Task 1

The GlastonBury graph is not a connected graph, to be able to compute the path length.

Therefore, we proceeded by extracting the largest connected components from all the graphs: Glastonbury as well as the 3 generated randomly using Erdos Renyi

Methodology

We used the functions in Networkx library to extract the largest connected components.

From the output, we constructed a list of edges of all possible combinations of pair of nodes in that list.

After sorting the list, we checked the existence of the edge in the initial graph before affecting it to a dictionary or a list of tuples in order to construct a new graph of the connected components.

Once our graph created, we computed the path length as well as the clustering coefficient.

The new Glastonbury 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 right one shows the clustering coefficient that defines the proportion of a node’s neighbors which are connected by an edge.

|  |  |
| --- | --- |
| GlastonBury\_AD vs Erdos renyi (p=0.00035) |  |
| GlastonBury\_AD vs Erdos renyi (p=0.000366) |  |
| GlastonBury\_AD vs Erdos renyi (p=0.00038) |  |

The average path length of the GlastonBury graphis ***l = 2.34*** and the clustering coefficient is

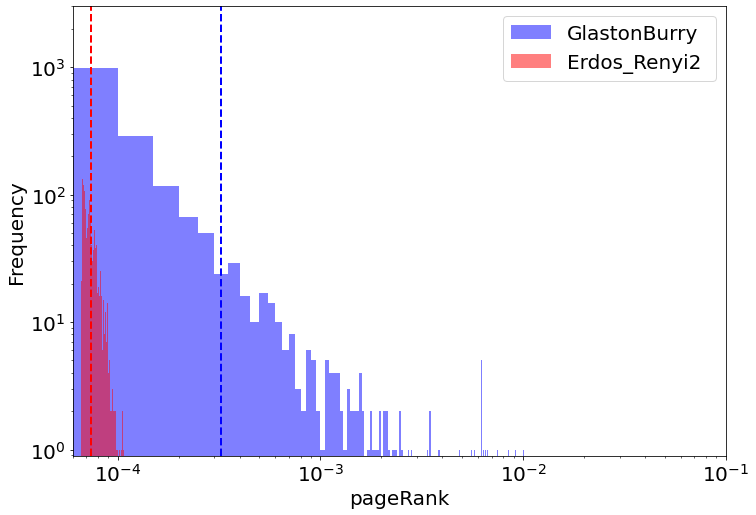
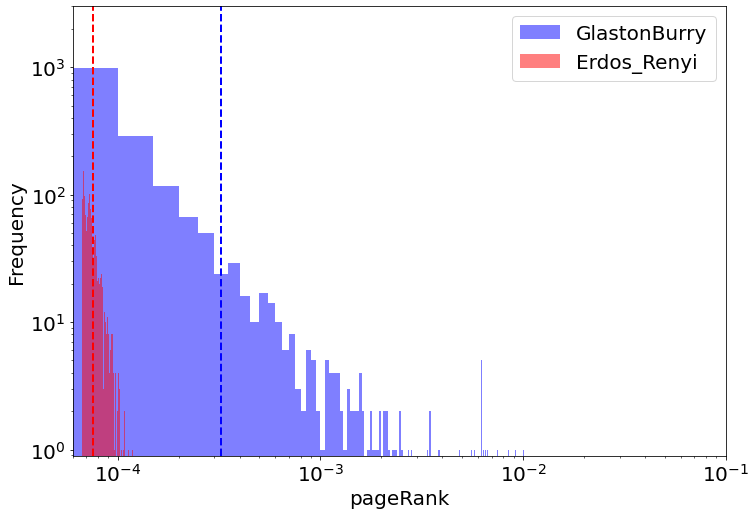
***C = 0.62***, that means nearly all hashtags are linked within a short number of steps to all other hashtags and are densely connected, while the average path length and the clustering coefficient of 3 ER-graphs of the same size and average degree as the GlastonBury are

***lrand* = 5.82** and ***Crand* = 4 × 10−**3, ***rand* = 5.12** and ***Crand* = 3 × 10−**3, ***rand* = 5.21** and ***Crand* = 4 × 10−**3 respectively.

ER-graphs have small average shortest path but higher compared to Glastonbury which states that nearly all nodes are linked within 5 steps and are quite sparsely connected and I think this is due to the large size of the data and their probability distribution, on the other hand they have a very small clustering coefficient.

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

Task3



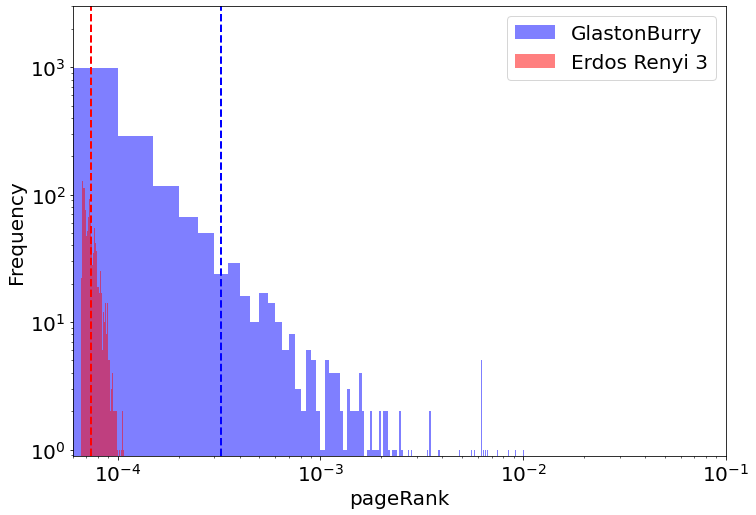


Fig : The graphs represent only the top 10 % of pageRank of Glastonbury and the Erdos-Renyi.

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

In the figure, we can observe that tail of Glastonbury curve is fatter than the 3 ER-graphs curve which is expected, since it some hashtags have a very large degree, being connected to other hashtags and that leads to 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 which leads to small scores.

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