PROJECT REPORT- PROJECT 1

Phase 1- Crawling

Social media network crawled is facebook, and the network is the friends network of a user. Due to the limitations of the official Graph API of facebook, Selenium Python API (https://selenium-python.readthedocs.org)was used which allowed access all functionalities of Selenium WebDriver in an intuitive way. With the help of selenium a total of 48332 nodes and 85518 edges were crawled, the resulting graph is an undirected graph.

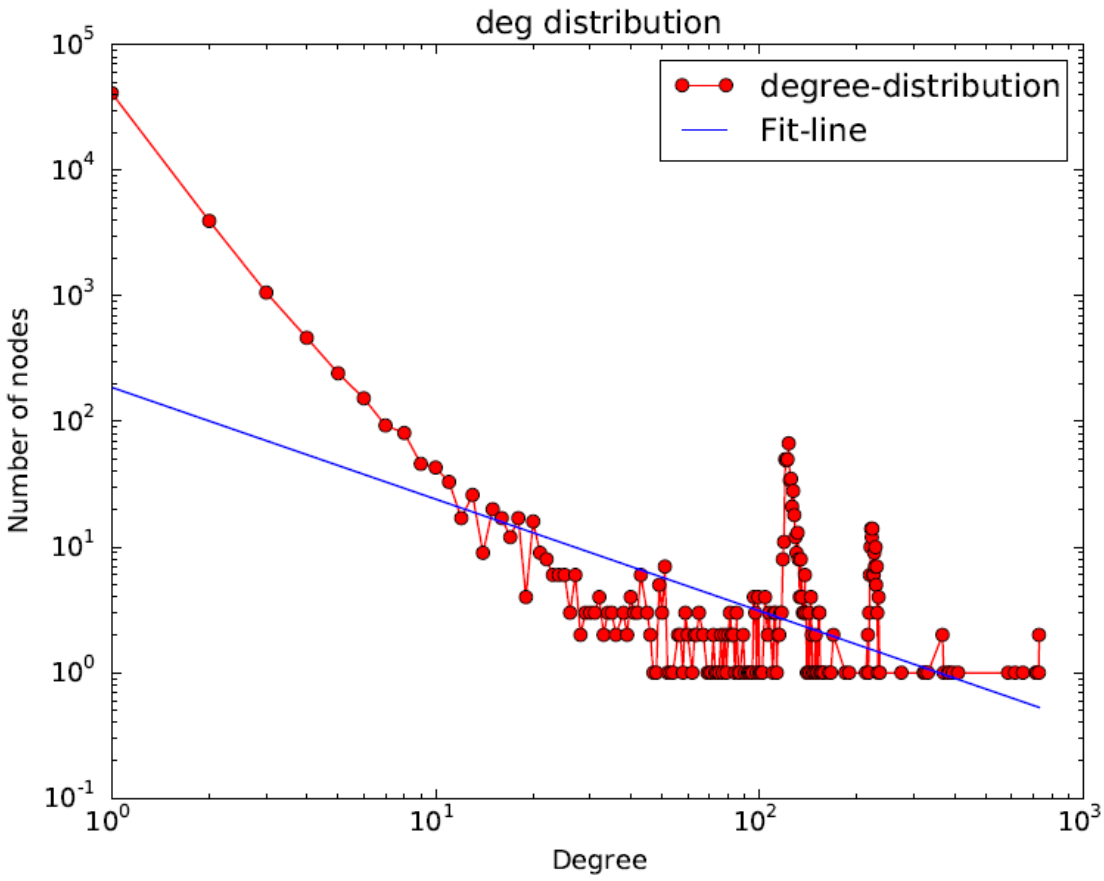
All the material for phase 1 is in the folder Phase 1. The files inside the folder are as follows

- Scrapping.py – It is the python script to scrape the data from facebook. To start with you will have to own a facebook account by which you will login through the script. Once you put in a valid id and password, the scrapping of your friends network start following a BFS algorithm with the person login in being the seed node. I have sampled the data by choosing just the first 100 friends of a person this was mainly done to constraint the amount of data I have to speed up the project. This sampling can be easily removed, which will then scrap all the friends in the friend list of a node. The Script is such that it can be stopped and continued from where it was last stopped, which is done with the help of object serialization using pickle. It also contains the code for pre-processing and formatting the data scrapped into the corresponding CSV files. If the script fails at first please restart the script it might be just fallacy in the web driver getting initialized of selenium.
- J.p,nodes.p,links.p- These are the pickels of object which will help in resuming where the scrapping last stopped.
- Nodes.csv- Anonymized file of nodes
- Nodes_name.csv- Node list with names and its corresponding anonymized value
- Edges.csv- Edges in the network in comma separated structure, with anonymized values.

Phase 2-Graph Essentials

For network analysis I have used the trio of networkx, matpolib, and numpy. My results are as follows

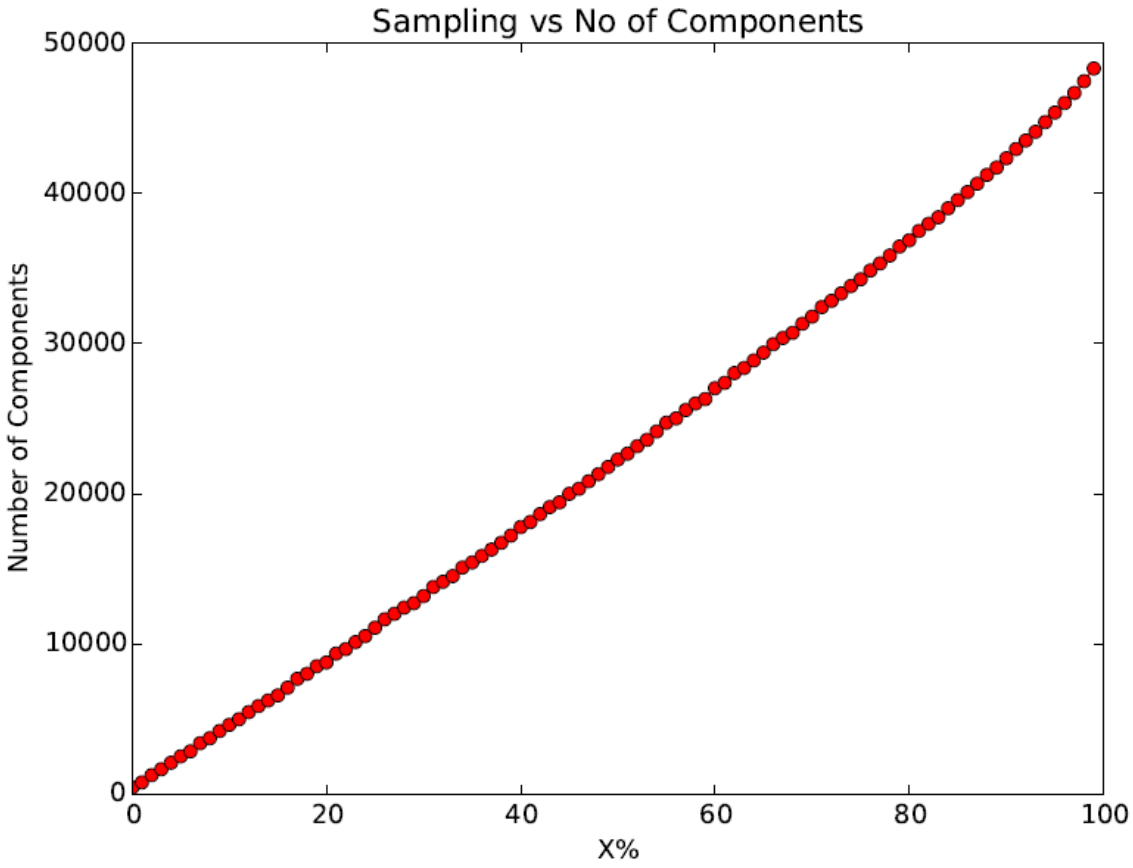
1. Degree Distribution



Exponents are b -0.890091413282 loga 5.23116522485, giving a power distribution.

2. Number of bridges in the graph are 41168
3. Number of 3 cycles found in the graph were 26148
4. Diameter of the graph is 6

5. Sampling vs No of componets graph



Phase 3 – Network Measures

1. Clustering coefficients are as follows :
    a. Local - 0542186951873
    b. Global - 0.14392364942

2. Centrality values of top 10 nodes were found as follows for the crawled graph :

    a. Eigenvector Centrality

Node	Value
1	0.13408756556606982
6	0.1157030444610927
28	0.11121016259529688
39	0.10648426528819417
173	0.10388963827034763
57	0.10292384721295351
55	0.10138130037324143
164	0.10115442379410566
11	0.10015657723022331
130	0.0993797126950918

    b. Pagerank Centrality

12	0.006245142995109705
15	0.0060807590129286
20	0.00515763603593942
18	0.004452990103870161
11	0.0036876339299335608
6	0.003402902026607273
2	0.0033604807970607172
8	0.0028662380706797534
22	0.0026925461730722268
29	0.0024697664759163614

c. Degree Centrality

11	0.0150834867890174
18	0.0150834867890174
6	0.015062796134985828
15	0.014835198940638514
12	0.014731745670480644
2	0.013448925120523059
20	0.012724752229417971
1	0.012104032608470754
22	0.008483168152945315
10	0.008172808342471706

d. Correlation between the centrality values of Pagerank and Eigenvector is (-0.2)

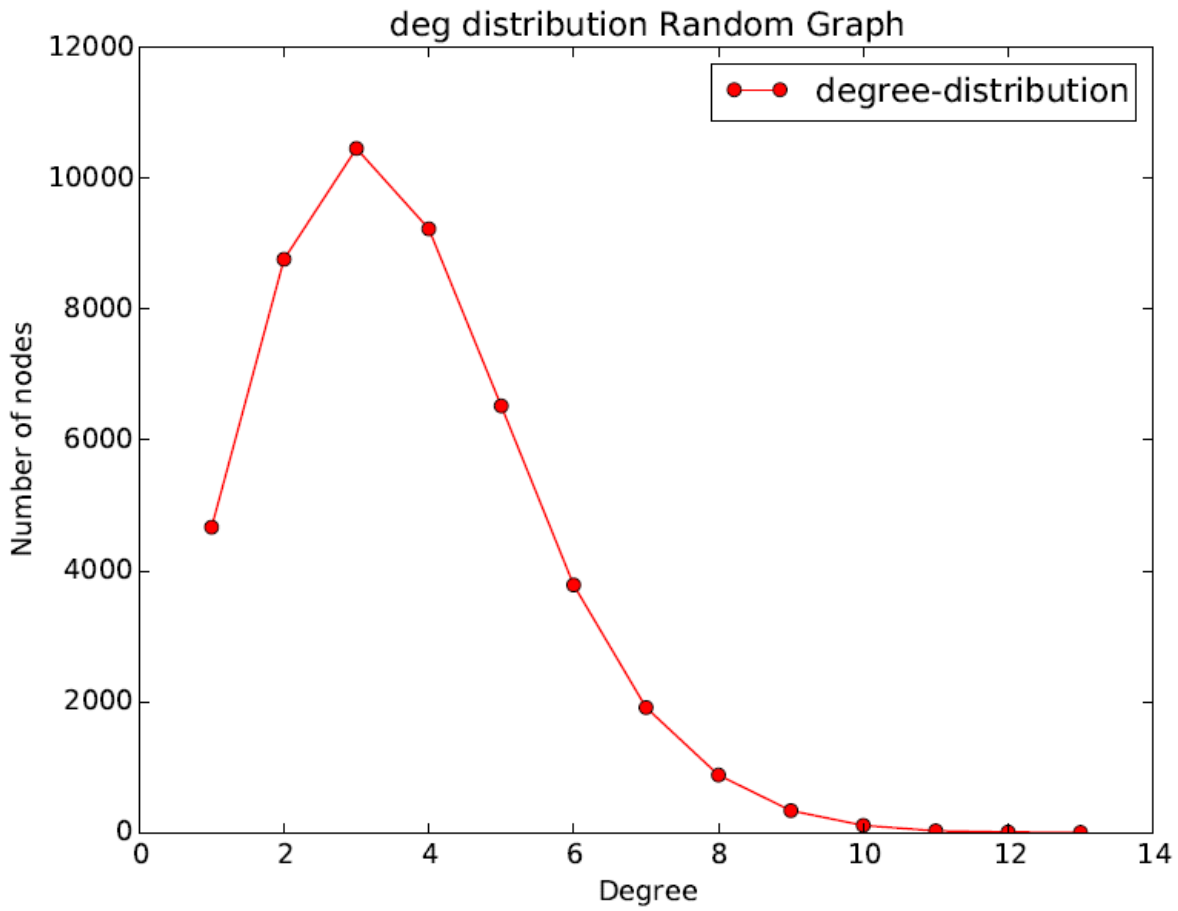
e. Correlation between the centrality values of Degree and Eigenvector is (0.0222222222222)

f. Correlation between the centrality values of Pagerank and Degree is (-0.644444444444)

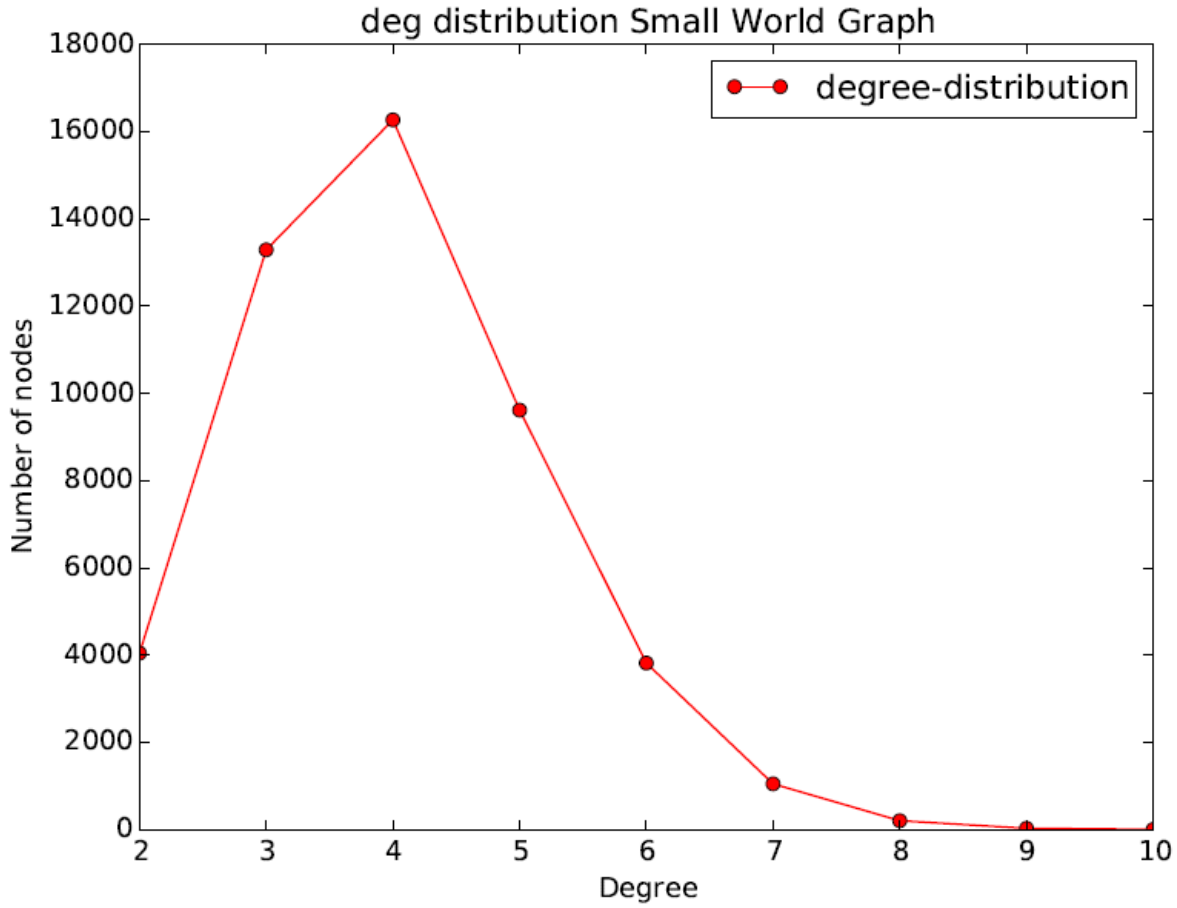
3. By Jaccard similarity two most similar nodes are 32822 32810 and its value is 1.0

Phase 4 – Network Models

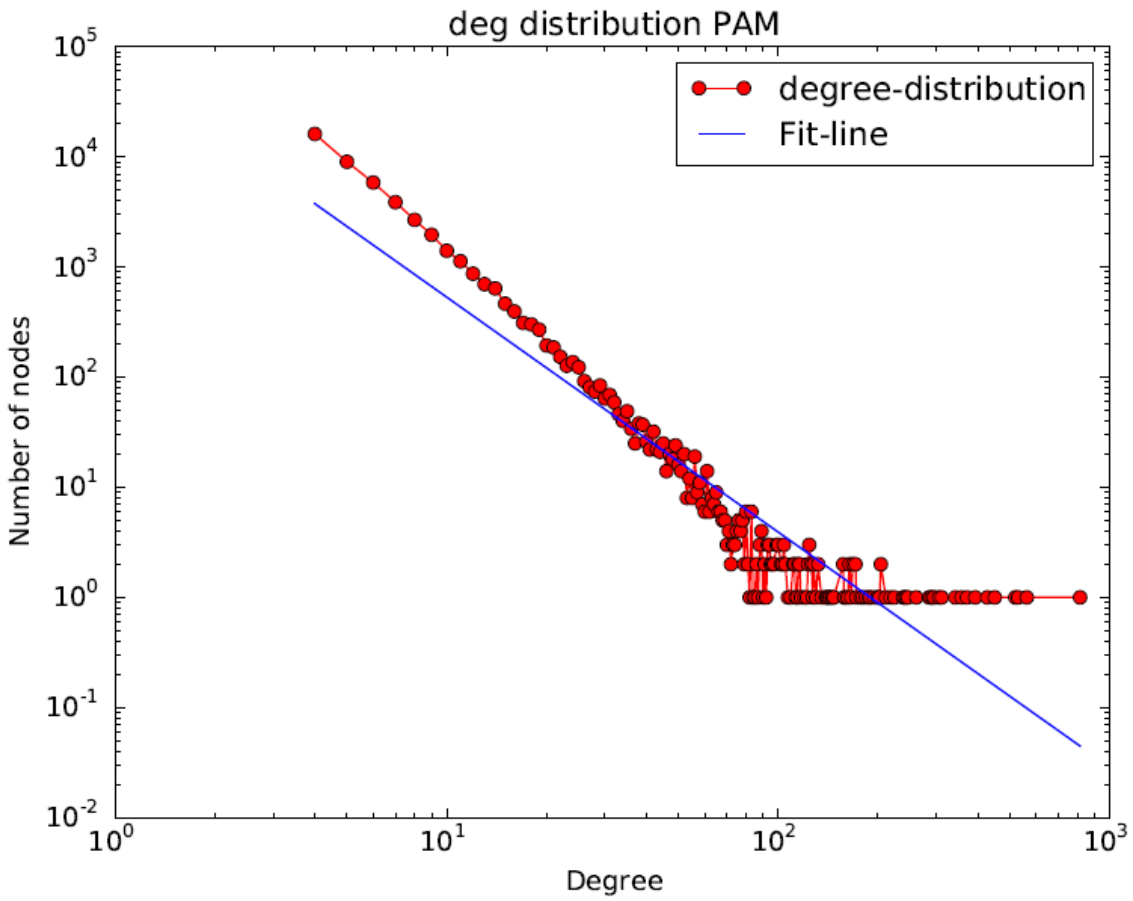
1. Random World Model
    a. Used G(n,m) model with n = 48332 and m = 85518
    b. Average path length: 5.342455
    c. Local clustering coefficients: 0.003868797586827
    d. Global clustering coefficients: 0.004720523650654
    e. Degree Distribution is as follows (doesn't follow power law)



2. For Small World data
    a. with p as : 0.454481912238
    b. Average path length : 9.14427701535
    c. ('Local clustering coefficients, 0.08701750721529827)
    d. ('Global clustering coefficients', 0.07205474716170415)
    e. Degree Distribution is as follows (doesn't follow power law)



3. For PAM data
    a. Average path length : 4.37092121441
    b. ('Local clustering coeffiecent', 0.001686007124919321)
    c. ('Global clustering coeffiecent', 0.0010511241540651708)
    d. Degree Distribution is as follows (follows power law)



4. Comparison

	Real	Random	Small World	PAM
Average path length	4.25215974305	5.342455	9.14427701535	4.37092121441
Local clustering coefficient	0.05421869518727754	0.003868797586827	0.08701750721529827	0.001686007124919321
Global clustering coefficient	0.14392364942004932	0.004720523650654	0.07205474716170415	0.0010511241540651708
Power Law Distribution	Yes	NO	NO	YES