LAB ASSIGNMENT - 1 REPORT

CREDITS

AUTHORS : RAKSHITH KUNCHUM
COURSE INSTRUCTOR : SRINIVASAN PARTHASARATHY
DATE : 2/1/2016
CONTACT : kunchum.1@osu.edu

INTRODUCTION

In this assignment, we are provided with different directed and undirected network datasets. For each node in the graph different network characteristics have been computed and generated into an output file. Relation between different network characteristics for different network datasets has been explored.

PACKAGES USED

The following python packages were used for the specified tasks:

- NetworkX package on Python.
- Matplot library to plot the scatter plots
- Numpy library to compute the correlation.

INPUT DATA

Input data files contain the information of edges between nodes. Each line listed corresponds to nodes between whom an edge exist. Except for the Facebook dataset all the other datasets are directed graphs. Data sets provided are:

1. Wiki-vote.txt
2. p2p-Gnutella08.txt
3. CA-GrQc.txt
4. facebook-ego-net.txt

NETWORK PARAMETERS

Following are the network parameters that are being studied in this project.

DEGREE CENTRALITY

Two kinds of degree centrality is being studied: In-degree-centrality and out-degree-centrality. This distinction only in the case of directed graphs. In case of undirected graph we only have degree-centrality.

CLOSENESS CENTRALITY
It is the measure which tells us about the reachability of every given node in the network from a given node. If the graph has disconnected components then we are computing harmonic closeness centrality. The networkx function for computing closeness centrality by default measures harmonic closeness. Higher values of closeness indicate higher centrality.

BETWEENNESS CENTRALITY
Node betweenness is the number of shortest paths that pass through a given node.

EIGENVECTOR CENTRALITY
Eigenvalue of a given node establishes the idea of importance of a node in the whole of network based on its incoming and outgoing nodes.

PAGERANK CENTRALITY
Pagerank of a given node establishes the idea of ranking of a node based on the neighbours they are connected to. Nodes will be ranked high if they are connected to nodes that are important themselves.

CLUSTERING COEFFICIENT
It is the measure of interconnectivity of a node's neighbours. This coefficient relevant only in the case of undirected graphs.

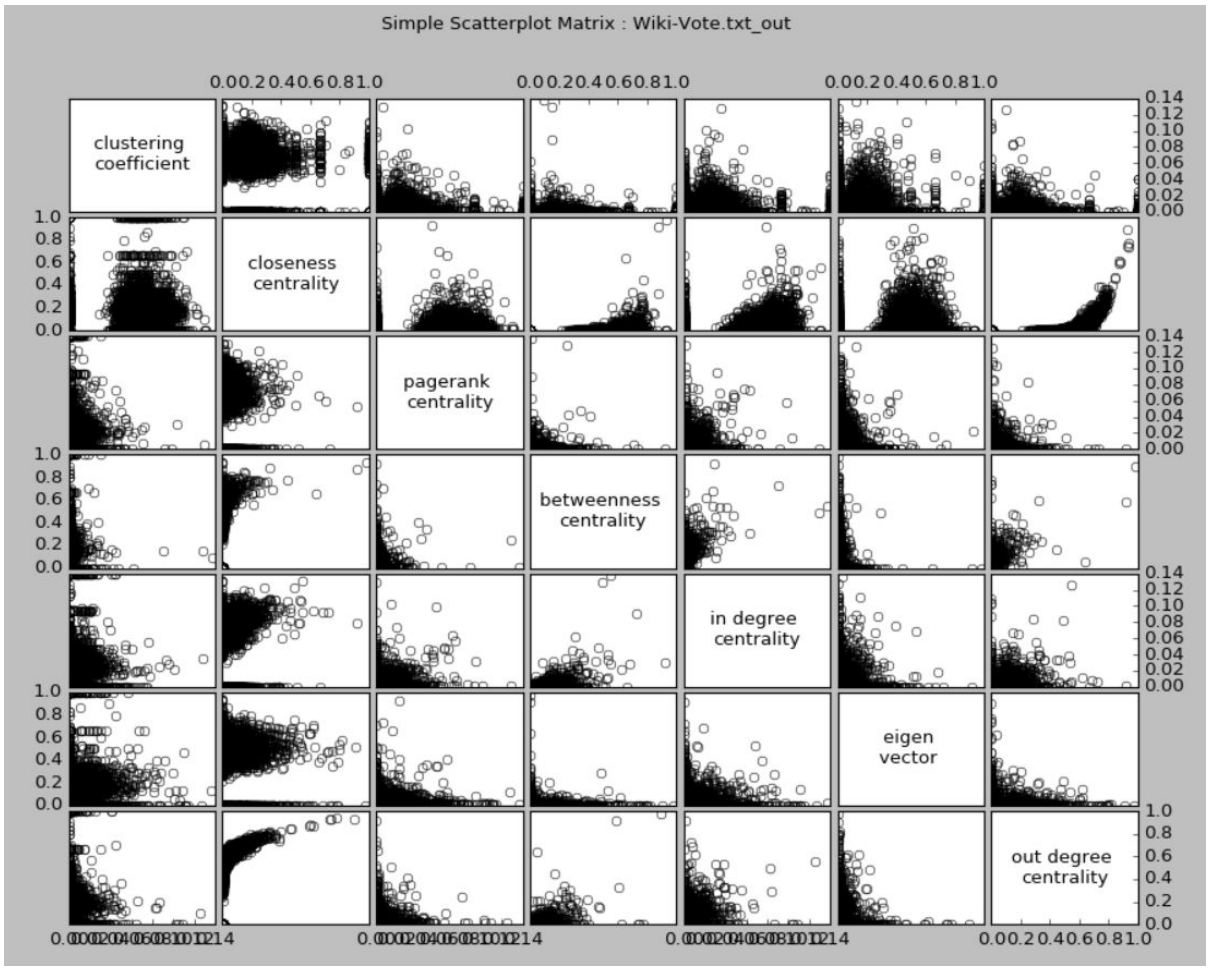
RESULTS AND OBSERVATION

DATASET : WIKI-VOTE

The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node i to node j represents that user i voted on user j. This is a directed graph.

The scattered plot for different characteristics is given on the next page.
From the correlation values and the scatter plot, we observe that :

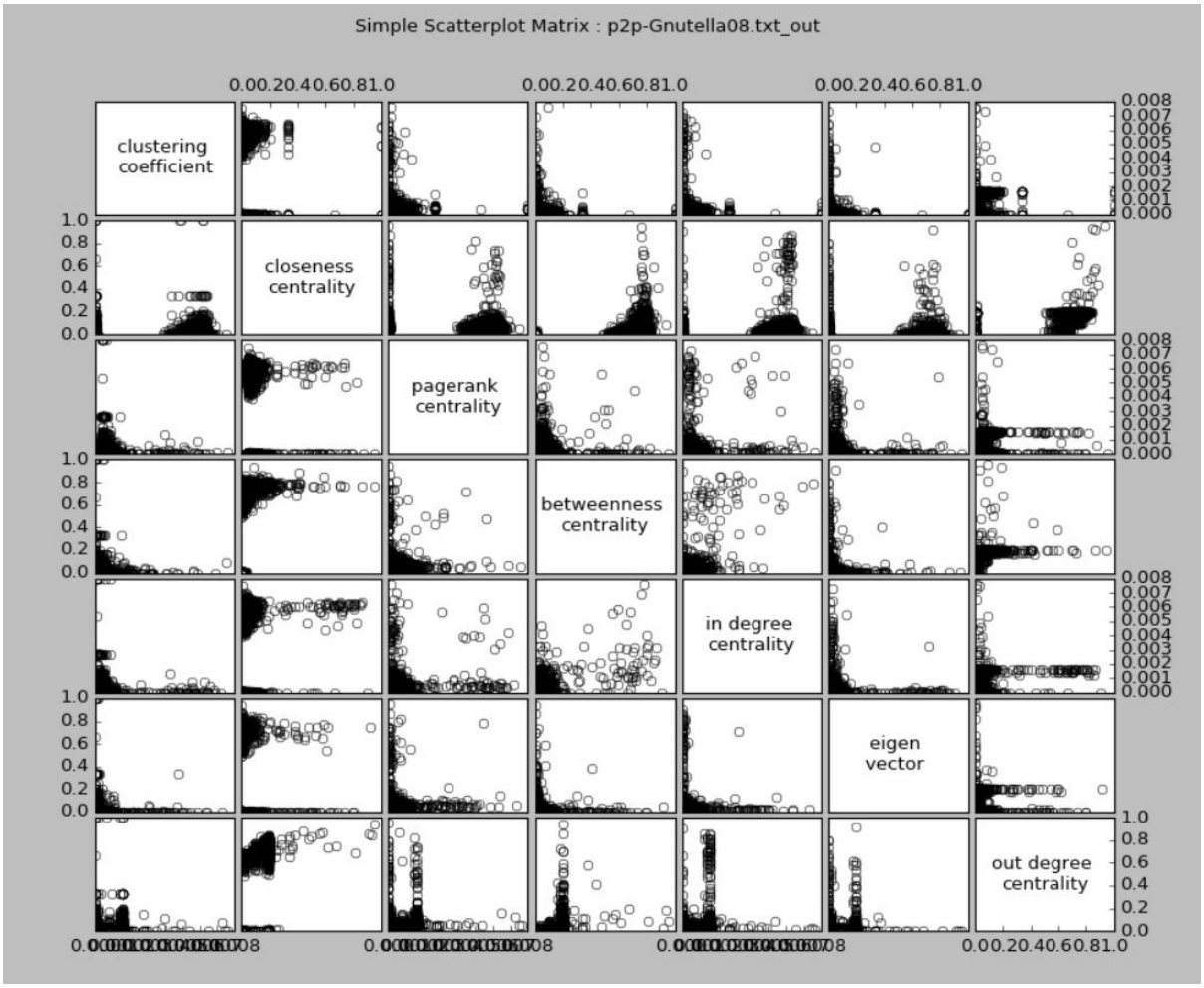
1. There is high correlation between in-degree centrality and eigenvalue centrality. This observation is almost implicit from the definition of eigenvalue centrality. Since higher the in-order of a given tends to increase the importance of the given node, it tends to say that higher in order implies higher eigenvalue centrality value.
2. Since pagerank centrality is also a type of eigenvector centrality, it eventually leads to the same inference as above. higher in-degree centrality implies higher pagerank centrality.
3. The third observation is almost a transitive inference of the above 2. eigenvector centrality and pagerank centrality are highly correlated.


Simple Scatterplot Matrix : Wiki-Vote.txt_out
clustering coefficient
closeness centrality
pagerank centrality
betweenness centrality
in degree centrality
eigen vector
out degree centrality

DATASET : GNUTELLA PEER TO PEER NETWORK

A sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. There are total of 9 snapshots of Gnutella network collected in August 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts.

The scattered plot for different characteristics.



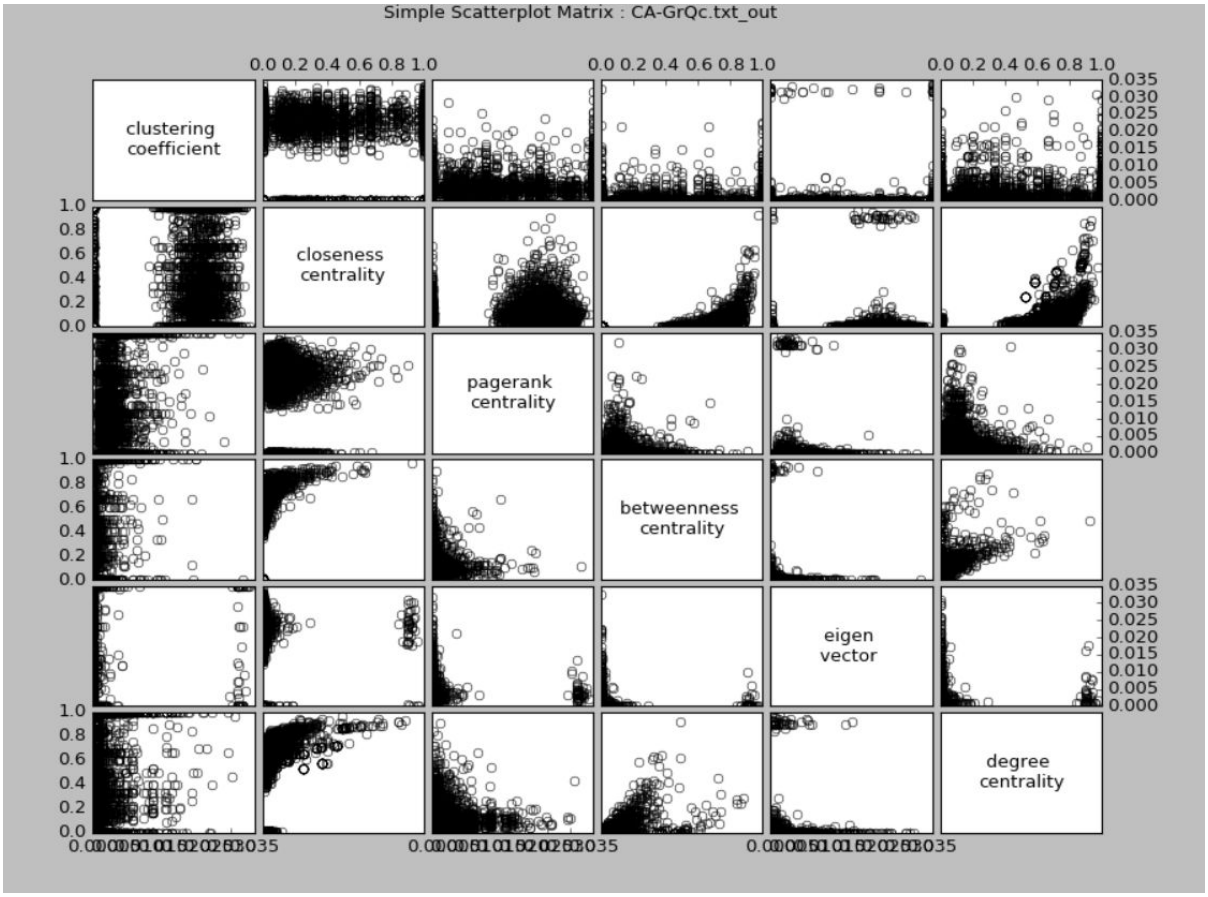
From the correlation values and the scatter plot, we observe that :

1. There is high correlation between closeness centrality and out-degree centrality. This tends to show that higher the out-degree of a given node, higher is its chances to reach entire network there by also having a higher value of closeness centrality.
2. There is high correlation between in-degree centrality and eigenvalue centrality. This observation is almost implicit from the definition of eigenvalue centrality. Since higher the in-order of a given tends to increase the importance of the given node, it tends to say that higher in order implies higher eigenvalue centrality value.
3. Since pagerank centrality is also a type of eigenvector centrality, it eventually leads to the same inference as above. higher in-degree centrality implies higher pagerank centrality.

4. The third observation is almost a transitive inference of the above 2. eigenvector centrality and pagerank centrality are highly correlated.
5. The above inferences are similar to that we had for the wiki-vote network.

DATASET : GENERAL RELATIVITY AND QUANTUM COSMOLOGY COLLABORATION NETWORK

Arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to General Relativity and Quantum Cosmology category. If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j.



From the correlation values and the scatter plot, we observe that :

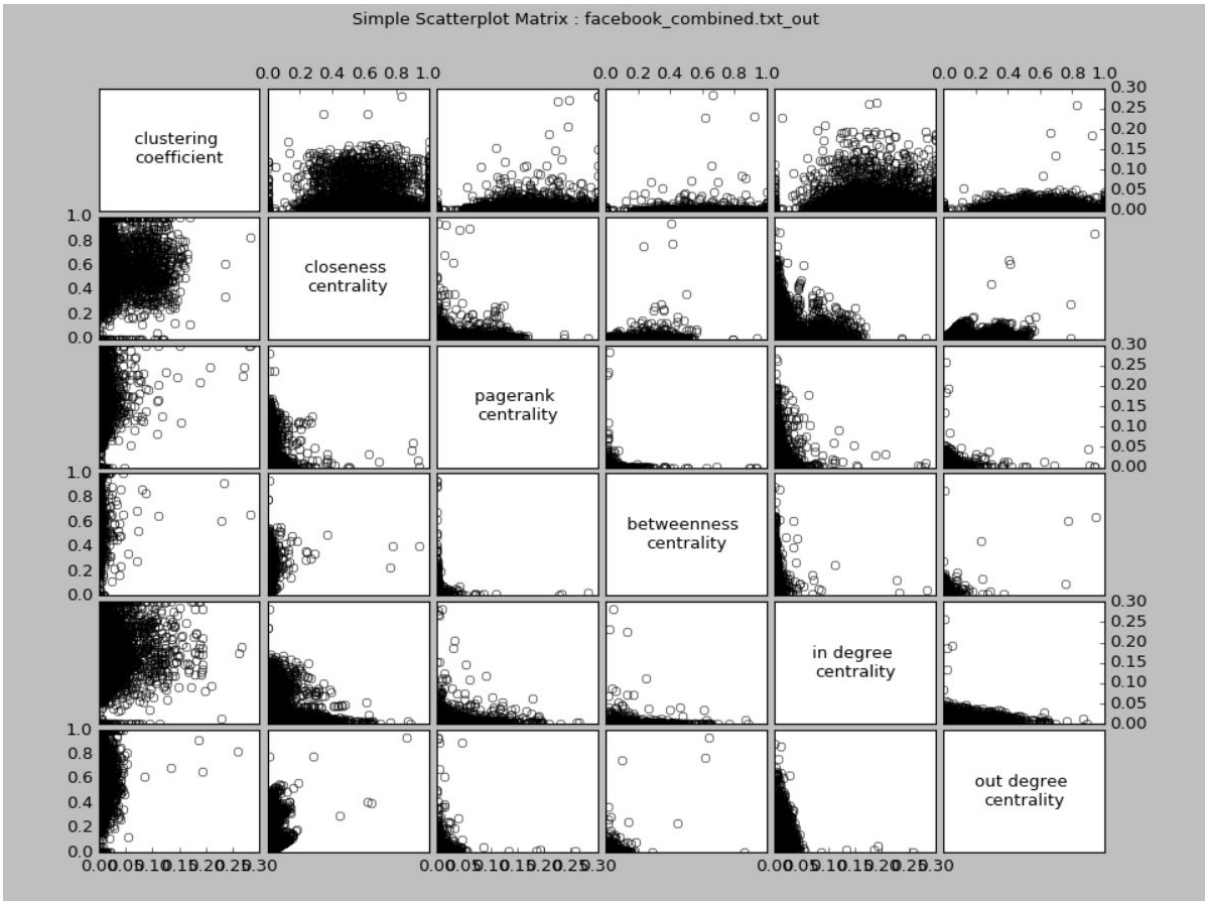
1. There is high correlation between betweenness-centrality and pagerank-centrality. This tends to show that nodes that are taking part in lot of shortest paths in the network are also the nodes with higher ranks. This is possible when we have a graph where there are lots of local clusters of nodes and these clusters are connected through few nodes which act as gateway. In such a scenario the nodes at the periphery of the cluster will have high betweenness centrality and also will have high pagerank centrality.

   When we see the data , we observe that the authors who have written a common paper are connected by a complete graph. Such complete graphs at the local level act as

clusters. Authors who have written papers with different set of people act as gateway between clusters. Such authors will have high betweenness centrality and high pagerank centrality.

2. Higher the degree centrality implies higher pagerank centrality.

DATASET : FACEBOOK EGO NETWORK



From the correlation values and the scatter plot, we observe that :

1. We don't observe strong correlations between columns as we saw for the previous data examples.
2. There is high correlation between in-degree centrality and pagerank centrality. This observation is almost implicit from the definition of pagerank centrality. Since higher the in-order of a given tends to increase the importance of the given node, it tends to say that higher in order implies higher pagerank centrality value.

GENERAL OBSERVATIONS

1. degree, eigenvector and closeness are all measure of an actor's prominence in a network. Hence it is almost evident that we find correlation among these values. [Wasserman & Faust, 1994]

ASSUMPTIONS

- ➢ While computing clustering coefficient we are assuming that the given network is undirected.

ISSUES

- ➢ While computing eigenvector coefficient for facebook data the iteration didn't go to convergence.
- ➢ In case of disconnected graphs we are computing harmonic closeness centrality.
- ➢ The other way to compute closeness centrality for disconnected graphs is by assigning N-1 value to distance between disconnected nodes.

FUTURE WORK

- ➢ Plotting heat map out of the correlation matrix values can give more insight for our meta level analysis.
- ➢ Apart from Meta-Level Analysis, analyzing some statistical parameters of the graph (say diameter, radius etc) might help us understand the graph structure more.
- ➢ For directed graphs we can compute in-closeness centrality and out-closeness centrality. This can be achieved just by reversing the direction of the edges.