IR Assignment 3

Abhey Kalia - 2020420
Bhavya Jain - 2020428
Utkarsh Arora - 2020143

Question 1)
For this assignment, we have chosen the soc-RedditHyperlinks dataset. It is a directed graph that represents hyperlinks between two subreddits (communities on reddit).
The network is directed, signed, temporal and attributed. The network is extracted from publicly available Reddit data of 2.5 years from Jan 2014 to April 2017.
A hyperlink originates from a post in the source community and links to a post in the target community. Each hyperlink is annotated with three properties: the timestamp, the sentiment of the source community post towards the target community post, and the text property vector of the source post.

The file is a tsv file (tab separated values), and is opened using the pandas library. For this part, no libraries other than the pandas and matplotlib are used.
This data is then converted into a graph, creating both an adjacency matrix and an edge list.

1) Number of nodes = 35776
2) Number of edges = 137821
3) Average in-degree = 3.852
4) Average out-degree = 3.852
5) Maximum in-degree = 2161
   Node with maximum in-degree = ‘askreddit’
6) Maximum out-degree = 1350

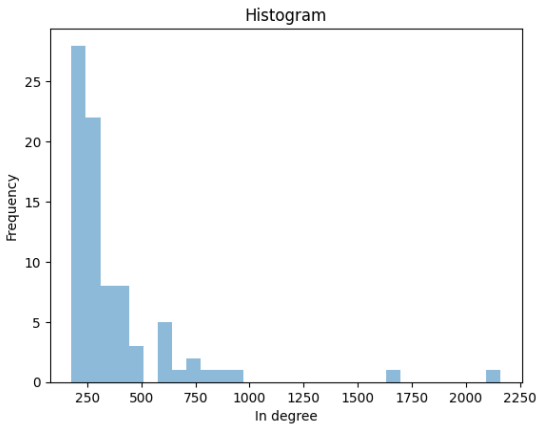
Node with maximum out-degree = ‘subredditdrama’

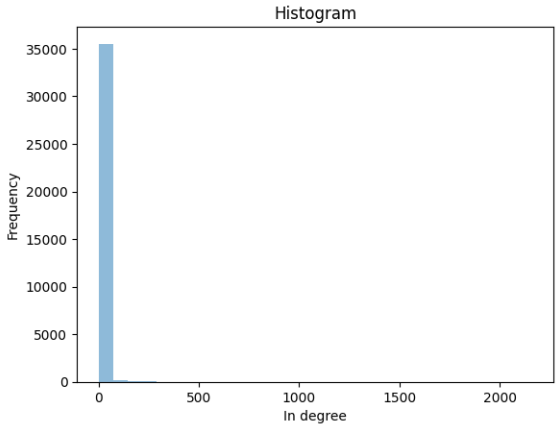
7) Density of the network = 0.00011

In degree is the number of edges directed to the node
Out degree is the number of edges directed from the node

Avg_in_degree = total_incoming_edges / total_nodes
Avg_out_degree = total_outgoing_edegs / total_nodes
Density = total_edges/maximum_possible edges
= total_edges/((total_nodes)*(total_nodes-1))

Degree distribution:





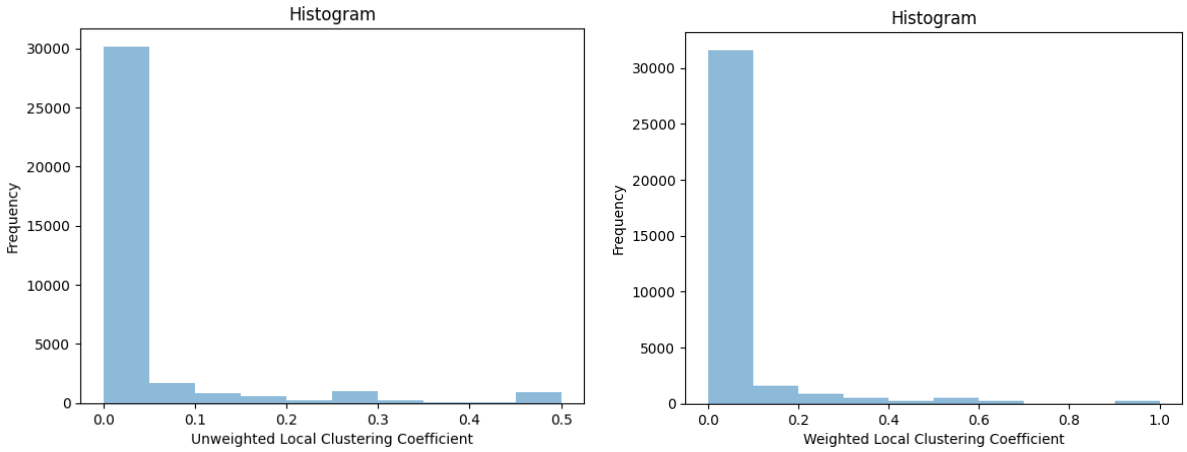
The local clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. It is defined as the proportion of the node's neighbors that are also neighbors of each other.
The neighbors of the nodes are taken in a set, and the density of that set is calculated
Ie (number of edges between those nodes)/(total possible edges between those nodes)
And the total possible edges = (total nodes)*(total nodes - 1)

Hence the LCC = (num_edges)/(num_nodes * (num_nodes-1))

As this is a directed graph, the LCC for in nodes and out nodes is calculated separately. Then these LCCs are averages

Unweighted LCC = (LCC_in + LCC_out)/2

Weighted LCC = ((LCC_in * num_nodes_in) + (LCC_out * num_nodes_out)) / (num_nodes_in + num_nodes_out)



Question 2)

PageRank is an algorithm used to measure the importance of nodes in a network, typically applied to web pages in a hyperlink graph. The basic idea is that a node is more important if it is pointed to by other important nodes. The algorithm assigns a score to each node based on the number and quality of the links pointing to it, as well as the scores of the nodes that are pointing to it. The PageRank score of a node is a measure of its relative importance within the network.

Authority score is a measure of the importance of nodes in a network based on the number and quality of incoming links they receive. Nodes with high authority scores are considered to be experts or authorities on a particular

topic, as they are pointed to by many other nodes. The authority score of a node is calculated as a function of the scores of the nodes that are pointing to it.

Hub score is a measure of the importance of nodes in a network based on the number and quality of outgoing links they have. Nodes with high hub scores are considered to be good at pointing to other nodes that are experts or authorities on a particular topic. The hub score of a node is typically calculated as a function of the scores of the nodes that it is pointing to.

Comparison between Page Rank Score, Authority Score and Hub Score:

	Subreddit	Page Rank Score	Authority Score	Hub Score
0	leagueoflegends	0.003620	1.163430e-03	2.498304e-03
1	theredlion	0.000009	2.271660e-04	5.831212e-05
2	inlandempire	0.000021	1.612344e-05	8.249749e-05
3	nfl	0.001360	3.280045e-04	1.745093e-03
4	playmygame	0.000134	1.694628e-04	4.655423e-05
5	dogemarket	0.000556	1.250484e-04	1.668185e-04
6	locationbot	0.000008	2.093679e-05	-4.413526e-20
7	indiefied	0.000008	8.394780e-05	-1.111755e-21
8	posthardcore	0.000129	2.054244e-05	9.417477e-05
9	gfycat	0.000054	1.144397e-04	8.389728e-05
10	metalcore	0.000241	1.279605e-04	2.458380e-04
11	suicidewatch	0.000285	1.136582e-03	6.733366e-04
12	dogecoin	0.001557	1.364454e-03	1.116510e-03

The correlation between Page Rank Score and Authority Score is 0.44667468261192805

The correlation between Page Rank Score and Hub Score is 0.8485383243837176
The correlation between Authority Score and Hub Score is 0.5551260436796563