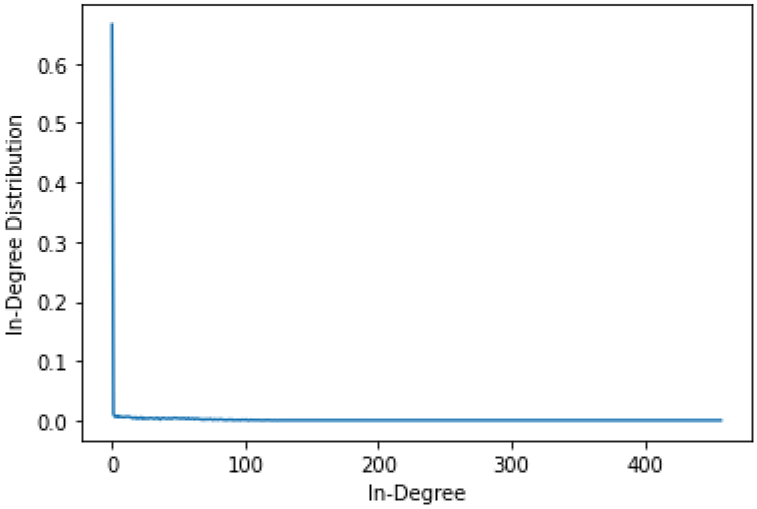
Assignment 3 Report
Group 47

Question 1

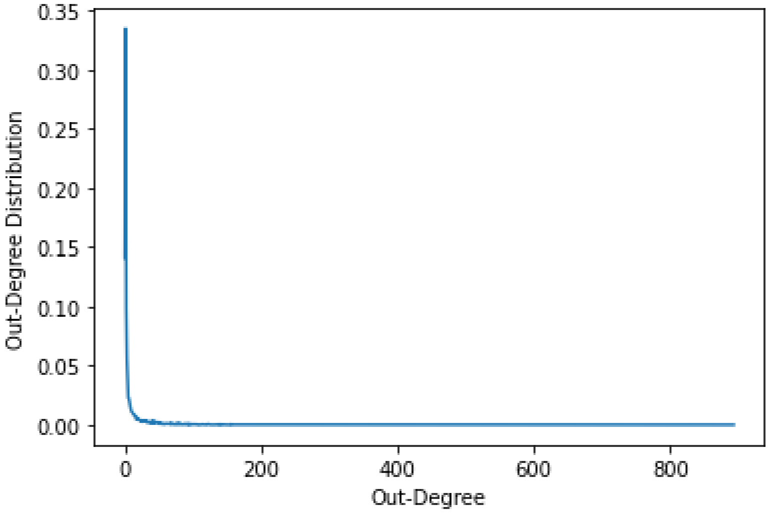
1. Number of Nodes: 7115
2. Number of Edges: 103689
3. Average In-Degree: $\frac{Sum\ of\ In-Degree\ of\ Nodes}{Number\ of\ Nodes}$ = 14.573295853829936
4. Average Out-Degree: $\frac{Sum\ of\ Out-Degree\ of\ Nodes}{Number\ of\ Nodes}$ = 14.573295853829936
5. Node with max In-Degree: Node(4037) with In-Degree(457)
6. Node with max Out-Degree: Node(2565) with In-Degree(893)
7. Density of the network:

$$\frac{Number\ of\ edges}{Number\ of\ possible\ edges} = \frac{\#Edges}{\#Nodes\times(\#Nodes-1)} = 0.0020485375110809584$$

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 n_k of them have in-degree k then $P(k) = \frac{n_k}{N}$

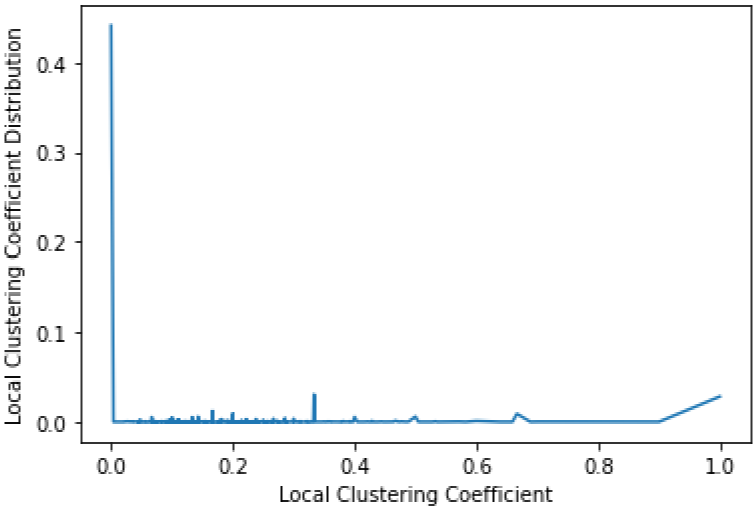


b. Out-Degree Distribution: Fraction of nodes with out-degree k. If there are N nodes in a network and n_k of them have out-degree k then $P(k) = \frac{n_k}{N}$



2. 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) [Source:Wikipedia]. The local clustering coefficient of a node v is given by $LCC(v) = \frac{Number\ of\ links\ between\ neighbours\ of\ v}{Number\ of\ possible\ connections} = \frac{Nv}{\frac{Kv(Kv-1)}{2}}$ where

   N_v is number of links between neigbours of v and K_v is the degree of node v



Question 2

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.

HITS(Hyperlink-Induced Topic Search) Algorithm It is a link analysis algorithm that computes the hub score based on outgoing links and the authority score based on the incoming links. It is also known as hubs and authorities algorithm. It ranks the web pages' relevant score for a particular search. Authority calculates the value of the content of the page and Hub calculates the value of its links to other pages.

Algorithm Here the most relevant web pages are retrieved through query search. Authority and Hub scores are defined as mutual recursion of one another. Authority score is calculated as the sum of the hub scores that point to that web page(incoming links). Hub score is calculated as the sum of the authority scores of the web pages that it points to(outgoing links).

Algorithm Steps

- Initialize the hub and authority of each node as 1.
- For each iteration, update the hub and authority of every node in the graph.
- The new authority is the sum of the hub of its parents.
- The new hub is the sum of the authority of its children.
- Normalize the new authority and hub.

The difference between PageRank and Hits are:

- PageRank calculates the ranks based on the proportional rank around the sites.
- HITS calculate the weights based on the HUBS and Authorities value.
- HITS algorithm is query dependent and PageRank is query independent.
- HITS algorithm is processed on a subset of relevant documents and PageRank is processed on the whole dataset.

Comparison between PageRank and HITS Algorithm

PageRank Algorithm

Top 10 Pagerank Score

```
[('4037', 0.0046127158911675485), ('15', 0.0036812207295292792), ('6634',
0.003524813657640259), ('2625', 0.0032863743692309023), ('2398',
0.0026053331717250192), ('2470', 0.0025301053283849546), ('2237',
0.002504703800483994), ('4191', 0.0022662633042363454), ('7553',
0.002170185049195958), ('5254', 0.0021500675059293235)]
```

HITS Algorithm

HITS Score Top 10 HUBS Score

```
[('2565', 0.00794049270807403), ('766', 0.007574335297444512), ('2688',
0.006440248991012525), ('457', 0.00641687049019565), ('1166',
0.006010567902433343), ('1549', 0.0057207540583986485), ('11',
0.004921182064008282), ('1151', 0.004572040701802756), ('1374',
0.004467888792672376), ('1133', 0.003918881732047633)]
```

HITS Score Top 10 Authorities Score

```
[('2398', 0.002580147178008918), ('4037', 0.002573241124142803), ('3352',
0.002328415091537902), ('1549', 0.0023037314804751075), ('762',
0.00225587485637424), ('3089', 0.0022534066884266454), ('1297',
0.00225014463679536), ('2565', 0.002223564103945871), ('15', 0.002201543492543811),
('2625', 0.0021978968035237852)]
```

Analysis Here we can observe that out of 10 highest scores of both PageRank and HITS algorithm only one node that is 4037 is similar between them. The results of PageRank highly vary from that of HITS algorithm because PageRank ranking is based on incoming links, while the HITS algorithm's ranking is based on hubs and authorities which are mutually recursive to each other. Hence, in HITS algorithm, the values are not bounded and the number of nodes with outgoing links are comparatively more than incoming links and hence influences the overall score. While, in the PageRank algorithm the score of a node is equally divided between all the outgoing links. In PageRank a node with enough number of incoming links has a comparatively higher score.