README

1. Link Analysis

Preprocessing steps

- Removing irrelevant contents of the file
- Separating NodeIDs using regex
- Finding the edges: from the separated node id list, a set of two consecutive elements represents an edge from the first element to the second element. i.e. [1,2,3,4] has two edges 1->2 and 3->4 storing node ids in from and to list
- Checking for self-loops

Methodology

Dataset Description:

- Gnutella peer-to-peer network, August 8 2002
- This dataset contains a set of snapshots of the Gnutella peer-to-peer file-sharing network from August 2002. In the graph network, the hosts in the Gnutella network topology are represented by the nodes and the connections between the Gnutella hosts are represented by the edges.
- The network contains 6301 nodes and 20777 edges.
- Creation of Adjacency Matrix
    - Storing all the nodes
    - Checking if node id is a set of consecutive whole numbers starting from 0 so they can be used as labels
    - Initializing the matrix with 0s
    - Filling the adjacency matrix list, if adjmat[a][b]=1, this means a->b is an edge
- Creation of Edge List
    - Adding all nodes to which a node has an edge in a list
    - We do this for all nodes and store the result in a dictionary
- Creation of graph representation
    - Creating dictionary for graph
    - Creating a dictionary corresponding to each node and inserted to the graph
    - Iterating over the list of nodes from which edges are arising and the list of nodes to which edges are directed and marking the connection as present

- Dataset properties:
  The following properties were calculated using the formulas.
    - Number of Nodes
        - Check the size of the adjacency matrix created
    - Number of Edges
        - Iterate over the graph adjacency matrix

    - Count all links
  - Avg In-degree
    - Find the in-degree of each node
    - Calculate the average of in degrees
  - Avg. Out-Degree
    - Find the out-degree of each node
    - Calculate the average of in degrees
  - Node with Max In-degree
    - Find the in-degree of each node
    - Compare the in degrees to find the node with max in degree
  - Node with Max out-degree
    - Find the out-degree of each node
    - Compare the out degrees to find the node with max out-degree
  - The density of the network
    - Formula: count of edges/total possible edges
    - Density = edgeCount / (nodeCount*(nodeCount-1))
- Degree Distribution of the graph
- Local Clustering Coefficient
  - Formula for directed graph: Nv / (Kv * (Kv - 1)) where V is a vertex, Nv denotes the number of links between neighbours of V and Kv denotes the number of neighbours of V
  - Iterating over each node in the graph, we find the number of neighbours using a modified adjacency matrix that considers all incoming and outgoing edges.
  - Traversing the adjacency matrix by looping over the neighbours of each node, we calculate the number of links between the neighbours of each node.
  - We calculate Local Clustering Coefficient using the formula and the values calculated above.

2. PageRank, Hubs and Authority

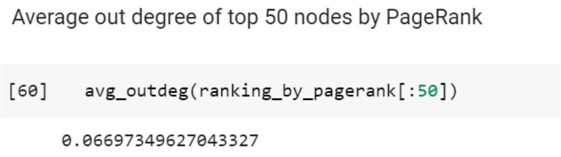
Methodology

- Iterating over the list of nodes with outgoing edges
- Adding edges for each pair to the graph structure using networkx
- Calculating PageRank using networkx.pagerank() function
- Calculating the hub score and authority score using networkx.hits() function

Comparison of Results from PageRank, Authority Score and Hub Score

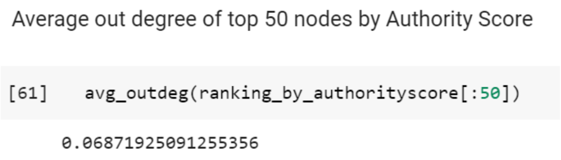
- PageRank, Authority Score and Hub Score are measures of the quality of nodes in a network based on the incoming and outgoing links.
- PageRank computes a ranking of nodes in the graph based on the structure of the incoming links, while the HITS algorithm computes the authority score for a node based on the incoming links and the hub score based on outgoing links.

- Good hub pages are defined as pages that point to many authoritative pages for a topic while a good authoritative page is pointed to by many good hubs for that topic.
- Therefore, higher hub score denotes more outgoing links while higher authority score and page rank denote higher incoming links.
- We can observe this in the plots and results below:
  - Average out-degree of top 50 nodes by Hub Score is greater than for Authority Score and PageRank while it is the opposite for in-degree:



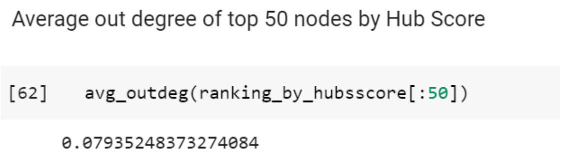
Average out degree of top 50 nodes by PageRank

```
[60]    avg_outdeg(ranking_by_pagerank[:50])

     0.06697349627043327
```



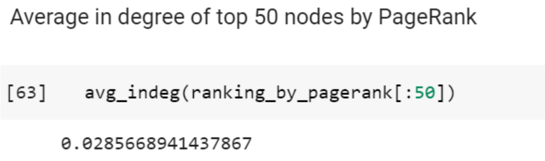
Average out degree of top 50 nodes by Authority Score

```
[61]    avg_outdeg(ranking_by_authorityscore[:50])

     0.06871925091255356
```



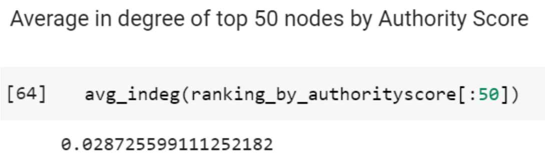
Average out degree of top 50 nodes by Hub Score

```
[62]    avg_outdeg(ranking_by_hubsscore[:50])

     0.07935248373274084
```



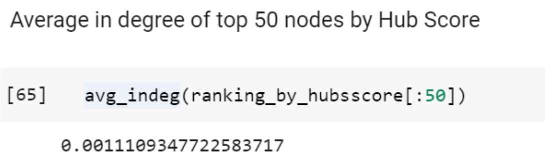
Average in degree of top 50 nodes by PageRank

```
[63]    avg_indeg(ranking_by_pagerank[:50])

     0.0285668941437867
```



Average in degree of top 50 nodes by Authority Score

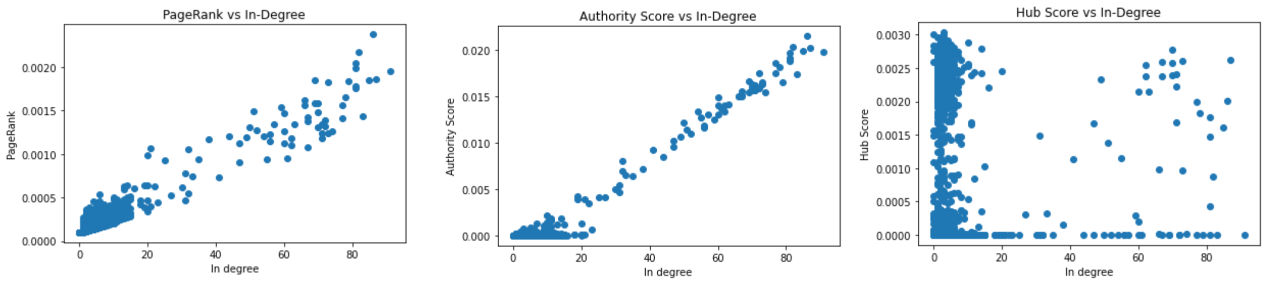
```
[64]    avg_indeg(ranking_by_authorityscore[:50])

     0.028725599111252182
```



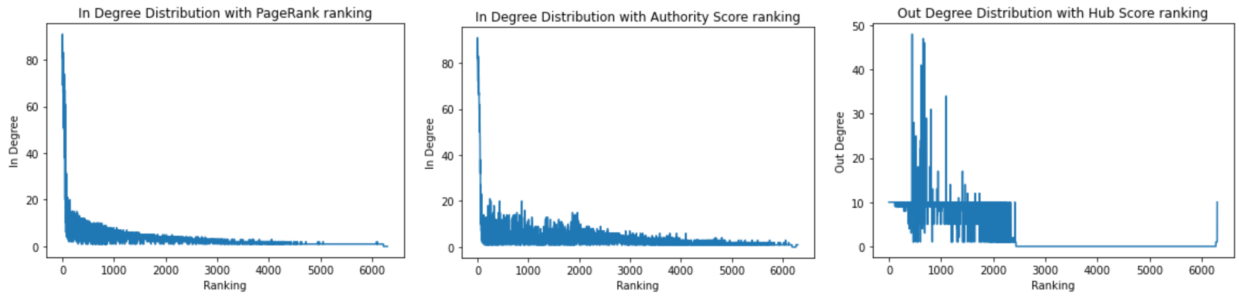
Average in degree of top 50 nodes by Hub Score

```
[65]    avg_indeg(ranking_by_hubsscore[:50])

     0.0011109347722583717
```

  - Plots against in degree showing increase of in degree with PageRank and Authority Score, while there is no such trend for Hub Score.



  - Plots showing the distribution of in/out-degree with the score:


In Degree Distribution with PageRank ranking
In Degree
Ranking
In Degree Distribution with Authority Score ranking
In Degree
Ranking
Out Degree Distribution with Hub Score ranking
Out Degree
Ranking