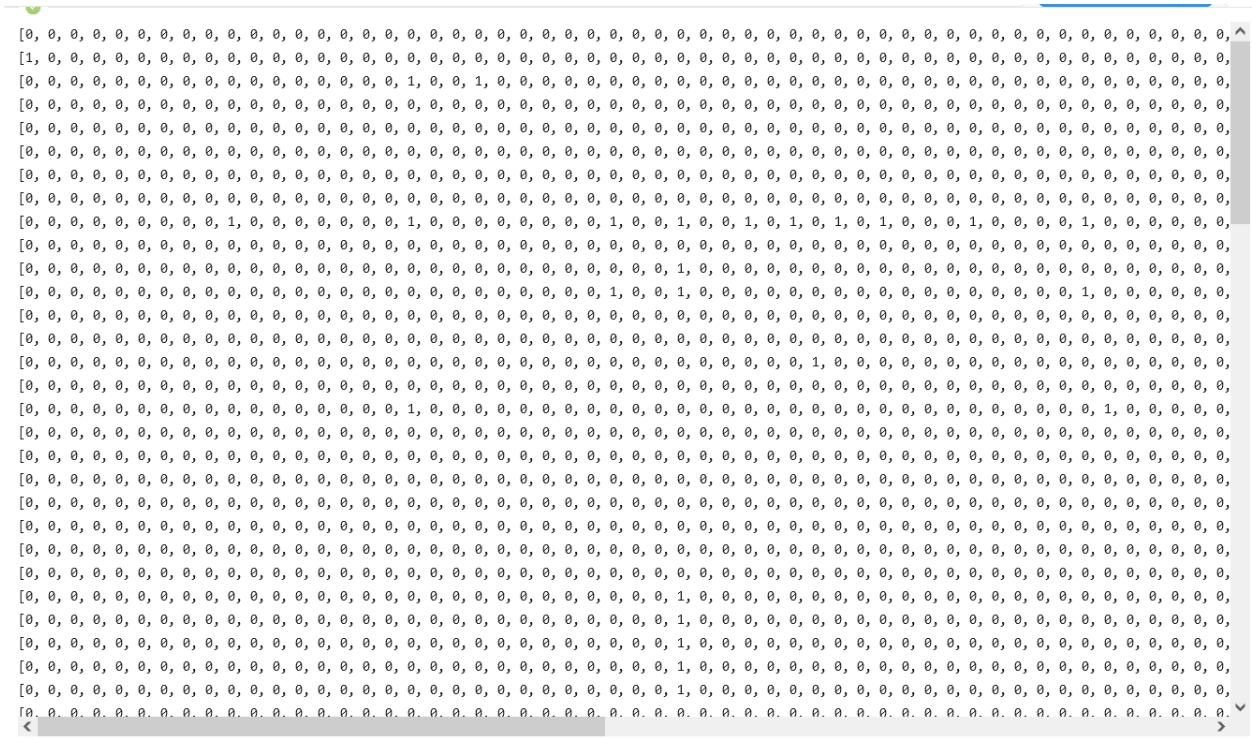
Information Retrieval - Assignment 3

Rishi Raj 2019090
Robin Garg 2019092

Q1 Network in form of adjacency matrix (small sample)



Network in form of edge list

7115 lines (7115 sloc) | 844 KB

```
30->['1412', '3352', '5254', '5543', '7478']
1412->[]
3352->['72', '285', '417', '465', '608', '762', '825', '827', '974', '1140', '1166', '1185', '1239', '1247', '1297',
5254->['840', '896', '2160', '2398', '2689', '3089', '4191', '4361', '4536', '4713', '4717', '5178', '5321', '5412',
5543->['15', '737', '1159', '1549', '2276', '2507', '2511', '2535', '2565', '2589', '2658', '2765', '2774', '2851',
7478->[]
3->['28', '30', '39', '54', '108', '152', '178', '182', '214', '271', '286', '300', '348', '349', '371', '567', '581'
28->['3', '6', '8', '19', '29', '33', '34', '35', '38', '49', '50', '54', '55', '56', '61', '75', '80', '86', '94',
39->['6', '8', '23', '28', '29', '33', '35', '50', '54', '80', '127', '135', '147', '152', '163', '168', '182', '216'
54->['3', '28', '214', '299', '356', '373', '394']
108->['95', '8283']
152->['28', '35', '50', '55', '56', '75', '182', '192', '214', '252', '514', '567', '581', '586', '589', '656', '665'
178->['3', '72', '182', '183', '228', '271', '282', '313', '371', '396', '405', '429', '560', '564', '590', '600', '6
182->['56', '86', '183', '204', '243', '271', '299', '310', '405', '407', '427', '432', '590', '600', '644', '647',
214->[]
271->[]
286->[]
300->[]
348->['86', '271', '325', '439', '665', '686', '697', '704', '769', '817', '878', '941', '981', '1022', '1128', '1156
349->['613']
371->['15', '29', '56', '72', '86', '204', '228', '255', '271', '282', '299', '304', '310', '313', '317', '373', '432
567->['371', '584']
```

1. Number of nodes
2. Number of edges

```
Number of nodes = 7115
Number of edges = 103689
```

3. Avg In-degree = $\frac{Total\ number\ of\ incoming\ edges}{Total\ number\ of\ nodes}$

```
Average in-degree = 14.573295853829936
```

4. Avg. Out-Degree = $\frac{Total\ number\ of\ outgoing\ edges}{Total\ number\ of\ nodes}$

```
Average out-degree = 14.573295853829936
```

5. Node with Max In-degree

```
Node with max in-degree - 4037
In-degree = 457
```

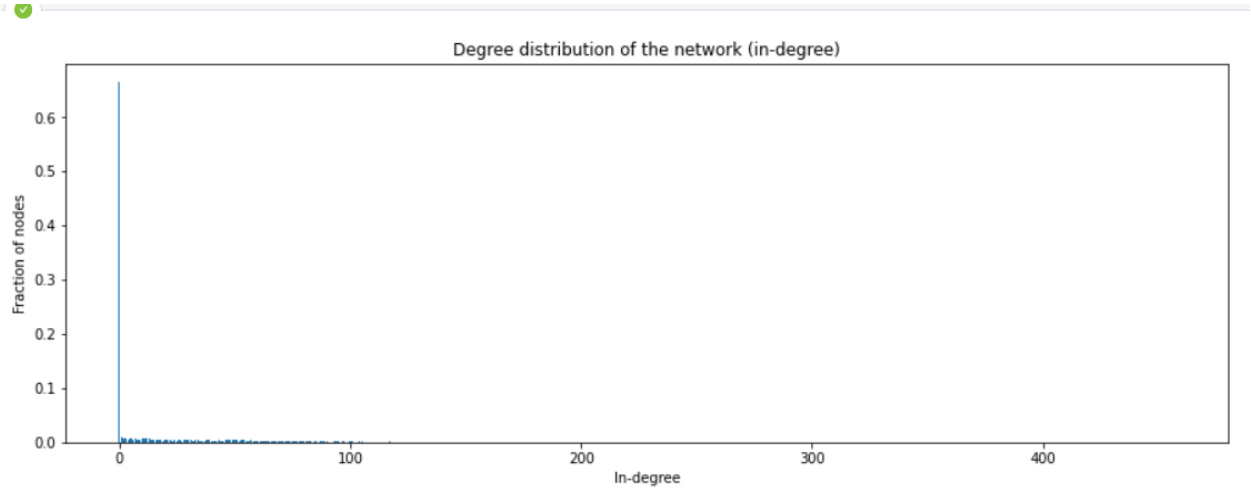
6. Node with Max out-degree

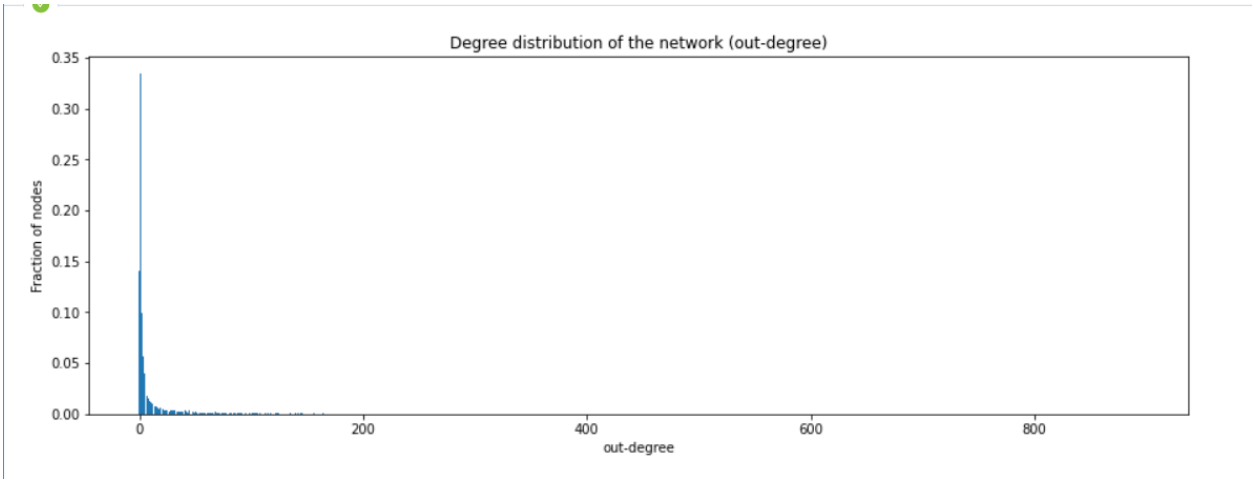
```
Node with max out-degree - 2565
Out-degree = 893
```

7. The density of the network

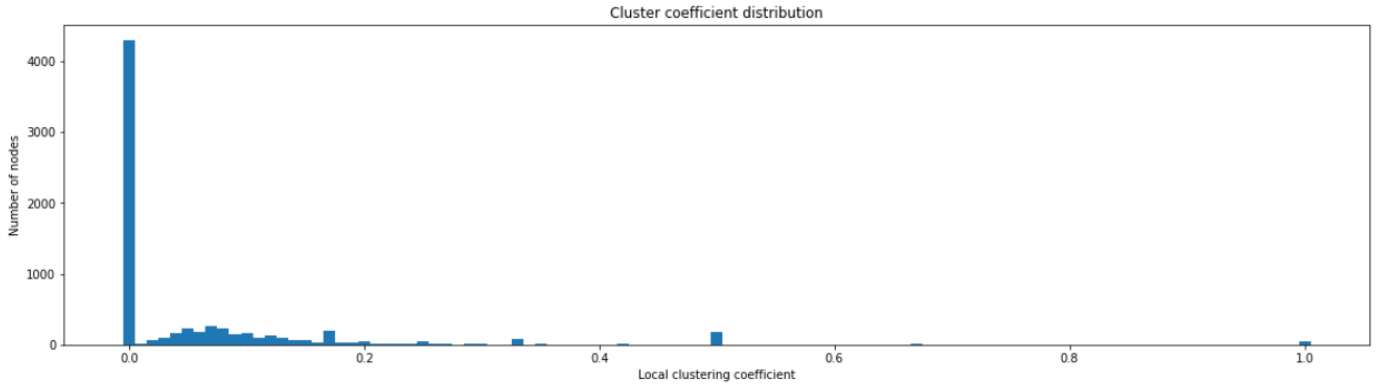
```
Density of the network = 0.0020485375110809584
```

1. Degree distribution of the network





2. Clustering-coefficient distribution of the network.



Cluster coefficient of node i = $\frac{Number\ of\ edges\ in\ neighbors\ of\ node\ i}{n^*(n-1)}$, where n is the number of neighbors of node i

Q2 In question -2 we had two tasks:
First I visualized the graph and saw the content of graph, then used networkx library to do the task.

```
Given graph have: DiGraph with 7115 nodes and 103689 edges
```

1. PageRank Score: Here I used networkx library to find the pagerank score of each node and then stored the score in a csv in decreasing order. 10 nodes having highest pagerank score are given below:

```
Top 10 nodes of Pagerank score are: [(4037, 0.004586879995073638), (15, 0.003675780093789909), (6634, 0.0033470695833168213),
(2625, 0.0032713153100210482), (2398, 0.002600193053403739), (2470, 0.002530134335199274), (2237, 0.0024754616361304556), (419
1, 0.0022518113639558167), (7553, 0.0021765417346990895), (5254, 0.002155060435801579)]
```

2. Authority and Hub Score: Here I used networks library to apply HITS algorithm on the given graph. HITS algorithm have two notion of page importance. I found Authority and Hub score of each node and then stored the score in a csv in decreasing order. 10 nodes having the highest Authority and Hub score are given below:

```
Top 10 nodes of Authority score are: [(2565, 0.007940492708143138), (766, 0.007574335297501241), (2688, 0.0064402489910298615),
(457, 0.006416870490261075), (1166, 0.006010567902411206), (1549, 0.005720754058269245), (11, 0.004921182063808111), (1151, 0.0
0457204070175641), (1374, 0.004467888792711107), (1133, 0.00391888173205735)]

Top 10 nodes of Hub score are: [(2565, 0.007940492708143138), (766, 0.007574335297501241), (2688, 0.0064402489910298615), (457,
0.006416870490261075), (1166, 0.006010567902411206), (1549, 0.005720754058269245), (11, 0.004921182063808111), (1151, 0.0045720
4070175641), (1374, 0.004467888792711107), (1133, 0.00391888173205735)]
```

In pagerank algorithm, the highest ranking node is 4037 which is the node with the maximum in-degree. In HITS algorithm, the ranking node is 2565 which is the node with the maximum out-degree.
Pagerank algorithm computes importance of nodes on the number of incoming links. Hence the node with maximum in-degree is highest ranking.
The HITS algorithm computes an authority score based on incoming links and hub score based on outgoing links. Here the node 2565 has out-degree 893 which is greater than in-degree of 4037 i.e. 457. Hence 2565 was ranked highest.