Assignment-3
IR

1.

Dataset: **Wikipedia vote network.**

Sample of the Edges:

```
Edges:

  [[  30 1412]
  [  30 3352]
  [  30 5254]
  ...
  [8150 8275]
  [8150 8276]
  [8274 8275]]
```

Sample of the Adjacency Matrix:

```
Adjacency Matrix:

  [[0. 1. 1. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]
  ...
  [0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]]
```

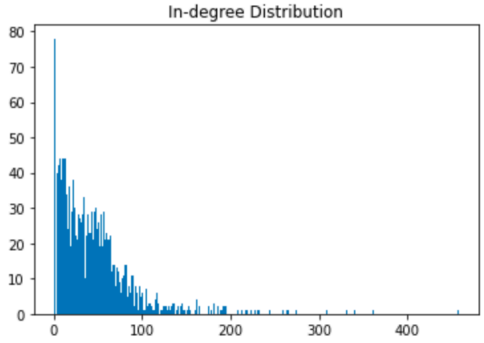
Wikipedia is run entirely by volunteers. Some of them, though, are administrators who assist with upkeep. To become an administrator, a request for administratorship must be submitted and voted on by the community.

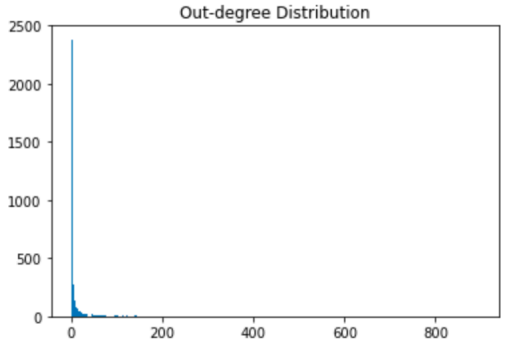
They were able to retrieve all of the admin elections as well as the voting history. There were 2794 elections with a total of 103,663 votes cast and 7,066 users voting.

The nodes are Wikipedia users; directed edges i -> j mean user *i* voted for user *j*.

1. Number of nodes: 7115
2. Number of edges: 103689
3. Average in-degree: 14.573295853829936
4. Average out-degree: 14.573295853829936
5. Node with max In-Degree: 457
6. Node with max Out-Degree: 893
7. Density of the network: 0.0020485375110809584

We plot indegree/outdegree distribution wherein the x-axis is the frequency of in/out degree and y-axis is the frequency of that degree. Below is the in degree distribution and out degree distribution of our graph:

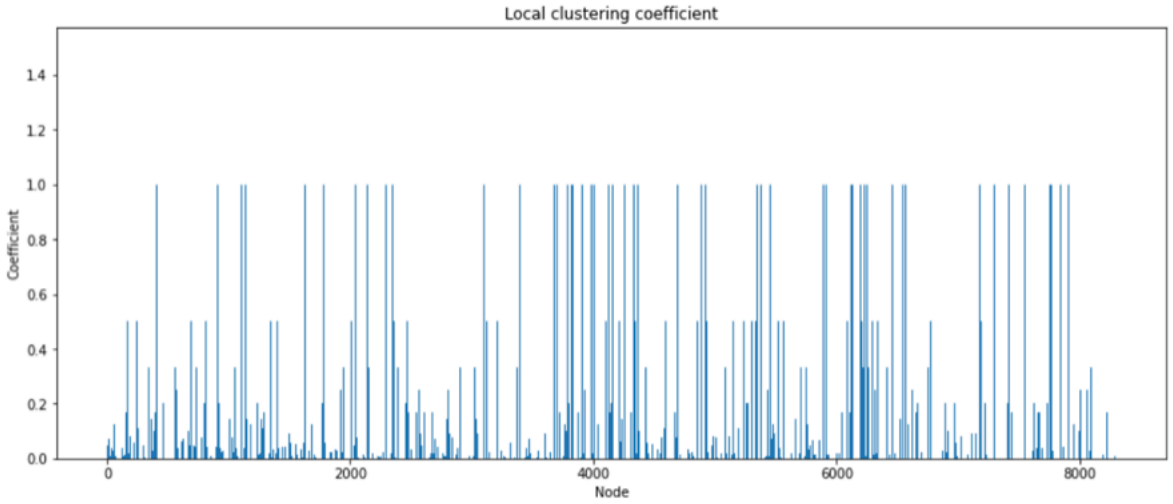




Local Clustering Formula

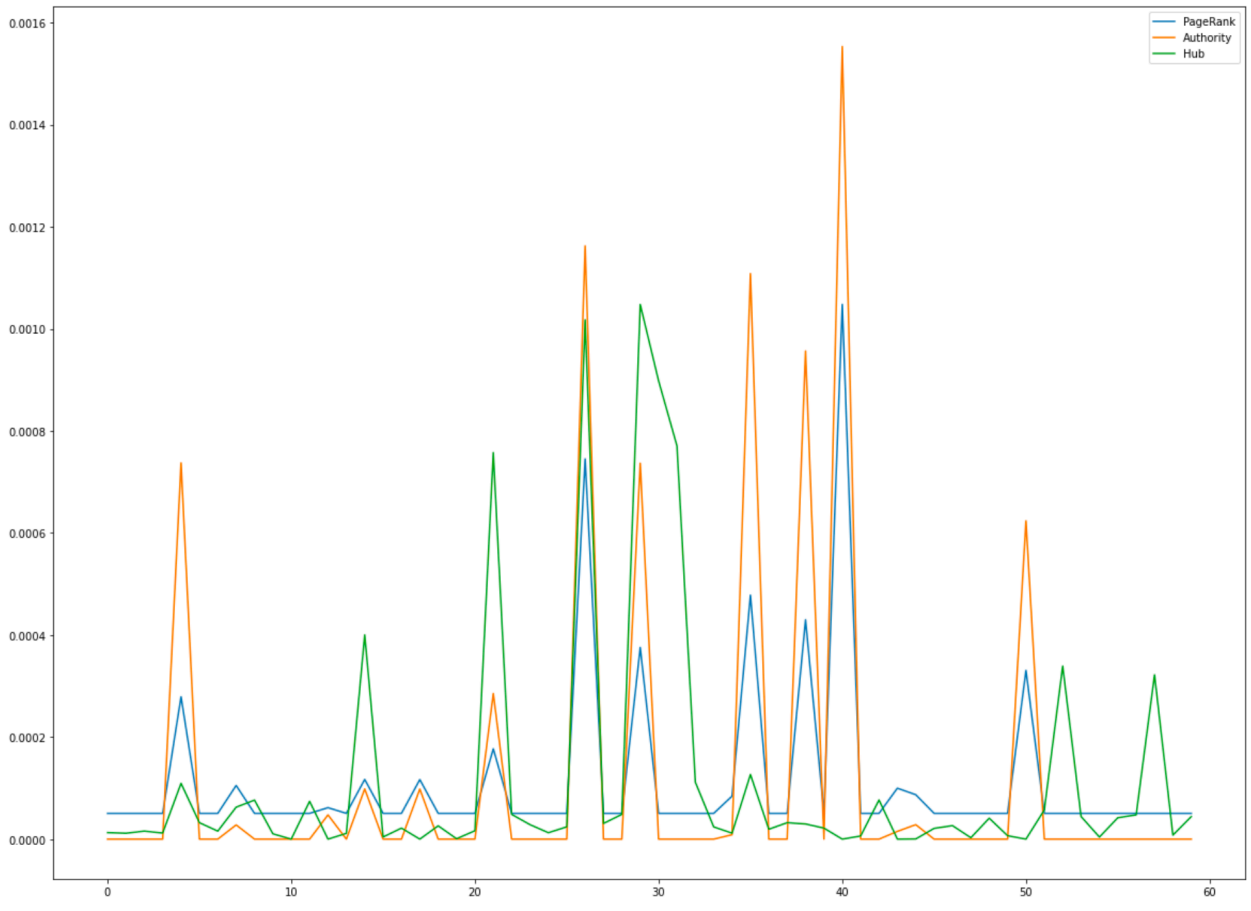
- The local clustering coefficient of a node in a graph is given by dividing the number of links between nodes within its neighborhood by the number of possible links that could exist between them.
- For a directed graph, for each neighborhood, there are K * (K - 1) where K is the number of nodes in its neighborhood.

  LCF = |Links between vertices in neighborhood| / (K * (K - 1))



2.

For this question, we used the networkx library to get the pagerank and hits (hub and authority score) for each node.



```
Top 20 in page rank
 [4037, 15, 6634, 2625, 2398, 2470, 2237, 4191, 7553, 5254, 1186, 2328, 1297, 4335, 7620, 5412, 7632, 4875, 3352, 2654]
Top 20 in Authority
 [2565, 766, 2688, 457, 1166, 1549, 11, 1151, 1374, 1133, 2485, 2972, 3449, 3453, 4967, 3352, 2871, 5524, 3642, 1608]
Top 20 in Hub
 [2398, 4037, 3352, 1549, 762, 3089, 1297, 2565, 15, 2625, 2328, 2066, 4191, 3456, 737, 3537, 2576, 4712, 5412, 2535]
```

- As we can see, there are common spikes between all three metrics, signifying some similarities and relations between them.
- Pagerank and hit scores differ a bit but have more than 5 nodes in common in top 20.
- Pagerank votes vertices 15 and 2565 higher signifying that a lot of important page have their links in them.
- These are voted slightly lower by authority score showing that they point to a lot of pages but there are more pointed at.
- Nodes that may be less scored by page rank are being more ranked by authority and hub scores

Abhinav Suresh Ennazhiyil - 2018003
Shaunak Pal - 2018098