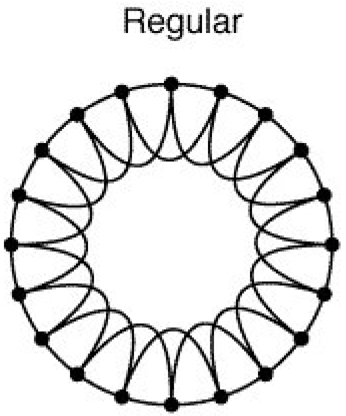
Small-world networks

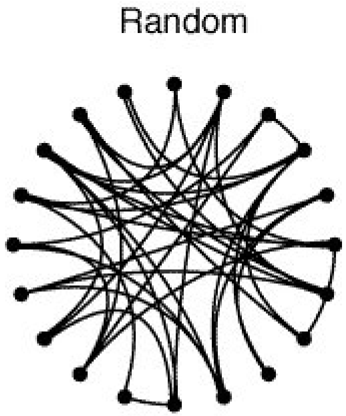
It’s a network after all

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393(6684), 440-442.

Small-world networks

It’s a network after all



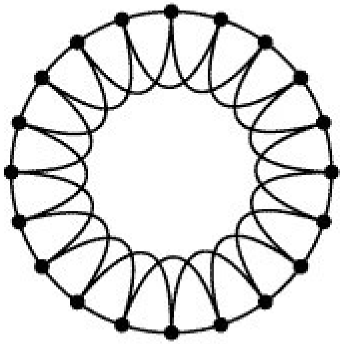


Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393(6684), 440-442.

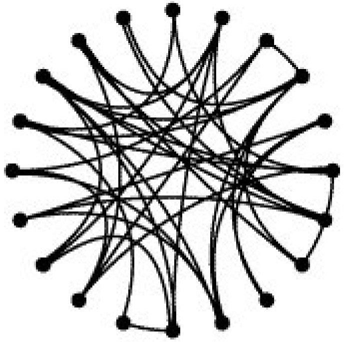
Small-world networks

It’s a network after all

Regular



Random



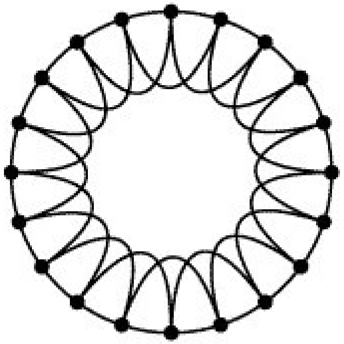
(triangles) High **clustering coefficient** Low **clustering coefficient**

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393(6684), 440-442.

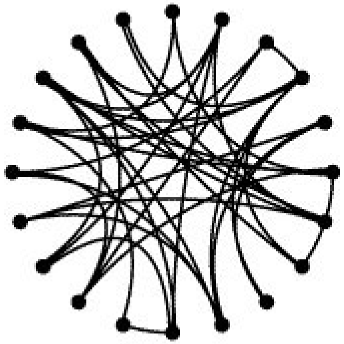
Small-world networks

It’s a network after all

Regular



Random



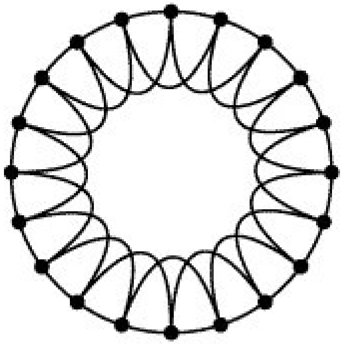
(triangles)	High **clustering coefficient**	Low **clustering coefficient**
(shortest paths)	High mean **geodesic path**	Low mean **geodesic path**

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393(6684), 440-442.

Small-world networks

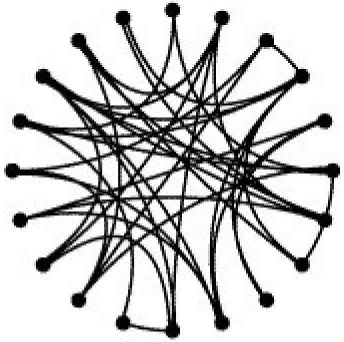
It's a network after all

Regular



Real-world networks

Random

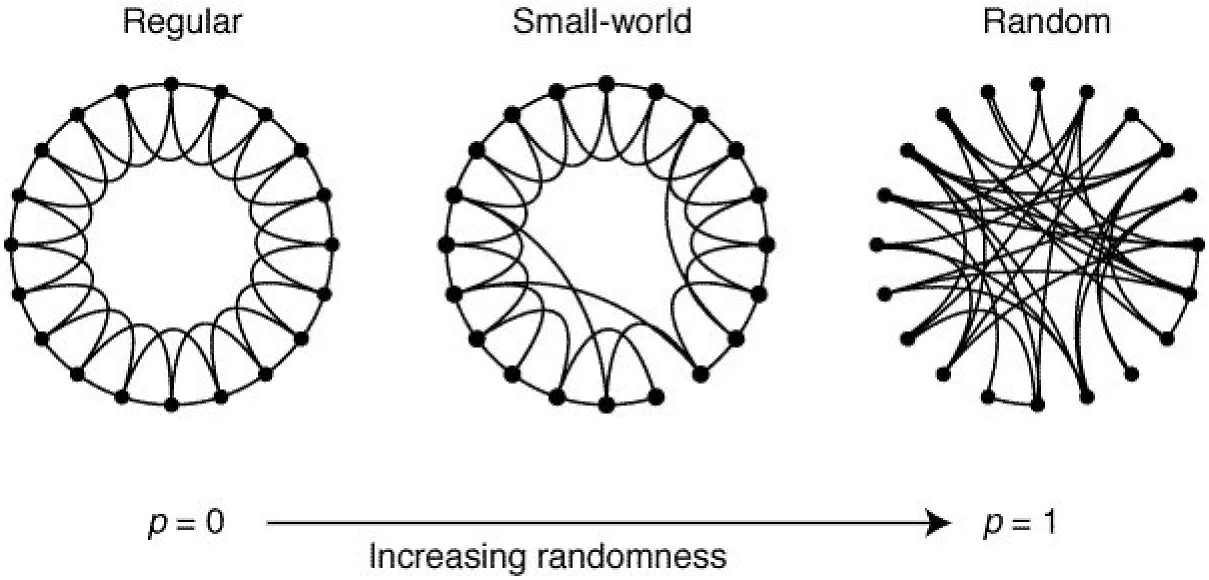


	Regular	Random
(triangles)	High **clustering coefficient**	Low **clustering coefficient**
(shortest paths)	High mean **geodesic path**	Low mean **geodesic path**

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440-442.

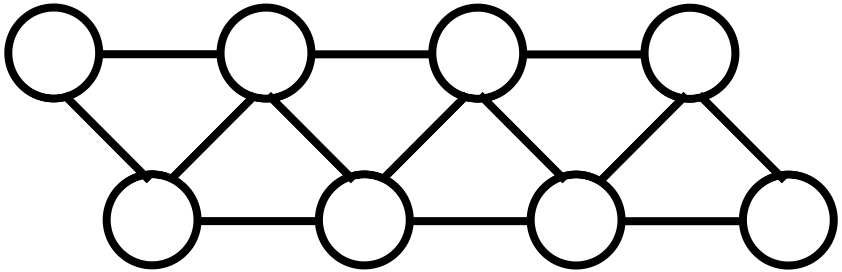
Small-world networks

It’s a network after all



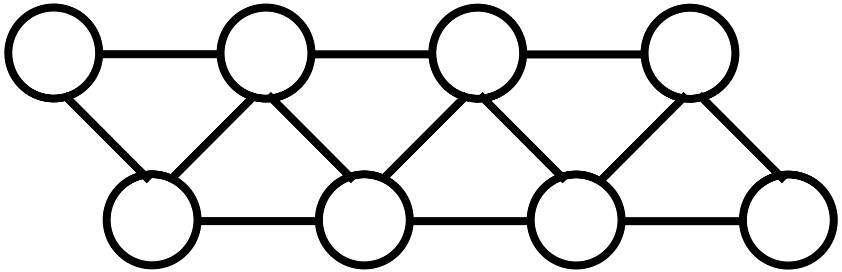
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393(6684), 440-442.

Exercise



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

Exercise



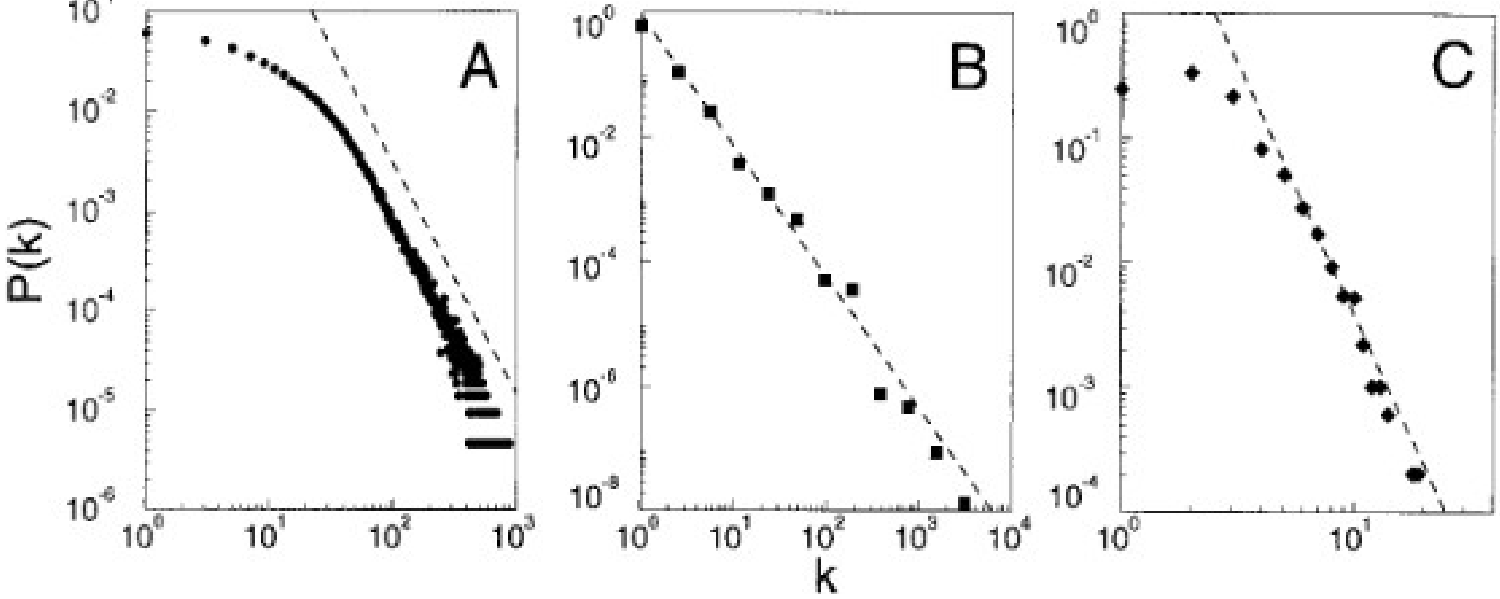
Calculate the **clustering coefficient** and mean **shortest path**

Now choose a pair of edges to randomly rewire and recalculate

"Scale-free" networks

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.

"Scale-free" networks



Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.

The Price model

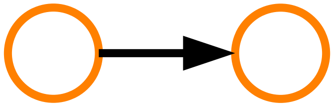
For undirected networks, this model is known as the Barabasi-Albert model

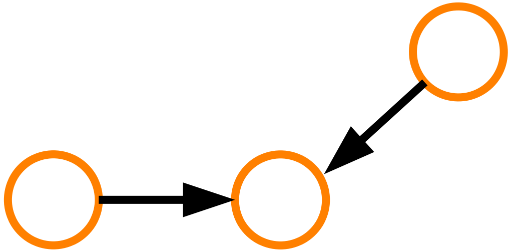
Add nodes to a network one at a time.

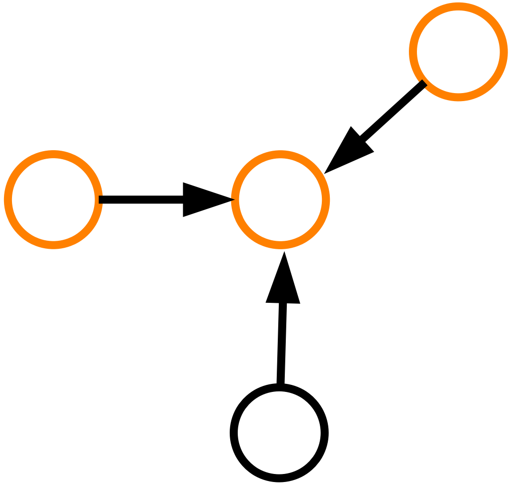
Connect to existing nodes with probability **proportional to their degree**

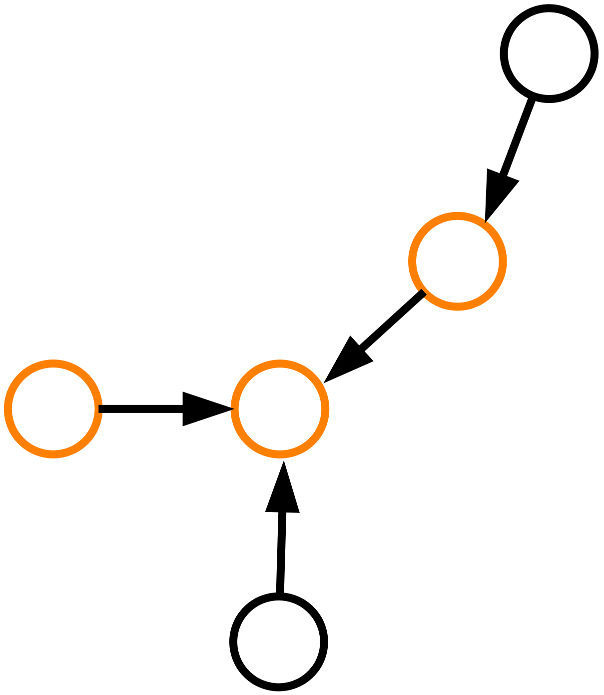
$$p_i = \frac{k_i}{\sum_j k_j},$$

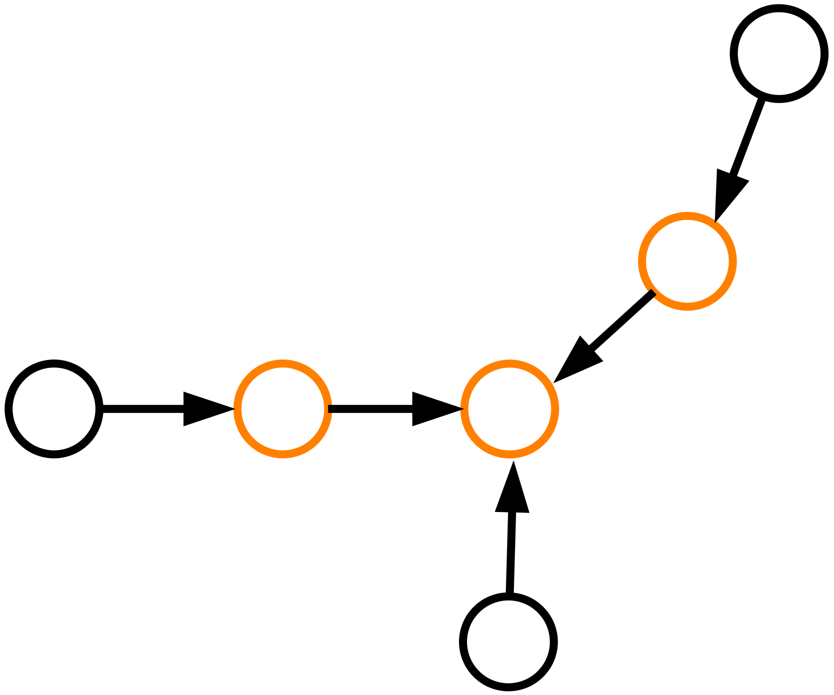
Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American society for Information science, 27(5), 292-306.

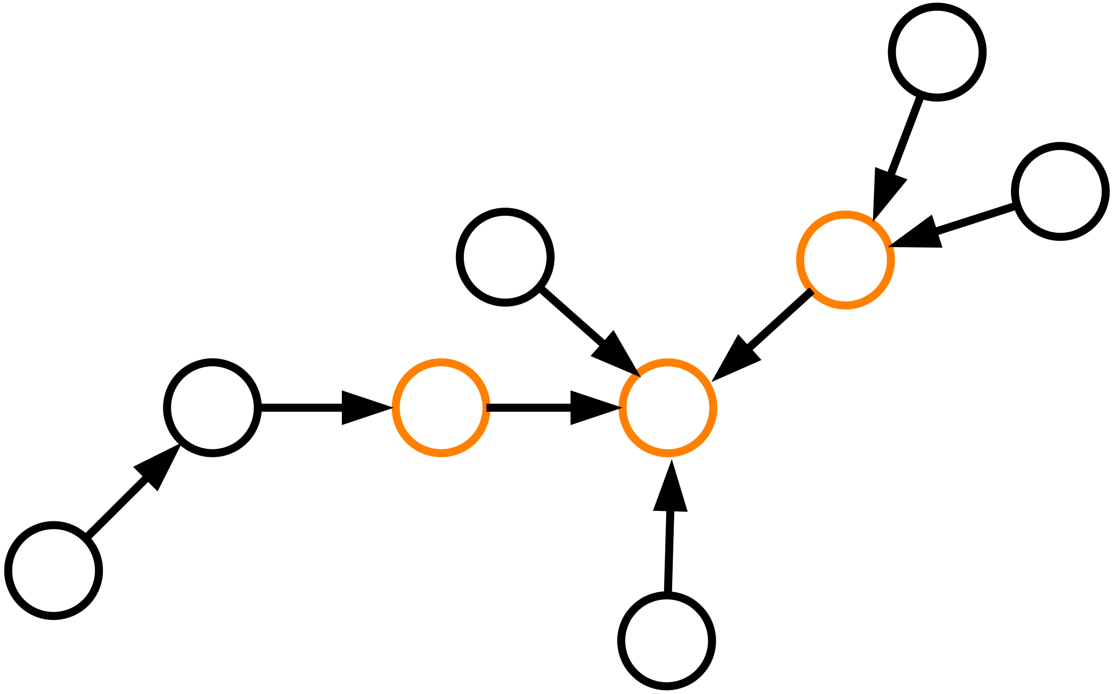


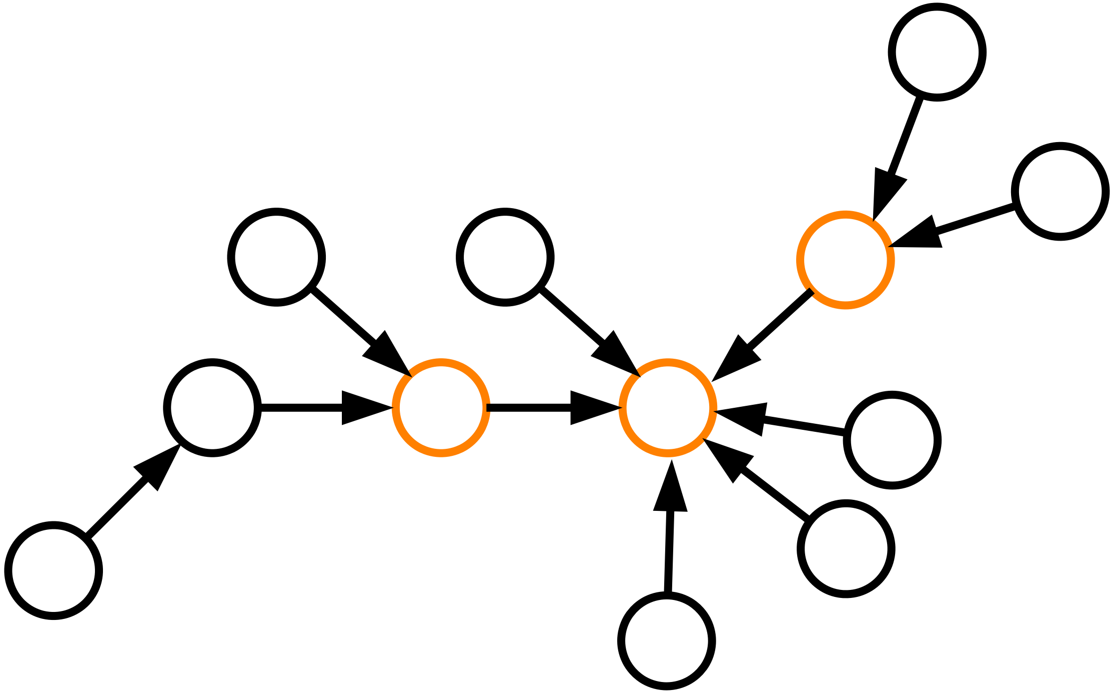


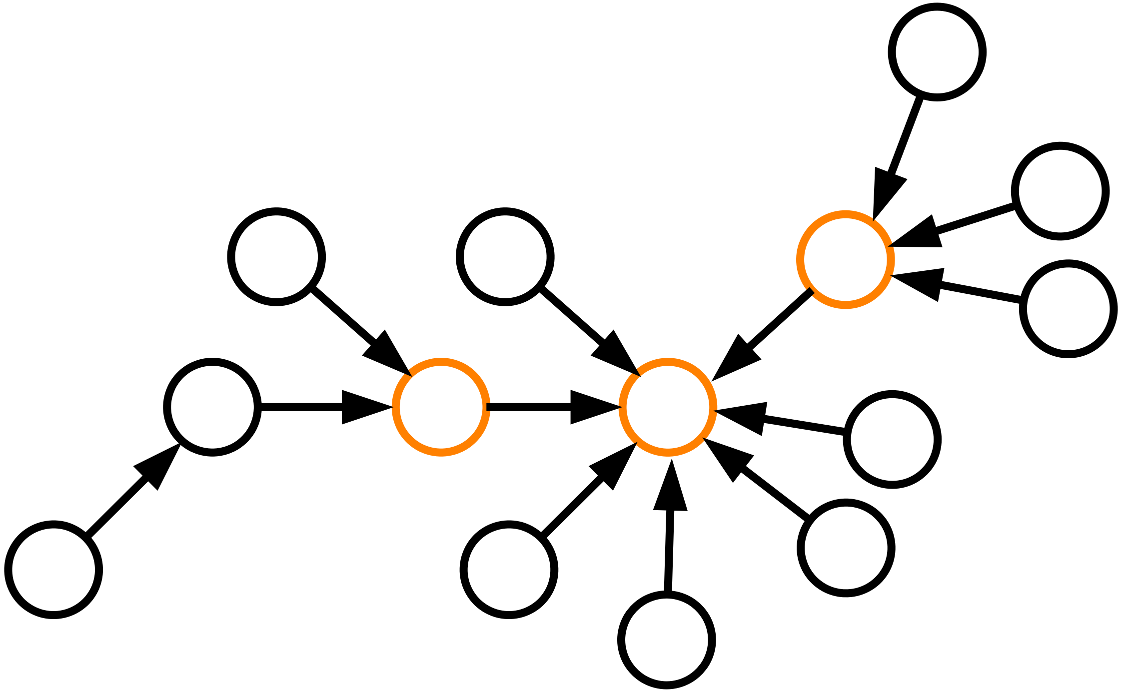




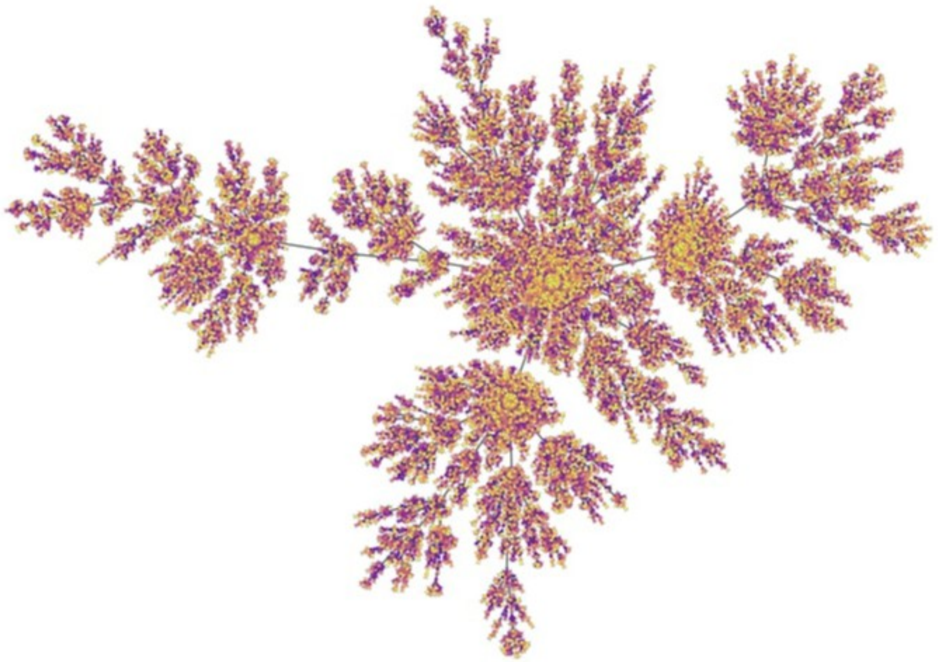


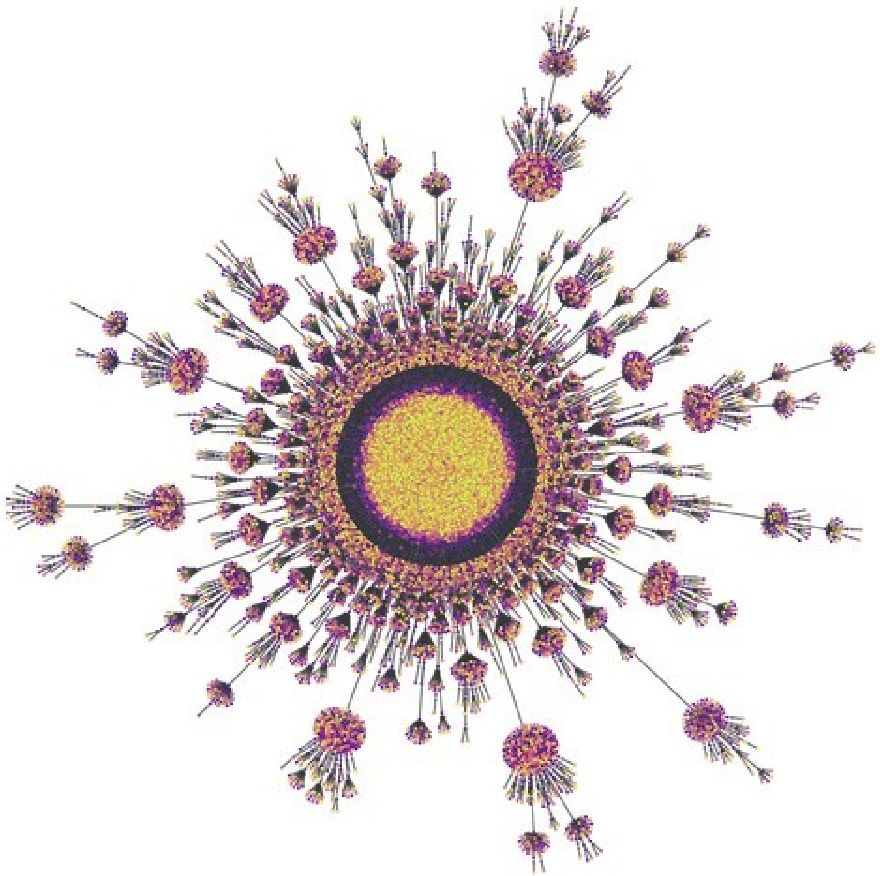






Nodes that join the network earlier have higher degree





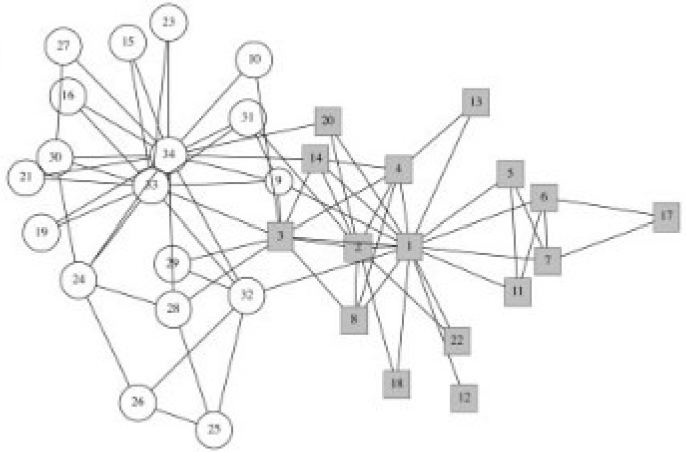
Community structure

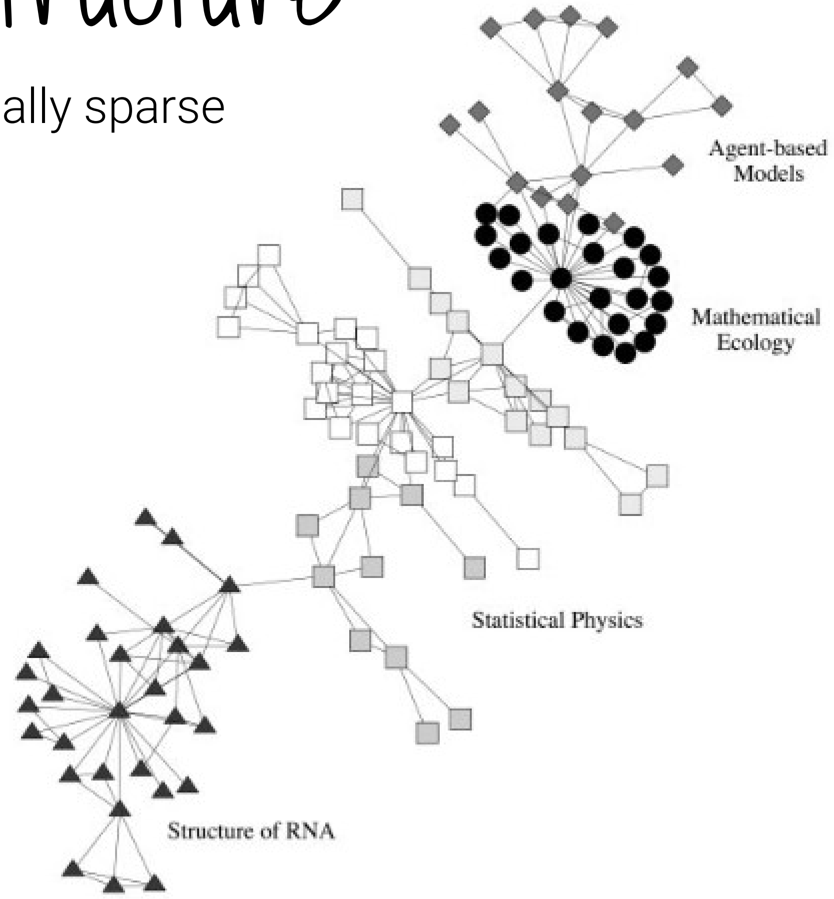
Locally dense, globally sparse

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. PNAS, 99(12), 7821-7826.

Community structure

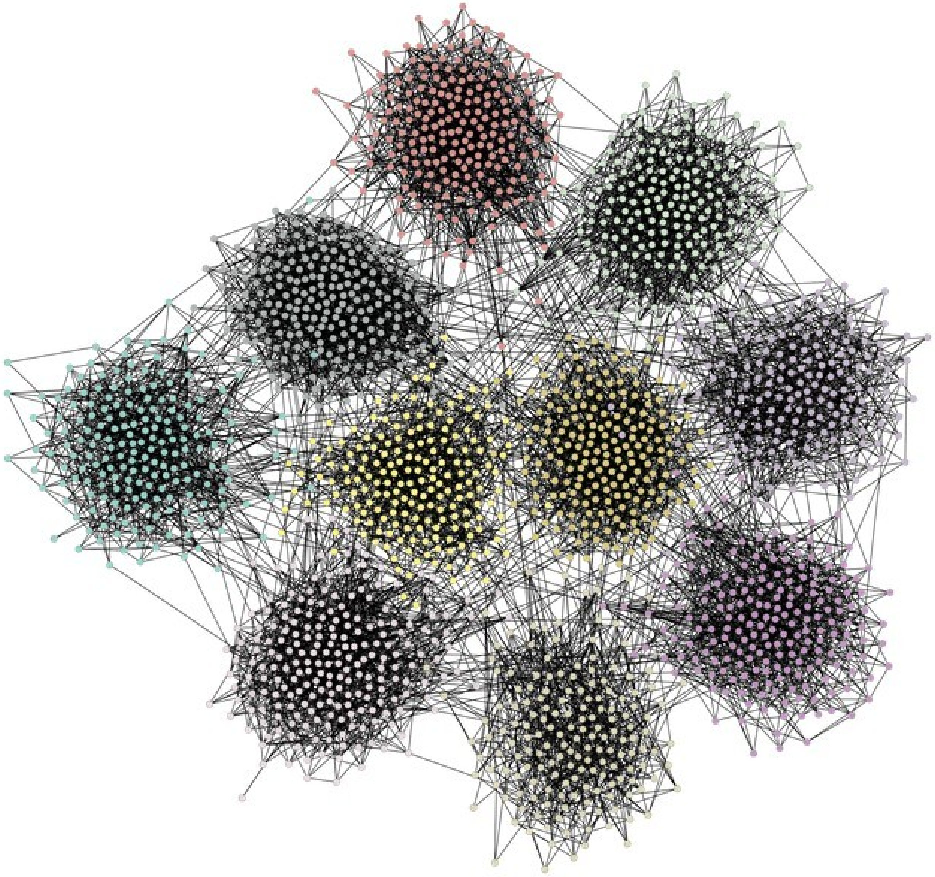
Locally dense, globally sparse





Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. PNAS, 99(12), 7821-7826.

Stochastic Block Models



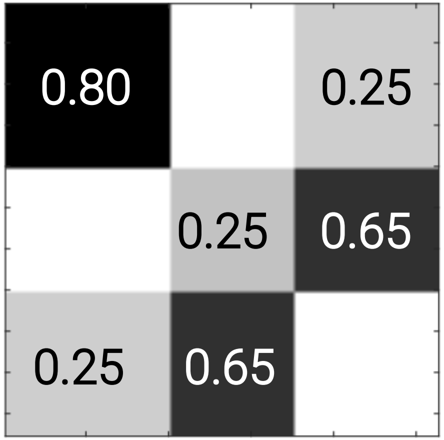
Generating a network using the SBM

Step 1 : Assign each node to a group

Generating a network using the SBM

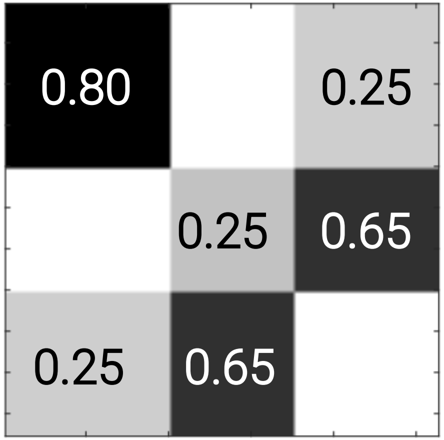
Step 1 : Assign each node to a group

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



Mixing Matrix

Generating a network using the SBM

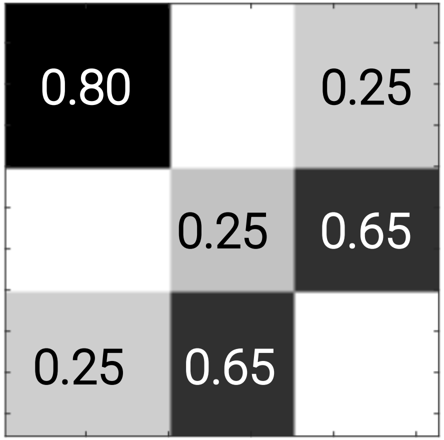


Mixing Matrix

Step 1 : Assign each node to a group

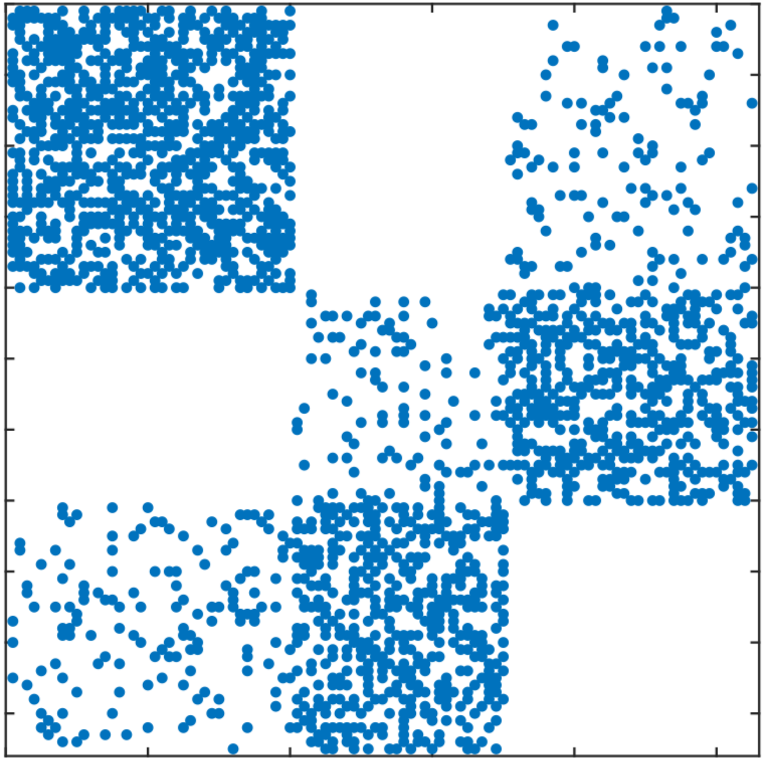
- **Step 2** : Select some connection probabilities (mixing matrix)
- **Step 3** : For each pair of nodes, add an edge with probability according to the group memberships

Generating a network using the SBM



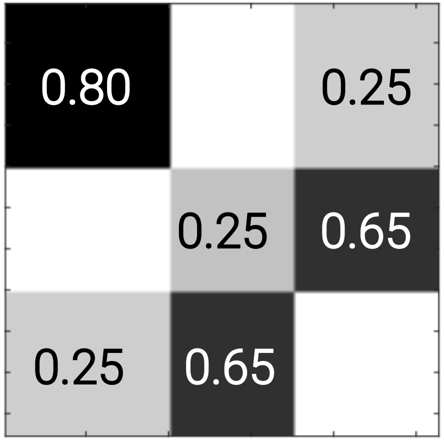
Mixing Matrix

generation



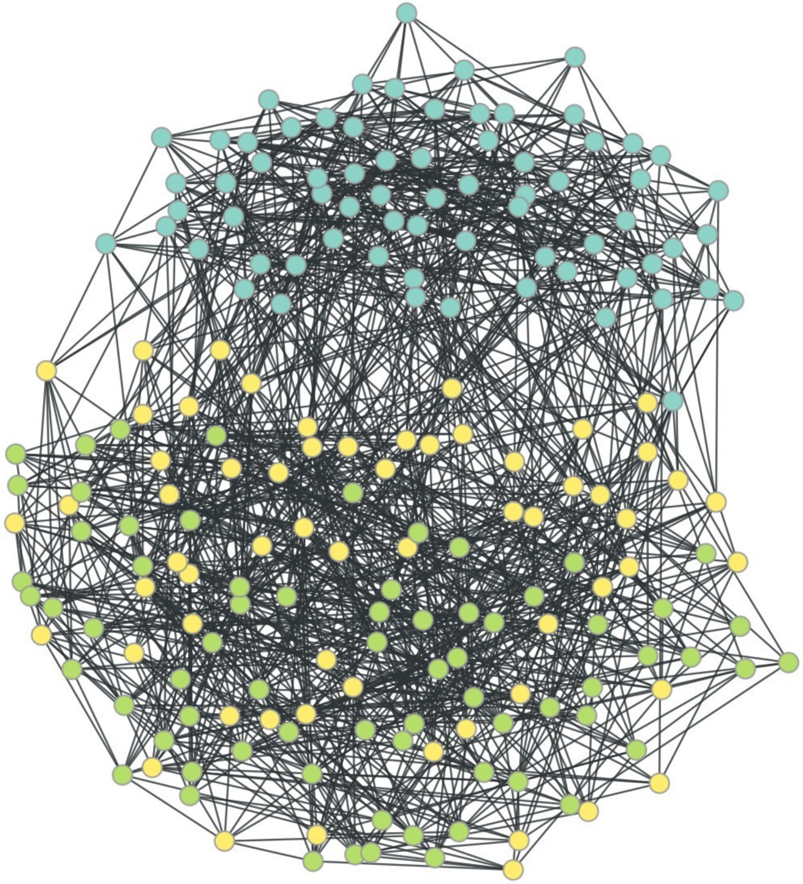
Adjacency Matrix

Generating a network using the SBM

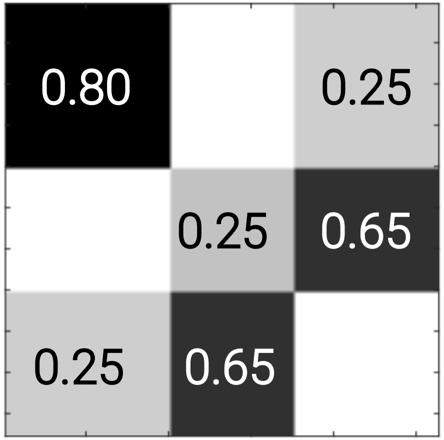


Mixing Matrix

generation



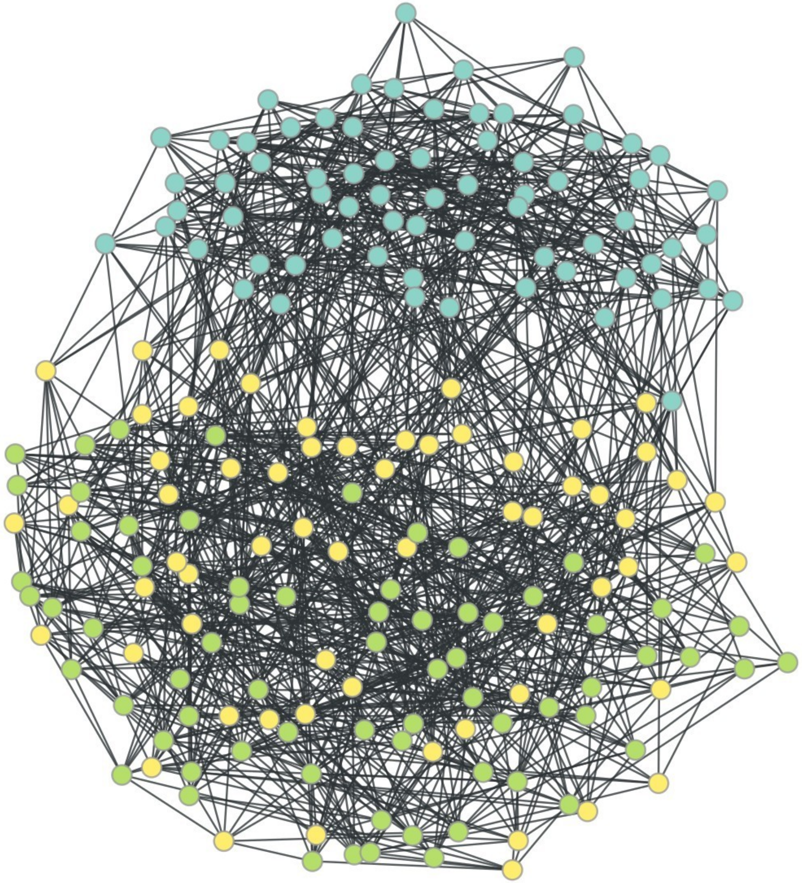
Generating a network using the SBM



Mixing Matrix

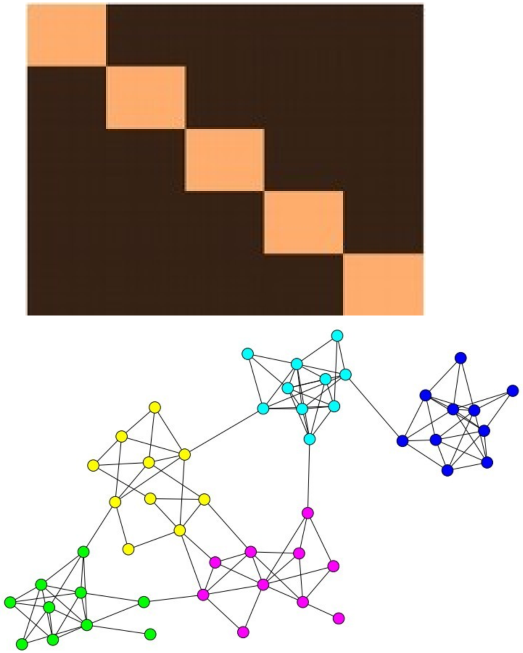
generation

inference
(later)



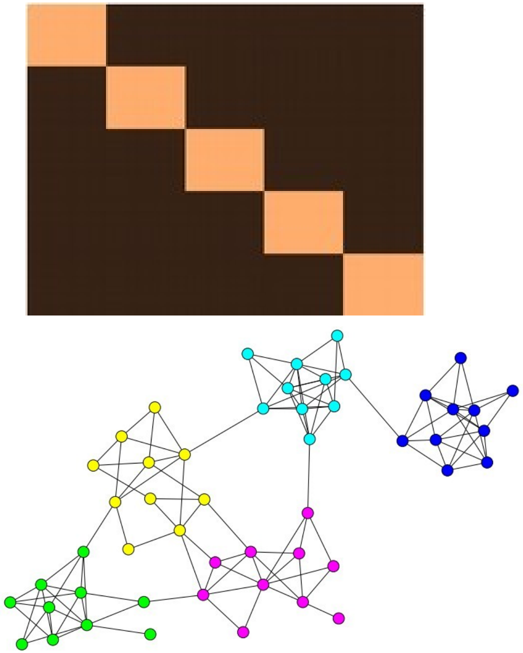
Different types of structure

assortative

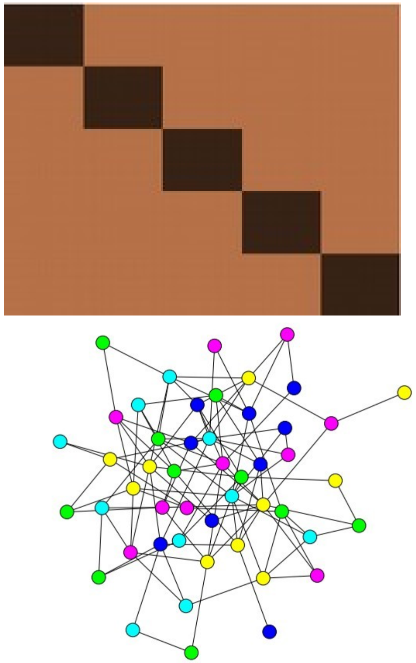


Different types of structure

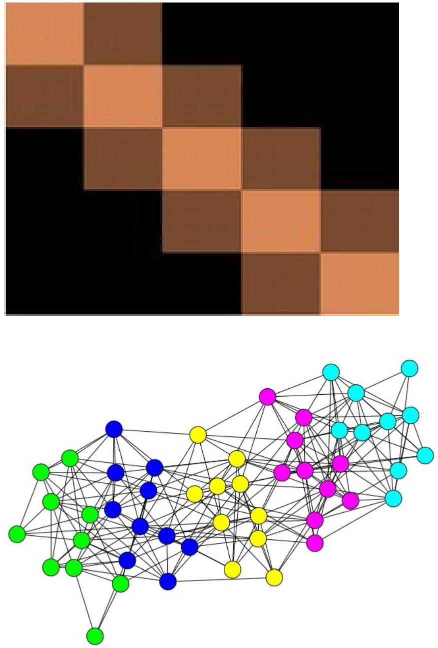
assortative



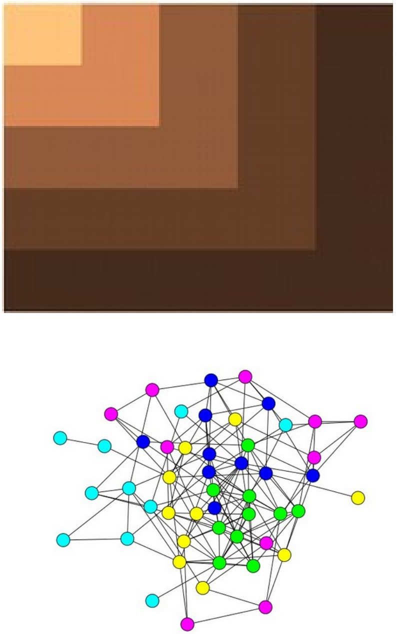
disassortative



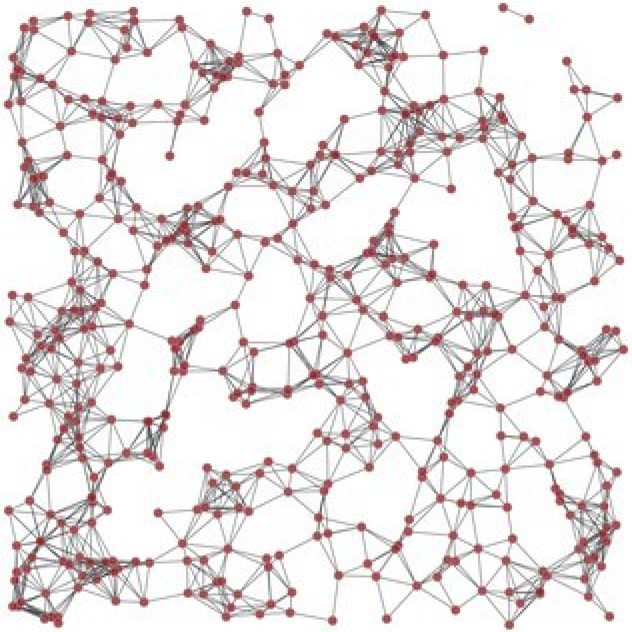
ordered

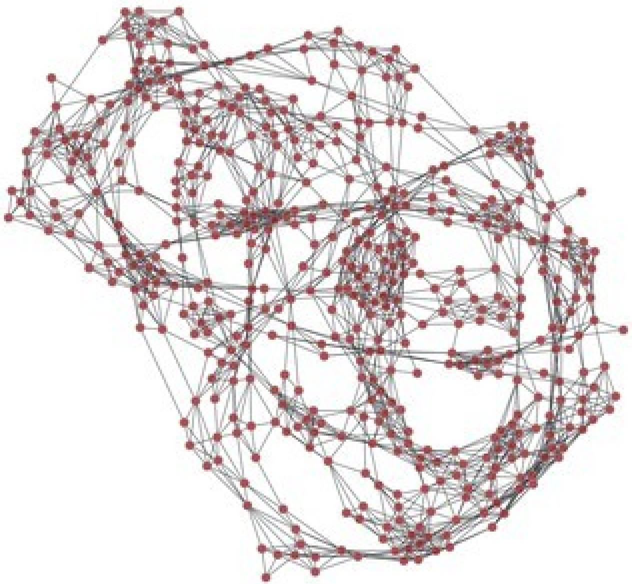


core-periphery



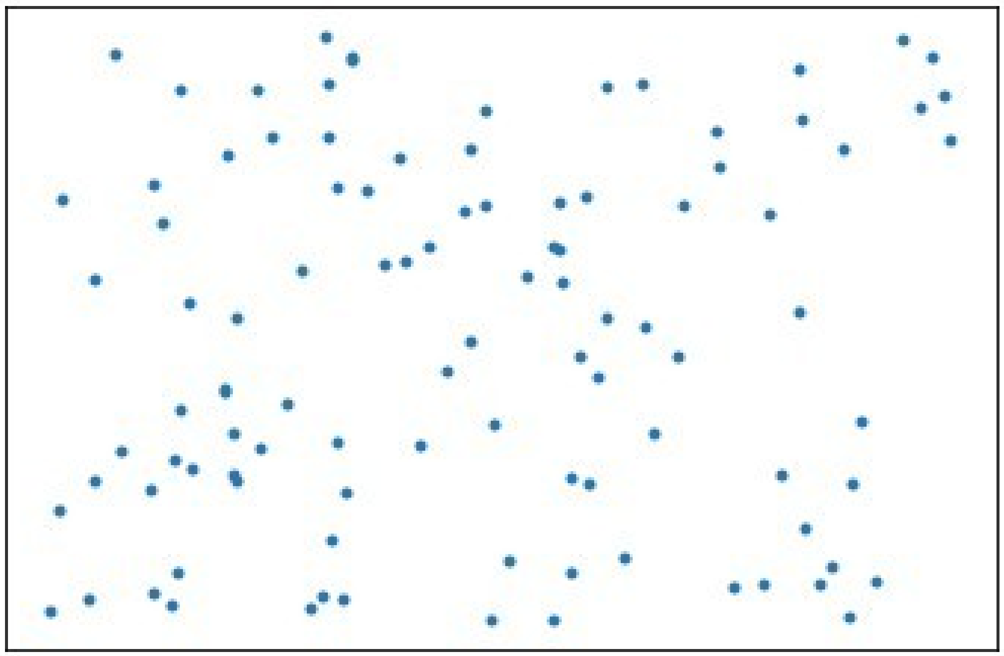
Geometric graphs





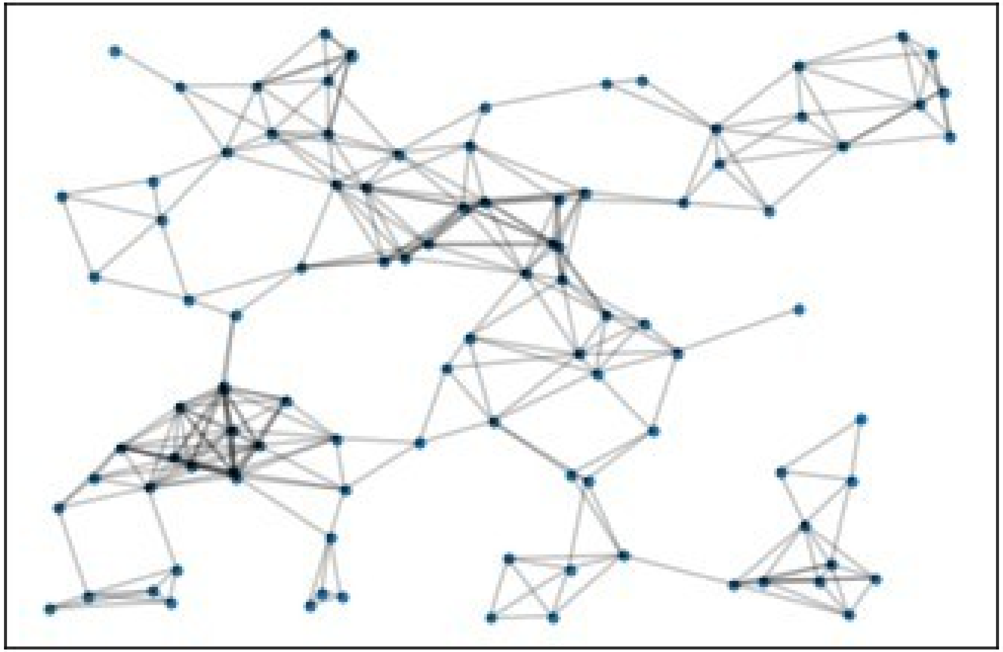
Generating a random geometric graph

Generating a random geometric graph



Step 1 : Assign each node to a random position in a 2D space

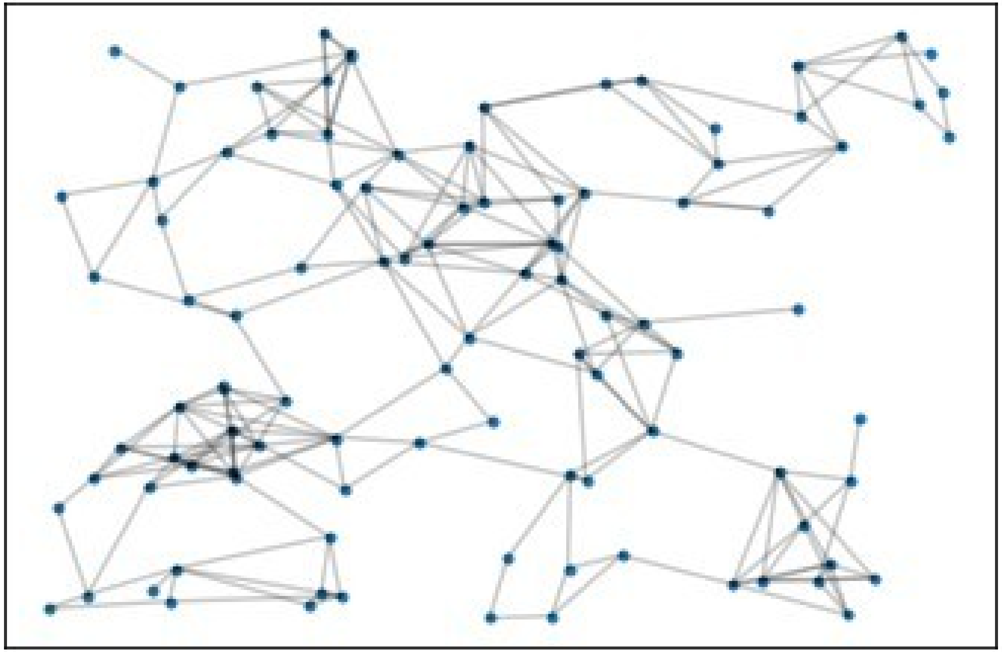
Generating a random geometric graph



Step 1 : Assign each node to a random position in a 2D space

Step 2 : Connect nodes if they are within a given radius, *r*

Latent space models



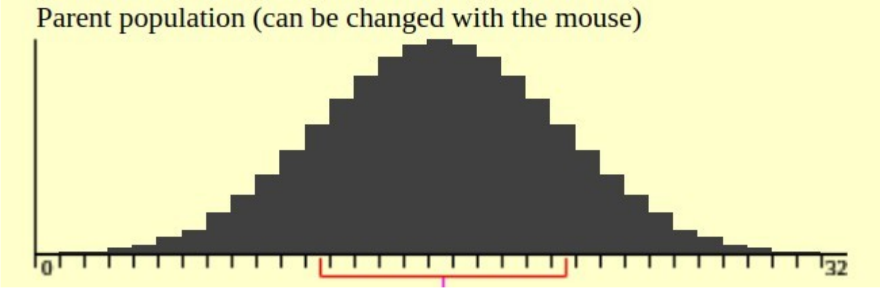
Similar to random geometric graphs...

except edges are assigned according to a probability as a function of the distance

Practical Part 1

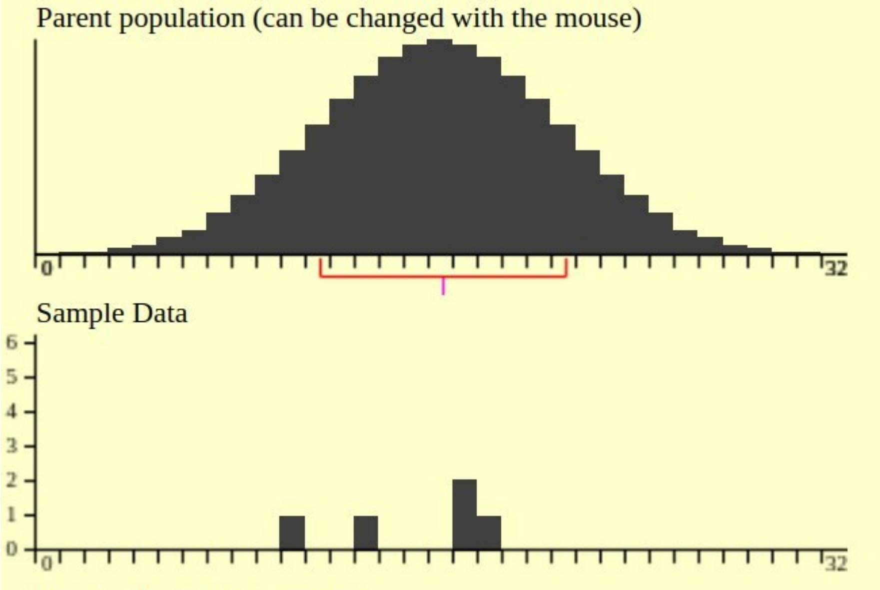
Null hypothesis testing

Population distribution



Mean = 16.00
Sd = 5.00

Population distribution



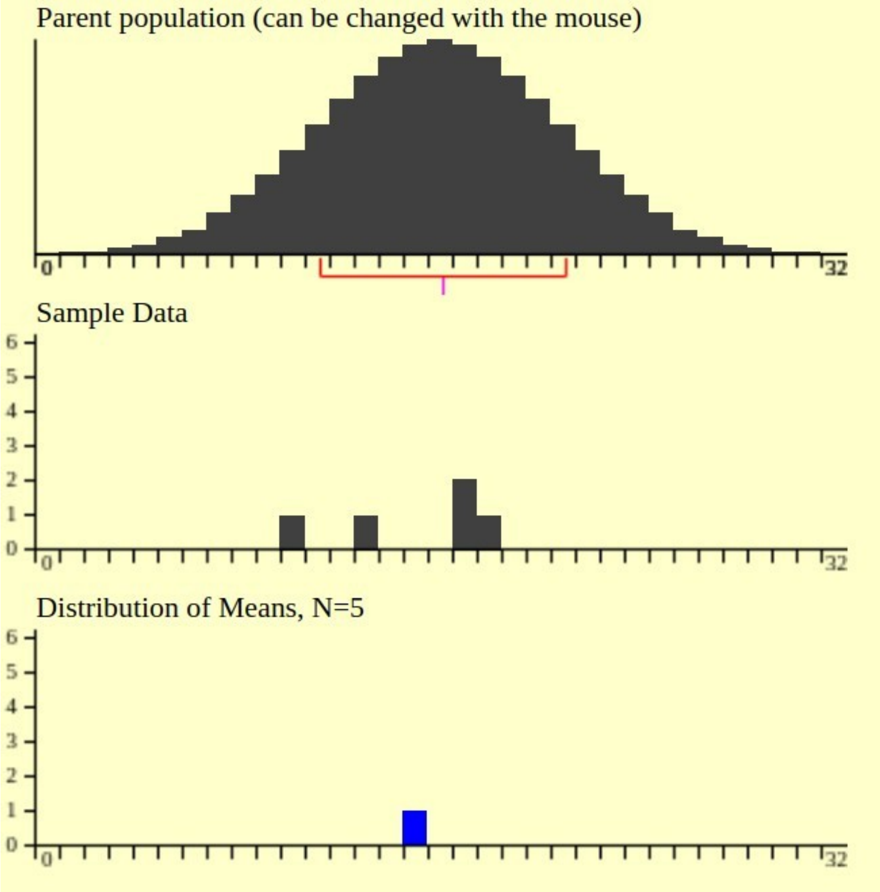
Mean = 16.00
Sd = 5.00

One sample
(size = 5)

Population distribution

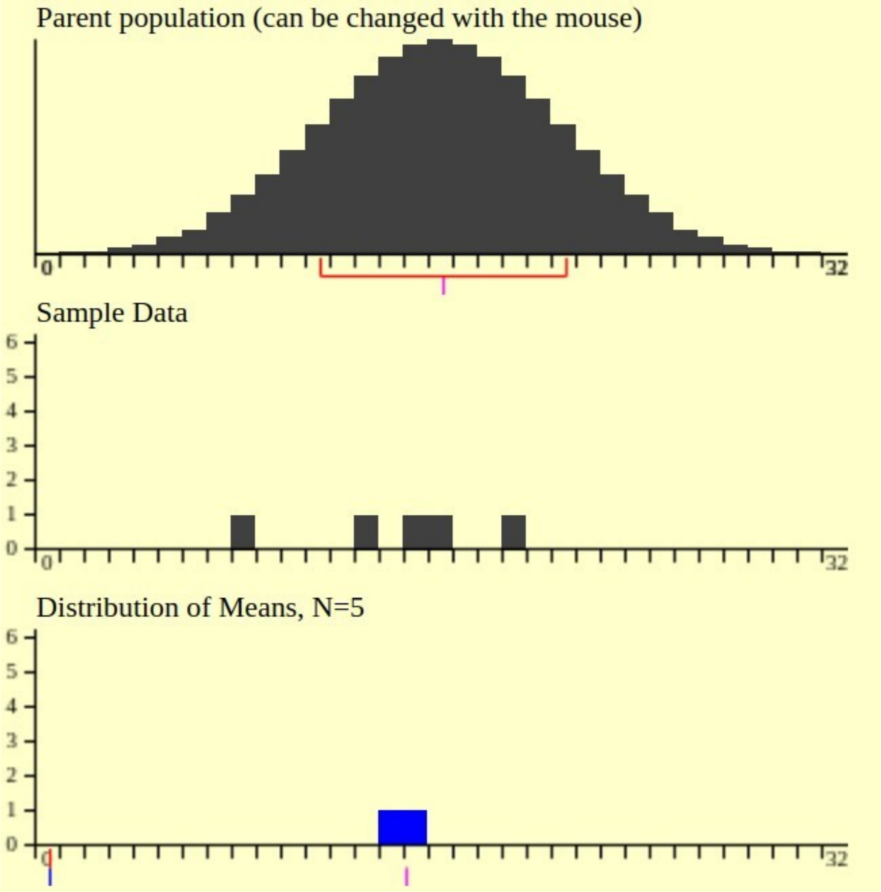
One sample
(size = 5)

Distribution of means
(1 sample)



Mean = 16.00
Sd = 5.00

Population distribution



Mean = 16.00
Sd = 5.00

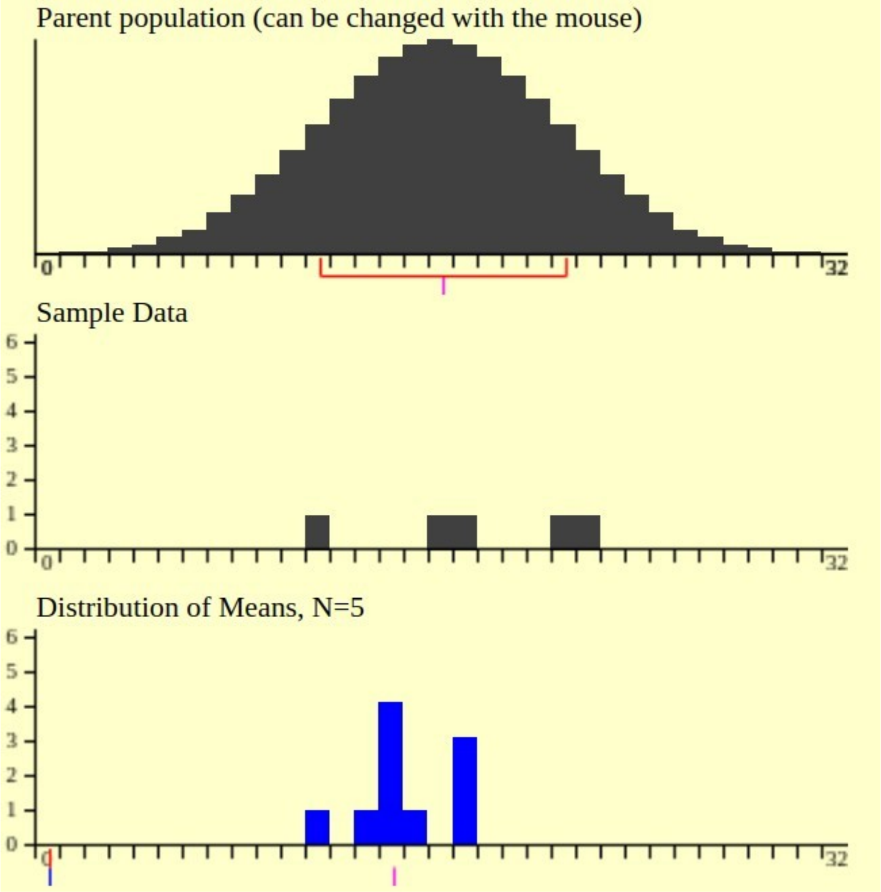
One sample
(size = 5)

Distribution of means
(2 samples)

Population distribution

One sample
(size = 5)

Distribution of means
(10 samples)



Mean = 16.00
Sd = 5.00

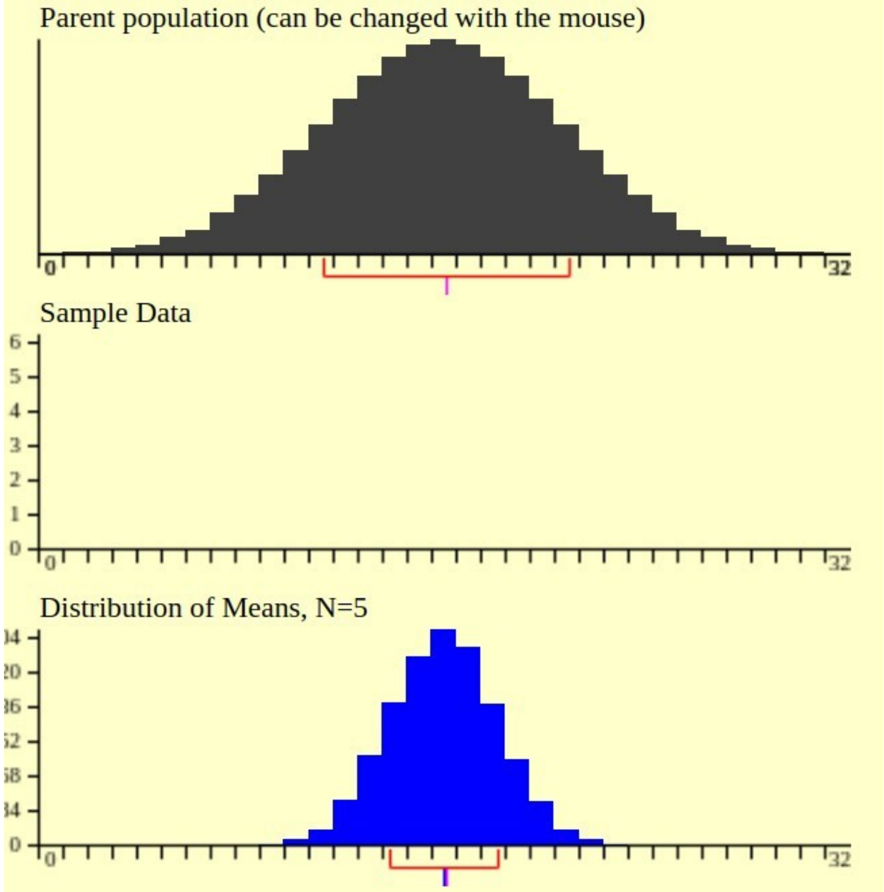
Population distribution

One sample

(size = 5)

Distribution of means

(10,000 samples)



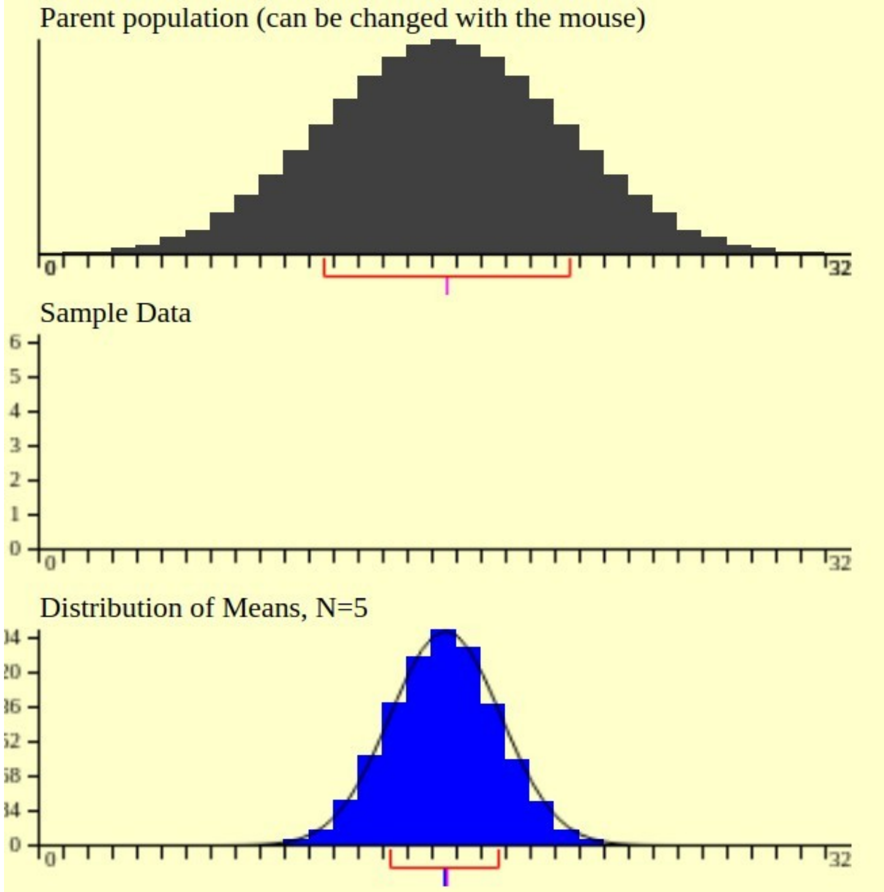
Mean = 16.00
Sd = 5.00

Mean = 15.99
Sd = 2.25

Population distribution

One sample
(size = 5)

Distribution of means
(10,000 samples)



Mean = 16.00
Sd = 5.00

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.

Definition, pg 405
Introduction to the Practice of Statistics, Fifth Edition

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

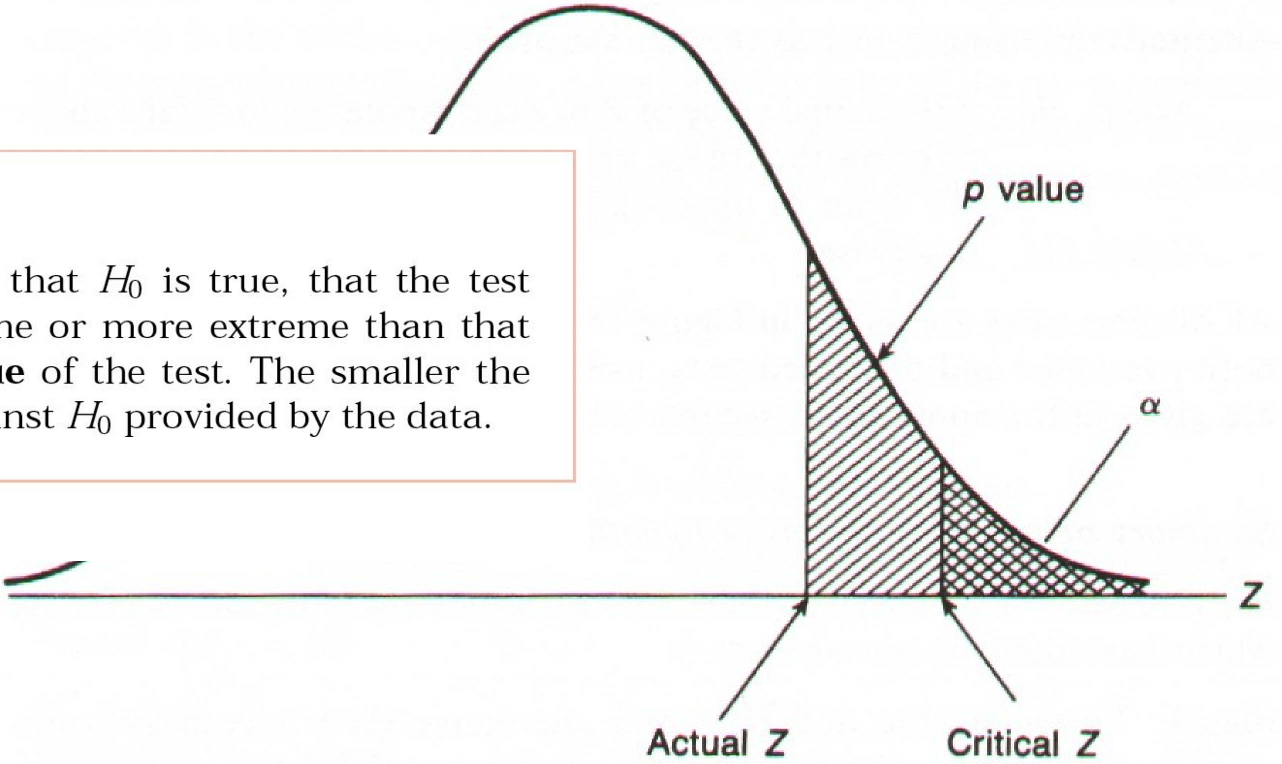


Figure 14.2 Comparison of *p* values and critical values of *Z* in a one-tailed test

The infamous P-value

If p-value < α,
then we reject the
null hypothesis

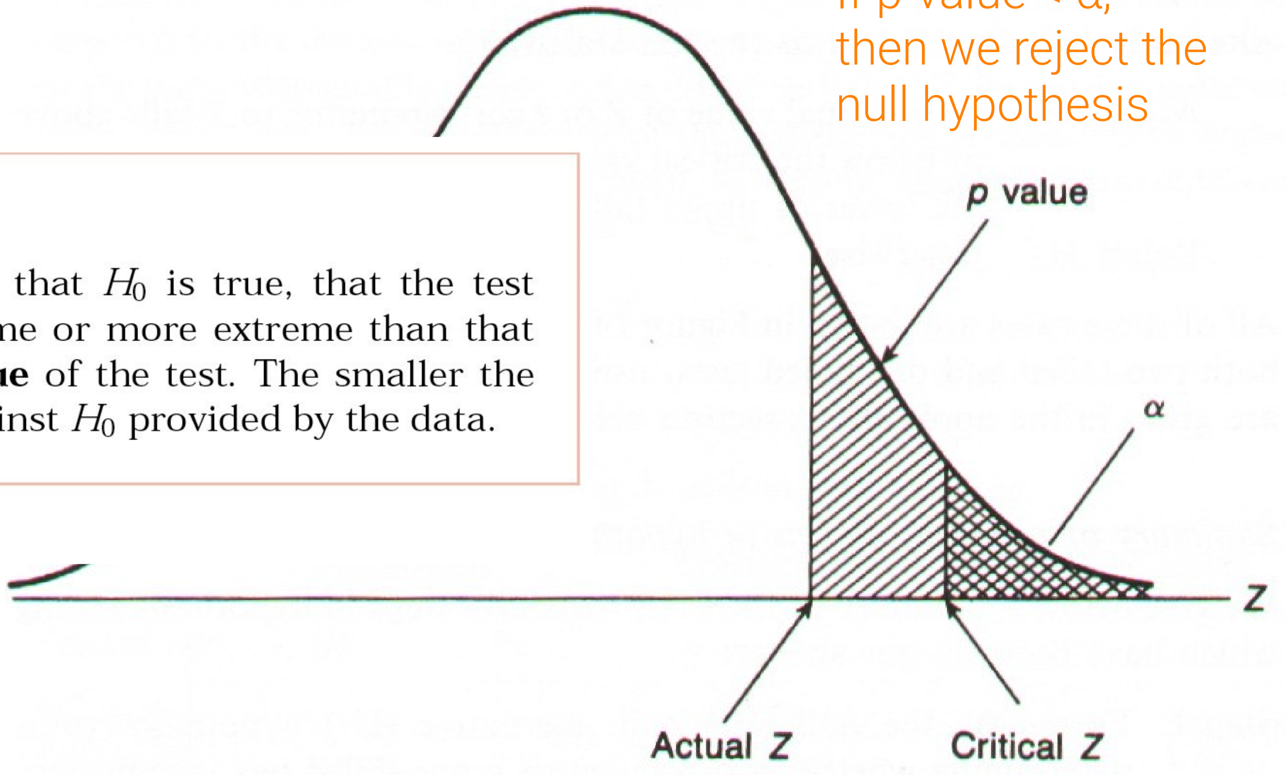


Figure 14.2 Comparison of *p* values and critical values of *Z* in a one-tailed test

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.

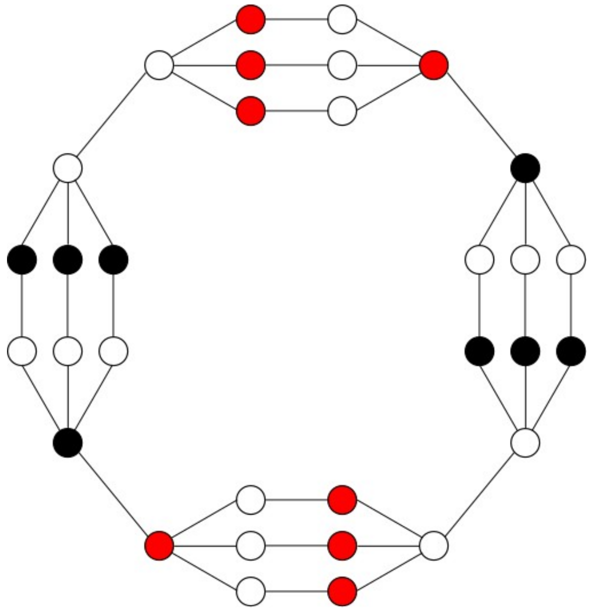
Definition, pg 405
Introduction to the Practice of Statistics, Fifth Edition

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 II

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?

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

Blockmodel Entropy Significance Test

How well do the metadata explain the network?

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

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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)

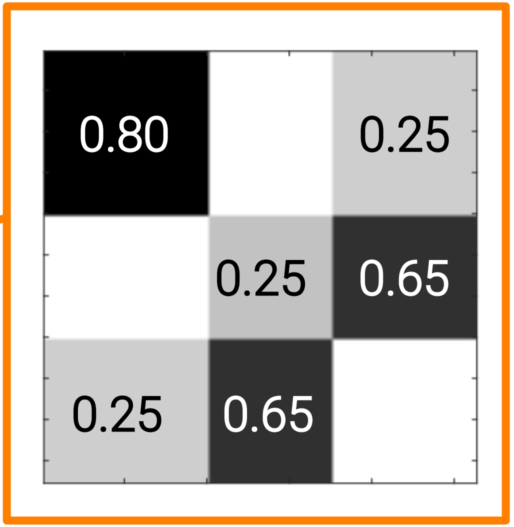
Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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)



Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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 $\boxed{\text{entropy } \mathbf{H}(G,M)}$

Test statistic

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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.

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

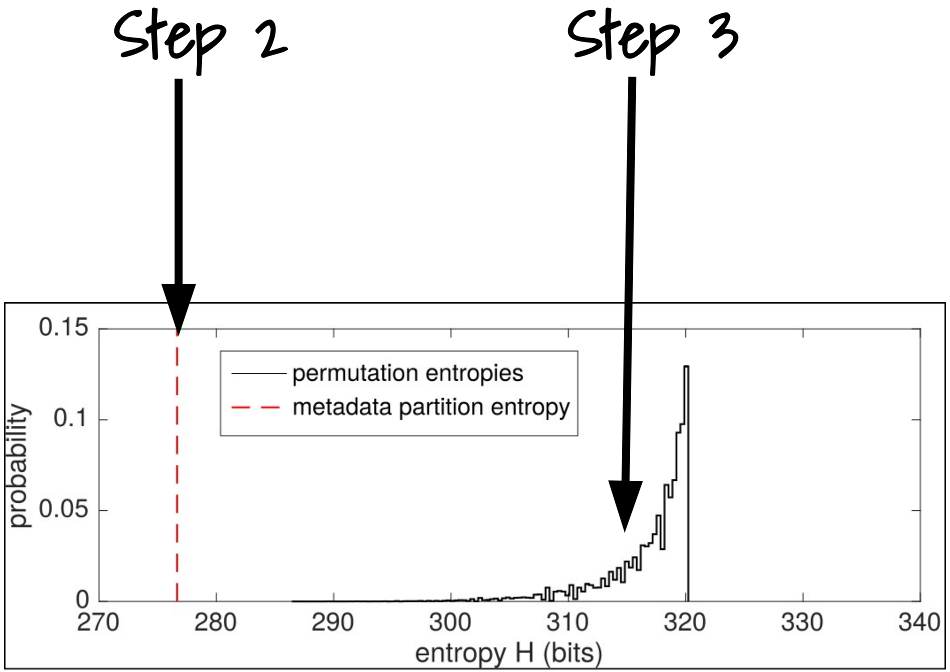
Blockmodel Entropy Significance Test

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.



Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

Blockmodel Entropy Significance Test

How well do the metadata explain the network?

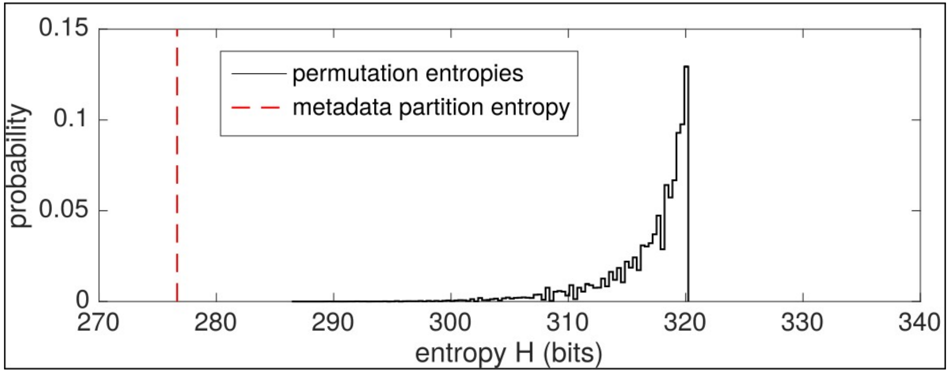
1. Divide the network G into groups according to metadata labels M.

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**



Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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.

Lazega, The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership, Oxford University Press (2001).

Practical Part III

Summary...

We can use graph models to simulate observed properties and form hypotheses

Null hypothesis tests are a popular way to test these hypotheses

Note: we can only accept/reject the null hypothesis

(more on this tomorrow)