CSE 5245 Introduction to Network Science

Lab Assignment 1

BY Anirban Gupta and Meraj Ahmed Khan

1. Problem Statement

Compute centrality measures and clustering coefficients for the provided graphs, and perform meta-analysis on computed measures to characterize the different graphs.

2. Proposed Solution

1. Use the networkx library to compute the required measures.
2. Calculate correlation between all pairs of measures, and create following plots - Scatterplot Matrix for all measures, Degree - Rank plot.
3. Analyze the plots, and measures for insights into graph characteristics for each graph.

3. Centrality Measures

Centrality Measures are the indicators of a node's importance/relevance in a Graph. The idea being, a more important node is defined to be more central to the graph. There are different measures to describe a node's centrality in a graph, with only one or a combination of few of them them relevant to the problem domain or the graph structure.

We compute the following centrality measures.

1. Degree Centrality - The number of edges incident upon a node is the degree centrality.
    a. Indegree centrality - For a directed graph, we can separately define Indegree centrality, and Outdegree centrality. Indegree centrality of a node is the number of arrowheads incident upon a node in a directed graph.
    b. Outdegree centrality - Outdegree centrality of a node in a directed graph is the number of edge or link tails adjacent to it.

2. Closeness centrality - Closeness centrality of a node is the inverse of its sum of shortest distances to all other nodes.

3. Betweenness Centrality - For a node, it is a measure of how many shortest paths between any two pair of nodes in the graph pass through it.
4. Eigen vector Centrality - It is the measure of influence of a node in a network.
5. Pagerank Centrality - This is a variation of Eigen vector centrality with a way to model behavior of a random surfer on web, using a damping factor.

4. Clustering Coefficient

Clustering coefficient on a global level is an indicator of how clustered are the database nodes in the graph. Clustering coefficient for a node is a measure of how close it is to form a clique with its neighbors.

5. Experiments

5.1 Dataset

We performed our computations, and analysis on the following four graphs
Directed Graphs - Gnutella-p2p(Nodes: 6301, Edges: 20777), Wiki-Vote(Nodes: 7115, Edges:103689)
Undirected Graphs - Facebook(Nodes:4039, Edges: 88234), Collaboration Network - Quantum Cosmology(Nodes: 5242, Edges: 14496)

5.2 Assumptions and Adjustments

1. We calculate clustering coefficients for all graphs by considering them as undirected.
2. In cases where the graph is not connected, we report harmonic centrality instead of closeness centrality.
3. For directed graphs, we report indegree and outdegree along with the degree centrality.

5.3 Results

We report the pairwise correlations between all graph measures for each graph, and plot them on scatterplot matrices.

	Degree	Page Rank	Closeness	Betweeness	Eigen	Clustering Coefficient
Degree	1.00	0.67	0.27	0.45	0.57	-0.13
Page Rank	0.67	1.00	0.15	0.77	0.08	-0.27
Closeness	0.27	0.15	1.00	0.14	-0.08	-0.18
Betweeness	0.45	0.77	0.14	1.00	0.02	-0.12
Eigen	0.57	0.08	-0.08	0.02	1.00	0.18
Clustering Coefficient	-0.13	-0.27	-0.18	-0.12	0.18	1.00

Table 1: Correlation between Graph Measures – Facebook

	Harmonic	Outdegree	Degree	Indegree	Page Rank	Betweeness	Eigen	Clustering Coefficient
Harmonic	1.00	0.20	0.53	0.74	0.63	0.29	0.69	0.20
Outdegree	0.20	1.00	0.87	0.32	0.25	0.58	0.30	0.01
Degree	0.53	0.87	1.00	0.75	0.66	0.69	0.71	0.05
Indegree	0.74	0.32	0.75	1.00	0.92	0.55	0.94	0.08
Page Rank	0.63	0.25	0.66	0.92	1.00	0.53	0.85	0.05
Betweenness	0.29	0.58	0.69	0.55	0.53	1.00	0.51	0.00
Eigen	0.69	0.30	0.71	0.94	0.85	0.51	1.00	0.09
Clustering Coefficient	0.20	0.01	0.05	0.08	0.05	0.00	0.09	1.00

Table 2: Correlation between Graph Measures – Wiki-Vote

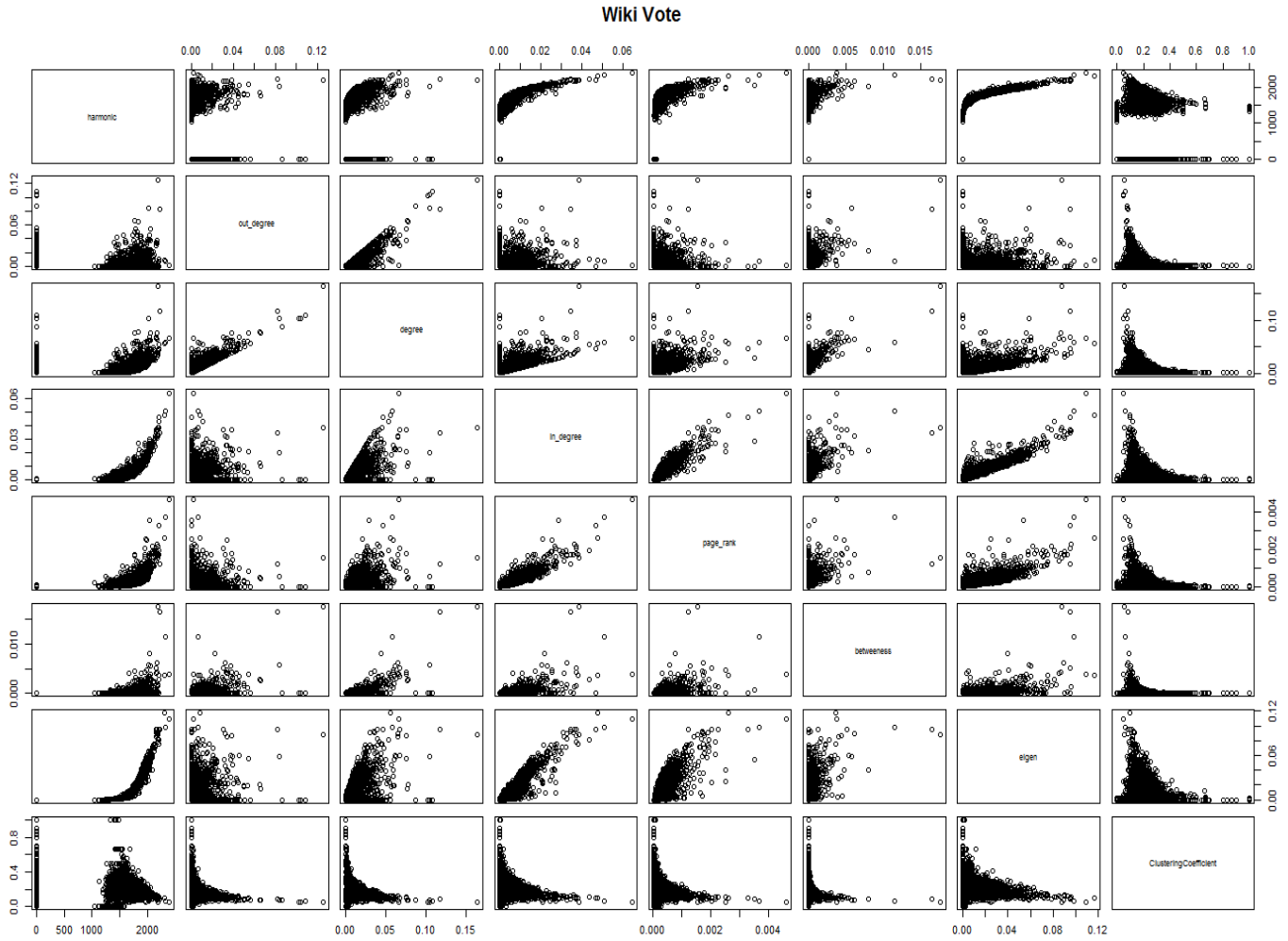
	Harmonic	Degree	Page Rank	Betweeness	Eigen	Clustering Coefficient
Harmonic	1.00	0.37	0.17	0.31	0.16	0.09
Degree	0.37	1.00	0.70	0.50	0.60	0.09
Page Rank	0.17	0.70	1.00	0.78	0.32	-0.16
Betweenness	0.31	0.50	0.78	1.00	0.14	-0.22
Eigen	0.16	0.60	0.32	0.14	1.00	0.06
Clustering Coefficient	0.09	0.09	-0.16	-0.22	0.06	1.00

Table 3: Correlation between Graph Measures – GR-QC

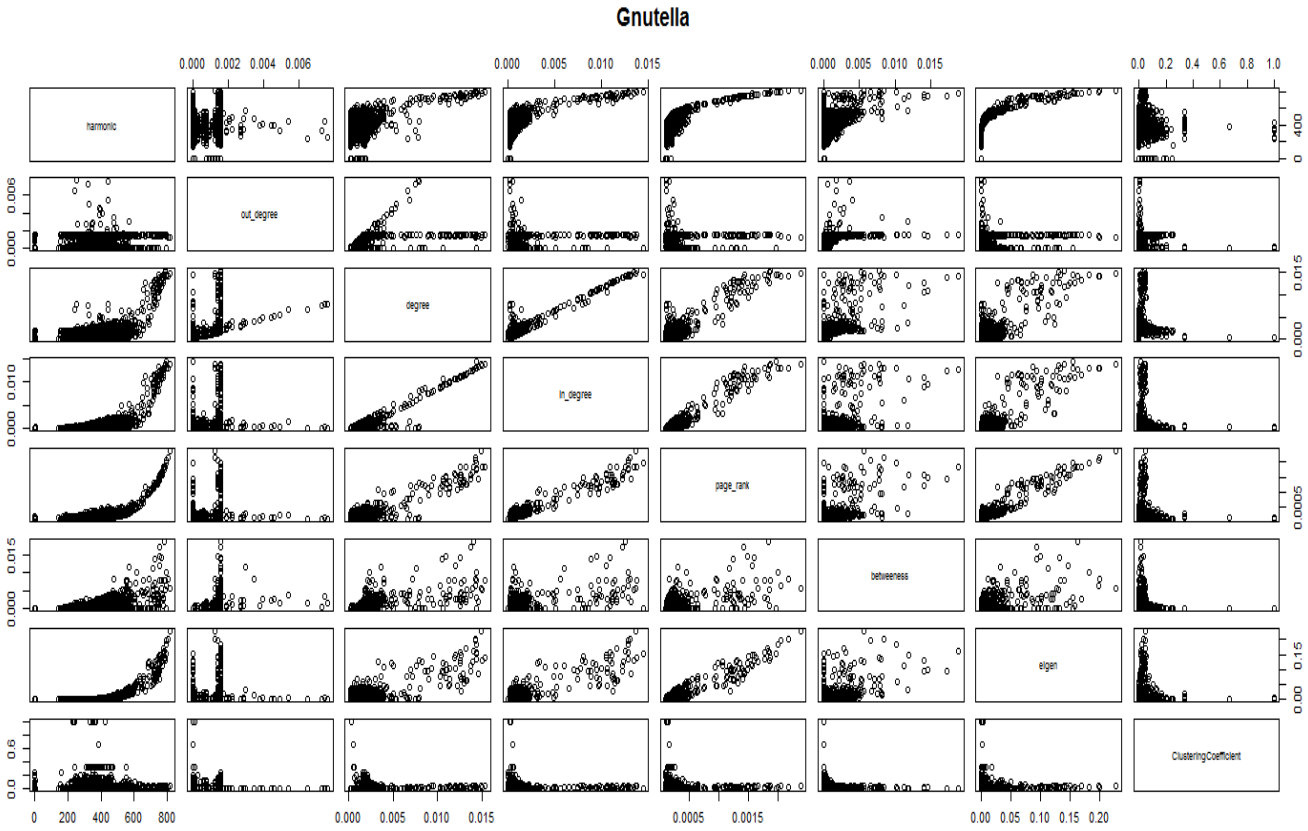
	Harmonic	Outdegree	Degree	Indegree	Page Rank	Betweeness	Eigen	Clustering Coefficient
Harmonic	1.00	0.07	0.43	0.51	0.54	0.38	0.52	0.05
Outdegree	0.07	1.00	0.65	0.13	0.11	0.47	0.09	0.11
Degree	0.43	0.65	1.00	0.84	0.79	0.70	0.69	0.09
Indegree	0.51	0.13	0.84	1.00	0.95	0.58	0.84	0.04
Page Rank	0.54	0.11	0.79	0.95	1.00	0.58	0.93	0.04
Betweenness	0.38	0.47	0.70	0.58	0.58	1.00	0.55	0.02
Eigen	0.52	0.09	0.69	0.84	0.93	0.55	1.00	0.04
Clustering Coefficient	0.05	0.11	0.09	0.04	0.04	0.02	0.04	1.00

Table 4: Correlation between Graph Measures – Gnutella

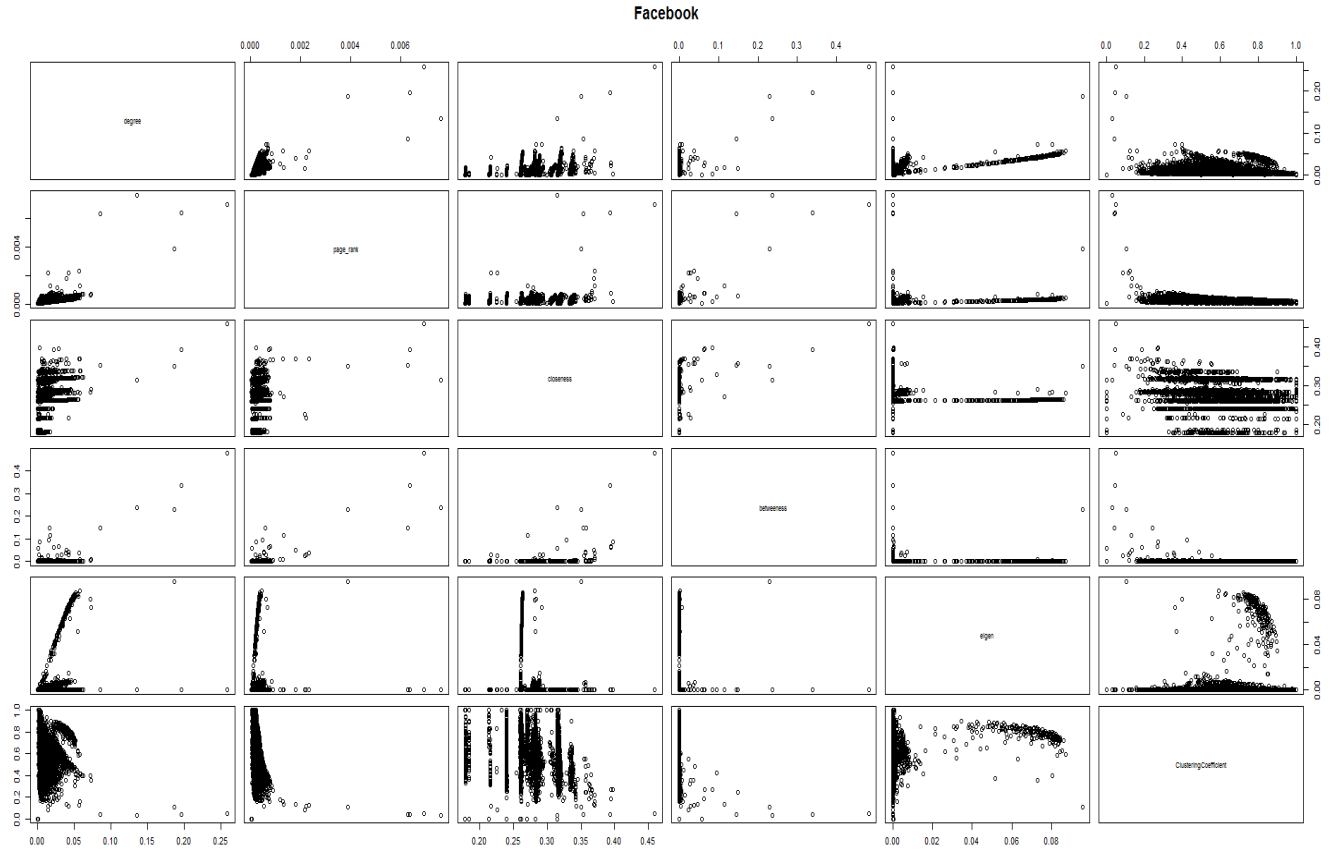
Scatterplot Matrices



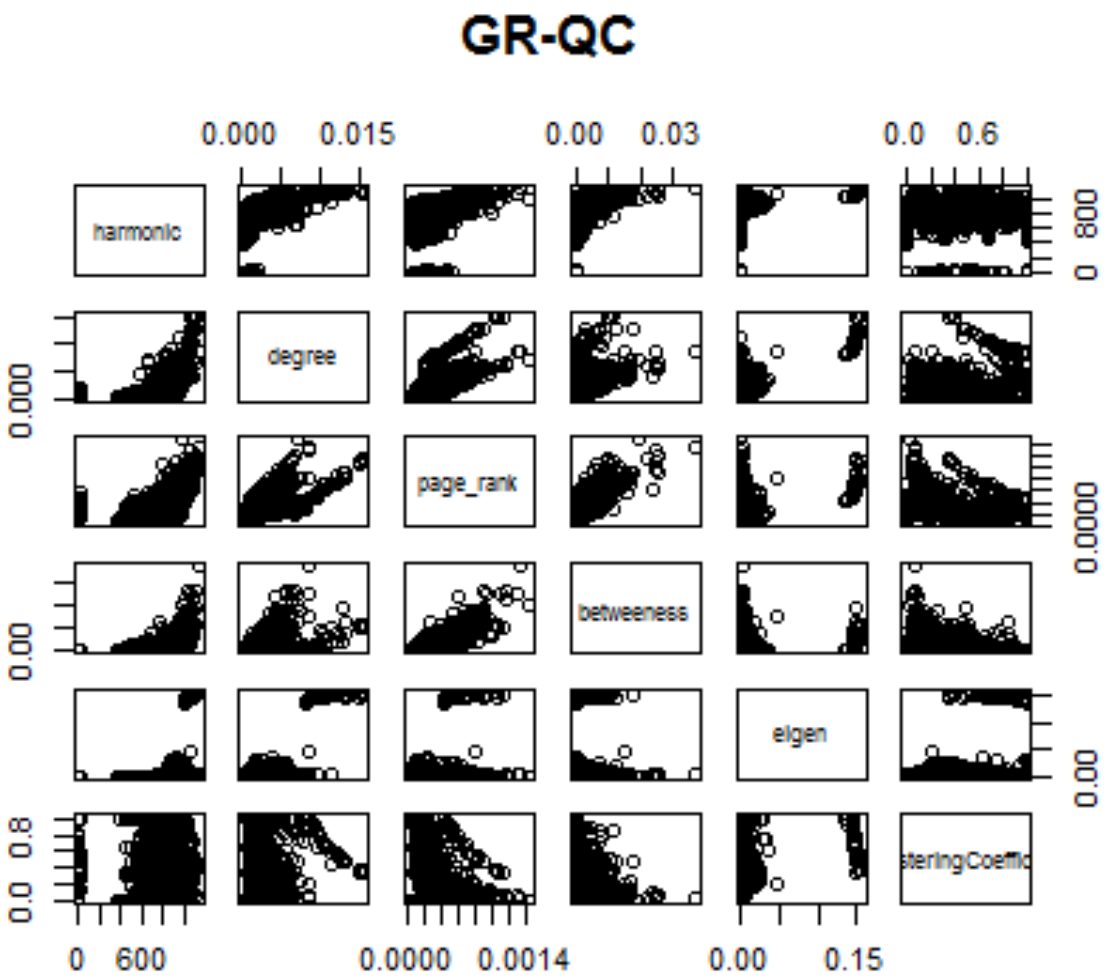
Axis Order - Harmonic Centrality, Outdegree, Degree, Indegree, Page Rank, Betweenness, Eigen, Clustering Coefficient



Axis Order - Harmonic Centrality, Outdegree, Degree, Indegree, Page Rank, Betweenness, Eigen, Clustering Coefficient



Axis Order – Degree, Page Rank, Closeness, Betweenness, Eigen, Clustering Coefficient



Axis Order – Harmonic, Degree, Page Rank, Betweenness, Eigen, Clustering Coefficient

6. Observations and Insights

6.1 Clustering Coefficients Analysis

The average clustering coefficients for the networks were

- Wiki-vote: `0.140897845893`
- Facebook: `0.60554671862`
- Gnutella: `0.0108679219358`

- GR-QC: `0.529635811052`

Observing the average clustering coefficients of the networks, one can see that the values are high for Facebook and GR-QC, but it's low for Wiki-vote and Gnutella. This shows that friends and authors are more likely to cluster in networks respectively, in comparison to voters and peers. There are several possible explanations for this, such as,

- Friends are more likely to rely on their interactions with other friends, in order to form new friends. This observation matches with the notion of friendship in real world. The co-author network shows how authors collaborate with same group of people for multiple works, and presence of sub-domains in the research area.
- As Gnutella is a p2p network, with it's low average clustering coefficient, one can guess it could be, due to it being more bipartite. Low clustering for wiki-vote shows the people could vote for anyone (and there does not exist any such party to vote for), thus leading to hardly any clustering. This would be the reverse, if the voting system during elections is observed.

6.2 Correlation Analysis

- PageRank has a high positive correlation with Degree, Betweenness, and Harmonic Centrality for all networks.
- PageRank has an enormously high positive correlation with indegree for directed networks.
- Clustering coefficient is pretty much independent of all centrality values.
- Eigen Vector shows very weak to almost no correlation with PageRank for the two undirected networks. Most likely this is just a coincidence.

6.3 Average Pearson Correlation Coefficient

The average degree Pearson Correlation coefficients for the networks were

- Wiki-vote: `-0.0832445577169`
- Facebook: `0.0635772291856`
- Gnutella: `-0.0285339698514`
- GR-QC: `0.659164032093`

This coefficient (ranging from -1 to +1), indicates if there exists correlation between two nodes. If the coefficient is positive, it indicates correlation between nodes of similar degree, and if it's negative, it indicates correlation between nodes of different degrees [1].

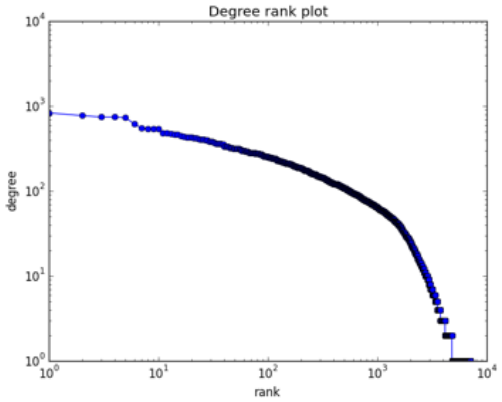
Observing the coefficients, one can notice that Facebook and GR-QC have positive degree Pearson correlation coefficients, but Gnutella and Wiki-vote have negative degree Pearson correlation coefficients. This is interesting to note as,

- This shows that there's a tendency of nodes with high degrees connect to other nodes with high degrees, and this is similarly found for nodes with low degrees, in the case of social networks (such as Facebook, or a co-author network).

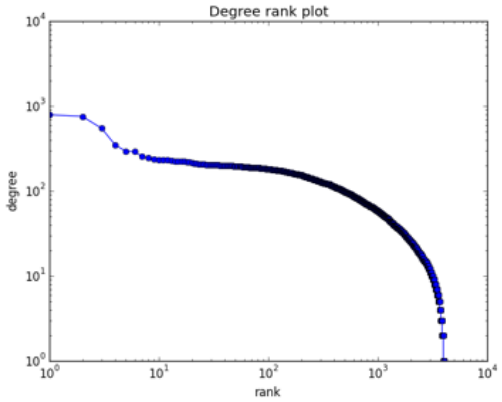
- However, there's a tendency of nodes with high degrees connecting to low degrees and vice versa in networks involving votes or a p2p network. This makes sense, as a large no of people (with low degrees), would vote for one person (who would tend to have a higher degree), thus making him popular, and ensuring more connections between him/her and others with lower degrees voting for that person. There also seems to be a correlation between the average degree coefficient and the average pearson degree correlation coefficient, but we aren't sure if this is true for all networks.

6.4 Degree Rank Plot

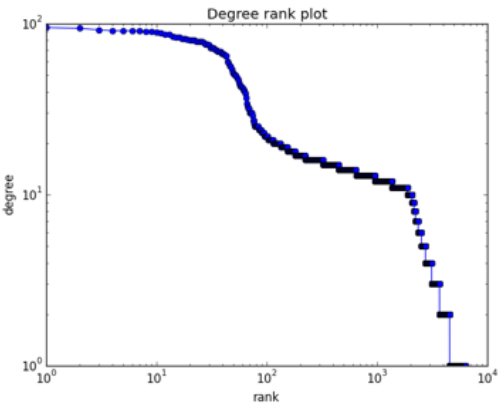
We plotted the degree histogram for the various networks (we converted the networks to an undirected network for the sake of simplicity).



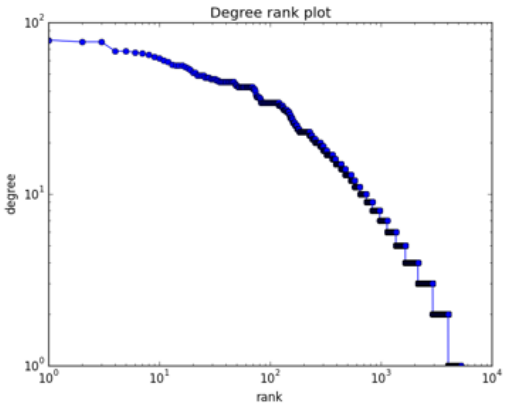
Wiki Degree Histogram



Facebook Degree Histogram



Gnutella Degree Histogram



GR-QC Degree Histogram

As one can observe,

- Nodes with smaller degrees are most frequent

- Fraction of highly connected nodes decreases (but is never zero)
- We used a log-log plot for easier scaling
- Power law degree distributions are observed in all networks.

7. Work Split up

Anirban Gupta – Graph Measure Calculations, Degree Rank Plot Analysis, Average Pearson Correlation Coefficient Analysis

Meraj Ahmed Khan – Graph Measure Calculations, Correlation Analysis

References

[1] - https://en.wikipedia.org/wiki/Assortativity

Packages used

The following packages were used:

- networkx: Extraction of graph information from data, performing calculations and drawing the graph
- matplotlib: Drawing the graph and scatterplots.
- csv: For storing the different centrality measures for comparison and correlation purposes.