Information Retrieval
Assignment 3

Group Members:

➢ Devansh Shukla (MT21023)

➢ Ramyanee Kashyap (MT21139)

Q1. Dataset: https://snap.stanford.edu/data/p2p-Gnutella06.html

Adjacency matrix and edge list implementations were made using this dataset.

Dataset description:

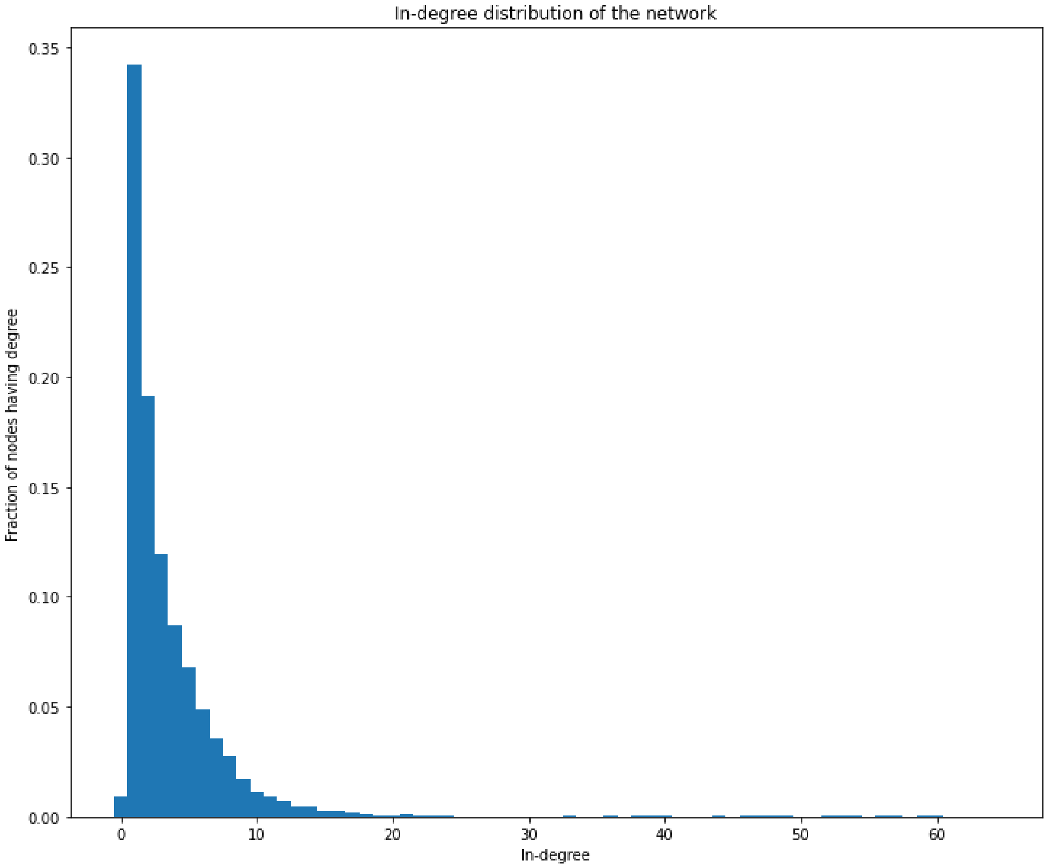
```
1. Number of nodes:  8717
2. Number of edges:  31525
3. Average In-degree:  3.6164965010898245
4. Average Out-degree:  3.6164965010898245
5. Node with Max In-degree:  356 (64)
6. Node with Max Out-degree:  6494 (113)
7. The density of the network:  0.0008298523407732502
```

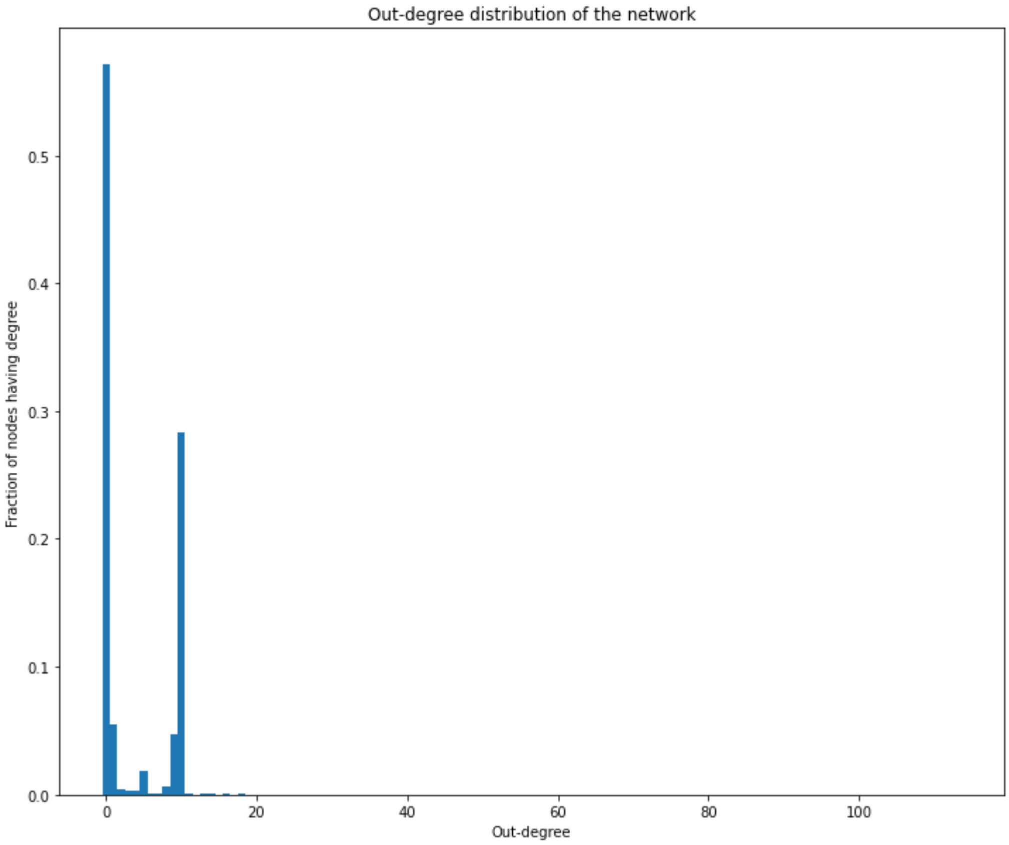
Formulas:

Density = No. of Edges/ ${}^{nodes}_{2}C$

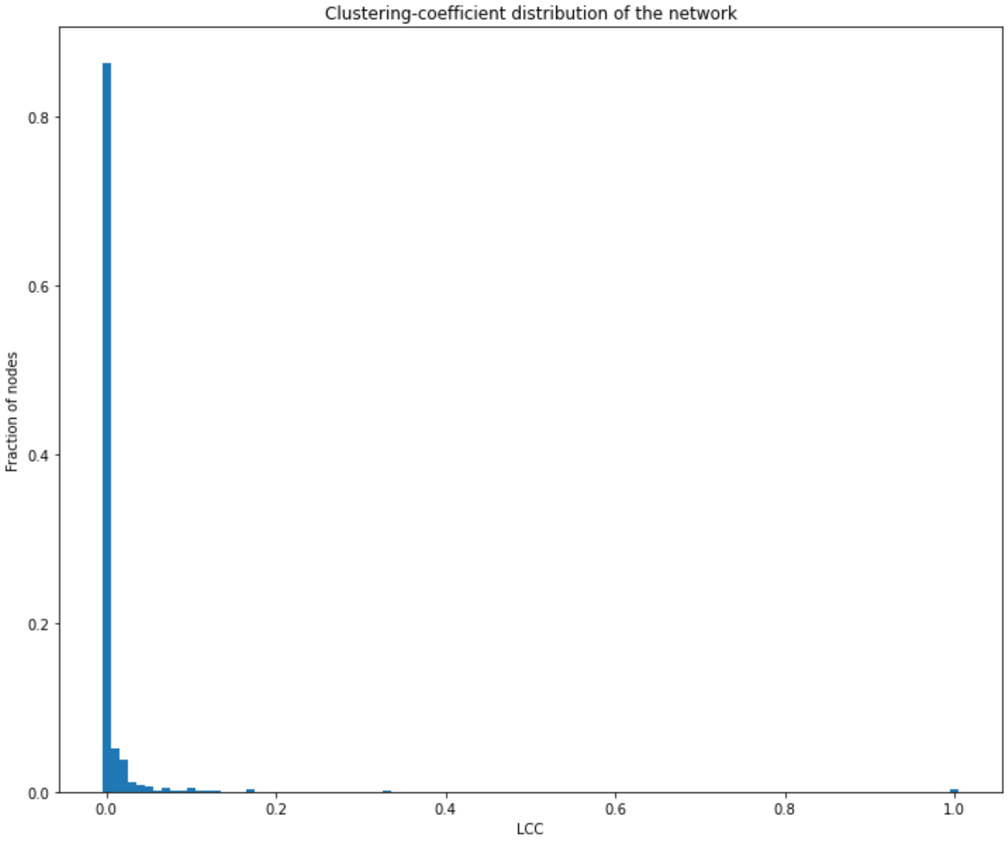
LCC = No. of neighbours having an edge between them / ${}^{neighbours}_{2}C$

1. Degree distribution of the network





2. Clustering-coefficient distribution of the network



Q2. Pre-processing:

Adjacency lists for both incoming edges as well as outgoing edges were created.

1. **PageRank:**

   PageRank score for a node is calculated as:

   $$\mathrm{PR(A)} = (1 - \mathrm{d}) + \mathrm{d}\,(\mathrm{PR(T1)/C(T1)} + \ldots + \mathrm{PR(Tn)/C(Tn)})$$

   where d is the dampening factor which we consider to be 0.85, and C(T_i) is the out degree of $T_{i.}$

2. **Hub & Authority:**

   Hub & Authority scores are calculated as:

   - Initialize hub and authority scores for each node with 1
   - For each iteration:
     - Update the hub and authority of each node as:
       - Authority(A) = Sum(Hub(Parents(A)))
       - Hub(A) = Sum(Authority(Children(A)))
   - Normalize Authority and Hub of each node

Results:

Node having max PageRank score: 556 (1.950733422083331)
Node having max Hub score: 8566 (0.015125446751113115)
Node having max Authority score: 8626 (0.008312806886373312)

Comparisons:

- When sorted over the three scores one at a time, there is no consistency of node pattern because of the different approaches to each scoring method.
- PageRank uses only the Authority score which considers only the incoming edges whereas Hub score is calculated by considering the outgoing edges.
- The time complexity of the Hits algorithm is O(kN^2) making it more time expensive when compared to PageRank.
- There are limitations to PageRank score such as rank sinks, spider traps and dangling links.
- HITS algorithm has limitations such as query dependency and irrelevant authorities problem.