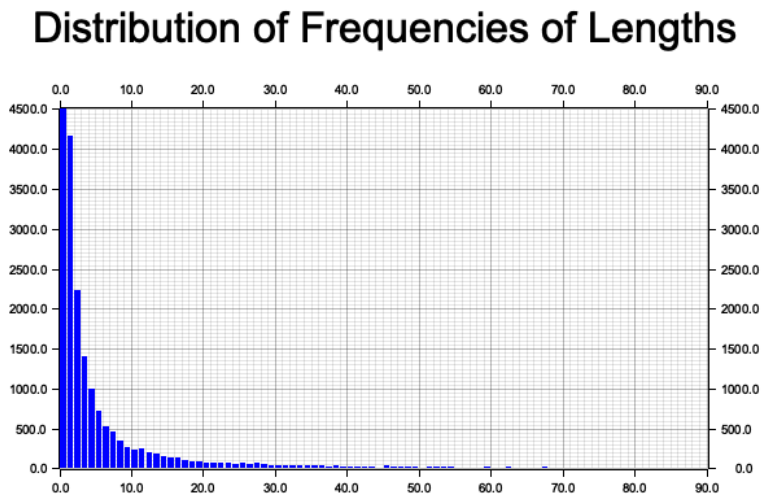
Final Project Report: Degree Distributions of Reddit Network

The dataset I used for the final project is the “Social Network: Reddit Hyperlink Network” dataset from Stanford Network Analysis Platform (SNAP). The goal of the dataset is to show intercommunity interactions across 36,000 communities (subreddits) on Reddit. This dataset represents a directed network, and each edge is a directed connections between two subreddits.

The main problem to be solved in this project is finding degree distributions of the Reddit graph. The columns in the dataset that represent edges are “SOURCE_SUBREDDIT”, the subreddit where the link originates, and “TARGET_SUBREDDIT”, the subreddit where the link ends. Therefore, the dataset can be represented as an adjacent list or an adjacent matrix. For faster performance, I chose to represent the graph using an adjacent list. For example, in the graph, vertice “robmains” has a list of neighboring vertices list [“littlemacmains”, “linkmains”]. To find the degree distributions of this graph, I can find the calculate of the number of vertices that each vertex (SOURCE_SUBREDDIT) is connected to and find the distribution of frequencies of the numbers.



After calculating the length of each vertex’s degree and plotting them frequencies of lengths in descending order, we can see that the graph has a characteristic power-law “long tail,” so we can conclude that the degree distributions follow a power–law distribution.