SMM Project 1

Submitted By: Bhuvnesh Singh

Phase1(Crawl):

I have crawled www.meetup.com using api provied by meetup and generated dataset using breadth first search algorithm. I got an undirected graph of :

Nodes: 8986

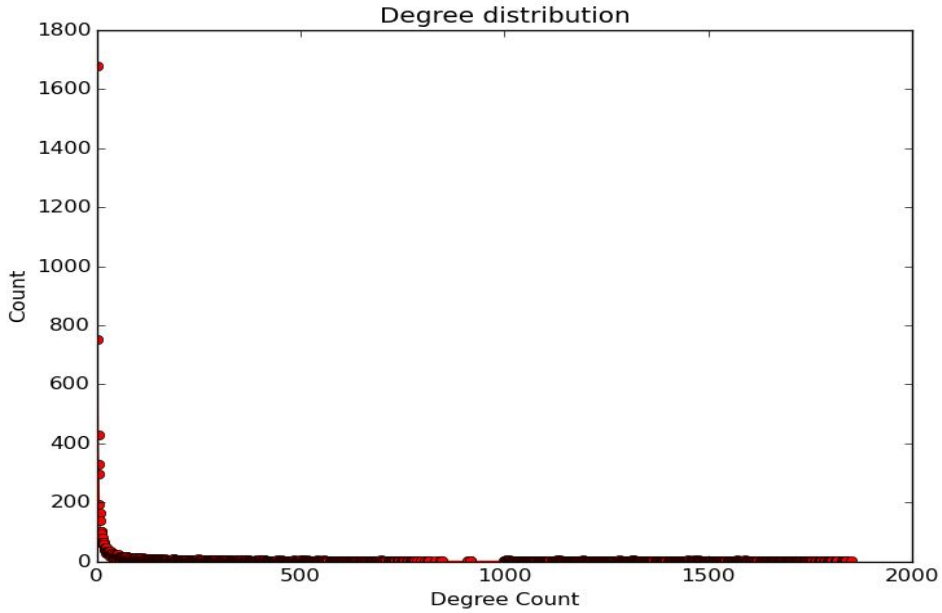
Edges 1000482

For phase 1 i have put all the code in p0 folder and it has all code with file name(phase1.py) related to crawling part and it also has all files related to data sets.Nodes data set is stored in anonymized.csv it has mapping of member_id to numbers(integer) **AND** edge_list.csv it has all edges which are explored yet. And then we have sampled.csv it has sampled.csv and sampled_edge_list.csv. And also it has anonymized_edge_list.csv it has anonymized edges information.And for each phase it has one folder containing python script and data sets.

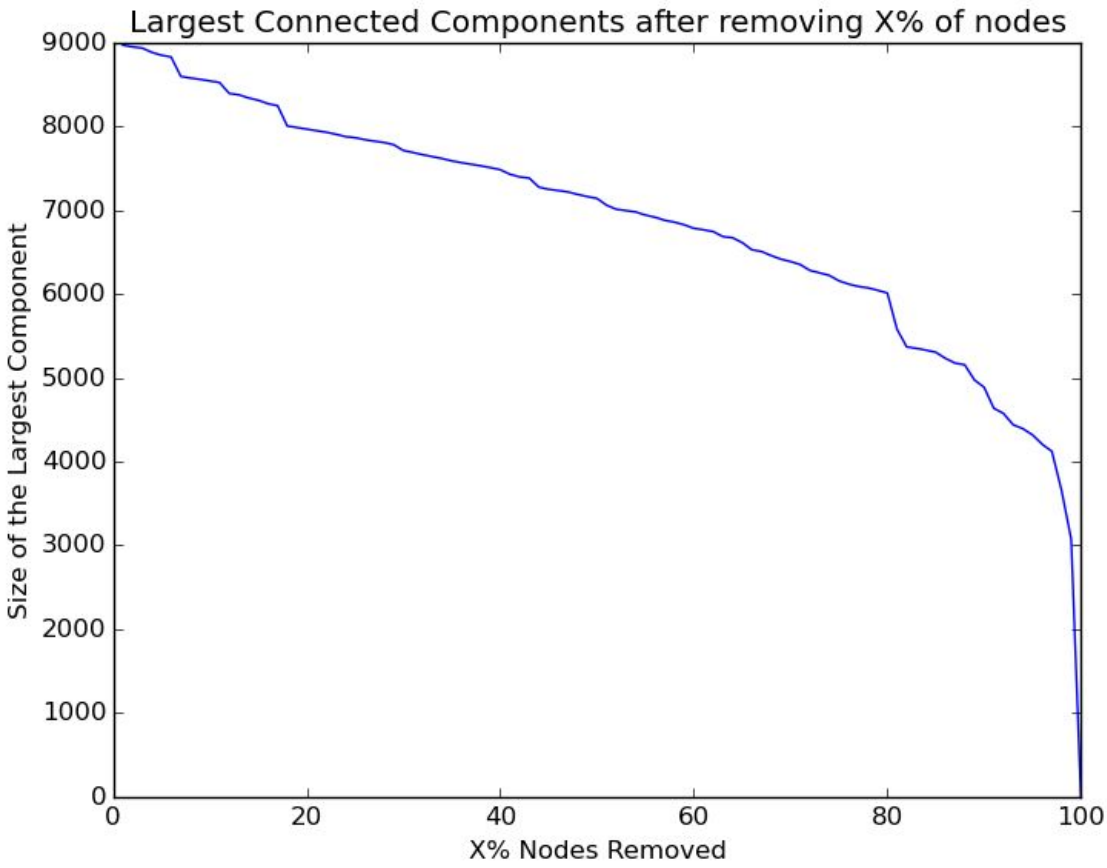
Summary:

phase1.py: code used to crawl meetup and generate datasets
Meetup output: non-anonymized dataset
anonymized.csv : member_id->numbers
anonymized_edge_list.csv: edge list i,j with i < j
edge_list.csv: edge list i,j and j,i {here i, j are member_id}
sampled.csv: it has member_id to degree
anonymized_sampled.csv

Phase2(Used Networkx)



1) From graph we can say it is following power law distribution .
   And for the power law exponent it y = 6000*x^(-1.3)
2) Number of bridges: 1678
3) The Number of 3-cycles : 479915088
4) Graph Diameter: 4
5) I have plotted graph using pyplot for the largest component to percentage of nodes removed.



Phase3:

- Average Global Clustering Coefficient: 0.35839030134
- Local Clustering Coefficient: 0.38756282015

- Page Rank:

Node	Page Rank	Member ID	Connections
2791	0.00645057848084	158803832	42

234	0.00581952808277	183386621	315
628	0.00468332045137	185540941	738
3509	0.00386021421907	185036848	1138
385	0.00382700960393	188359159	1109
3132	0.00381144477408	176950922	1158
1516	0.00324091257013	182636602	1057
2813	0.0032145162858	190876563	1154
2221	0.00239668940849	10139950	1126
156	0.00223866236221	190963503	1008

- Eigen Vector Centrality

Node	Eigen Vector Centrality	Member ID	Connections
1822	0.0368691231406	12312982	1448
1827	0.0367992722694	183335500	1376
1809	0.0366318076407	184568519	180
1727	0.0365423747959	184250864	40
1924	0.0364288360677	190801121	52
2102	0.0364213817814	179526552	49
1714	0.0364052701579	10422111	12
1744	0.0363859297065	182560136	1309
2095	0.0363540250916	11490266	1349
2083	0.0363413618063	191268019	372

- Degree Centrality

Node	Degree Centrality	Member ID	Connections
2191	0.205787423484	153172902	1355

2239	0.204563160824	184942385	1448
2163	0.204229271007	14036016	1405
2241	0.204006677796	6519555	182
2224	0.203895381191	77400662	1571
2208	0.203784084585	180292612	270
2235	0.203450194769	192059349	1475
2083	0.202114635504	191268019	372
2263	0.201892042293	192570486	470
2132	0.201669449082	191251279	1553

Upon calculating these above mentioned values degree centrality works best in terms of connections as it represent maximum number of connections or friends.Degree centality also works best in terms of importance as it has highest centrality among all.Page rank also has very good number of connections as it nodes has soo many connections.And eigen vector centrality is more than page rank but it has few number of connections.

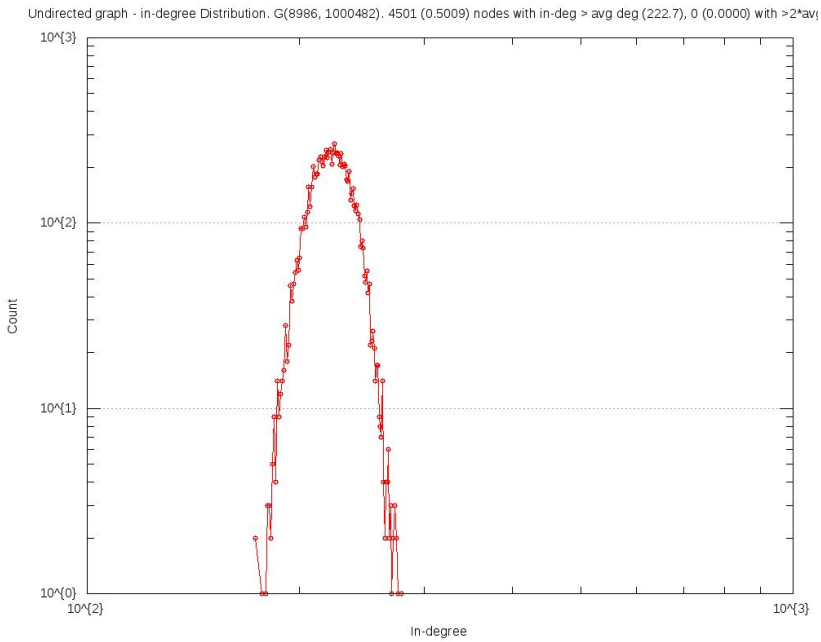
- I have selected **Jaccard similarity:**

  Max comes for nodes **1305** and **1403**
  and value of jaccard similarity is **0.8947368421052632**

Phase4(Used SNAP):

1) Random Graph:
   number of nodes : 8986
   number of edges : 1000482
   - Average Path Length : 1.9010415463806052
   - Global Clustering : 0.0247766985353

- Local clustering : 0.0284695839782



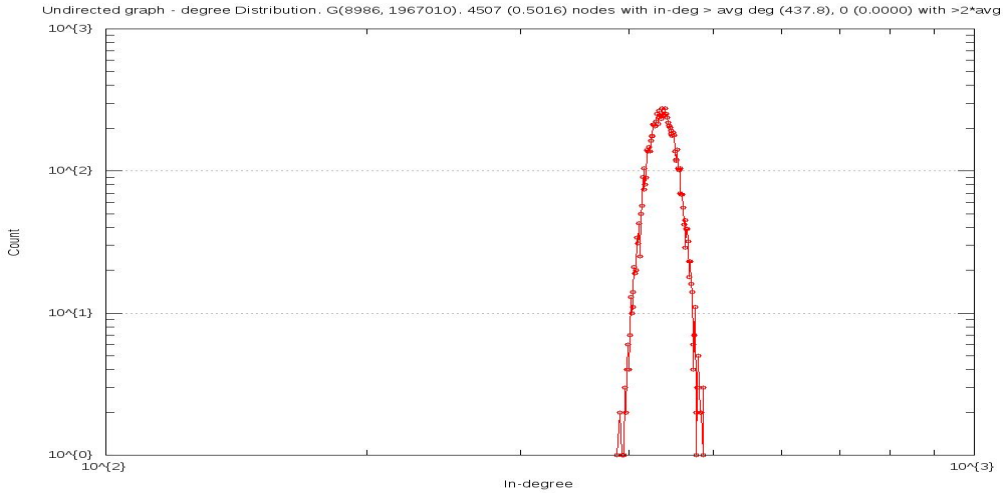
Above graph is of degree distribution and we can say that random graph also doesn't follow power law and it has small average path length.

2) Small World Model:

Number of nodes : 8986

For small world i have took 8986 nodes and as per formula a) C(p) = (1-p)^3C(0) and with average degree of 222, , C(0) = 3/4 *(c-2)/(c-1), so p comes out be 0.74

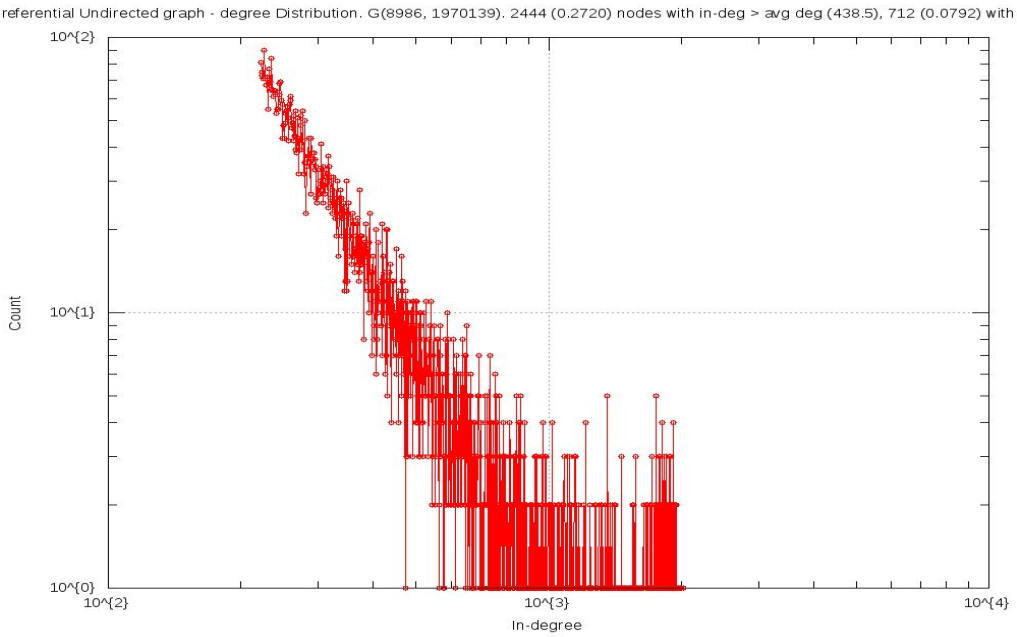
- Average path length: 1.89486622454
- Global clustering coefficient : 0.0599413159786
- Local clustering coefficient : 0.0599413159786



For small world in terms of degree distribution it also doesn't follow power law as we can see from above figure but it has clustering coefficient similar to real world and low average path length.

3) Preferential attachment : In this i have selected number of nodes to be 8986 and degree to be 222

- Average Path Length : 1.89486220194
- Global Clustering Coefficient : 0.119032756683
- Local Clustering Coefficient : 0.171284433056



As we can see from above figure that it follows power law distribution as real world model follows and it also doesn't have similar clustering coefficient and it also has small average path length.

	Average Path Length	Global Clustering Coefficient	Local Clustering Coefficient	Average Degree
Original	4	0.35839030134	0.38756282015	222
Random	1.901041546380	0.02477669853	0.028469583978	222
Small	1.89486622454	0.05994131597	0.059941315978	222
Preferential	1.89486220194	0.11903275668	0.171284433056	222