Assignment - 03

“Web and Social Computing (IT752)”

Submitted by
Tarushi Jat (202IT029)

Submitted to
Dr. Sowmya Kamath

MASTER OF TECHNOLOGY
in
INFORMATION TECHNOLOGY

Department Of Information Technology

National Institute Of Technology, Surathkal

Task Description

Part (a): Calculate and compare PageRank of the top 100 pages, and generate a ranked list based on the calculated scores.

Part (b): Which of the Web Pages have the highest and lowest PageRank? What are your observations regarding the scale-free nature of real-world graphs and its impact on PageRank?

Part (c): Correlate your calculated scores with reference to the experiments performed in Assignment 1 (w.r.t various network properties) and write a detailed analysis of your observations, along with plots of the rank/score distribution.

Datasets Used for the Task

Dataset 1 : Collaboration Networks Dataset: ca-GrQc

This dataset covers scientific collaborations between authors papers submitted to the General Relativity and Quantum Cosmology category. In this collaboration Network, nodes are the authors in the network and If an author *i* co-authored a paper with author *j*, the graph contains an undirected edge from *i* to *j*.

Number of Nodes	5242
Number of Edges	14496

Dataset 2 : Communication Network : email-Eu-core

This dataset contains the information of incoming and outgoing email between members of a European research institution. There is an edge (u, v) in the network if person u sent person v at least one email.

Number of Nodes	1005
Number of Edges	16706

Dataset 3 : Social Networks Dataset : soc-sign-bitcoin-otc

Nodes are the people who trade using Bitcoin on a platform called Bitcoin OTC. Since Bitcoin users are anonymous, there is a need to maintain a record of users' reputation to prevent transactions with fraudulent and risky users. Members of Bitcoin OTC rate other members in a scale of -10 (total distrust) to +10 (total trust).
Therefore, there will be an edge between any 2 nodes if one user gives rating to the other user.

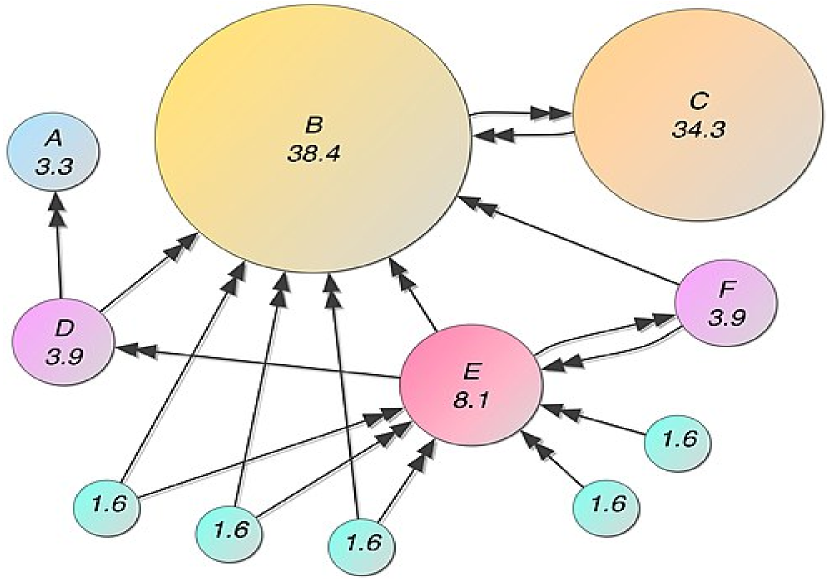
Number of Nodes	5881
Number of Edges	21492

PageRank Algorithm

PageRank is a way of measuring the importance of website pages. It is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google.

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

PageRank can be interpreted as a frequency of webpage visits by a random surfer, and thus it reflects the popularity of a webpage. The algorithm outputs the *probability distribution* used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. This algorithm assigns a numerical weighting to every node of a connected network. This measure represents the relative importance of a node within the graph (its rank).



To compute Page Rank a random walk is performed. This random walk is defined as follows:
1. The walker starts at a random node in the graph.
2. At each iteration, the walker follows an outgoing edge to one of the next nodes with a probability α or jumps to another random node with a probability 1-α.

Damping Factor: The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor *d*.

Calculation of PageRank for Different Networks

1. Page Rank Score

PageRank score tells the importance of the page in the network. For all the three datasets, PageRank of each node is being calculated with a damping factor of 0.85 and 1000 number of iterations are used for the PageRank algorithm to get converged. Below are the PageRank scores of 6 nodes in the tabulated form of the output generated after calculation of PageRank of each node in the network.

Dataset 1: ca-GrQc

Node Id	***PageRank Score***
3466	0.00023865912661567162
937	0.00018851339942181232
5233	0.0001366220742603169
26	0.0001647272972101157
1407	0.00019590370201218394
1488	0.0009176610229244632

Dataset 2: email-Eu-core

Node Id	***PageRank Score***
0	0.0010582741646366091
1	0.0011696800025691682
2	0.001848884531385519
3	0.0012990612406463993
4	0.0016544516227495472
5	0.0032996822751076506

Dataset 3: soc-sign-bitcoin-otc

Node Id	***PageRank Score***
6	0.0004160222338936658
2	0.0004611200512749037
5	0.00011373805459229972
1	0.0024134847467929027
15	0.00021983626930781772
4	0.0007710624954816695

2. Ranked List of Top 100 Web Pages

After calculations of the PageRank score of each node in the network, I have extracted out the top 100 nodes which are having the highest PageRank score than the other nodes in the network and assigned a rank based on the pageRank score a node has. Node with the highest PageRank score is assigned rank 1 and so on. Following is the tabular format where I have listed out some nodes with their rank and PageRank score. Ranks of the entire top 100 nodes with the highest PageRank score can be seen in the python notebook submitted.

Dataset 1: ca-GrQc

Rank	***Node Id***	***PageRank Score***
5	**2710**	0.00115308
12	**449**	0.00102037
13	**4956**	0.000972954
16	**5052**	0.000950909
19	**1488**	0.000917661
31	**1217**	0.00082341

Dataset 2: email-Eu-core

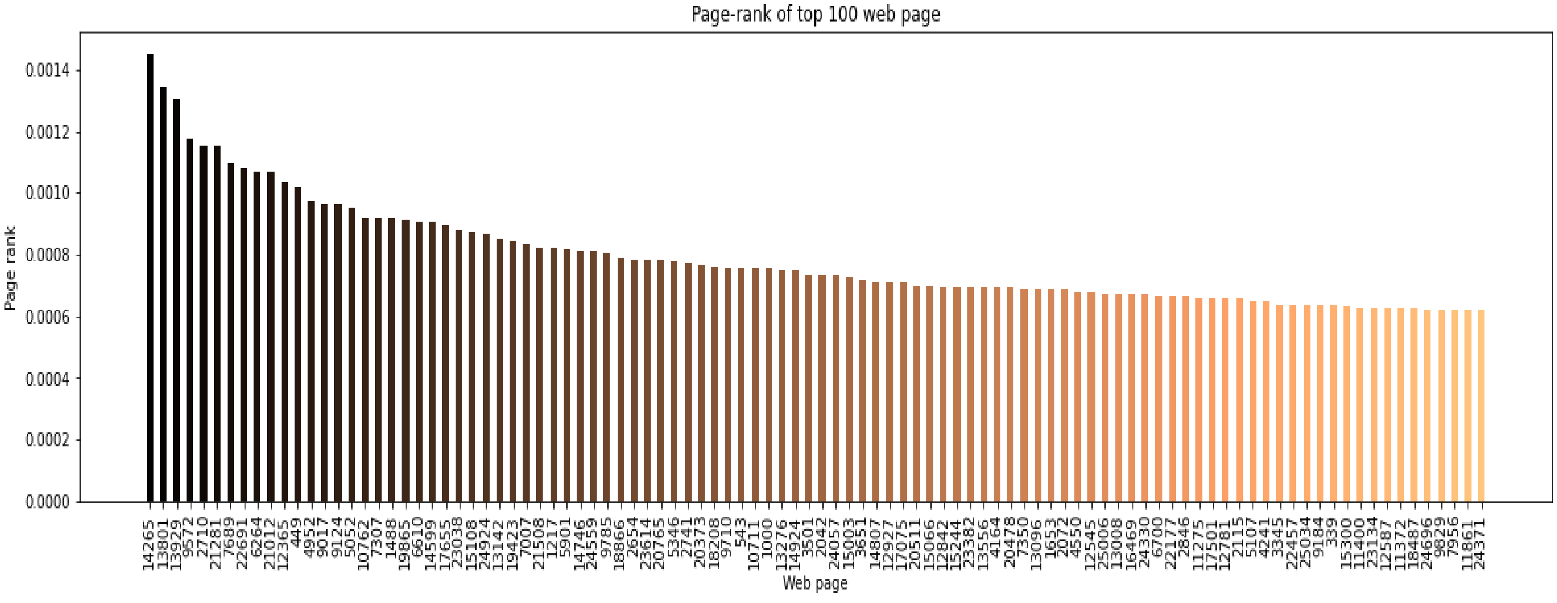
Rank	Node Id	PageRank Score
1	160	0.00505865
2	337	0.00351913
3	121	0.00350089
4	107	0.00348118
5	82	0.00347807
6	5	0.00329968

Dataset 3: soc-sign-bitcoin-otc

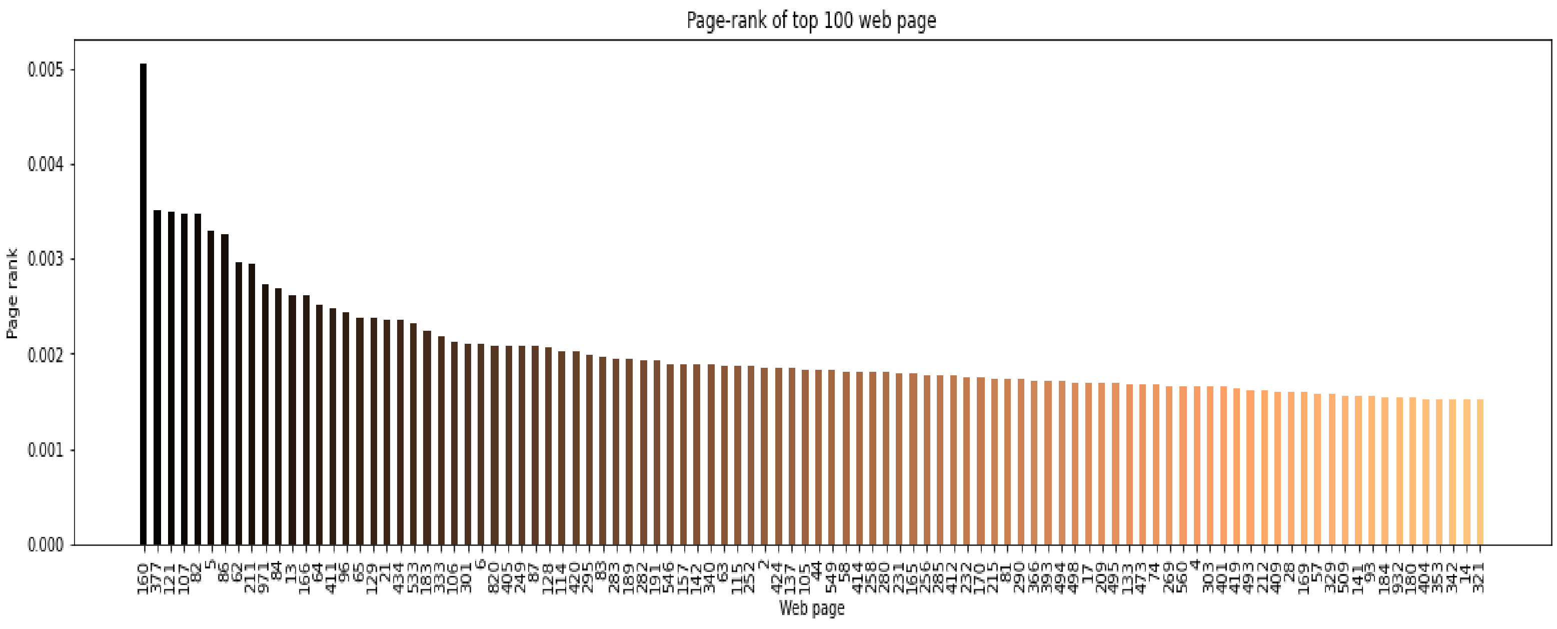
Rank	Node Id	PageRank Score
1	35	0.0199641
2	2125	0.00825667
3	2642	0.00622455
4	1810	0.00514806
5	3129	0.00490398
6	2028	0.00404358

3. Visualization of Top 100 Web Pages

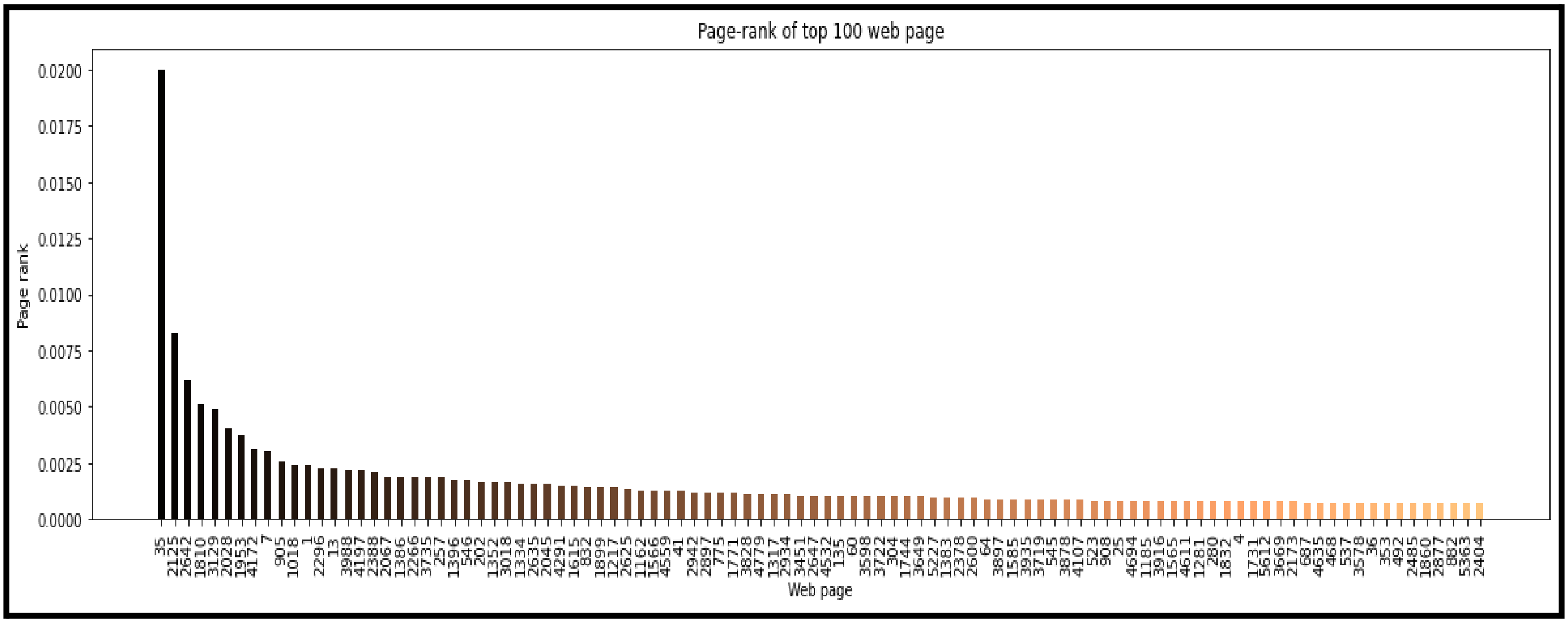
Dataset 1: ca-GrQc



Dataset 2: email-Eu-core



Dataset 3: soc-sign-bitcoin-otc



4. Page with Highest & Lowest PageRank

PageRank is only a score that represents the importance of the page in the network. The higher the PageRank score of the page, the more authoritative it is. And if the PageRank of a web page is lower then there is a very less chance that any person who is surfing by clicking randomly on the webpages will land up on the page. I have listed out the “Node_Id” and its respective “score” for the web page with highest PageRank score and with the lowest PageRank in the entire network.

Dataset 1: ca-GrQc

PageRank Score	***Node Id***	***Score***
Highest	1425	0.0014491
Lowest	4382	3.757399510637949e-05

Dataset 2: email-Eu-core

PageRank Score	***Node Id***	***Score***
Highest	160	0.00505865
Lowest	853	0.0006030113804405378

Dataset 3: soc-sign-bitcoin-otc

PageRank Score	***Node Id***	***Score***
Highest	35	0.0199641
Lowest	4704	0.00010456196480210277

Scale-Free Nature of Real-World Graphs & its Impact on PageRank

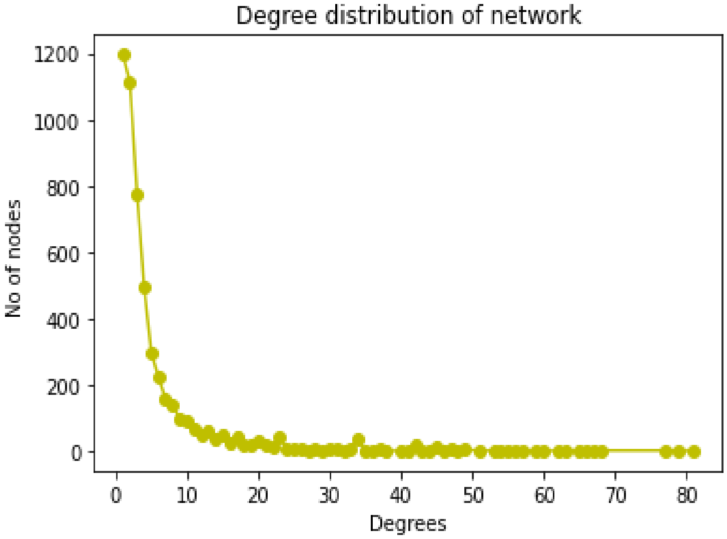
Real world networks grow over time. Nodes that already have a high number of edges are more likely to see new edges to them established compared with nodes with a lower number of edges. That is the idea of preferential attachment or the rich-get-richer principle. It is attractive to be connected to people who are already highly connected like celebrities, sporters and politicians in social networks.

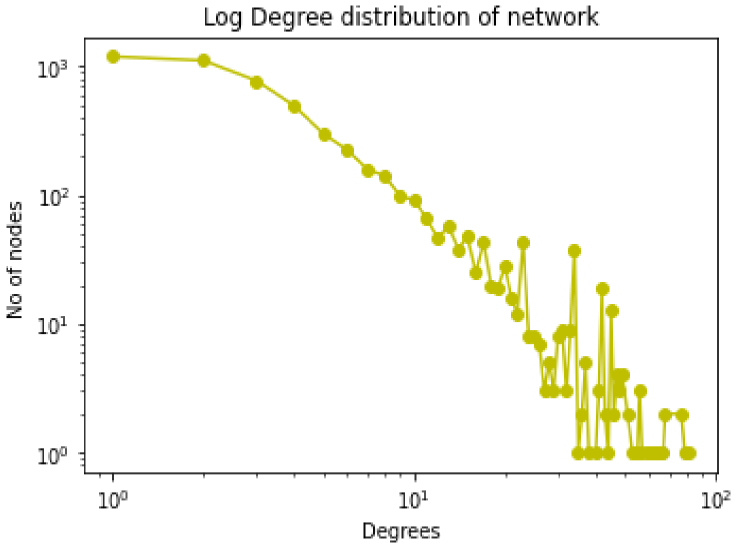
A network is scale-free if the characteristics of the network are independent of the size of the network, i.e. the number of nodes. That means that when the network grows, the underlying structure remains the same. A scale-free network is defined by the distribution of the number of edges of the nodes following **"power law distribution"**.

The number of nodes with really high numbers of edges is much higher in the power-law distribution. This means that in networks we will often find a small number of very highly connected nodes. Let us have a look at **degree distribution** for our networks to see the phenomena.

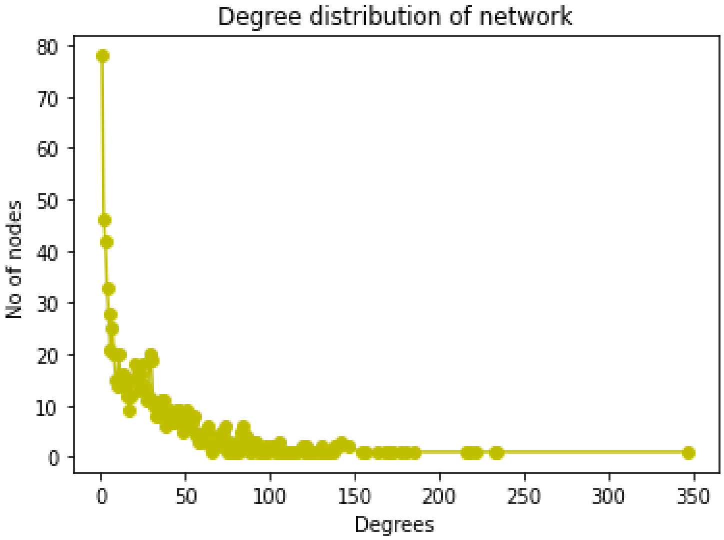
1. Degree Distribution

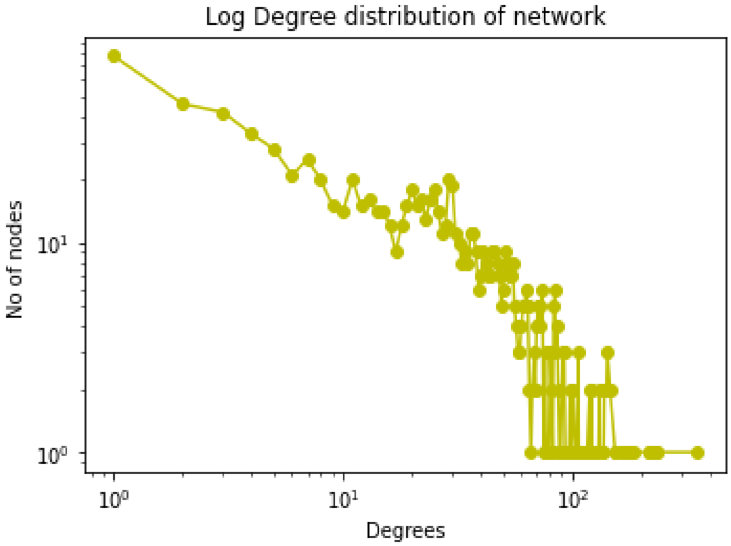
Dataset 1: ca-GrQc



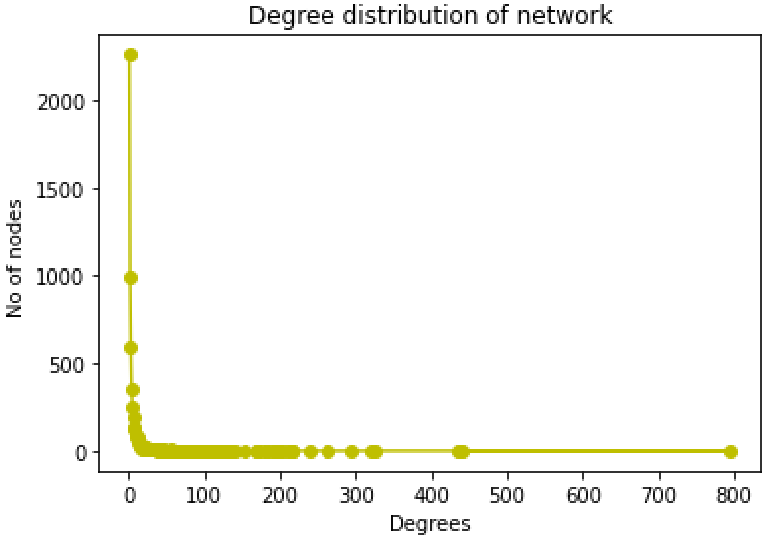


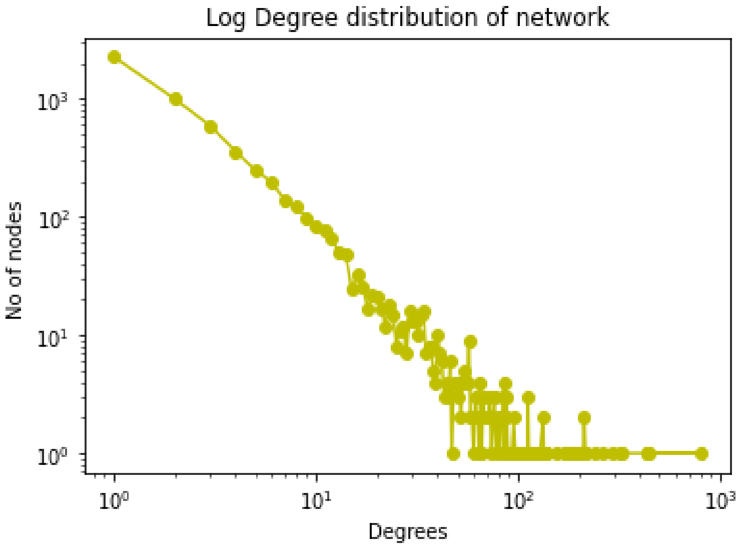
Dataset 2: email-Eu-core





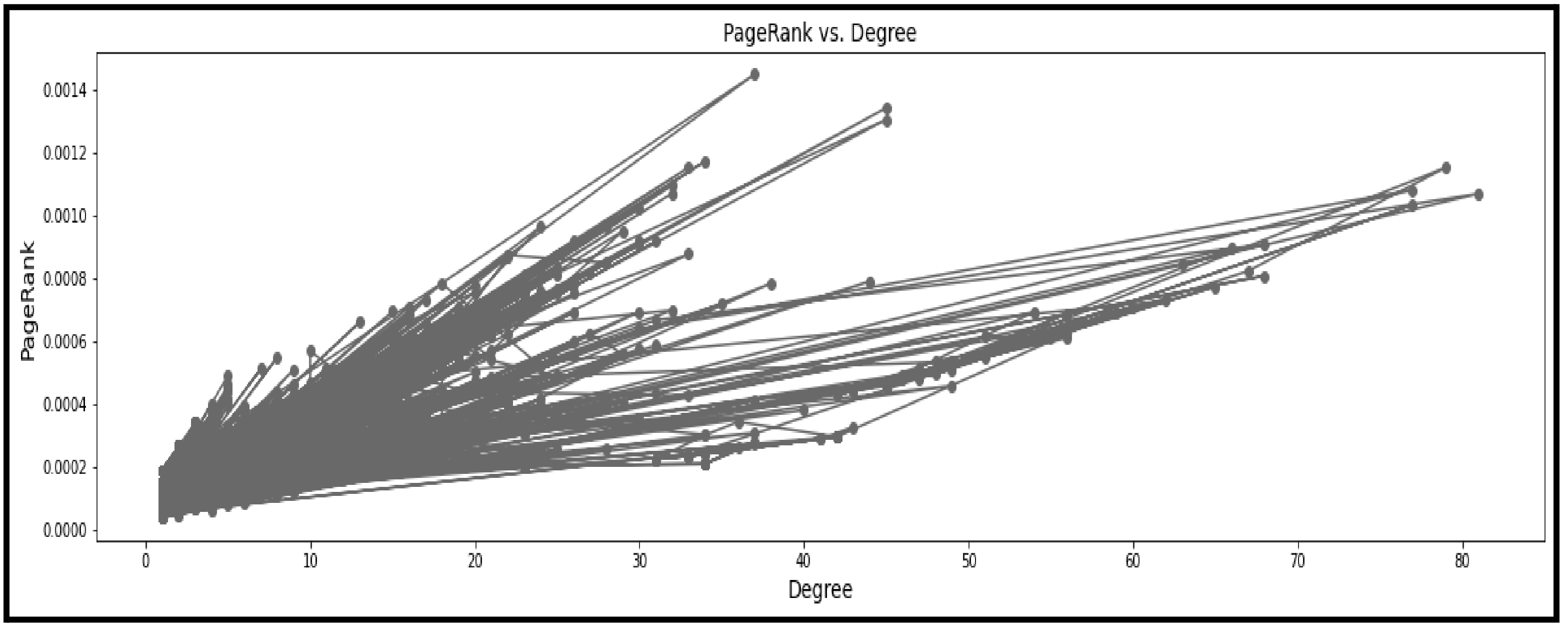
Dataset 3: soc-sign-bitcoin-otc



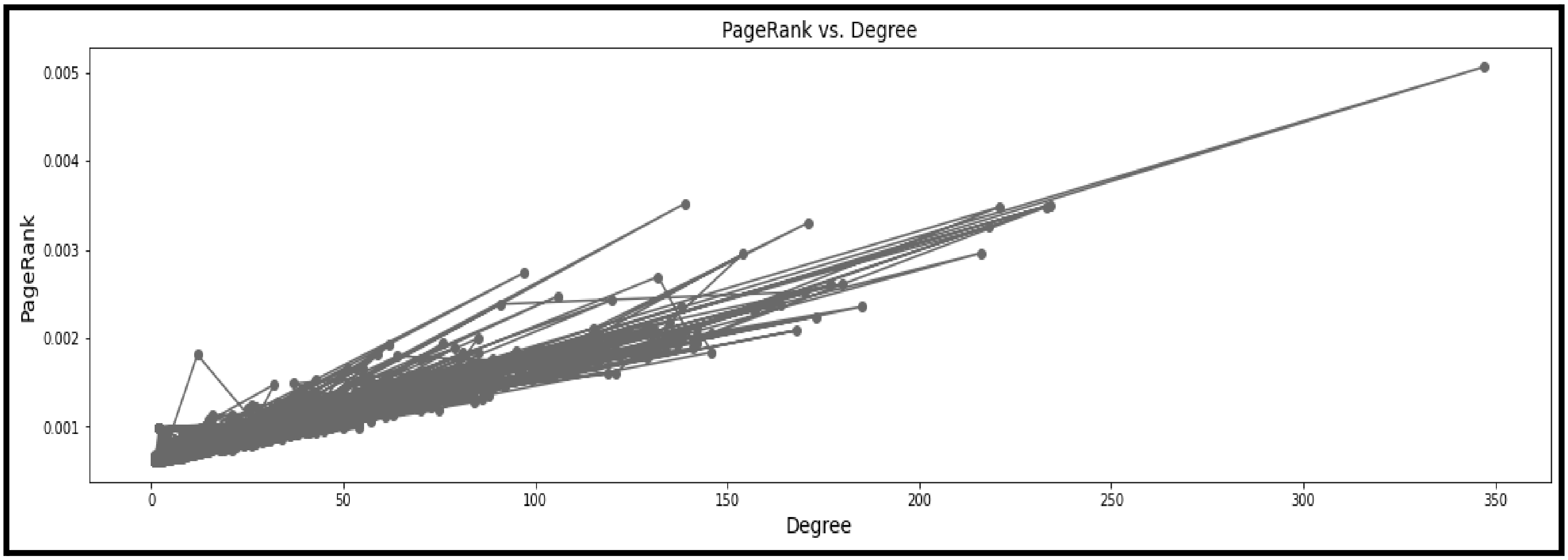


2. Visualization of PageRank and Degree of Nodes

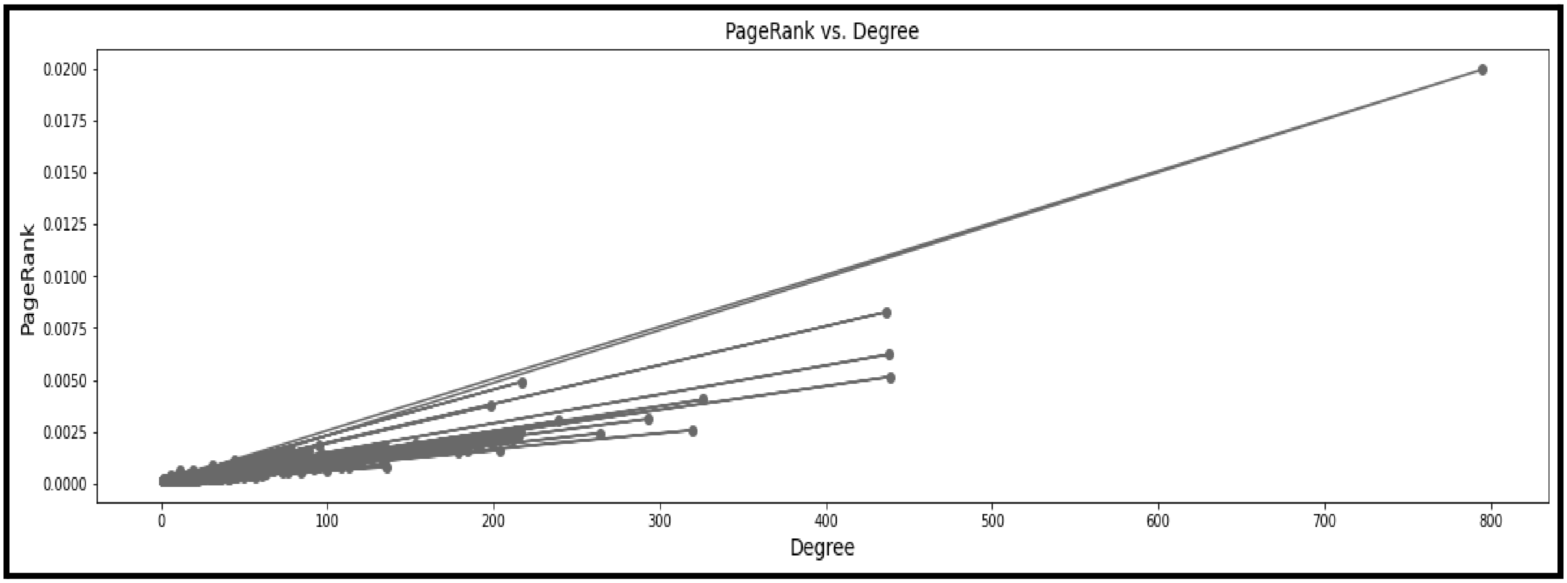
Dataset 1: ca-GrQc



Dataset 2: email-Eu-core



Dataset 3: soc-sign-bitcoin-otc



Observations: It is observed from the above visualization of PageRank score and degree of nodes for all the three networks that nodes having higher degree also have higher PageRank score and nodes with small degree also having low PageRank score. Also, pages with really high degree (or PageRank score) are very very less in all the three networks taken for this experiment. And the number of pages with relatively low degree are very high.

3. Average Degree and Average Clustering Coefficient of the network

The average number of edges per node in a graph is the average degree and a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. A high clustering coefficient for a network is another indication of a small world network. Following table shows the average degree and average clustering coefficient for all three networks.

Dataset 1: ca-GrQc

Number of Nodes	Number of Edges	Avg. Degree	Avg. Clustering Coefficient
5242	14496	5.5307	0.529635811052136

Dataset 2: email-Eu-core

Number of Nodes	Number of Edges	Avg. Degree	Avg. Clustering Coefficient
1005	16706	33.2458	0.3993549664221539

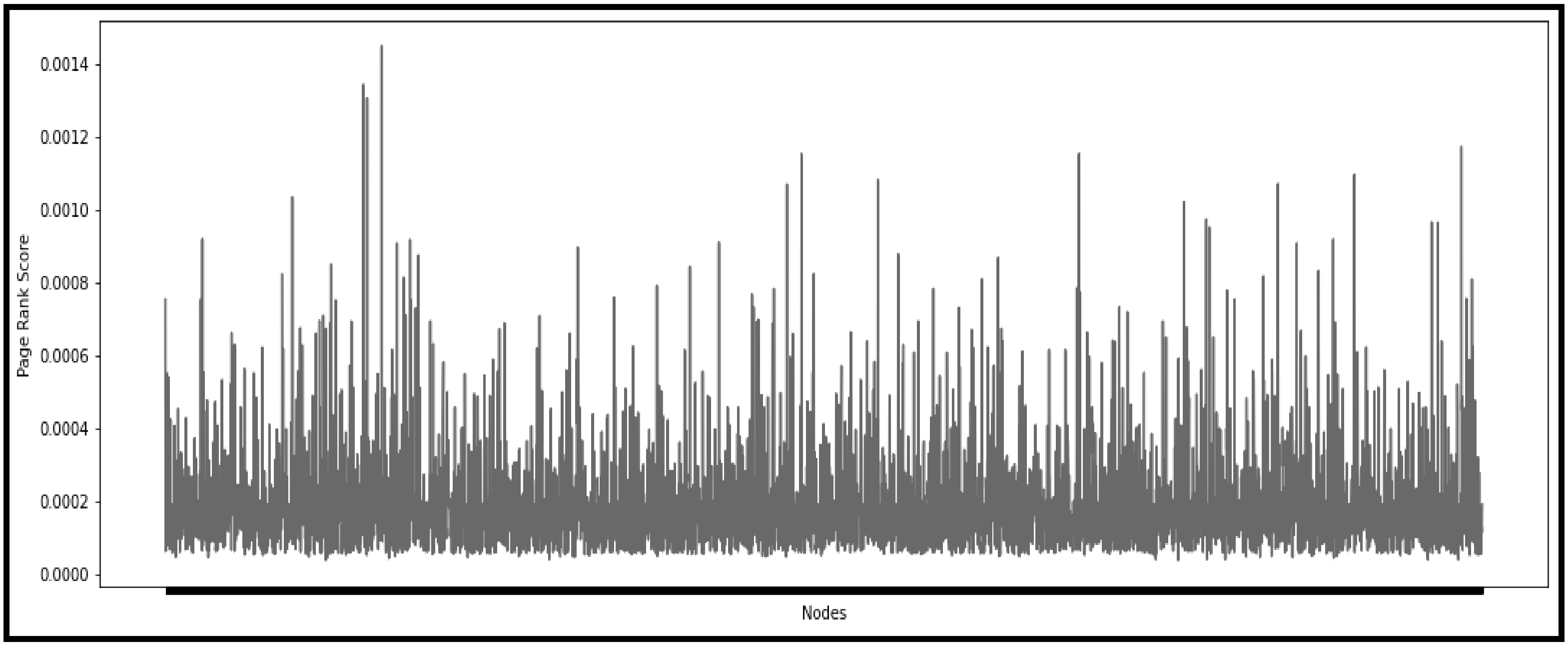
Dataset 3: soc-sign-bitcoin-otc

Number of Nodes	Number of Edges	Avg. Degree	Avg. Clustering Coefficient
5881	21492	7.3090	0.17750449405289517

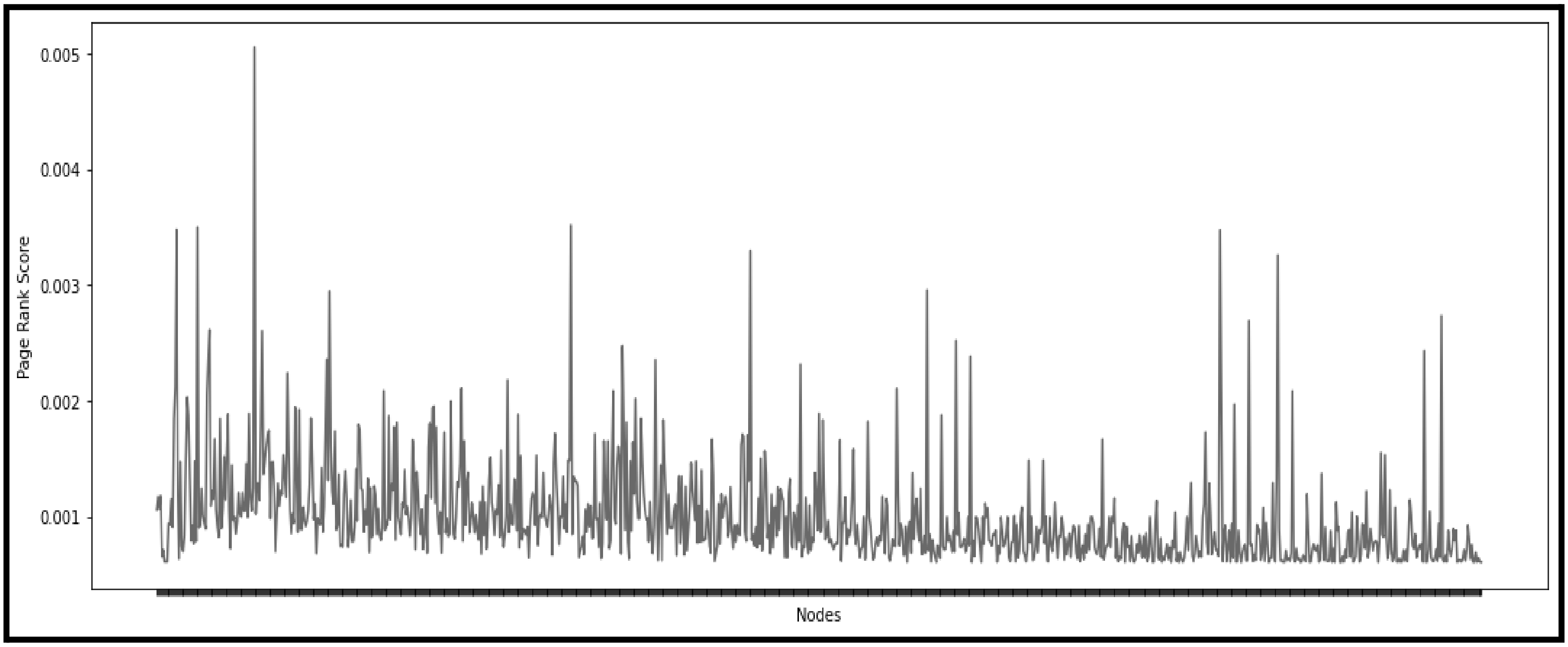
Observations: Nodes with higher PageRank Score will always have a higher chance that a random surfer will land up on that page and more web pages will get connected to the page with higher score. Therefore, a web page with higher PageRank will have even more high PageRank score with time as the network grows and a web page with low PageRank will have even low PageRank score. It can be concluded that the real world network follows the idea of preferential attachment which says that rich gets richer and poor gets poorer.

4. PageRank Distribution Visualization of all the Web Pages of the Network

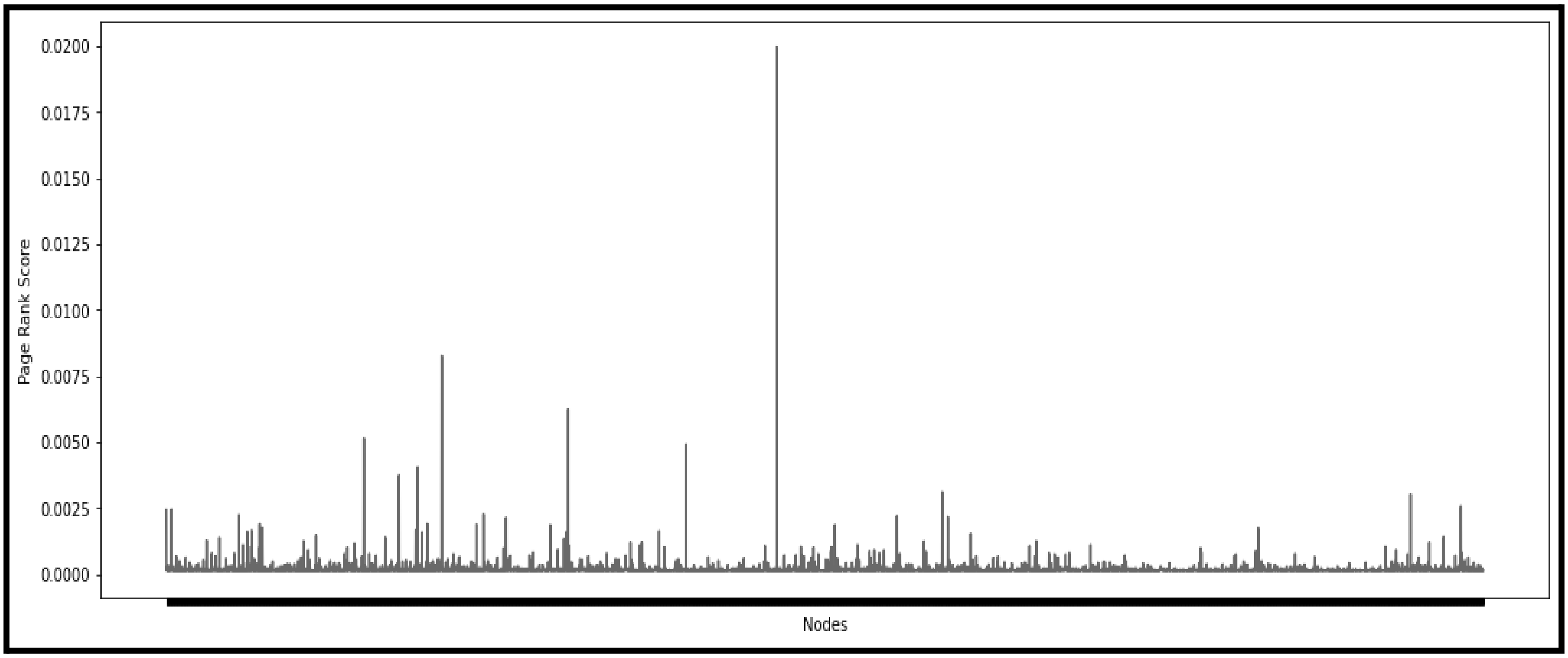
Dataset 1: ca-GrQc



Dataset 2: email-Eu-core



Dataset 3: soc-sign-bitcoin-otc



Code Implementation Notebook :-

https://colab.research.google.com/drive/10Kks59ZnwL8wwv-Gz3aP13jDx27Tbjn9?usp=sharing