Task 1

The Glastonbury graph is not a connected graph, to be able to compute the path length we proceeded by extracting the largest connected components from all the graphs: Glastonbury as well as the 3 generated randomly using Erdos Renyi function since they are also not connected graphs.

**Methodology**

We used the functions in Networkx library.

The aim was to construct a new connected graph from the initial graphs.

we started by extracting the largest connected components,then we used the output returned to construct a list of edges of all possible combinations of pair of nodes in that list. Then we sorted the elements inside the tuples representing the edges and checked the existence of the edge in the initial graph before affecting it to a dictionary if is a weighted graph or to a list otherwise.

Once our connected graphs were created, we computed the path length as well as the clustering coefficient.

The Glastonbury connected Graph contains *20901* hashtags and *79974* edges.

**Findings:**

In the illustrated figure, we have two graphs, the left one represents the shortest path length between every two hashtags and Y-axis represent the frequency of those shortest path length.

While the other one shows the clustering coefficient that defines the fraction of possible interconnections between the neighbors of a hashtags

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Fig: The graphs shows the path lengths as well as the clustering coefficients

The average path length of the Glastonbury graph is ***l = 2.34*** and the clustering coefficient is ***C = 0.62***, that means nearly all hashtags are linked within a short number of steps to all other hashtags and that they are densely connected , while the average path length and the clustering coefficient of 3 ER-graphs of the same size as the GlastonBury are ***lrand* = 5.12** and ***Crand* = 3 × 10−**3, ***rand* = 5.12** and ***Crand* = 4× 10−**3, ***rand* = 5.13** and ***Crand* = 4 × 10−**3 respectively.

ER-graphs have small average shortest path however still higher compared to Glastonbury path length which indicates that nearly all hashtags are linked within 5 steps, and I think this is due to the large size of the data and the random graphs probability distribution, on the other hand they have a very small clustering coefficient indicating that hashtags are sparsely connected.

The Networks that meet the criteria of both high clustering and small average path length are referred to as small world networks.

Task3





Fig: The graphs represent only the top 10 % of pageRanks.

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a random hashtag will refer to a particular hashtag.

The average pageranks of glastonBurry is 0.003 while the average pageRanks of the 3 random graphs is approximately 7.45e-05.

As we can observe from the plots, the tail of Glastonbury curve is fatter than the 3 ER-graphs curves which is expected, since Glastonbury is a real world network it contains some hashtags with a very large degree being connected to other hashtags which leads to a higher scores and therefore a higher probability.

Also, in the ER-graphs as we can see nodes with small degree are the most frequent and the ones with high degree are very rare and that leads to small scores

A highly connected node will have a remarkably higher degree and will generate a higher page rank score. We Noticed that the top 10% of our pageRank are the hashtags with highest degree in the Glastonbury connected graph.