Data to Network: Facebook friendships

In a previous exercise with movies, we have seen how, from real data, several similarity scores could be designed, and how those scores could affect the resulting network. In this exercise, we will see that, even for a single score, several ways of creating a network can sometimes be meaningful and lead to different networks.

1. Getting familiar with the data
   (a) Go to https://dataforgood.fb.com/docs/social-connectedness-index-methodology/ and read the basic description of the dataset.

   You can find on my website a pre-processed version of the dataset in 2 files:

   1. The connectedness index as a .csv file 2. The barycenter of each country as a .csv file (lat,lon) coordinates

   (b) Load the connectedness index using pandas.

   (c) Check the number of pairs with non-zero values, the number of different countries, and thus compute (manually) the density of the network.

   (d) Load the barycenter using pandas

   (e) Plot the distribution of connectedness scores. Check the highest and lowest values.

2. Modeling choice
   (a) One possibility to create a graph is to filter values lower than a threashold. But what threashold to use? One possibility is to use the distribution of connectedness score to make a choice. Another possibility is to keep only the X largest values, where X is chosen to reach a reasonable average degree. Try this and plot the networks (plotting as a spatial network helps)

   (b) We can observe that many of the largest values are between small countries. As a consequence, by picking only the largest values, we capture only part of the story. Another approach is to go back to the definition of the score in term of probability of observing a frienship. We know that we want to reach a number of edges of approximately X. Can you design a random process such as, for each pair of node, there is an edge with a probabilty which depends on X and on the Social Connectedness index, such that the expected total number of edges is X ?

   (c) A variant of this approach consists in picking exactly X edges at random among all pairs, with a probability which is biased by the connected index. (`np.random.choice`)

3. Summing up
   (a) Make a parallel between those approaches and the two variants of the Erdos Renyi model (n,p), (n,L)

   (b) Compare the network properties of the graphs obtained with the 3 different approaches, and a same number of edges.

4. Sharing
   (a) Share your result and observations with other students.