IR Assignment-3
Report

Group-4
Team Members:

1. Uttkarsh Singh
2. Darsh Parikh
3. Kartik Jain

Q1.
The dataset we have used consists of only edges where the first node is source and second node is destination.
nodeId.edges : The edges in the ego network for the node 'nodeId'. Edges are undirected for facebook, and directed (a follows b) for twitter and gplus. The 'ego' node does not appear, but it is assumed that they follow every node id that appears in this file.
We have used the twitter data.

```
# create a dictionary that assigns a unique id to each node in the order they occur
def get_node_id_dict(source_list, destination_list):
    node_dict = {}
    for source, destination in zip(source_list, destination_list):
        if source not in node_dict:
            node_dict[source] = len(node_dict)
        if destination not in node_dict:
            node_dict[destination] = len(node_dict)
    node_id_dict = {}
    for node, node_id in node_dict.items():
        node_id_dict[node_id] = node
    return node_id_dict, node_dict

source, destination = get_source_destination('12831.edges')
unique_nodes = count_unique_nodes(source, destination)
node_dict, node_id_dict = get_node_id_dict(source, destination)
print(node_dict,'\n', node_id_dict)
```

```
{0: 398874773, 1: 652193, 2: 18498878, 3: 14749606, 4: 14305022, 5: 8479062, 6: 22253, 7: 12741, 8: 15540222, 9: 14809096, 10: 7415132, 11: 14172562,
 {398874773: 0, 652193: 1, 18498878: 2, 14749606: 3, 14305022: 4, 8479062: 5, 22253: 6, 12741: 7, 15540222: 8, 14809096: 9, 7415132: 10, 14172562: 11
```

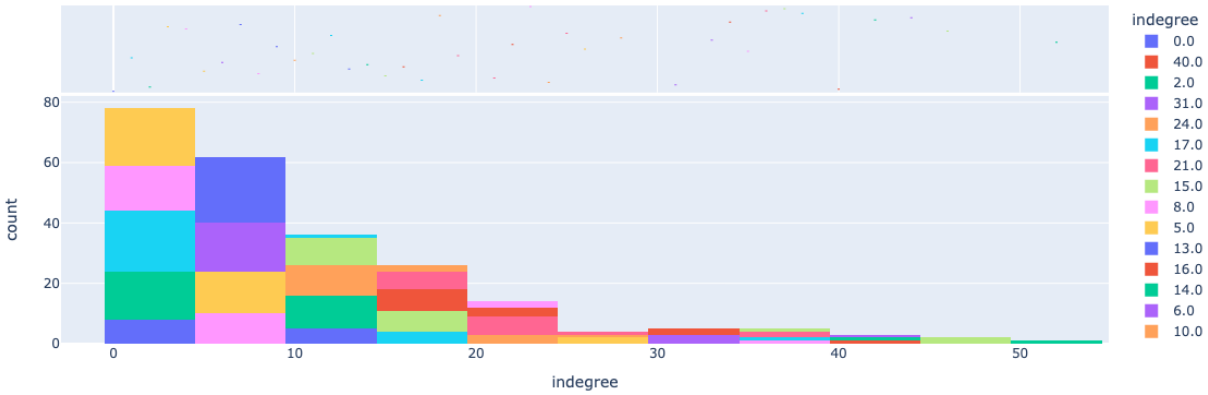
Since the values of nodes were so large, we used a dictionary to map values of nodes starting from 0.

Number of unique nodes: 236
Number of edges: 2478.0
Average indegree: 10.5
Average outdegree: 10.5
Node with max indegree: 180505807

Node with max outdegree: 1186
Density of network 0.04449152542372881

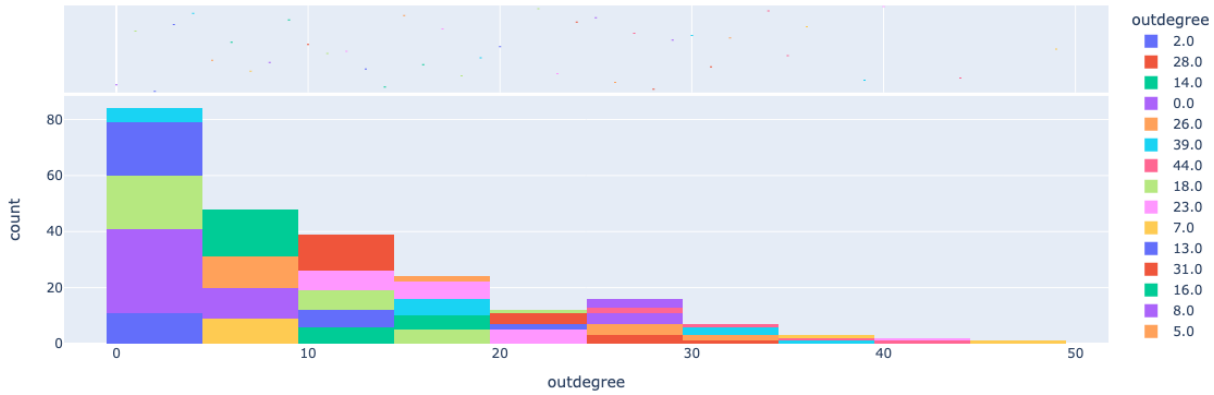
In-degree plot

The indegree of a node in a directed graph is the number of incoming edges to that node. It is a measure of how many other nodes are pointing towards the given node.



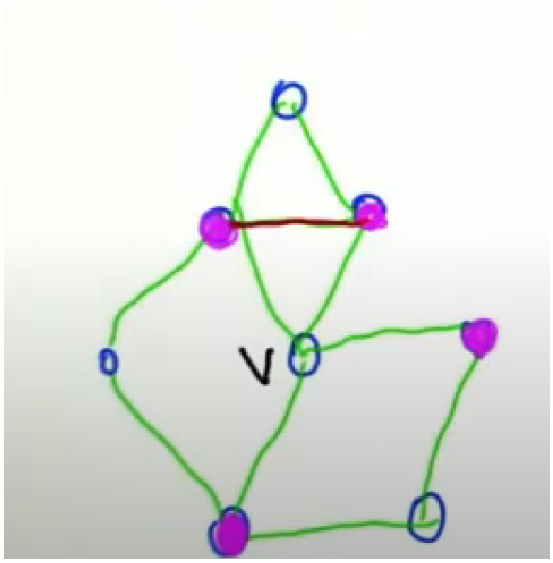
Out-degree plot

The outdegree of a node in a directed graph is the number of outgoing edges from that node. It is a measure of how many other nodes are pointed to by the given node.



The local clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster or form triangles with their neighbors. It is calculated as the number of triangles that include a particular node divided by the maximum number of triangles that could possibly include that node.

We calculated lcc by finding the number of neighbors of a particular node and how many of those neighbor nodes are connected by an edge. Thus giving us the number of triangles.



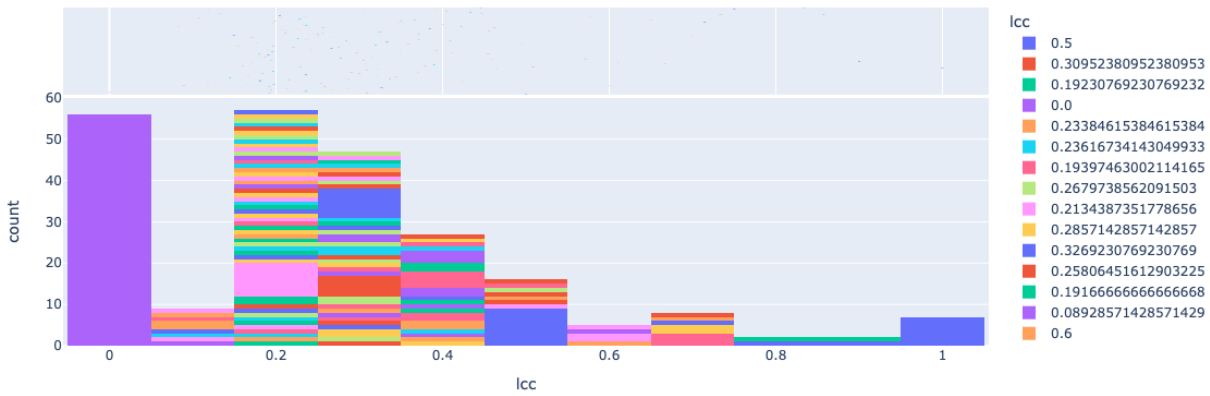
Here pink nodes are neighbors of v and red edge is which connect 2 neighbors.
Formula used

$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}}.$$

Now because our graph is directed to find the number of possible triangles around the vertex we find the number of neighbors (n) and use 2*nC2 to find the number of possible edge connections between the neighbors. And the e be the total real edges between neighbors.
The formula comes out to be.

$$Lcc = e/n(n-1)$$

lcc vs frequency of lcc

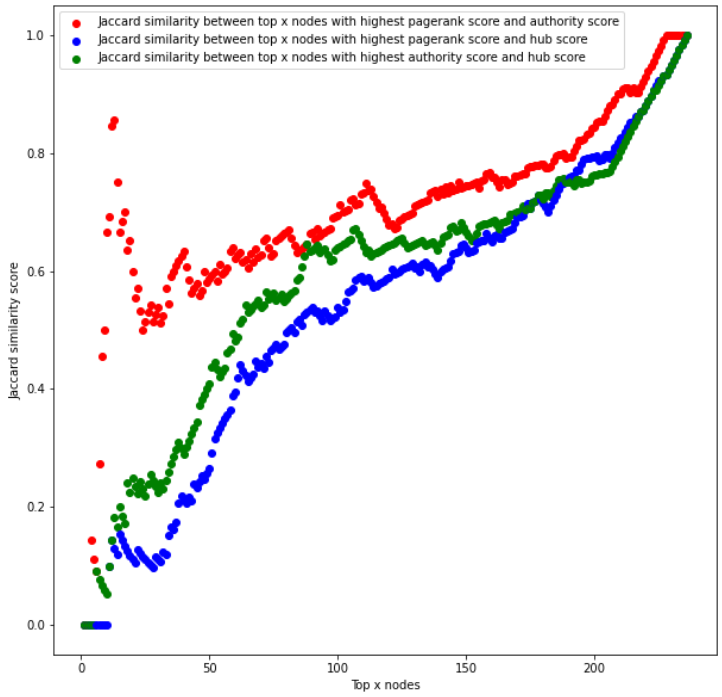


Q2.
Use networkx library to make a graph, plot the graph to visualize it. Networkx has an implementation to get the pagerank score of each node, as well as hub and authority score.
The pagerank score and authority score are based on the incoming links, whereas the hub score is based on the outgoing links.

Comparison:
We used the jaccard coefficient for top x nodes where x ranges from 1 to the number of nodes, this shows how common are the results for the mutual score rankings by authority, pagerank and hub score.

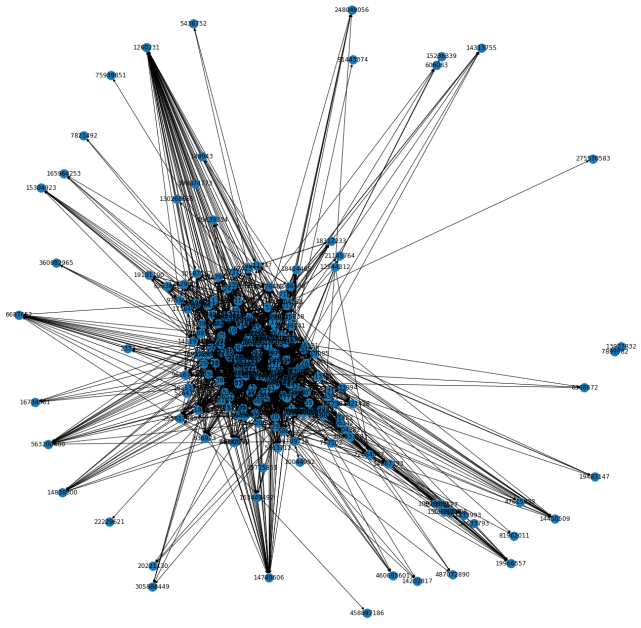
For low x values the pagerank and the authority score give a high level of common top nodes, this is due to the fact of underlying dependence on incoming nodes for pagerank and authority. In contrast to the opposite dependence of hub score on outgoing nodes the jaccard score for the hub vs authority and hub vs pagerank is low for low x values.

Same can be seen using the correlation between the node score for hub, authority and pagerank score.



topx similarity using jaccard

```
Correlation between normalized pagerank score and hub score:  0.4076778442854586
Correlation between normalized pagerank score and authority score:  0.847895690607122
Correlation between normalized hub score and authority score:  0.5742361441822412
```



the graph network.