CS529 Data Mining

Assignment-I Phase-II

Group 11

Ayush Jain 150101012

Mukul Verma 150101038

Shubhanshu Verma 150101073

Introduction

This report shows coefficient analysis of the well-known centrality metrics (degree centrality, eigenvector centrality, betweenness centrality, closeness centrality, and clustering coefficient) for network analysis studies on real-world network graphs representing diverse domains (ranging from 61 nodes to 10410 nodes).

-We use the Spearman correlation for correlation analysis which is a nonparametric measure of the monotonicity of the relationship between two datasets. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

Correlation Analysis

- The centrality metrics obtained and the metrics from the library(Networkx) functions, as expected, have correlation values close to 1.0.

  These are the corresponding comparisons of the correlation of implemented centralities with the library functions:

  DBLP Co-authorship network:

Centrality 1	Centrality 2	Correlation factor
Degree Centrality	Degree Centrality	1.0
Closeness Centrality	Closeness Centrality	0.99999999999999978
Betweenness Centrality	Betweenness Centrality	1.0
Eigen value Centrality	Eigen value Centrality	1.0
Clustering Coefficient	Clustering Coefficient	1.0

Higgs Twitter:

Centrality 1	Centrality 2	Correlation factor
Degree Centrality	Degree Centrality	0.99999999999999978
Closeness Centrality	Closeness Centrality	1.0
Betweenness Centrality	Betweenness Centrality	0.99999952668061065
Eigen value Centrality	Eigen value Centrality	0.997553679440359
Clustering Coefficient	Clustering Coefficient	1.0

9/11 Hijackers:

Centrality 1	Centrality 2	Correlation factor
Degree Centrality	Degree Centrality	1.0
Closeness Centrality	Closeness Centrality	1.0
Betweenness Centrality	Betweenness Centrality	0.99999999999999978
Eigen value Centrality	Eigen value Centrality	0.999847036327977
Clustering Coefficient	Clustering Coefficient	0.9984057001929072

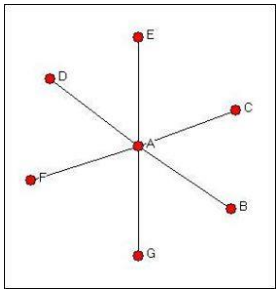
Critical Analysis

- We observe the degree and betweenness centrality measures to be highly correlated across all the networks studied. We get the following values for the analysed networks

  Higgs Twitter: 0.7339

  DBLP Co-authorship network: 0.8375

  9/11 Hijackers: 0.7777



To see this correlation, we can look at the star network where the central node has the highest degree also has the highest betweenness as more shortest paths pass through it whereas the adjacent nodes having degree 1 also have low betweenness.

- We expect that measures of degree and closeness centrality will be more highly correlated with each other than with other measures, because they are both based on direct ties. For the Higgs Boson twitter dataset, we got correlation value 0.83759

- There is a moderate level of correlation between the shortest paths-based centrality metrics (betweenness and closeness) and such a correlation is consistently observed across all the networks.

DBLP Co-authorship network:	0.1571
Higgs Twitter:	0.6962
9/11 Hijackers:	0.2779

- For large graphs, i.e. the Higgs Twitter and the DBLP CO-authorship dataset, we observe that degree and clustering are highly correlated. This can be seen by the fact that as degree of a node increases, it is more likely to form a triangle with the adjacent nodes than a node with less degree. To support this correlation, we can see in DBLP dataset, out of 10410 nodes around 9500 nodes have degree <4 most of these nodes have clustering coefficient zero.

- Though we observe a poor correlation between a degree based centrality metric and a shortest-path based centrality metric for regular random networks, as the variation in the degree distribution of the vertices increases (i.e., as the network gets increasingly scale-free), the correlation coefficient between the two classes of centrality metrics increases.

- We can observe that for small networks like the 9/11 Hijackers dataset, most correlation values are very low and for large and dense networks like the Higgs Twitter network, the correlation values are large. From this

is, we can say that as a network increases in size, the correlation between values increases.

- We observe a moderate level of correlation between Eigenvector and degree centrality The values calculated were:

  DBLP Co-authorship network: 0.1581

  Higgs Twitter: 0.5765

  9/11 Hijackers: 0.6630

  This can be seen as eigenvector centrality indicates how likely a node is to be visited in a network which can be correlated with the in-degree of a node.

- There is very low correlation observed in sparse graphs between closeness centrality and eigenvector centrality.

  DBLP Co-authorship network: 0.0452

  9/11 Hijackers: 0.0436

  Where as in relatively denser graph, the correlation was

  Higgs Twitter: 0.4831

  Although there is no intuitive relation between closeness and eigenvector centrality as closeness is path based and eigenvector is degree based centrality but the density of the given network maybe a correlating factor.

- For these datasets, we did not get any negative correlation values which means that for most networks, if some correlation value increases, other values are also likely to increase or remain constant.

Observations

DBLP Co-authorship network:

Centrality 1	Centrality 2	Correlation factor
Degree Centrality	**Closeness Centrality**	0.16766680183235003
Degree Centrality	**Betweenness Centrality**	0.83754188223398818
Degree Centrality	**Eigen value Centrality**	0.15812818181346877
Degree Centrality	**Clustering Coefficient**	0.72284111750208457
Closeness Centrality	**Betweenness Centrality**	0.15711988065151725
Closeness Centrality	**Eigen value Centrality**	0.045213934823175754
Closeness Centrality	**Clustering Coefficient**	0.091669376827837074
Betweenness Centrality	**Eigen value Centrality**	0.15335950057710057
Betweenness Centrality	**Clustering Coefficient**	0.34824911483989379
Eigen value Centrality	**Clustering Coefficient**	0.18995648397541381

Higgs Twitter network:

Centrality 1	Centrality 2	Correlation factor
Degree Centrality	Closeness Centrality	0.83755996090929663
Degree Centrality	Betweenness Centrality	0.73391097920909942
Degree Centrality	Eigen value Centrality	0.57656736543758513
Degree Centrality	Clustering Coefficient	0.62978940241619985
Closeness Centrality	Betweenness Centrality	0.69628759537438234
Closeness Centrality	Eigen value Centrality	0.48310260329297289
Closeness Centrality	Clustering Coefficient	0.5647841689897477
Betweenness Centrality	Eigen value Centrality	0.65178886868183183
Betweenness Centrality	Clustering Coefficient	0.62738847682046894
Eigen value Centrality	Clustering Coefficient	0.63434071484841459
Clustering Coefficient	Clustering Coefficient	1.0

9/11 Hijackers network:

Centrality 1	Centrality 2	Correlation factor
Degree Centrality	Closeness Centrality	0.25483432389777749
Degree Centrality	Betweenness Centrality	0.77776400086618147
Degree Centrality	Eigen value Centrality	0.6630708483835337
Degree Centrality	Clustering Coefficient	0.069185272920547117
Closeness Centrality	Betweenness Centrality	0.27797744250701983
Closeness Centrality	Eigen value Centrality	0.04363069900113356
Closeness Centrality	Clustering Coefficient	0.12323946498054709
Betweenness Centrality	Eigen value Centrality	0.55378812979036207
Betweenness Centrality	Clustering Coefficient	0.24431667555658701
Eigen value Centrality	Clustering Coefficient	0.060252801618782062