MS6032 Networks and Complex Systems

Assignment 1

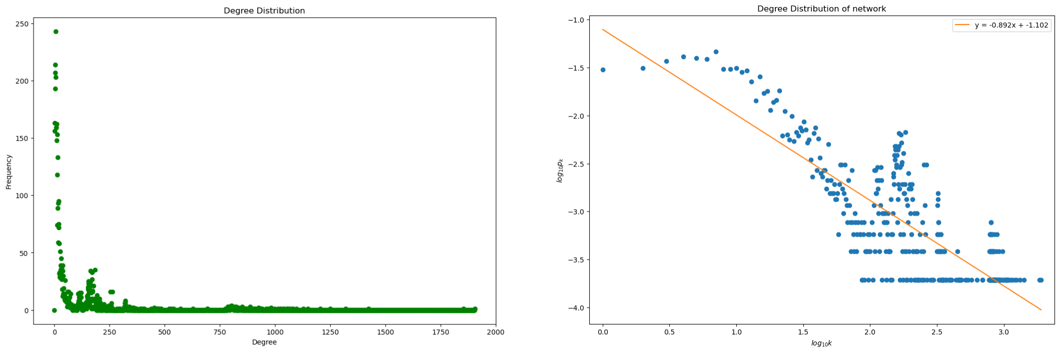
Dara Corr - 22275193

March 2, 2023

The network I chose for this project is an undirected page-page network on the topic of squirrels on the English language version of Wikipedia from December 2018. Nodes represent Wikipedia pages and the edges are mutual links between them. This is a connected network. This network dataset represents a page-page network on a particular topic on Wikipedia, and also gives some insight into the network structure of the page-page network of Wikipedia as a whole.

This network dataset consists of 5201 nodes and 198493 edges. It is a relatively small subset of Wikipedia's 6,625,293 articles and 57,665,461 pages but sufficiently large to get some insight into how pages on a specific topic are related to one another.

Since this is an undirected graph I calculated the mean degree $< k >$ of each node by using $< k >= \frac{2L}{N}$ where L is the total number of links/edges in the network and N is the total number of nodes. I found that the mean degree in this network is $< k >= 76.33$ which means that on average we expect each page to link to about 76 other related pages on this topic on Wikipedia. I also found that the average clustering coefficient is equal to 0.42. This means that 42% of nodes' neighbours are connected to eachother. I calculated this value using python's NetworkX's `average_clustering()` function which returns the average of all the local clustering coefficients in the network. I found that the diameter of this network is 10, meaning that the shortest path distance (number of edges) between the two furthest away points in this network is 10.



normal scale

log scale with line of best fit

Figure 1: Degree Distribution of the Wikipedia Squirrel page-page network on normal and log scales

I investigated the degree distribution of the network by plotting the degree frequency against degree both on a normal scale and on a log-log scale. The resulting plots indicate that the network may follow a power-law model (scale-free network). In this network, the most frequent pages are ones with degrees in the range 1-50. Then nodes(pages) with higher degrees (number of links) are less probable but they are still very numerous. The nodes with highest degrees are called hubs and there seems to be many of them in this network as the bottom right of the plot appears to contain a lot of points and some noise.

I applied linear regression to fit a power law to the loglog plot of $log(p_k)$ vs $log(k)$. The fit gave a slope value of -0.89 when I expected a slope $>$ -2. The line of best fit did not fit the data the way I expected and I believe this is because this network is not well approximated by a power-law model and it is not a scale-free network as a result. It appears like maybe it may be better modelled by a different distribution such as log-normal or stretched exponential distribution.

I estimated the average distance/path length between nodes in the network using $< l >= \frac{logN}{log<k>}$. I found the average path length to be equal to 1.97 $\approx$ 2. This means that on average any page related to the topic of squirrels is 2 page links away from any other page related to the topic of squirrels on Wikipedia. This shows that the small world phenomenon occurs in this network. I also find it interesting that despite there being a total of over 5000 nodes, that the diameter of this network is only 10. The 20 nodes with highest degrees all have degrees of over 1000, making it easy for nodes to travel to each other with small path lengths.

I also found that the average nearest neighbour degree $< k_{nn} >= 308.76 \approx 309$. I obtained this value by computing the nearest neighbour degree for all the nodes using `nx.average_neighbor_degree()`, summing these values and dividing by the total number of nodes. This number is greater than the average degree $< k >= 76.33$. What I observed here is an example of the friendship paradox. Here, I interpreted this as if I were to take a page on a certain topic on Wikipedia and then navigate to a page linked to that page, then I would expect the second page to have more links to other pages than the first page has.

I then looked at degree assortativity of the network. The assortativity of a network is a measure of how nodes tend to attach to nodes of similar degree in the network or not. For a network to be assortative, the hub nodes with large degree tend to have links to other hub nodes in the network. And for disassortative networks, large degree hub nodes tend to link to small degree nodes in the network. The value for assortativity is calculated as a Pearson correlation coefficient and ranges from -1 to +1. by using networkX's `degree_assortativity_coefficient()` function I found that the degree assortativity of this network is -0.227. This means that this is a disassortative network - meaning that hub nodes tend to connect to nodes with smaller degrees. This is consistent with what I expected since technological networks tend to be disassortative.

In conclusion, I saw that the Wikipedia squirrel page-page network appears to have an exponential or log-normal degree distribution. Each page has a average number of $\approx$ 76 links to other pages in this network. This network exhibits the friendship paradox, where on average, a neighbouring page has more links than the first page. There are several hub nodes of large degree which tend to connect to nodes of low degree since it is a disassortative network. Because of this, the average path length is quite low with a value of 2 and a diameter of 10, showing that the topic of squirrels on Wikipedia is a very well connected network.