EECS 4415 Project 2 Report

Graph Data Analytics

**Anika Prova - 216474306**

1. **Computing the largest WCC and SCC**

In the initial Web graph from the given data, there are 281903 nodes and 2312497 edges.  
The dangling and isolated nodes are then removed for accurate computation and analysis in the later steps.   
The cleaned-up graph after removing dangling and isolated nodes has 281115 nodes and 2309731 edges.

In the next step, we calculate the largest Weakly Connected Component and the largest Strongly Connected Component in the graph using the NetworkX API. The number of the edges and nodes are as follows:

* Number of nodes in WCC: 254532  
  Number of edges in WCC: 2231860
* Number of nodes in SCC: 150532  
  Number of edges in SCC: 1576314

Then, the PageRank’s for both WCC and SCC are computed using the default parameters in NetworkX’s PageRank API

1. **Random Graph vs Barabasi-Albert graph**

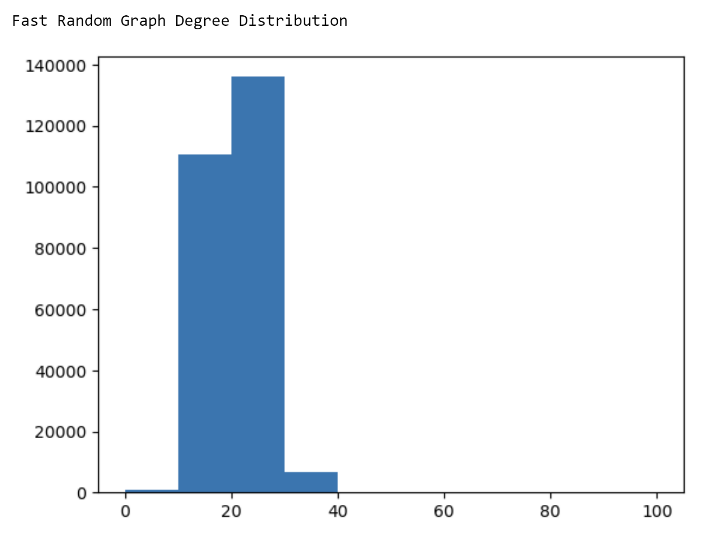
A Fast Random Graph and a Barabasi-Albert graph are generated using the NetworkX API.

The parameters used for generating the random graph are shown in the working code.

When generating the Barabasi-Albert graph, the value of the m parameter is found using trial and error. We do a trial and error using different m values starting from 0 until the number of edges in the Barabasi-Albert graph exceeds the number of edges in the original largest WCC. We found that when the m value is 9, the number of edges in the BA graph is 2290707. Since the number of edges in the original WCC graph is 2231860, the BA graph has 58847 edges in excess. So, we proceed with m = 9

For each graph the following properties are computed - Degree distribution, clustering coefficient and transitivity. These features are some of the empirical evidences of a graph’s feature. Therefore 2 graphs with similar values for any of these features will imply that the graphs are similar as well.

***Chart, histogram

Description automatically generated2.1 Degree Distribution:***

Chart, histogram

Description automatically generated

From the Degree Distribution plots, we can see that the BA graph’s degree distribution is very similar to that of the original largest WCC. However, the degree distribution plot of Fast Random Graph is different than that of WCC.

***2.2 Transitivity:***

The transitivity for each graph was calculated and the following results were obtained:

Transitivity of the largest WCC graph: 0.3547436522786994  
Transitivity of the fast Random Graph: 7.914945270314145e-05  
Transitivity of the BA graph: 0.0006090497087519401

The difference between the transitivity of the largest WCC and the BA graph is smaller (0.35413460257) than the difference between the transitivity of the WCC and the random graph (0.35466450282). Therefore, this property also implies that the BA graph is more similar to the original Web graph.

***2.3 Clustering coefficient:***

Before calculating the clustering coefficient values, the WCC graph was converted to undirected graph for minimizing time of computation. For each graph, the average of the clustering coefficients of the nodes was calculated. So, we ended up with one average clustering coefficient value for each graph as follows:

* Clustering Coefficient of all the nodes in largest WCC: 0.6191262078694576
* Clustering Coefficient of all the nodes in fast Random Graph: 7.825047385982156e-05
* Clustering Coefficient of all the nodes in BA Graph: 0.0007309407107381252

The difference between the clustering coefficient of WCC Graph and BA Graph is: 0.61839526715  
The difference between the clustering coefficient of WCC Graph and Random Graph is:   
0.61904795739

The clustering coefficient of the BA Graph is closer to the WCC graph than the Random Graph’s clustering coefficient. This observation also implies that the BA graph is more similar to the original Web graph.

From analysing the 3 features, we can conclude that the BA graph is more similar to that of the original largest WCC graph and is therefore superior to the Fast Random Graph. *Fast Random Graph is the inferior one.*

1. **PageRank Comparison**

The similarity between the WCC and each of the generated graph was further analysed by computing cosine similarity between the PageRank vectors of the largest WCC graph and each of the generated graph. The nodes for each graph is sorted according to their PageRank values. Then, the cosine similarities are calculated using the following formula:

Similarity = (A.B) / (||A||.||B||) - where A and B are the PageRank vectors of the two graphs being compared.

* Cosine similarity between WCC and the Fast Random Graph is: 0.10621941156062113
* Cosine similarity between WCC and the BA Graph is: 0.6462315815038534

The similarity between the PageRank vectors of the WCC and BA graph is higher than that of the similarity between the WCC and Random Graph. This further confirms our suggestion made in the previous section that the BA graph is superior to the Random Graph because it better approximates the Web Graph structure.

1. **Plotting the Graphs:**

After the graph properties and the page rank vectors were compared, each type of graph was plotted to visualize and support the analysis in the previous steps. Since the number of nodes is too high, we select the top k nodes with the highest PageRank values for each graph.   
  
The K was selected using trial and error. K values of 500, 1000, 2000 and 3000 were tried.

For k = 500, WCC graph was not being represented completely and the nodes in fast Random Graph was also not being represented completely as there were several missing nodes in the middle, and it was hard to get a complete idea of the graph’s form.

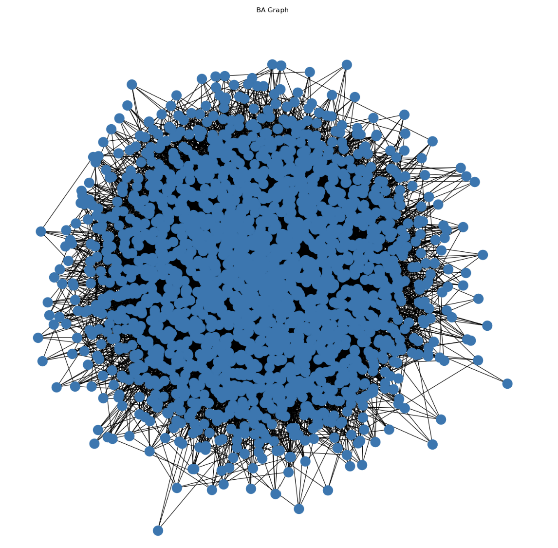
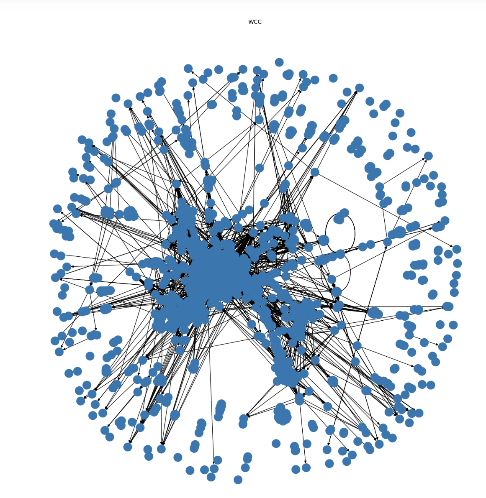
When k = 1000, the fast Random and BA graphs were being completely represented but the WCC graph was missing some important nodes and edges required for complete visualisation.

When k = 2000, all the Graphs are being completely represented. However, the BA graph has some overlapping nodes and edges making the graph slightly clumsy. However, the graph form was clear and easy to interpret the overall appearance of the graph.

When k = 3000, nodes and edges in all the graphs are overlapping, and the graphs did not have a discrete shape or feature.

Therefore, the optimal k value is 2000. When the graphs are visualised with their top k value, we can conclude that the BA graph and the WCC graphs are similar because both the graphs have a cluster of nodes in the centre and a ring of nodes surrounding the cluster at the centre. For the fast Random graph however, there is a dense ring of nodes, and there are multiple nodes evenly scattered inside that ring.

The graph drawing for the 3 graphs using k = 2000 are given below. The other plots for trial and error using k = 500, k = 1000, k = 3000 are shown in the python notebook.

Shape, circle

Description automatically generated

System Specifications

Processor 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz 2.80 GHz  
Installed RAM 16.0 GB (15.7 GB usable)  
System type 64-bit operating system, x64-based processor