Mining in Large Networks

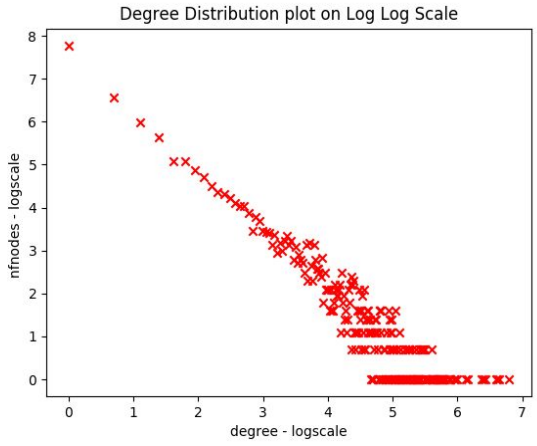
Assignment 1 Report

Utilized the **networkx module** in python for all the computations wherever required and mentioned to do so in question -

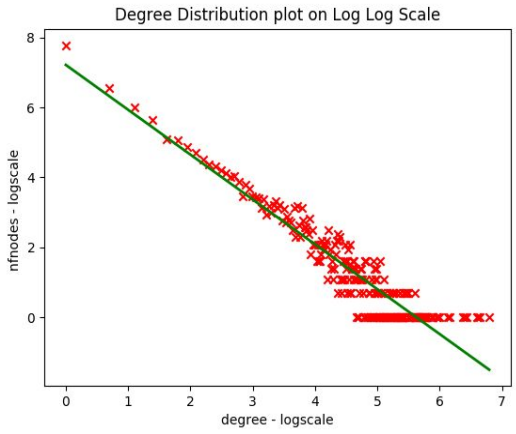
1a

1.Number of Vertices = 7115
Number of Edges = 103689
2. Number of Self Loop Edges = 0
3.Number of NON - self Loop Edges = 103689
4a. Number of Unique Combinations of edges such that (a,b) is considered the same as (b, a) = 100762
4b. Number of Unique Combinations of edges such that a,b are NOT considered the same as (b, a) = 103689
5. Number of Edges such that if a,b edge exists then so does (b, a) edge = 2927
6.Number of nodes with in-degree zero = 4734
7.Number of nodes with outdegree zero = 1005
8.Number of weakly connected components = 24
9.Number of nodes with in-degree > 10 = 1906
10.Number of nodes with outdegree > 10 = 1612

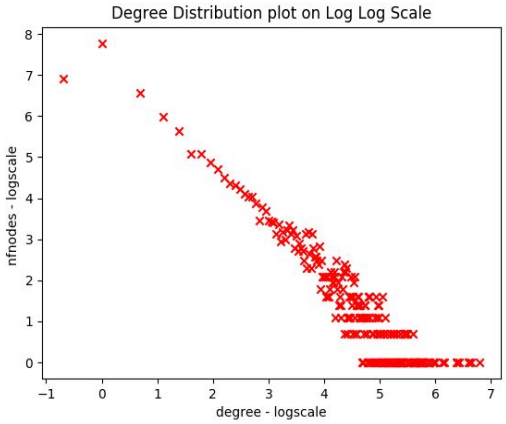
1b



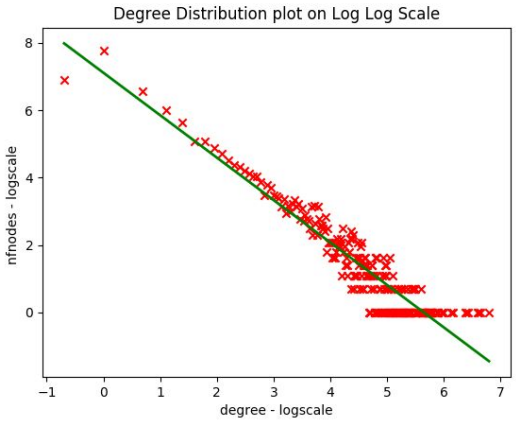
(a)



(b)



(c)

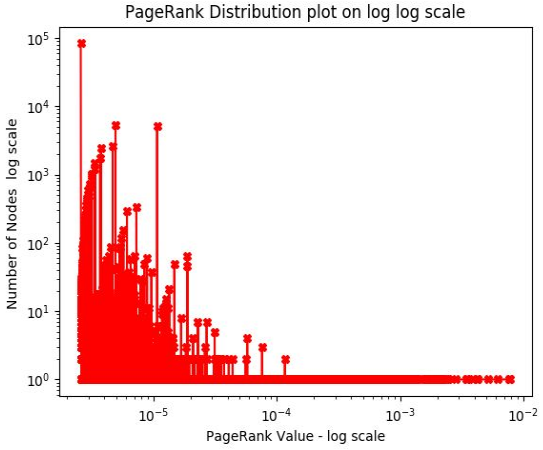


(d)

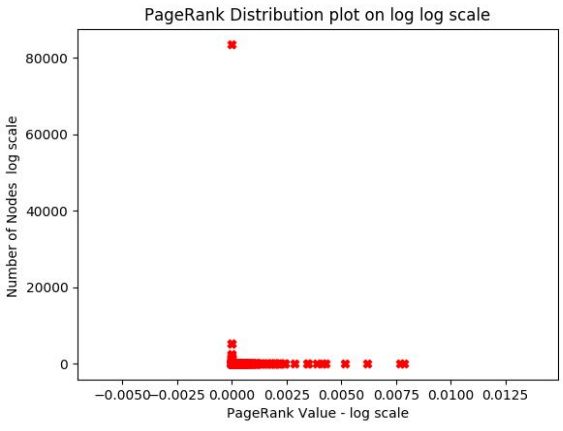
(a)and (c) is the degree distribution plots of the graph in log-log scale. (a) includes all degrees except zero plotted whereas (c) includes zero degrees assumed as 0.5 in order to fit into the log-log scale.
(b) and (d) are the corresponding best line fits of the of (a) and (c).

2a

Number of weakly connected components = 10143
Number of strongly connected components = 142474
Number of Vertices in largest_weak_connected_graph = 131188
Number of Edges in largest_weak_connected_graph = 322486
Node with the highest Pagerank Score is = 992484



(a)



(b)

(a) - log log scale

(b) - NOT on log log scale

Library implementation

----------TOP 5 Hubs----------
1 Node = 892029 Hub_Score = 0.00060387438361127
2 Node = 1194415 Hub_Score = 0.0004901759190513973
3 Node = 359862 Hub_Score = 0.00046815658534133905
4 Node = 648138 Hub_Score = 0.00045759932879617875
5 Node = 470184 Hub_Score = 0.0004377321680723338
----------TOP 5 Authority----------
1 Node = 22656 Authority_Score = 0.02700343024139879
2 Node = 157882 Authority_Score = 0.01333812448310903
3 Node = 571407 Authority_Score = 0.012677599370634961
4 Node = 57695 Authority_Score = 0.012025562401490466
5 Node = 139985 Authority_Score = 0.01113675806528405

2b

Self Made Implementation

-------TOP 5 HUB SCORES---------------
(892029, 0.0006038743852630347)
(1194415, 0.0004901759201988988)
(359862, 0.0004681565862466529)
(648138, 0.0004575993298920696)
(470184, 0.0004377321689986313)
-------TOP 5 AUTHORITY SCORES---------------
(22656, 0.027003429637642858)
(157882, 0.013338124678972127)
(571407, 0.012677599488078015)
(57695, 0.012025562440537648)
(139985, 0.011136758110801325)

Max_iters = 500
Tol = 1e-10
Converges in 27 iterations

The L2 normalization Difference is b/w Library and Self Made Implementation

Hub Score = 5.731583414622844e-16

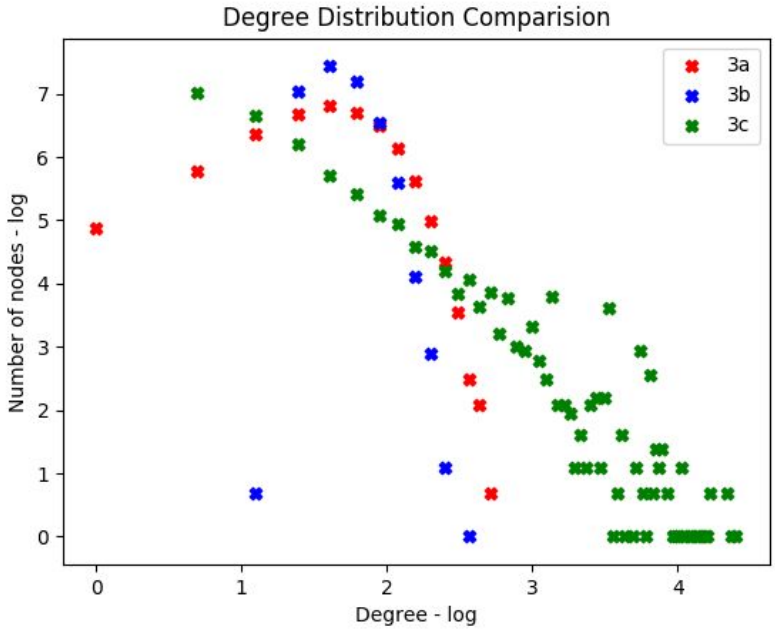
Authority Score = 4.437316966875576e-15

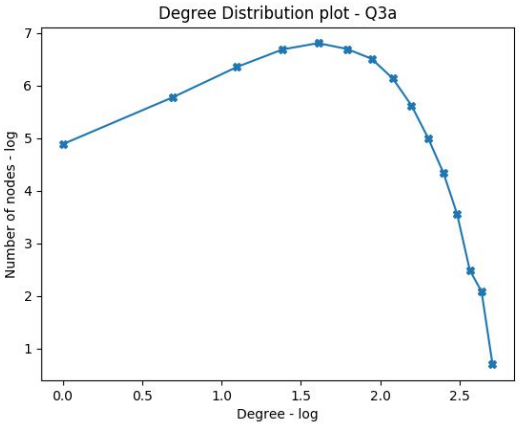
3a - Pickle FIle included in the folder

3b - Pickle FIle included in the folder

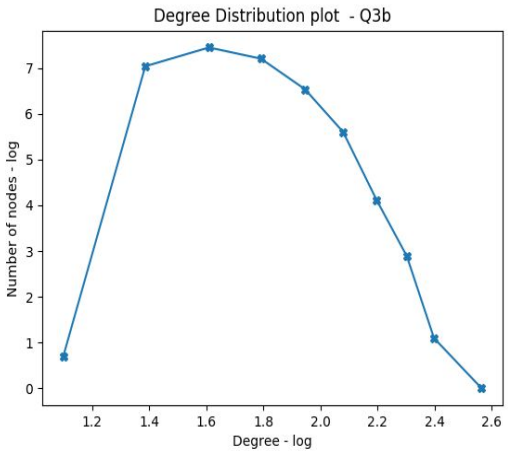
3c

3d

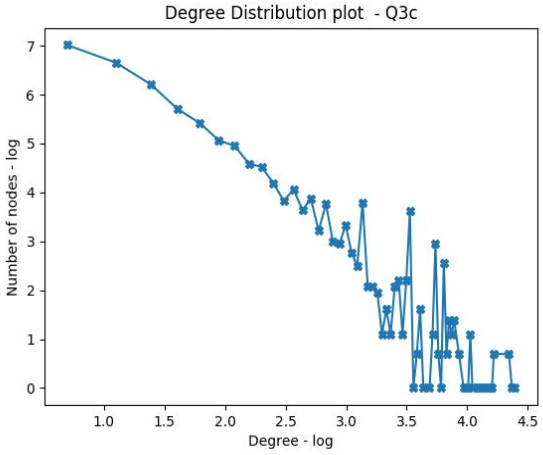




(a)



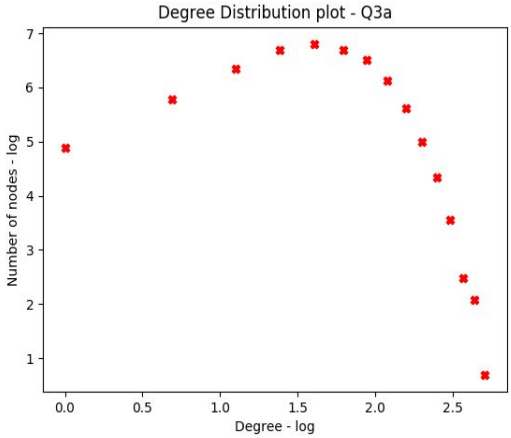
(b)

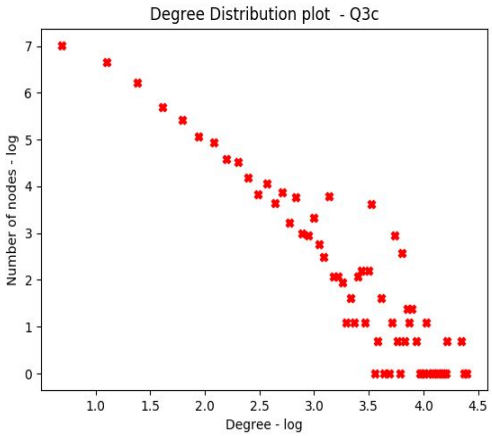


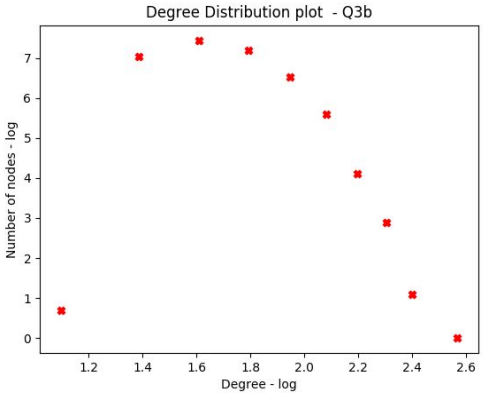
(c)

Observations

1. In graphs (a) and (b) - Random graphs and Wattz Strogatz Models have a degree distribution in which there are very few nodes with less degree and very few nodes with a high degree. However, in a real graph - (c), there are a high number of nodes with a low degree which makes the degree distribution a decreasing curve with some variances.
2. There is an incredible amount of variation in the Real graph(3c), we do NOT see such variation in Random Graphs.
3. The random graph (3a) has a has unrealistic Poissonian degree distribution.
4. The watts Strogatz model(3b) has very less variation in degrees i.e #unique degrees. We attribute this to the construction of the graph all nodes are expected to have the same degree, however, the variation exists because of the 4000 random edges.







3e

Graph 3a

Average Clustering Coefficient 3a self-made= 0.0009345752050310508
Average Clustering Coefficient 3a library = 0.0009345752050310508

Graph 3b

Average Clustering Coefficient 3a self-made= 0.2840983205659885
Average Clustering Coefficient 3a library = 0.2840983205659885

Graph 3c

Average Clustering Coefficient 3a self-made= 0.5338735707109258
Average Clustering Coefficient 3a library = 0.529635811052136

Observations

1. The average clustering coefficient follows:

   3c > 3b > 3a

2. Our Real graph which is an authorship network has a **higher clustering coefficient than a random graph**. This gives the inference that **on average a particular author's co-authors also collaborate with each other very often**.
3. The clustering coefficient gives an insight into the type of graph, most real-life networks have a high avg clustering coefficient while random networks have a low clustering coefficient.
   This result is consistent with other real-life networks.

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

4. A high clustering coefficient gives an indication to the formation of communities within the networks. In more real-life networks this may correspond to an interesting phenomenon like echo -chambers.