(c) $n = 4$ and $\lambda = 2/3$ give

$$\vec{p} = \left(\frac{7}{34}, \frac{25}{68}, \frac{7}{34}, \frac{15}{68}\right) = (0.206, 0.368, 0.206, 0.221)\,.$$

Exercise 3

(1.5 points) The Search for Extra Terrestrial Intelligence (SETI) Institute has three long sequences of numbers

1. One that shows a straight line with a -1 slope when the frequency of every number is plotted in double logarithmic scale (the y-axis indicates the logarithm of that frequency and the x-axis indicates the logarithm of the corresponding number).
2. One that shows a straight line with a -3 slope when the frequency of every number is plotted in double logarithmic scale as before.
3. Another one that shows a straight line when the frequency of every number is plotted taking logarithms on the frequencies only (the y-axis indicates the logarithm of that frequency and the x-axis indicates the corresponding number).

These sequences are being used by SETI to evaluate candidates for a new job in one of its research programs. Candidates are told that every number of the sequence can be

a The number of die rolls until a 6 is produced by a fair die.

b The rank of a word type from human language.

c A vertex degree from a network obtained simulating the Barábasi-Albert (BA) model.

d None of the above.

1. Indicate the likely source of every sequence (die rolls, human language, BA model or none of the previous ones).
2. Give a mathematical argument linking each of the straight lines of the plots with the corresponding source.

Answer of exercise 3

1. Sequence 1 is consistent with human language, Sequence 2 is consistent with the BA model and Sequence 3 is consistent with die rolls (further statistical properties would be necessary to stablish a strong connection).
2. In sequences 1 and 2,
$$\log y = a \log x + b$$
with $a = -1$ for sequence 1 and $a = -3$ for sequence 2. We get rid of the logarithms by exponentiating, i.e.
$$e^{\log y} = e^{a \log x + b},$$
which finally gives
$$y = cx^a$$
with $c = e^b$. Therefore, sequence 1 follows the distribution of ranks of English (Zipf's rank-frequency law) while sequence 2 follows the power-law degree distribution of the BA model. The probability of x, the number of die rolls until a six is obtained, is
$$p(x) = \pi(1 - \pi)^x$$
with $\pi = 1/6$. In a sequence of T numbers, one expects that the frequency of x is $y = Tp(x)$. Taking logarithms on y one obtains the straight line of System 3, i.e.
$$\log y = a \log x + b$$
with $a = \log(1 - \pi)$ and $b = \log(T\pi)$.

Exercise 4

(1 point) Define the paradox of choice and explain why it originates.

Answer of exercise 4

The paradox of choice occurs when customers end up buying less (and are less satisfied if they buy) when the are offered more choices. *A priori*, one expects that customers buy more when they are offered more choices (if there are not enough products available, the customer will not buy because he/she does not find what he/she is looking for). However, having many choices overloads the customer. More possibilities require increasing time and effort and they can eventually lead to anxiety, regret or excessively high expectations.

Exercise 5

(1.5 points) We have a matrix with two columns, i.e., x and y, and n rows of the form (x_i, y_i) (with $1 \leq i \leq n$ and $n \geq 2$). We wish to calculate the Kendall τ correlation between x and y. This measure of correlation is analogous to the Pearson correlation ($-1 \leq \tau \leq 1$) but is defined on the notion of concordance. The points (x_i, y_i) and (x_j, y_j) are concordant if $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$. (x_i, y_i) and (x_j, y_j) are discordant if $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$. Then the Kendall correlation between x and y is defined as

$$\tau(x, y) = \frac{n_c - n_d}{\binom{n}{2}},$$

where n_c is the number of concordant pairs of points and n_d is the number of discordant pairs of points.

Solve the problem of calculating $\tau(x, y)$ using the MapReduce programming model. Give pseudocode for the *map* and *reduce* functions and, if appropriate, for the *combine* function.

Answer of exercise 5

Warning: discordant is not equivalent to not concordant. Therefore,

$$n_c + n_d \leq \binom{n}{2}$$

but

$$n_c + n_d = \binom{n}{2}$$

is not generally true.

The sign function is

$$sign(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 < x_2 \\ 0 & \text{if } x_1 = x_2 \\ -1 & \text{if } x_1 > x_2 \end{cases}$$

Solution 1 (the maps produce at most 2 distinct keys):

Each instance of *map* receives a distinct (unordered) pair of elements $\{(x_i, y_i), (x_j, y_j)\}$. The *reduce* produce n_c and n_d.

map($\{(x_1, y_1), (x_2, y_2)\}$)
    s = $sign(x_1, x_2)$*$sign(y_1, y_2)$
    if s $\neq$ 0 output(s, 1)

reduce(k,L) = combine(k, L)
    output(k, sum(L))

Solution 2 (the maps produce at most 1 distinct key):

Each instance of map receives a distinct (unordered) pair of elements $\{(x_i, y_i), (x_j, y_j)\}$. The *reduce* produces $n_c - n_d$.

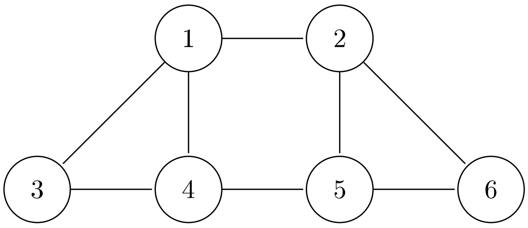
map($\{(x_1, y_1), (x_2, y_2)\}$)
    s = $sign(x_1, x_2)$*$sign(y_1, y_2)$
    if s $\neq$ 0 output(0, s)

reduce(k,L) = combine(k, L)
    output(k, sum(L))

Other correct solutions can be built generalizing the *map* so as to receive more than one pair of points. However, solutions where the instances of *map* receive the whole matrix are discouraged (the matrix could be very large).

Exercise 6

(1.5 points) We wish to analyze various statistical properties of the following network:



1. The diameter (in edges).
2. The average vertex clustering, namely

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i,$$

   where n is the number of vertices of the network and C_i is the vertex clustering of the i-th vertex.
3. T, the transitivity coefficient, namely the proportion of connected triples that belong to a triangle.
4. If the network was an Erdős-Rényi graph, what would be the expected clustering coefficient?

Answer of exercise 6

The network consist of two triangles joined by two edges. By symmetry, there are two kinds of vertices:

A Vertices 1, 2, 4 and 5.

B Vertices 3 and 6.

1. Vertices of type A are at distance ≤ 2 from from the remainder of the vertices while vertices of type B are at distance ≤ 3. Thus, the diameter is 3.
2.

$$C_A = \frac{1}{\binom{3}{2}} = 1/3$$
$$C_B = 1$$

   and then

$$C = \frac{1}{6}(4C_A + 2C_B) = \frac{5}{9} = 0.\overline{5}$$

3. There are 2 triangles and

$$4\binom{3}{2} + 2$$

   connected triples of vertices. Then

$$T = \frac{3 \cdot 2}{4\binom{3}{2} + 2} = \frac{3}{7} \approx 0.428$$

4. In an Erdős-Rényi graph, the expected clustering matches the density of links,

$$\delta = \frac{m}{\binom{n}{2}}$$

$m = 8$ and $n = 6$ give $\delta = 8/15 = 0.5\overline{3}$.

Exercise 7

(1.5 points) We wish to study that community structure of the network of the previous exercise assuming that there are only two communities. We consider two partitionings: $\alpha = \{\{1, 2\}, \{3, 4, 5, 6\}\}$ and $\beta = \{\{1, 3, 4\}, \{2, 5, 6\}\}$.

1. Given what you have learnt about principles to define what a community is, argue which of the two partitionings looks *a priori* better (before applying any concrete community detection algorithm).

2. Calculate Newman's modularity, i.e.

$$Q = \frac{1}{2m} \sum_{ij} W_{ij} \delta(C_i, C_j)$$

for each of the two partitionings. Recall

$$W_{ij} = a_{ij} - \frac{k_i k_j}{2m}.$$

What is the best partitioning?

Answer of exercise 7

1. A common notion of good partitioning is that the number of intracommunity edges should exceed the number of intercommunity edges. In α, the density of links of the first community is maximum while the density of links of the second community is rather low (the edges of the 2nd community define a tree, namely, their number number is the minimum to keep the community connected). Besides, the two communities are linked by 4 edges. In β, both communities have maximum density of links while the number of edges connecting the two communities is only 2. Therefore β looks *a priori* better. Other principled answers are possible.

2. Notice that Q is defined over the whole matrix of W_{ij}^2 (including the diagonal!). This can be inferred from the proof of normalization ($Q \leq 1$) or the definition of the configuration model. Since $W_{ij} = W_{ji}$, we obtain

$$Q_\alpha = \frac{1}{m}(W_{12} + W_{34} + W_{45} + W_{56} + W_{35} + W_{36} + W_{46}) + \Delta$$

$$Q_\beta = \frac{2}{m}(W_{13} + W_{14} + W_{34}) + \Delta,$$

where

$$\Delta = \frac{1}{2m} \sum_{i=1}^{n} W_{ii}. \quad (3)$$

Thanks to the previous exercise, we know that there are only two kinds of vertices, A and B. The definition

$$W_{ij}(a) = a - \frac{k_i k_j}{2m}$$

yields

$$Q_\alpha = \frac{1}{m}(2W_{AA}(1) + 2W_{AB}(1) + 2W_{BB}(0) + 2W_{AB}(0) + 2W_{AA}(0))$$

$$Q_\beta = \frac{1}{m}(2W_{AA}(1) + 4W_{AB}(1) + 2W_{AA}(0) + W_{BB}(0)).$$

The fact that

$$
\begin{aligned}
k_A &= 3 \\
k_B &= 2 \\
m &= 8 \\
W_{AA}(a) &= a - \frac{k_A k_A}{2m} = a - \frac{9}{16} \\
W_{AB}(a) &= a - \frac{k_A k_A}{2m} = a - \frac{3}{8}
\end{aligned}
$$

eventually gives

$$
\begin{aligned}
Q_\alpha &= \frac{1}{8}\left(2(1-9/16) + 2(1-3/8) - 2(1/4) - 2(3/8) - 2(9/16)\right) = -\frac{1}{32} \approx -0.03125 \\
Q_\beta &= \frac{1}{8}\left(2(1-9/16) + 4(1-3/8) - 2(9/16) - 1/4\right) = \frac{1}{4}.
\end{aligned}
$$

Therefore, β is a better partitioning.