P0 - Crawling the website

I crawled the website meetup.com using meetup APIs provided on the site (http://www.meetup.com/meetup_api/). I got an undirected graph with number of nodes 1547 and number of edges 367109.

Inside Folder phase_1, phase_1.py is the code used to crawl meetup.com. The following summary lists the output files obtained as part of crawling the site.

Summary:

phase_1.py - code used to crawl meetup.com

Inside phase_1_output,

anonymized_edge_list.txt - Contains anonymized list of edges separated by commas.

anonymized.txt - Contains the mapping between actual member ids and anonymized ids.

edge_list.txt - Contains the edge list obtained through crawling.

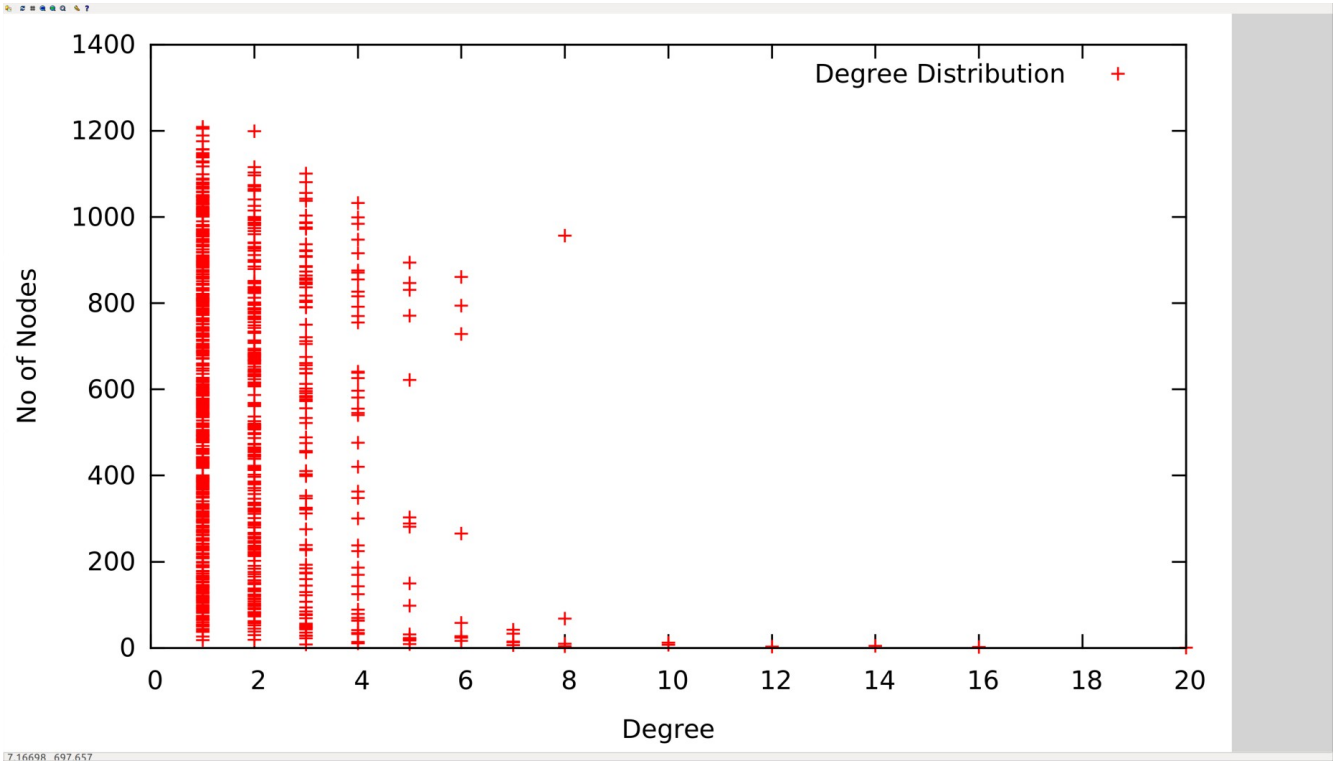
anonymized_sampled.txt - Contains anonymized sampled nodes.

sampled.txt - Contains sampled nodes.

For P1 to P3, I used Snappy/NetworkX for constructing and analyzing graphs using python.

Phase - 2 (Graph Essentials)

Degree Distribution (Plotted using gnuPlot)



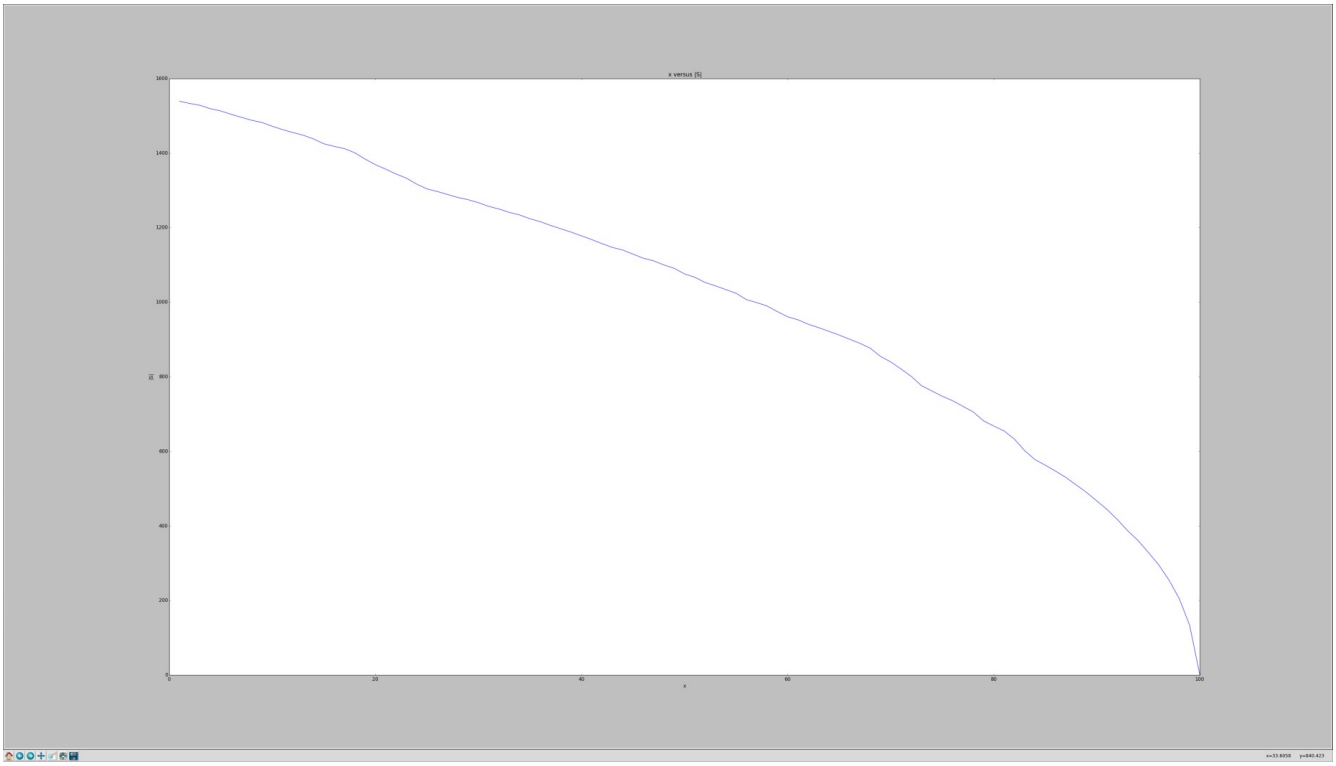
From the graph, we can see that the fitted line is: y = 1200*x^(-1.556).

(ii) The number of bridges in the graph is 20

(iii) The number of 3-cycles is 182428455

(iv) The graph diameter is 5

(v) The following graph has x(decrease of edges in %) along the x axis, and |S|, the size of the largest connected component on the y-axis. The graph shows that the size of the largest connected component decreases with the decrease of edges and becomes zero when x attains 100%. The graph is plotteed using pyplot.



Phase – 3 (Network Measures)

P2-Network Measure

(i) Average Local Clustering Coefficient = 0.796833600871 . The average clustering coefficient is calculated using the Snappy tool, computing the local coefficients of all nodes and finding the average.

Global Clustering Coefficient = 0.796833600871 . This is also calculated using Snappy for the graph.

(ii) Page Rank (Top 10 nodes): Listing node id and centrality value.

1789	0.00167878794644
698	0.0015847141404

1674	0. 00158471414049
978	0. 00157942732126
1549	0. 001578848835
748	0. 00154370158164
1262	0. 00153061253476
1830	0. 00152771313445
1368	0. 00152227237629
2260	0.00149938440648

(iii) Eigen Vector Centrality: Listing node id and centrality value of top ten nodes.

978	0. 0444720802666
1789	0. 0443570037503
698	0. 0443330984806
1674	0. 0443330984806
1830	0. 0441259243327
1549	0. 0441151407691
1123	0. 0438447050678
748	0. 043840619322
2058	0. 0437975288598
1368	0. 0436487925712

(iv) Degree Centrality: Top ten node ids and their centrality values

978	0. 784605433376
1789	0. 783311772316
698	0. 778783958603
1674	0. 778783958603
1549	0. 771668822768
1830	0. 763906856404
748	0. 752263906856

1368	0. 752263906856
2058	0. 747089262613
1123	0. 743855109961

(iv) Nodes having maximum Jaccard similarity: 698,1674

These nodes share the same neighbors and have Jaccard similarity as 1.

P4 – Network Models

(i) random graph

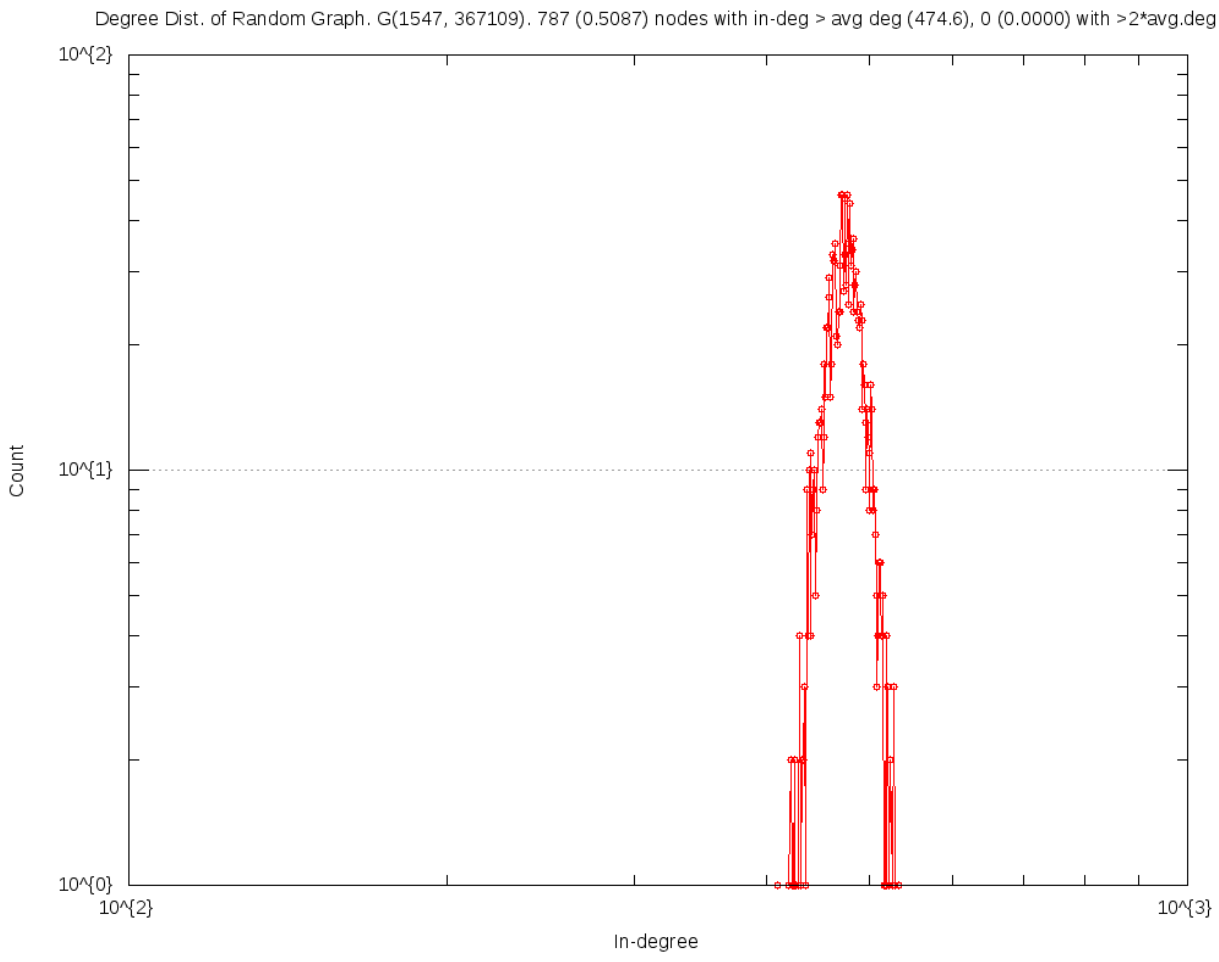
For random graph, Snap provide the Gnm model, n =1547, m = 367109

Average path length: 1.85560845495

Global clustering coefficient: 0.307078116755

Local clustering coefficient: 0.307078116755

The degree distribution is shown below. Since it is an undirected graph, the indegree distribution is the degree distribution. As we can see, it doesn't follow power-law distribution.



(ii) small world model

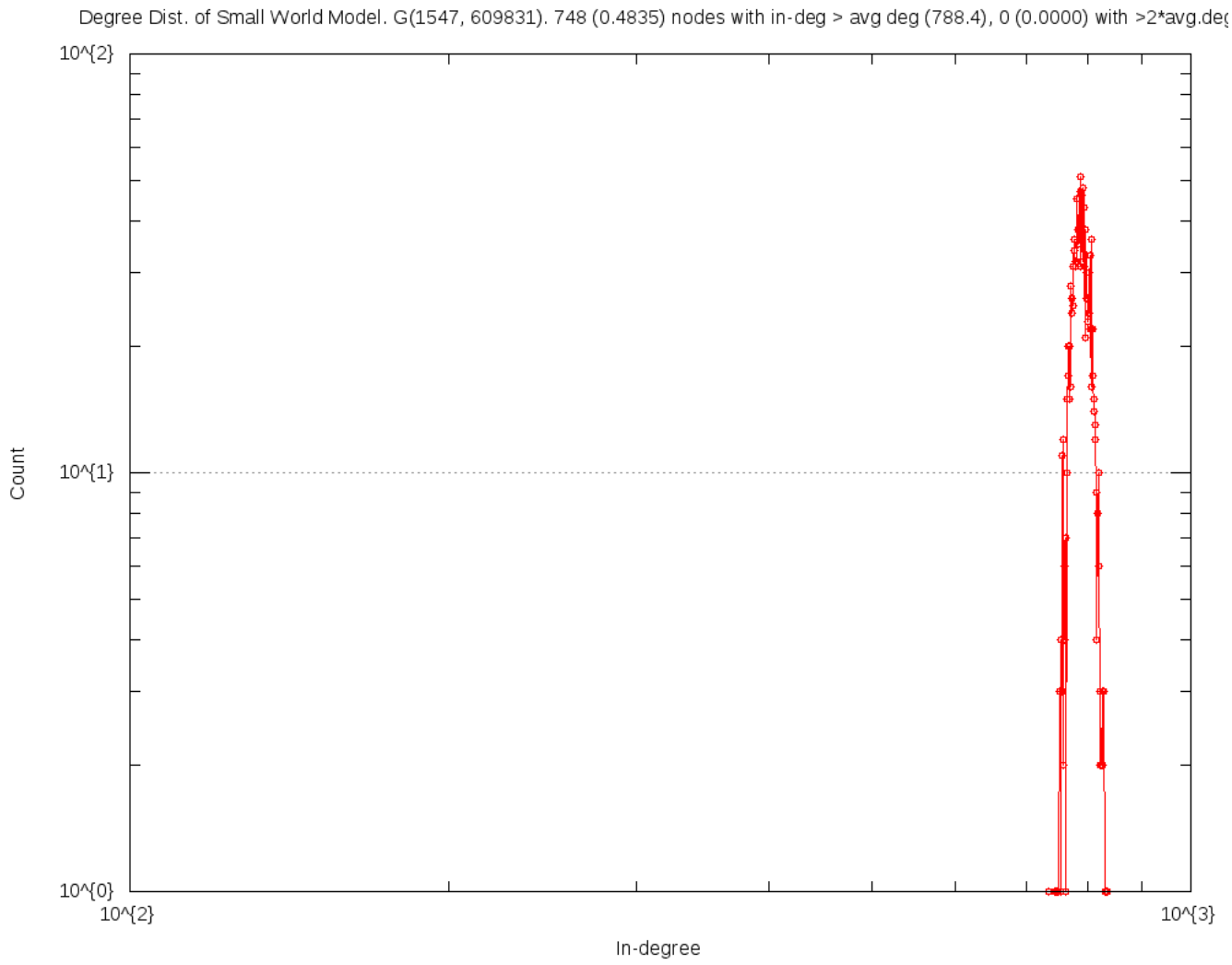
For small world model, I set n to be 1547 ,C(0) = 3/4 *(c-2)/(c-1), and I get p = 0.748414376321. Average degree(c) is 474 in the input graph. Typically, p is between 0.01 to 0.1.

Average Path Length = 1.79580127986

Average Local Clustering Coefficient = 0.510348538672

Global Clustering Coefficient = 0.510348538672

The degree distribution is as follows for the small world model graph and it doesnt follow power law.



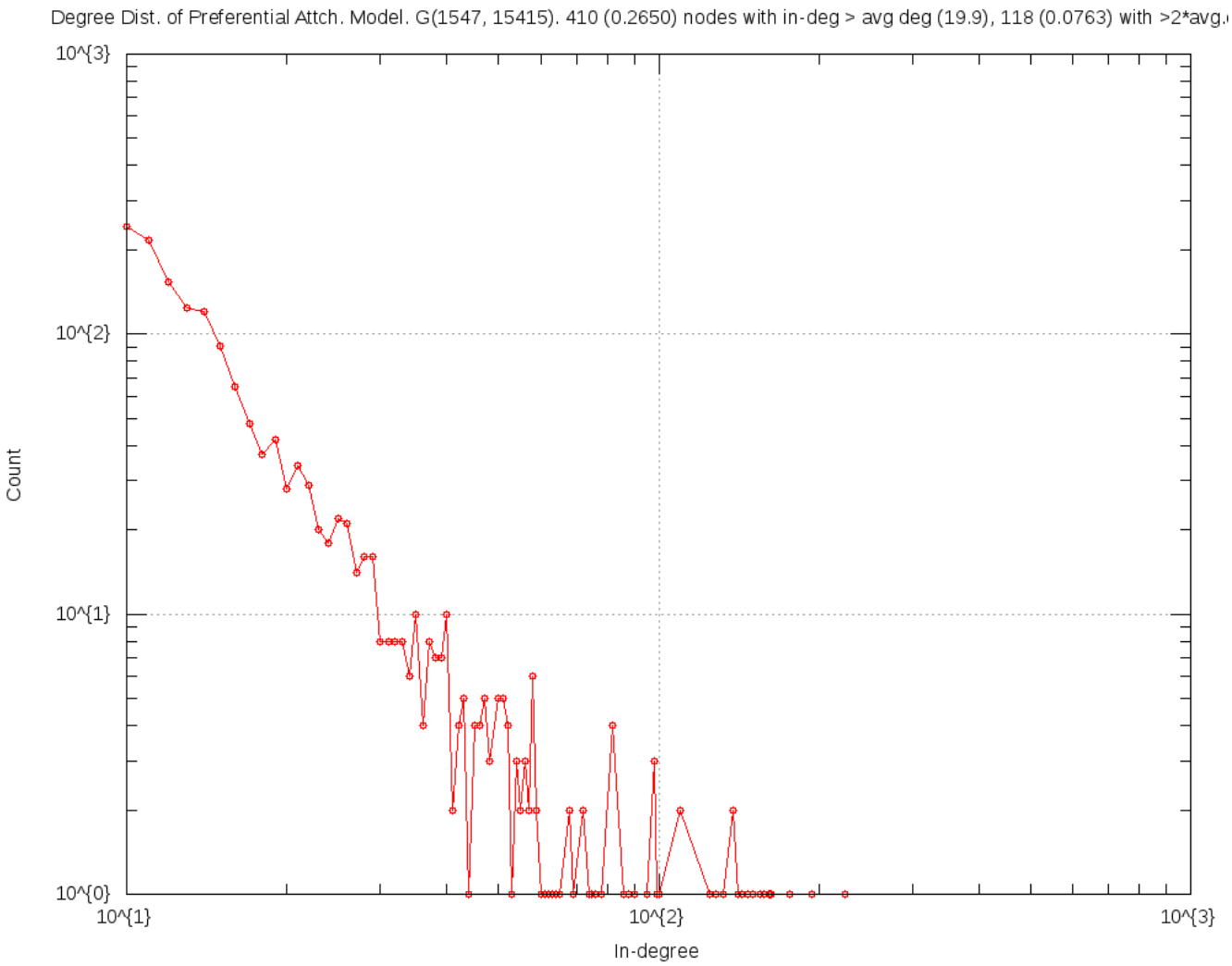
(iii) Preferential Attachment Model

Average Path Length = 2.85467590883

Average Local Clustering Coefficient = 0.0471289887629

Global Clustering Coefficient = 0.0471289887629

The average degree of the input graph is 474 and this is used as a parameter to the model generated. The degree distribution of the preferential attachment model is as follows: This follows power law distribution.



Comparison:

	Average Path Length	Clustering Coefficient	Degree Distribution
Original graph	1.93121254624	0.796833600871	Power-law
Random graph	1.85560845495	0.307078116755	Not power-law
Small world graph	1.79580127986	0.510348538672	Not power-law
Preferential Attachment	2.85467590883	0.0471289887629	Power-law

It can be seen from the table that all the models(random, small world and preferential attachment) have

low average path length similar to original graph. The small world graph has similar clustering coefficient to that to original graph though the random graph and preferential attachment model differ widely from the original group with respect to clustering coefficient.