Graph mining - lesson 3
Tests and random graphs

Nathalie Vialaneix

nathalie.vialaneix@inra.fr
http://www.nathalievialaneix.eu

M2 Statistics & Econometrics
January, 30th 2019

Sketch of this lesson

Issue at stake:

- short overview of different random graph models
- tests based on these models to assess the significance of numerical characteristics or of attributes

Notations for this class

Notations

In the following, a graph $\mathcal{G} = (V, E, W)$ with:

- V: set of vertices $\{x_1, \ldots, x_n\}$;
- E: set of (undirected) edges. $m = |E|$;
- W: weights on edges s.t. $W_{ij} \geq 0$, $W_{ij} = W_{ji}$ and $W_{ii} = 0$ (also called, *adjacency matrix*).

Notations for this class

Notations

In the following, a graph $\mathcal{G} = (V, E, W)$ with:

- V: set of vertices $\{x_1, \ldots, x_n\}$;
- E: set of (undirected) edges. $m = |E|$;
- W: weights on edges s.t. $W_{ij} \geq 0$, $W_{ij} = W_{ji}$ and $W_{ii} = 0$ (also called, *adjacency matrix*).

If needed, attributes for the nodes will be denoted by $f_j(x_i)$ (jth attribute for node i) and attributes for the edges (other than the weights) by $g_j(x_i, x_{i'})$ (jth attribute for the edge $(x_i, x_{i'})$).

Outline

Some standard random graph models

Permutation tests

Permutation tests for testing vertex labels

Some well known models studied in this class...

- Erdös-Rényi model [Erdös and Rényi, 1959] first and simplest random graph model in which the probability to observe an edge between two vertices is uniform over all pairs of vertices
- scale free model [Barabási and Albert, 1999] network in which the degree distribution fits a power law
- Stochastic Block Model [Snijders and Nowicki, 1997] random graphs with a community structure

Erdös-Rényi model

2 closely related Erdös-Rényi models:

- $G(n, m)$ the first method consists to set a number of vertices, n, and a number of edges, m, and to choose one graph uniformly at random among the set of graphs with n vertices and m edges;

- $G(n, p)$ the second method consists to set a number of vertices, n, and the probability, p, to decide, for every pair of vertices, if an edge exists between the two vertices, independantly at random with probability p

Some properties of ER graphs $G(n, p)$

- $\mathbb{E}(\text{number of edges in } G(n, p)) = \left(\begin{array}{c} n \\ 2 \end{array}\right) p$
- degree distribution is binomial:

$$\mathbb{P}(d_i = k) = \left(\begin{array}{c} n-1 \\ k \end{array}\right) p^k (1-p)^{n-1-k}$$

Some properties of ER graphs $G(n, p)$

- $\mathbb{E}(\text{number of edges in } G(n,p)) = \binom{n}{2} p$

- degree distribution is binomial:

$$\mathbb{P}(d_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

- diameter $\sim \frac{\log n}{\log(np)}$ [Albert and Barabási, 2002]

Some properties of ER graphs $G(n,p)$

- $\mathbb{E}(\text{number of edges in } G(n,p)) = \begin{pmatrix} n \\ 2 \end{pmatrix} p$
- degree distribution is binomial:

$$\mathbb{P}(d_i = k) = \begin{pmatrix} n-1 \\ k \end{pmatrix} p^k (1-p)^{n-1-k}$$

- diameter $\sim \frac{\log n}{\log(np)}$ **[Albert and Barabási, 2002]**
- depending on whether $p < \frac{(1-\epsilon)\log n}{n}$ or $p > \frac{(1+\epsilon)\log n}{n}$, $\mathbb{P}(G(n,p) \text{ contains isolated vertices}) \xrightarrow{n\to+\infty} 1$ or $\mathbb{P}(G(n,p) \text{ is connected}) \xrightarrow{n\to+\infty} 1$

scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2,3[$ (in general).

scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2,3[$ (in general).

Widely used model: Barabási-Albert model [Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.

scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model [Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.

Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

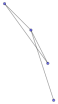
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model [Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

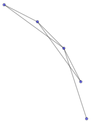
scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).
Widely used model: Barabási-Albert model
[Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

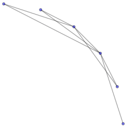
$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model

[Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

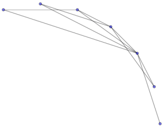
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

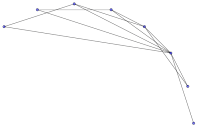
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

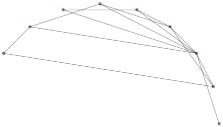
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

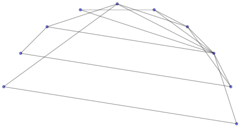
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2,3[$ (in general).

Widely used model: Barabási-Albert model [Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

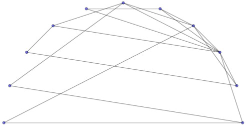
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

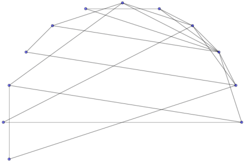
$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model

[Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

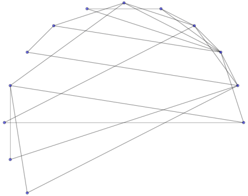
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model [Albert and Barabási, 2002]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

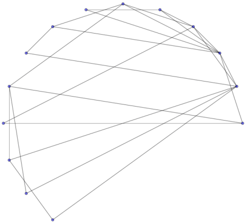
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

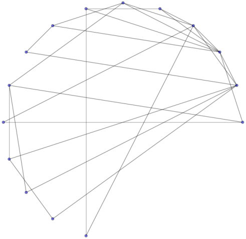
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

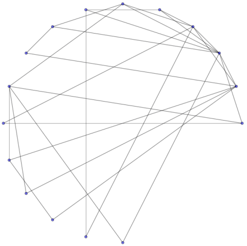
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

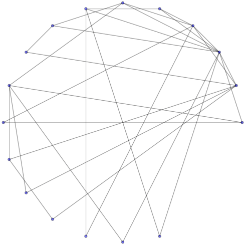
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2, 3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

Network is a network whose degree distribution (asymptotically) follows a power law:

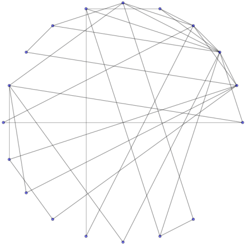
$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2,3[$ (in general).

Widely used model: Barabási-Albert model

[**Albert and Barabási, 2002**]:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

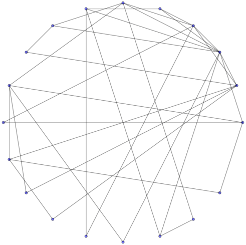
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2,3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

scale free model

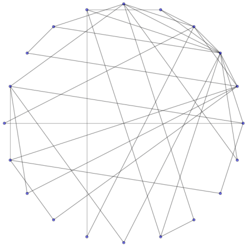
Network is a network whose degree distribution (asymptotically) follows a power law:

$$\mathbb{P}(d_i = k) = k^{-\gamma}$$

for a $\gamma \in]2,3[$ (in general).

Widely used model: Barabási-Albert model **[Albert and Barabási, 2002]**:

- vertices are added one at at time;
- the new vertex x_{n+1} has a probability to be connected to vertex x_i equal to $\frac{d_i}{\sum_j d_j}$.



Source: By Horváth Árpád, CC BY-SA 3.0

Some properties of BA and scale free graphs

- Degree distribution of BA graph: scale free with $\gamma = 3$

Some properties of BA and scale free graphs

- Degree distribution of BA graph: scale free with $\gamma = 3$

- [Cohen and Havlin, 2003] depending on whether $\gamma < 3$, $\gamma = 3$ and $\gamma > 3$, the average shortest path length is of order $\log \log n$ ("ultra-small worlds"), $\frac{\log n}{\log \log n}$ or $\log n$

SBM

SBM with n vertices are random graphs for which a partition of the vertices, $\mathcal{C}_1, \ldots, \mathcal{C}_K$ is given, for which the probability that a vertex $x_i \in \mathcal{C}_k$ and a vertex $x_j \in \mathcal{C}_{k'}$ are connected, has a probability $\pi_{kk'} \in [0, 1]$.

SBM

SBM with n vertices are random graphs for which a partition of the vertices, $\mathcal{C}_1, \ldots, \mathcal{C}_K$ is given, for which the probability that a vertex $x_i \in \mathcal{C}_k$ and a vertex $x_j \in \mathcal{C}_{k'}$ are connected, has a probability $\pi_{kk'} \in [0, 1]$.

Some results describe the conditions for which the partition can be recovered (for a number of vertices that tends to $+\infty$), according to the relation between the intra/inter-block probabilities.

Comparison with random graphs...

Erdos-Renyi model with the same number of nodes and the same number of edges than the original graph $G(n, m)$

Comparison with random graphs...

Erdos-Renyi model with the same number of nodes and the same number of edges than the original graph $G(n, m)$

Method: compare the observed values with those of a large number of randomly generated random graphs (with no loop, only connected graphs are kept)

Results of the comparison with random graphs...

For $B = 500$ graphs (only connected graphs are kept), we have:

	density	transitivity	diameter	radius	girth	cohesion
X	Min. :0.0725	Min. :0.0523	Min. :4.00	Min. :3	Min. :3	Min. :1.00
X.1	1st Qu.:0.0725	1st Qu.:0.0679	1st Qu.:4.00	1st Qu.:3	1st Qu.:3	1st Qu.:2.00
X.2	Median :0.0725	Median :0.0721	Median :4.00	Median :3	Median :3	Median :3.00
X.3	Mean :0.0725	Mean :0.0722	Mean :4.11	Mean :3	Mean :3	Mean :2.52
X.4	3rd Qu.:0.0725	3rd Qu.:0.0762	3rd Qu.:4.00	3rd Qu.:3	3rd Qu.:3	3rd Qu.:3.00
X.5	Max. :0.0725	Max. :0.0971	Max. :5.00	Max. :3	Max. :3	Max. :4.00

which has to be compared to 0.072, 0.56, 18, 9, 3 and 1.

- the transitivity is much larger;
- the diameter and the radius are much larger also;
- the cohesion is smaller.

The first remark indicates a stronger local connectivity in my NVV network than in ER models with same number of nodes and edges. Large radius and diameter are explained by an isolated branch in the network.

Comparison with random graphs...

Scale free model: Barabási and Albert model is used with a number of edges added at each step which is chosen so that the final number of edges resembles that of the original graph (4 edges, which gives 478 edges in the final graph, compared to 535 in the real NVV network.

Comparison with random graphs...

Scale free model: Barabási and Albert model is used with a number of edges added at each step which is chosen so that the final number of edges resembles that of the original graph (4 edges, which gives 478 edges in the final graph, compared to 535 in the real NVV network.

Method: compare the observed values with those of a large number of randomly generated random graphs

Results of the comparison with random graphs...

For $B = 500$ graphs, we have:

	density	transitivity	diameter	radius	girth	cohesion
X	Min. :0.0648	Min. :0.105	Min. :4	Min. :2.00	Min. :3	Min. :4
X.1	1st Qu.:0.0648	1st Qu.:0.117	1st Qu.:4	1st Qu.:3.00	1st Qu.:3	1st Qu.:4
X.2	Median :0.0648	Median :0.121	Median :4	Median :3.00	Median :3	Median :4
X.3	Mean :0.0648	Mean :0.121	Mean :4	Mean :2.98	Mean :3	Mean :4
X.4	3rd Qu.:0.0648	3rd Qu.:0.125	3rd Qu.:4	3rd Qu.:3.00	3rd Qu.:3	3rd Qu.:4
X.5	Max. :0.0648	Max. :0.138	Max. :4	Max. :3.00	Max. :3	Max. :4

which has to be compared to 0.072, 0.56, 18, 9, 3 and 1.

- the transitivity is still much larger;
- the diameter and the radius are much larger also;
- the cohesion is much smaller.

The first remark indicates a stronger local connectivity in my NVV network than in BA models with same number of nodes and edges.

Outline

Some standard random graph models

Permutation tests

Permutation tests for testing vertex labels

Limits of the previous approaches

Until now, we have compared the real graph to graphs randomly generated according to a given random model but:

- this approach only gives information about global characteristics of the observed graph;
- none of the distributions of the current characteristics is preserved during the process, especially not the degree distribution which is central for controlling local/global connectivity, counts of specific patterns...

A null model closer to the real graph...

Sketch of statistical tests on graphs

1. sample at random within the set of graphs with the same degree distribution than the observed graph (*B* times)

2. compute a numerical statistics for each of these randomly generated graphs

3. compare the observed value of the statistics and its distribution over the random graphs, a p-value can be derived (for *B* large enough)

A null model closer to the real graph...

Sketch of statistical tests on graphs

1. sample at random within the set of graphs with the same degree distribution than the observed graph (*B* times)
2. compute a numerical statistics for each of these randomly generated graphs
3. compare the observed value of the statistics and its distribution over the random graphs, a p-value can be derived (for *B* large enough)

Two main approaches to sample at random with fixed degrees:

- configuration model **[Bender and Canfield, 1978]**
- permutation approach **[Rao et al., 1996, Roberts Jr., 2000]**

Sampling at random within the set of graphs with a given degree distribution

Aim:

- all graphs can exhaustively be sampled
- all graphs have the same probability to be sampled

⇒ MCMC approach

Sampling at random within the set of graphs with a given degree distribution

Aim:

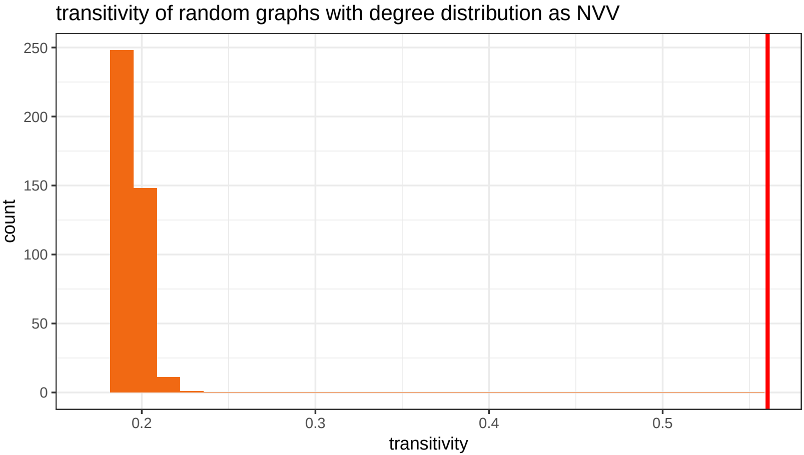
- all graphs can exhaustively be sampled
- all graphs have the same probability to be sampled

⇒ MCMC approach

Method:

```
Start from the observed graph $\mathcal{G}$
for $t = 1 \rightarrow T|E|$ do
    Select uniformly at random two edges $e^1 = (x_i^1, x_j^1)$ and $e^2 = (x_i^2, x_j^2) \in E$
    $E' \leftarrow E \setminus \{e^1, e^2\} \cup \{e_s^1, e_s^2\}$ with $e_s^1 = (x_i^1, x_j^2)$ and $e_s^2 = (x_i^2, x_j^1)$
    if $\mathcal{G}' = (V, E')$ is simple and connected then
        $\mathcal{G} \leftarrow \mathcal{G}'$
    end if
end for
return $\mathcal{G}$
```

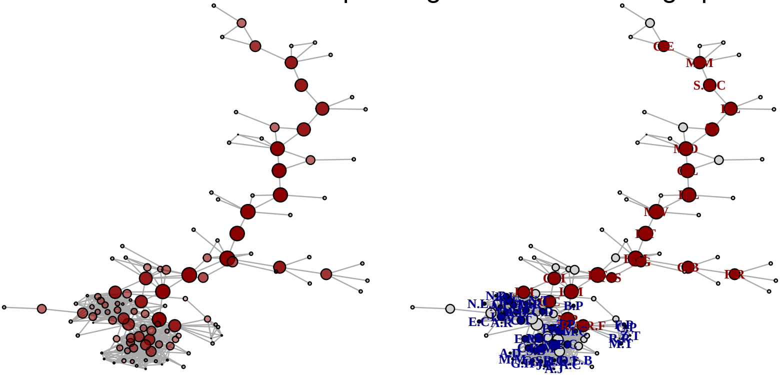
In practice... (for the transitivity)



Again, this is evidence for a strong local connectivity in the network.

In practice... (for the vertex characteristics)

Find a(n empirical) p-value for all vertices which indicates if its betweenness is higher or lower than expected with respect to its degree: ratio of random graphs for which the observed betweenness is higher (resp. lower) than 95% of the betweennesses for the corresponding vertex in random graphs.

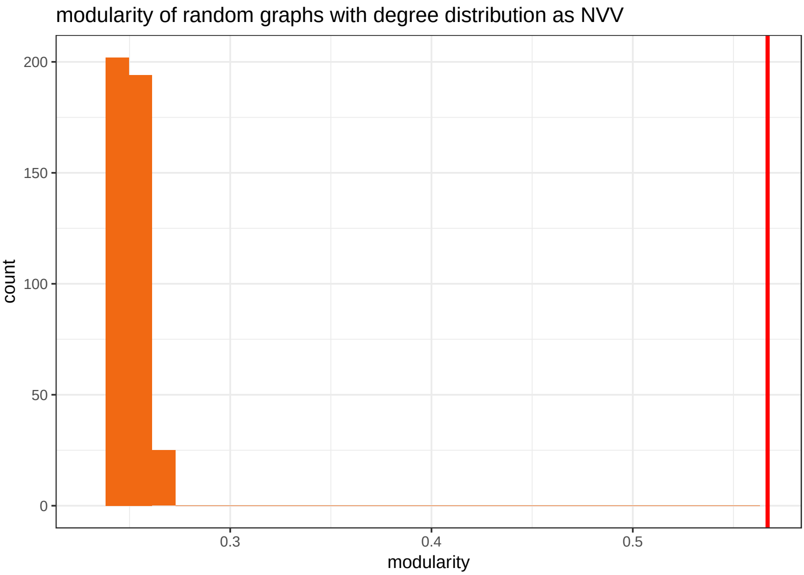


Using random graphs to assess the relevance of the clustering

Method:

- Find the clustering associated to the maximum modularity.
- Compare this modularity with the distribution of maximum modularities over *B* random graphs obtained by edge permutation.

Relevance of the clustering of NVV



More on random graphs generation

Sometimes, one wants to compare the observed graph with a more sophisticated (constrained) null model (taking into account some additional information on edges or nodes for instance):

- This can be achieved using the same principle and throwing away the random graphs which do not satisfy the constrains.

More on random graphs generation

Sometimes, one wants to compare the observed graph with a more sophisticated (constrained) null model (taking into account some additional information on edges or nodes for instance):

- This can be achieved using the same principle and throwing away the random graphs which do not satisfy the constrains. Warning: The more sophisticated the model is, the more costly the simulation would be. For instance, only removing graphs with multiple edges and graphs which are not connected leads to throw away 53 graphs during the previous generation process.

More on random graphs generation

Sometimes, one wants to compare the observed graph with a more sophisticated (constrained) null model (taking into account some additional information on edges or nodes for instance):

- This can be achieved using the same principle and throwing away the random graphs which do not satisfy the constrains. Warning: The more sophisticated the model is, the more costly the simulation would be. For instance, only removing graphs with multiple edges and graphs which are not connected leads to throw away 53 graphs during the previous generation process.
- Possible solution: [Tabourier and Cointet, 2011] use multiple edge switching to improve such simulations.

Outline

Some standard random graph models

Permutation tests

Permutation tests for testing vertex labels

General setting

A label $f(x_i)$ is given for all vertices in the graph.

Question: Are the labels “related” to the graph structure? *e.g.*

- connected nodes tend to have similar labels;
- OR connected nodes tend to have opposite labels.

Join Count Statistics

First case: binary labels $f(x_i) \in \{0, 1\}$

Join Count Statistics

First case: binary labels $f(x_i) \in \{0, 1\}$

General form:

$$JC = \frac{1}{2} \sum_{i \neq j} W_{ij} \xi_i \xi_j$$

where ξ_i is either $f(x_i)$ or $1 - f(x_i)$.

Join Count Statistics

First case: binary labels $f(x_i) \in \{0, 1\}$

Derived statistics:

- Number of "1" labels in the neighbor of a vertex labelled "1"

$$JC_1 = \frac{1}{2} \sum_{i,j:\ f(x_i)=f(x_j)=1} W_{ij}$$

- Number of "0" labels in the neighbor of a vertex labelled "0"

$$JC_0 = \frac{1}{2} \sum_{i,j:\ f(x_i)=f(x_j)=0} W_{ij}$$

- Number of "1" labels in the neighbor of a vertex labelled "0" (and the opposite)

$$JC_{0-1} = \sum_{i,j:\ f(x_i)=0,\ f(x_j)=1} W_{ij}$$

Interpretation

Basic interpretation: If JC_1 is “large” (“small”) then vertices labelled “1” in the network tend to be connected with vertices labelled the same way (or tend not to be related to vertices labelled “1”).

Interpretation

Basic interpretation: If JC_1 is “large” (“small”) then vertices labelled “1” in the network tend to be connected with vertices labelled the same way (or tend not to be related to vertices labelled “1”).

Statistical significance: When is JC_1 significantly large or small?

- Method 1: **[Noether, 1970]** proves the asymptotic normal distribution of JC_1: requires additionnal assumptions on the network and not valid for small networks;

Interpretation

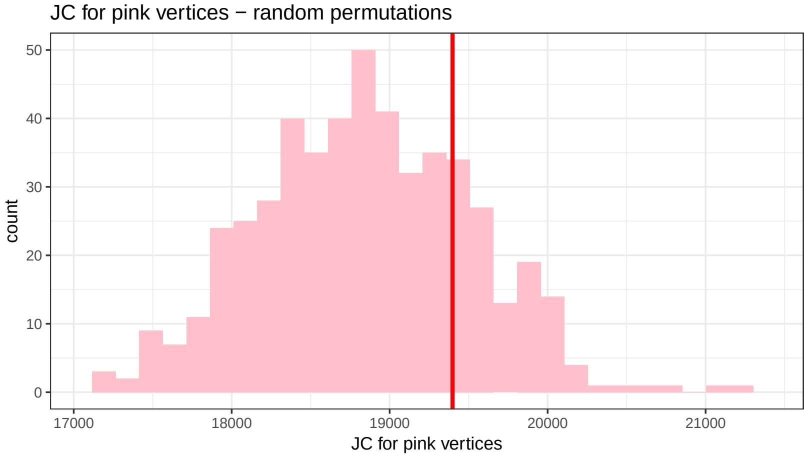
Basic interpretation: If JC_1 is "large" ("small") then vertices labelled "1" in the network tend to be connected with vertices labelled the same way (or tend not to be related to vertices labelled "1").

Statistical significance: When is JC_1 significantly large or small?

- Method 1: [Noether, 1970] proves the asymptotic normal distribution of JC_1: requires additionnal assumptions on the network and not valid for small networks;
- Method 2: Monte Carlo approach: Randomly permute the values $f(x_i)$ over the vertices, B times (where B is large) and obtain the empirical distribution of JC_1. Compare with the true JC_1.

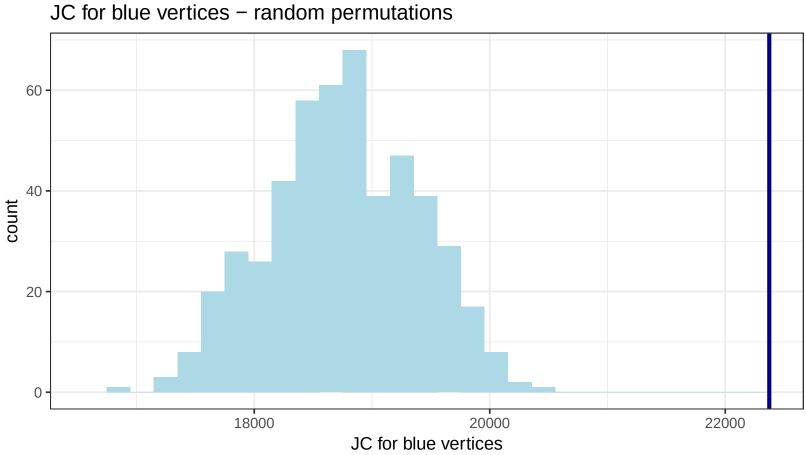
  $\Rightarrow$ Estimation of the distribution of JC_1 given the network and the numbers of labels "1" and "0".

FB network and connection gender



girls tend not to be particularly more connected with other girls that what could be expected by random choice

FB network and connection gender



boys tend to be more connected with other boys that what could be expected by random choice

Continuous labels with Moran's I

[Moran, 1950] proposes to measure spatial correlation with the *I* statistics:

$$\mathbf{I} = \frac{\frac{1}{2m}\sum_{i\neq j} W_{ij}\overline{f(x_i)f(x_j)}}{\frac{1}{n}\sum_i \overline{f(x_i)}^2}$$

where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\overline{f(x_i)} = f(x_i) - \bar{f}$ with $\bar{f} = \frac{1}{n}\sum_i f(x_i)$.

Continuous labels with Moran's I

[**Moran, 1950**] proposes to measure spatial correlation with the *I* statistics:

$$\mathbf{I} = \frac{\frac{1}{2m}\sum_{i\neq j} W_{ij}\overline{f(x_i)f(x_j)}}{\frac{1}{n}\sum_i \overline{f(x_i)}^2}$$

where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\overline{f(x_i)} = f(x_i) - \bar{f}$ with $\bar{f} = \frac{1}{n}\sum_i f(x_i)$.

Interpretation: When *I* is "large", vertices tend to be connected to other vertices having similar labels; when *I* is "small", vertices tend to be connected to other vertices having opposite labels. Average *I* means that there is no special relation between labels and the graph structure.

Continuous labels with Moran's I

[Moran, 1950] proposes to measure spatial correlation with the *I* statistics:

$$\mathbf{I} = \frac{\frac{1}{2m}\sum_{i\neq j} W_{ij}\overline{f(x_i)f(x_j)}}{\frac{1}{n}\sum_i \overline{f(x_i)}^2}$$

where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\overline{f(x_i)} = f(x_i) - \bar{f}$ with $\bar{f} = \frac{1}{n}\sum_i f(x_i)$.

Interpretation: When *I* is "large", vertices tend to be connected to other vertices having similar labels; when *I* is "small", vertices tend to be connected to other vertices having opposite labels. Average *I* means that there is no special relation between labels and the graph structure.

Deriving a test for *I*: again, asymptotic normality can be proved or using a Monte Carlo simulation is also possible.

Albert, R. and Barabási, A. (2002).
Statistical mechanics of complex networks.
Reviews of Modern Physics, 74:47–97.

Barabási, A. and Albert, R. (1999).
Emergence of scaling in random networks.
Science, 286:509–512.

Bender, E. and Canfield, E. (1978).
The asymptotic number of labeled graphs with given degree sequences.
Journal of Combinatorial Theory, Series A, 24(3):296–307.

Cohen, R. and Havlin, S. (2003).
Scale-free networks are ultrasmall.
Physical Review Letters, 90(5):058701.

Erdös, P. and Rényi, A. (1959).
On random graphs. i.
Publicationes Mathematicae, 6:290–297.

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M., and Alon, U. (2004).
On the uniform generation of random graphs with prescribed degree sequences.
eprint arXiv: cond-mat/0312028v2.

Moran, P. (1950).
Notes on continuous stochastic phenomena.
Biometrika, 37:17–23.

Noether, G. (1970).
A central limit theorem with non-parametric applications.
Annals of Mathematical Statistics, 41:1753–1755.

Rao, A., Jana, R., and Bandyopadhyay, S. (1996).
A markov chain monte carlo method for generating random (0, 1)-matrices with given marginals.

Sankhyã: The Indian Journal of Statistics, Series A (1961-2002), 58(2):225–242.

Roberts Jr., J. (2000).
Simple methods for simulating sociomatrices with given marginal totals.
Social Networks, 22(3):273 – 283.

Snijders, T. and Nowicki, K. (1997).
Estimation and prediction for stochastic block-structures for graphs with latent block structure.
Journal of Classification, 14:75–100.

Tabourier, L.and Roth, C. and Cointet, J. (2011).
Generating constrained random graphs using multiple edge switches.
ACM Journal of Experimental Algorithmics, 16(1):1.7.