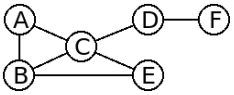
Homework #1
Key Network Properties, Graph Models and Gephi
Due: April 11th, 2021

- The assignment should be delivered digitally by email. Your message should be sent to **pribeiro@dcc.fc.up.pt** with subject *"[ARSI HWK1] FirstName LastName StudentNumber"*
- Your delivery should be a **zip file**, containing a **PDF report with the answers** and all additional files that were used for producing those answers
- This homework is to be made **individually**. You can collaborate with more students by talking about the exercises, but **you should not copy answers or code** and you should do you own writeup.
- Please **acknowledge any help you got** and state any references you consulted (including internet pages) and any other students with whom you talked about the exercises.
- Answers should be **submitted until 23:59 of the due date**. Up to 24h of delay will get you a 25% penalty. 24h to 48h of delay will get you a 50% penalty. After 48h your work will not be counted.

Key Network Properties

(for all answers explain how you arrived to the requested values and show any intermediate calculations)

1. Consider the simple (undirected) network represented by the following graph:



   (a) Show the **degree** of each node and make a **plot** of its (normalized) **degree distribution**.

   (b) Calculate the **diameter** and the **average path length** of the network

   (c) Calculate the **local clustering coefficient** of each node and the **average local clustering coefficient** of the entire network.

   (d) Calculate the (normalized) **betweenness centrality** and **closeness centrality** of each node.

2. Another possible clustering metric is the **global clustering coefficient** (also known as *transitivity*) which can be defined as in the following way:

$$C_{global} = \frac{3 \times triangles}{number\ of\ triplets}$$

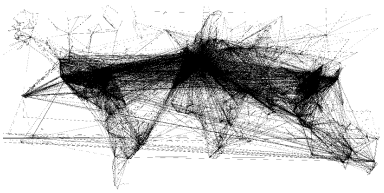
   A *Triplet* (or *triad*) are 3 nodes that are connected by either 2 (open triplet) or 3 (closed triplet) links. A triangle graph therefore includes three closed triplets, one centered on each of the nodes.

   Which network clustering metric (global or average cluster coefficient) gives **more weight to high degree nodes**? Why?

Using Gephi

For this exercise you will be asked to analyze a **flight network** obtained from the **openflights dataset**. Start by downloading a zip file containing the network in gexf format.

http://www.dcc.fc.up.pt/~pribeiro/aulas/arsi2021/homework/flights.zip



Open the network in **Gephi** and use **Geo Layout** to visualize (you may need to install that plugin). The network is directed and corresponds to a multigraph (multiple edges between the same airports, corresponding to different airlines). When loading the network you will be asked about edge merge strategy: not merging will keep the multigraph, merging with "sum" will create a weight attribute corresponding to the number of flights). Different questions might require different merging strategies.

3. Answer to the following questions. For each one give a brief explanation of the steps you took. For top-k like questions you need to show the name and city of the airports and also the respective numbers (ex: number of flights or centrality value). If there are ties, show all tied answers on the k first positions.

   You might need to use various features of Gephi (appearance, layouts, filters and statistics). If needed, you can just reload the network to have a fresh restart.

   (a) What is the number of **airports** and **flights** in the network?

   (b) On average, an airport has how many **outgoing flights**? And to how many **different airports**?

   (c) What is the **diameter** and **average path length** of the network?

   (d) What is the **pair of airports with more flights** between each other (and how many flights)?

   (e) List the top-3 of the airports that have **flights to the highest number of other airports**.

   (f) List the top-3 of the airports with highest normalized **betweenness centrality**.

   (g) Consider **Ted Stevens Anchorage International Airport**. What is its global ranking in terms of betweenness centrality and out-degree? Can you explain the discrepancy? Indicate another airport with the same kind of behavior (high betweenness centrality but relatively low out-degree).

   (h) List the top-3 of **countries with the highest number of airports**.

   (i) List the top-3 of **airlines with the highest number of flights**. *(info about airline codes)*

   (j) What is the number of **domestic flights** inside USA ?

   (k) How many **airports in China** fly to at least 50 other airports?

   (l) How many **flights are there between Brazil and Portugal**?

   (m) Consider a network formed only by **Ryanair flights**. What is the number of nodes and edges of its **giant component**? Considering only this giant component, what is the most important airport in terms of **closeness centrality**?

   (n) How many airport are **reachable** from Francisco de Sá Carneiro Porto Airport in **1 flight**? And in at most **2 flights**? And in at most **3 flights**?

   (o) Create an image showing the **flight network between american and canadian aiports** with more than 100 destinations in the global network. The size of the nodes should reflect the global **betweenness centrality**, and their colors should be different for each **time zone**. Nodes should be labeled with the **city name**. Try to make your image as comprehensible and aesthetically pleasing as possible.

This exercises in this page will involve a series of **small programs that you should code**. You can use any programming language of your choice but you must include any source code you developed (or adapted) both as an appendix of your PDF and as source files in the zip file you submit.

For all the generated **(undirected)** networks you should produce the following (text file) format:

- The first line should contain n, the number of nodes in the network
- The following lines should each contain two integers a b (separated by a single space) indicating there is an edge between nodes a and b (nodes should be integers between 1 and n).

Erdos-Renyi Model

4. Write **code for generating a random network** following the $G_{n,p}$ **Erdős-Rényi model**), that is, a network with n nodes, where each pair of nodes has probability p of being connected.

   Include as attached files, two random networks **random1.txt** and **random2.txt** generated respectively with $n = 2000, p = 0.0001$ and $n = 2000, p = 0.005$ (in the described format).

5. Write **code for computing the size of the giant component** of a given network.

   Use it to compute the size of giant component of the two random networks you generated (random1.txt and random2.txt), and show the obtained results.

   Hint: you can use any graph method traversal to compute the giant component, such as breadth-first search (BFS) or depth-first search (DFS). Take care to implement an $\mathcal{O}(|V| + |E|)$ algorithm that only passes trough each node once.

6. The following is an exercise to study the emergence of a **giant component**.

   Combining your previous code, generate a series of random networks with $n = 2000$ and p varying from 0.0001 to 0.005 (with steps of 0.0001).

   Show a plot of your results, with the X axis representing p and the Y axis representing the size of the giant component.

   Was the plot what you were expecting? What is the shape of it? At what average degree values do you notice something happening?

Barabási-Albert Model

7. Write **code for generating a random network** following the $BA_{n,m_0,m}$ **Barabási-Albert model**). This model uses a preferential attachment mechanism and it works in the following way (**some hints**):

   - You begin with a fully connected network containing m_0 nodes
   - In each iteration (until you reach a total n nodes) you add one new node connected to m existing nodes, with a probability proportional to the number of already existing connections of previous nodes. Formally, the probability p_i that the new node is connected to node i is:
     $p_i = \frac{k_i}{\sum_j k_j}$
     k_i is the degree of node i and the sum is made over the degrees of all existing nodes.

   Include as attached files, two barabási-albert networks **ba1.txt** and **ba2.txt** generated respectively with $n = 2000, m_0 = 3, m = 1$ and $n = 2000, m_0 = 5, m = 2$ (in the described format).

8. The previous process should generate a scale-free network with a power law degree distribution with exponent $\alpha = 3$.

   Plot the degree distribution of both your generated networks using **cumulative binning** (**see slides 102 and 103**) and try to fit with the corresponding power law function (showing it in the plot).
