CSE508 IR ASSIGNMENT – 2

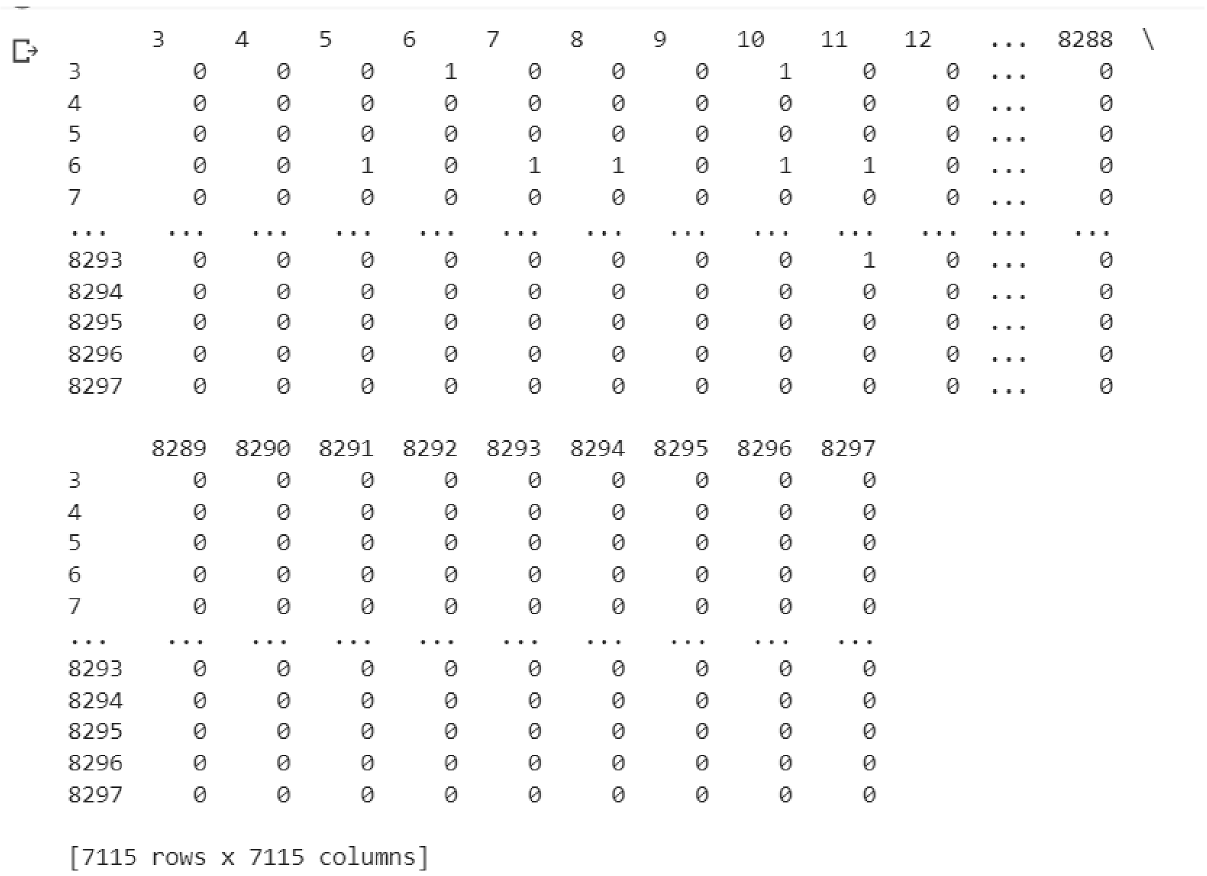
Mohammed Taasir Fruitwala MT22029
Pranshu Patel MT22117
Amey Pawar MT22010

Question 1:

Pick a real-world directed network dataset (with number of nodes > 100) from here. [2points] Represent the network in terms of its ‘adjacency matrix’ as well as ‘edge list’.

We have chosen ‘wiki-Vote.txt’.

Adjacency matrix

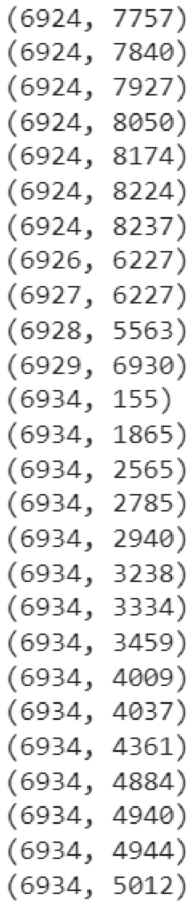


```
        3     4     5     6     7     8     9     10    11    12    ...  8288  \
3       0     0     0     1     0     0     0     1     0     0   ...     0
4       0     0     0     0     0     0     0     0     0     0   ...     0
5       0     0     0     0     0     0     0     0     0     0   ...     0
6       0     0     1     0     1     1     0     1     1     0   ...     0
7       0     0     0     0     0     0     0     0     0     0   ...     0
...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...
8293    0     0     0     0     0     0     0     0     1     0   ...     0
8294    0     0     0     0     0     0     0     0     0     0   ...     0
8295    0     0     0     0     0     0     0     0     0     0   ...     0
8296    0     0     0     0     0     0     0     0     0     0   ...     0
8297    0     0     0     0     0     0     0     0     0     0   ...     0

      8289  8290  8291  8292  8293  8294  8295  8296  8297
3       0     0     0     0     0     0     0     0     0
4       0     0     0     0     0     0     0     0     0
5       0     0     0     0     0     0     0     0     0
6       0     0     0     0     0     0     0     0     0
7       0     0     0     0     0     0     0     0     0
...   ...   ...   ...   ...   ...   ...   ...   ...   ...
8293    0     0     0     0     0     0     0     0     0
8294    0     0     0     0     0     0     0     0     0
8295    0     0     0     0     0     0     0     0     0
8296    0     0     0     0     0     0     0     0     0
8297    0     0     0     0     0     0     0     0     0

[7115 rows x 7115 columns]
```

Edge List



```
(6924, 7757)
(6924, 7840)
(6924, 7927)
(6924, 8050)
(6924, 8174)
(6924, 8224)
(6924, 8237)
(6926, 6227)
(6927, 6227)
(6928, 5563)
(6929, 6930)
(6934, 155)
(6934, 1865)
(6934, 2565)
(6934, 2785)
(6934, 2940)
(6934, 3238)
(6934, 3334)
(6934, 3459)
(6934, 4009)
(6934, 4037)
(6934, 4361)
(6934, 4884)
(6934, 4940)
(6934, 4944)
(6934, 5012)
```

(i) Briefly describe the dataset chosen and report the following:

1. Number of Nodes

2. Number of Edges

3. Avg In-degree

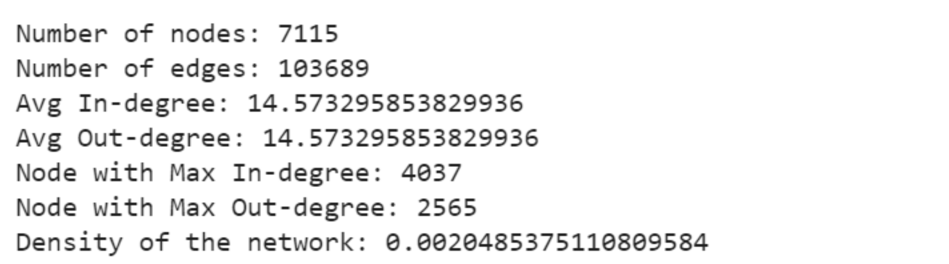
4. Avg. Out-Degree

5. Node with Max In-degree

6. Node with Max out-degree

7. The density of the network

Answer:



```
Number of nodes: 7115
Number of edges: 103689
Avg In-degree: 14.573295853829936
Avg Out-degree: 14.573295853829936
Node with Max In-degree: 4037
Node with Max Out-degree: 2565
Density of the network: 0.0020485375110809584
```

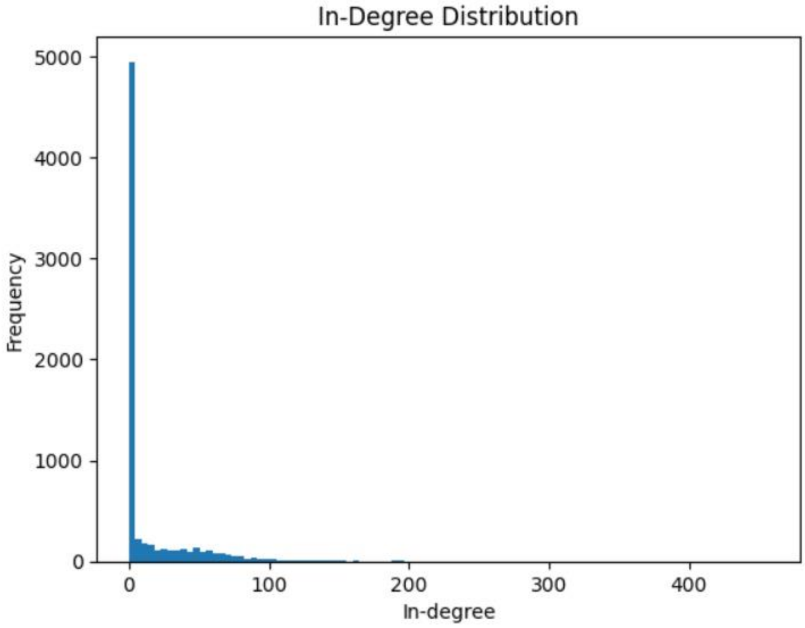
Node with max indegree: 457

Node with max outdegree: 893

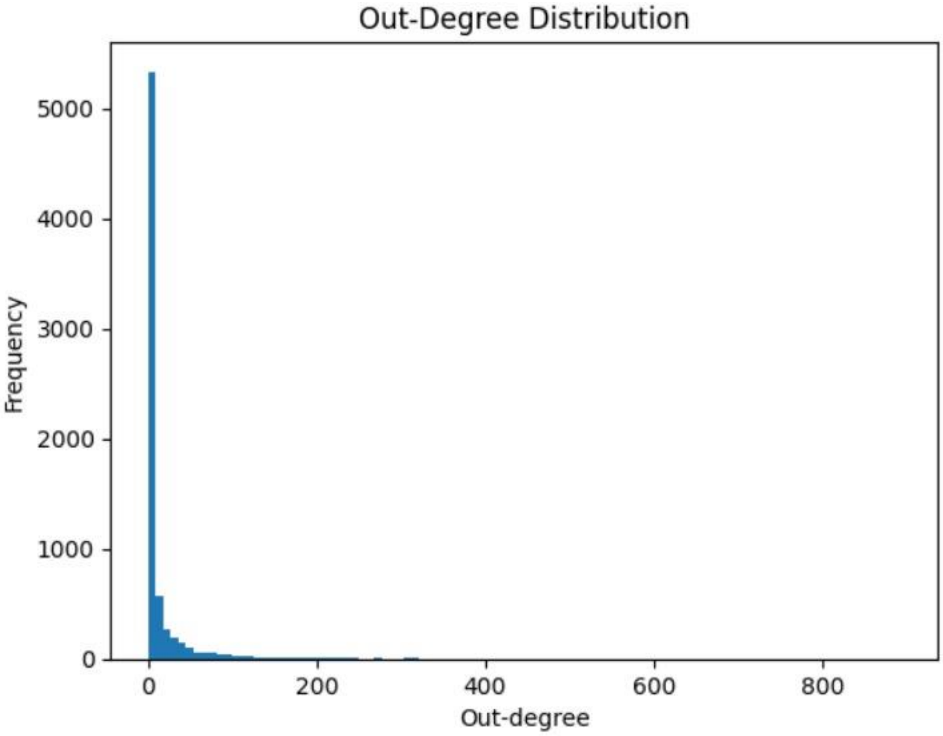
(ii) Further, perform the following tasks:
1. Plot degree distribution of the network (in case of a directed graph, plot in-degree and out-degree separately).
2. Calculate the local clustering coefficient of each node and plot the clustering-coefficient distribution (lcc vs frequency of lcc) of the network.

1. Degree Distribution of Network

a. In-Degree Distribution: Fraction of nodes with in-degree k. If there are N nodes in a network and nk of them have in-degree k then ⍰(⍰) = n_k/N.

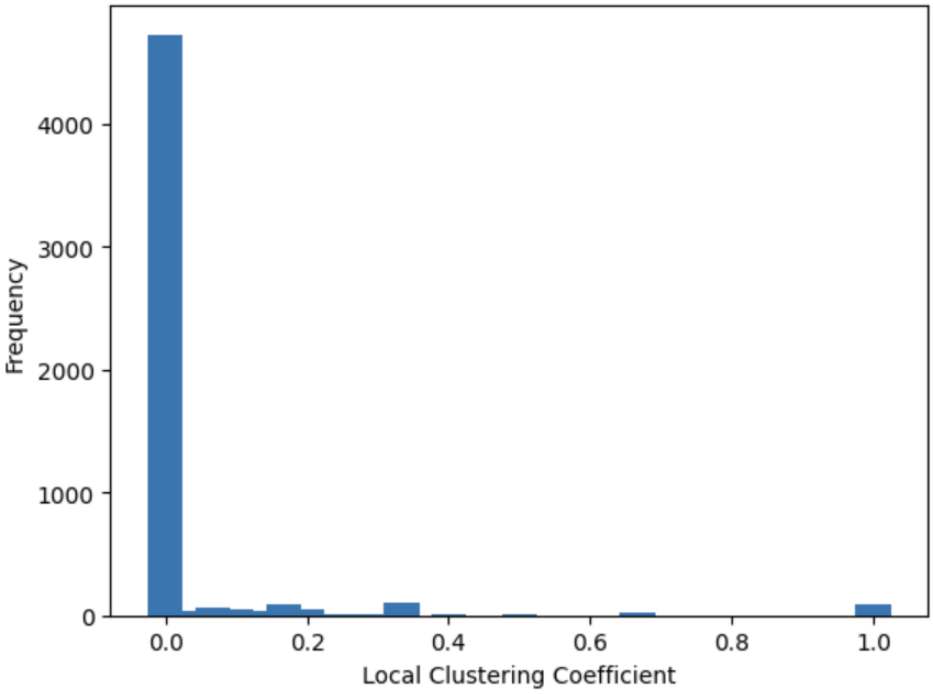


b. Out-Degree Distribution: Fraction of nodes with out-degree k. If there are N nodes in a network and nk of them have out-degree k then $P(k)$ = n_k/N.



The image part with relationship ID rId12 was not found in the file.

c. Local Clustering Coefficient Distribution: Considering undirected version of network, local clustering coefficient of a node quantifies how close its neighbours are to being a clique (complete graph).



Question 2 - PageRank, Hubs and Authority

For the dataset chosen in the above question, calculate the following:
1. PageRank score for each node
2. Authority and Hub score for each node
Compare the results obtained from both the algorithms in parts 1 and 2 based on the node scores.

PageRank(PR) Algorithm It is a link analysis algorithm used by Google Search to rank web pages.

According to Google PageRank works by counting the number and quality of links to a page to determine an estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

The PageRank computes a ranking of nodes in the graph based on the structure of the incoming links.

Algorithm It outputs a probability distribution used to represent the likelihood when a user clicks on links that will take him to any page on the website.

Algorithm Steps
● Initialize the PageRank of every node with a value of 1.
● For each iteration, update The PageRank of every node in the graph.
● The new PageRank is the sum of the proportional rank of all of its parents.
● PageRank value will converge after several iterations.

Damping Factor When a user is clicking on the links and goes to a different page on the website, he will ultimately stop clicking on the links. The probability that a user will continue clicking at any point is named as damping factor d.

PageRank Equation

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where p1, p2, ..., pN are the pages, M(pi) is the set of pages that link to pi,L(pj) is the number of outbound links on page pj , N is the total number of pages and d is the damping factor.Here, the damping factor d is subtracted from 1 and divided by the total number of documents N in the dataset collection and this term is added to the sum of the incoming PageRank scores.

```
Nodes according to PageRank score:
4037 0.004612715891167545
15 0.0036812207295292714
6634 0.003524813657640258
2625 0.0032863743692308997
2398 0.002605333171725021
```

```
2470 0.0025301053283849502
2237 0.002504703800483991
4191 0.0022662633042363433
7553 0.0021701850491959583
5254 0.0021500675059293226
1186 0.0020438936876029136
2328 0.0020416288860889173
1297 0.001951860821612229
4335 0.0019353014475784864
```

HITS(Hyperlink-Induced Topic Search) Algorithm The algorithm being referred to is a link analysis algorithm that calculates the hub score based on outgoing links and the authority score based on incoming links. It is commonly known as the hubs and authorities algorithm and is primarily used to rank web pages based on relevance to a particular search query. The authority score reflects the value of the content on a page, whereas the hub score reflects the value of its links to other pages.

In this algorithm, relevant web pages are identified through a query search. The scores for Authority and Hub are mutually recursive. Authority score is computed as the sum of hub scores from web pages that point to a particular web page (i.e., incoming links). Hub score is calculated as the sum of authority scores from web pages that are pointed by the particular web page (i.e., outgoing links).

Algorithm Steps:

- **Initialize the hub and authority of each node to 1.**
- **Update the hub and authority of every node in the graph for each iteration.**
- **Calculate the new authority as the sum of the hub of its parents.**
- **Calculate the new hub as the sum of the authority of its children.**
- **Normalize the new authority and hub values.**

Differences between PageRank and HITS:

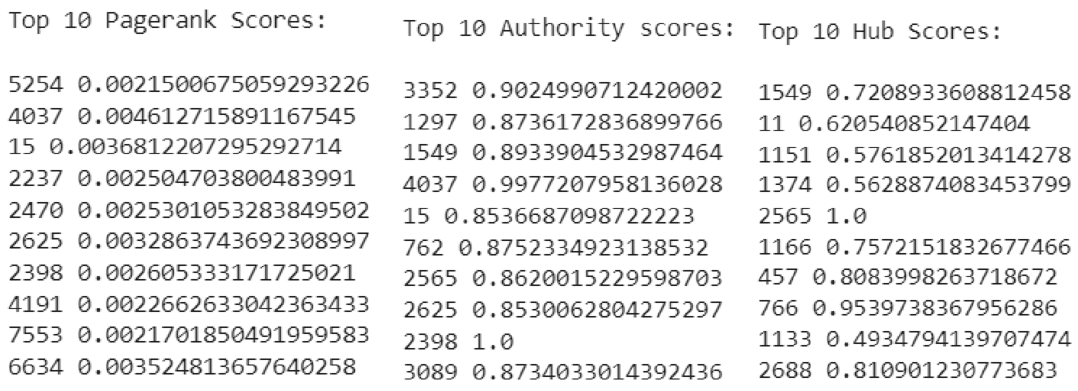
- **PageRank computes the ranks based on the relative importance of the websites, while HITS calculates the weights based on the HUBS and Authorities values.**
- **HITS algorithm is query-dependent, meaning it is performed on a subset of relevant documents, whereas PageRank is query-independent and processed on the entire dataset.**
- **HITS algorithm considers both the hubs and authorities, while PageRank only considers the authority of a website.**
- **HITS algorithm is suitable for smaller datasets, while PageRank is more efficient for larger datasets.**

```
Node Number  :  HUB Score
30  :     0.00998179932694693
1412  :     0.0
```

```
3352  :      0.42573918623360957
5254  :      0.04750055792326323
5543  :      0.17590560962380986
7478  :      0.0
3  :      0.00508778113384111
28  :      0.045127947887486315
39  :      0.013485426941127372
54  :      0.003195859318214718
108  :      0.00032640956457402566
152  :      0.007575360797951532
178  :      0.05503223958138495
182  :      0.0840078883781553
214  :      0.0
271  :      0.0
286  :      0.0
300  :      0.0
```

```
Node Number  :  Authority Score
30  :          0.03707041191889022
1412  :          0.04735802530176851
3352  :          0.9024990712420002
5254  :          0.7075491553162044
5543  :          0.4981085394963819
7478  :          0.295706551449484
3  :          0.03706006574782208
28  :          0.09927335397023307
39  :          0.023934452266701815
54  :          0.054655751098838704
108  :          0.0018980375662156956
```

Comparison between Pagerank, Authority and Hub Scores



```
Top 10 Pagerank Scores:          Top 10 Authority scores:  Top 10 Hub Scores:

5254 0.0021500675059293226       3352 0.9024990712420002   1549 0.7208933608812458
4037 0.004612715891167545        1297 0.8736172836899766   11 0.620540852147404
15 0.0036812207295292714         1549 0.8933904532987464   1151 0.5761852013414278
2237 0.002504703800483991        4037 0.9977207958136028   1374 0.5628874083453799
2470 0.0025301053283849502       15 0.8536687098722223     2565 1.0
2625 0.0032863743692308997       762 0.8752334923138532    1166 0.7572151832677466
2398 0.002605333171725021        2565 0.8620015229598703   457 0.8083998263718672
4191 0.0022662633042363433       2625 0.8530062804275297   766 0.9539738367956286
7553 0.0021701850491959583       2398 1.0                  1133 0.4934794139707474
6634 0.003524813657640258        3089 0.8734033014392436   2688 0.810901230773683
```

The comparison between the top 10 scores obtained from PageRank and HITS algorithms shows that only one node (4037) appears in both rankings. The significant difference in the results can be attributed to the fundamental difference in the way these algorithms compute scores. PageRank relies on incoming links to rank nodes, whereas HITS algorithm calculates the ranking based on the mutually recursive relationship between hubs and authorities. As a result, the scores produced by HITS algorithm are not bounded and are more influenced by the number of nodes with outgoing links, which tend to be higher. On the other hand, PageRank algorithm distributes the score of a node equally among all its outgoing links. Therefore, a node with a high number of incoming links will have a higher score in PageRank.