CS529
Topic and Tools on Social Media Data Mining

(Assignment #1)

Understanding Complex network through different Network Models and Centrality Measures

Created by:-
Kevin Savsani
(160101063)

Datasets Introduction

DBLP Co-authorship network
DBLP dataset, is a bibliographic information network extracted from DBLP , which involves three types of nodes: conference, paper and author, connected by two types of relations/links: authoredBy link and publishedIn link. We treat authors as our target instances, with the research area of the authors as the instances labels.

IMDB Movie Dataset
IMDB dataset contains four types of nodes: movie, director, actor and actress, connected by two types of relations/links: directed link and actor/actress staring link. The target instance type is movie instance, which assigned with a set of class labels, indicating genres of the movie.

SLAP Gene Dataset
SLAP Dataset: The SLAP dataset contains integrated data related to chemical compound, gene, disease, tissue, pathway etc. We treat genes as our target instances. Each gene can belong to one of the gene family.

Higgs Twitter Dataset
There are a number of ways users can interact on Twitter. Users can “follow” other users to receive regular updates of their tweets, and users may “mention” other users in their own tweets. Here we are going to analyse Twitter mention dataset.

Foursquare Restaurant Review Dataset
This dataset includes long-term (about 22 months from Apr. 2012 to Jan. 2014) global-scale check-in data collected from Foursquare, and also two snapshots of user social networks before and after the check-in data collection period. The check-in dataset contains 22,809,624 checkins by 114,324 users on 3,820,891 venues. The social network data contains 363,704 (old) and 607,333 (new) friendships.Dataset contains four column User ID (anonymized), Venue ID, UTC time, Timezone offset in minutes.
File dataset for friends contains Each row indicating a friendship between two users.

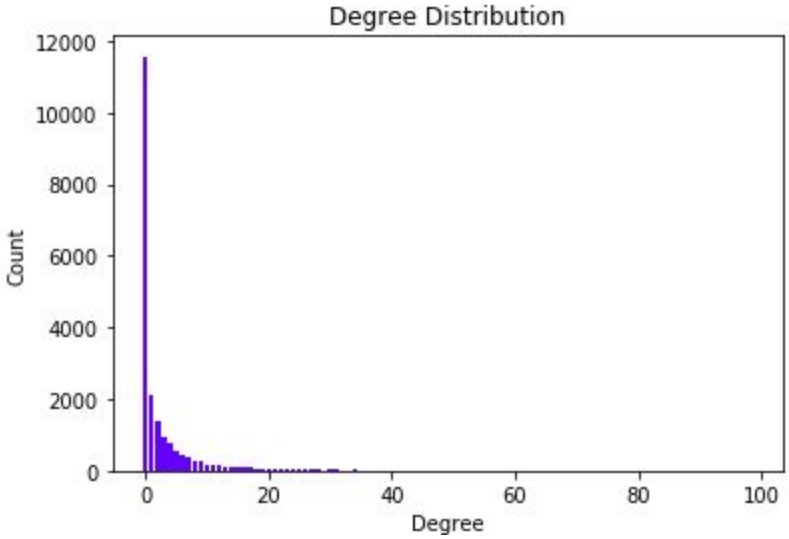
DataSet Analysis

1. SLAP Gene Dataset

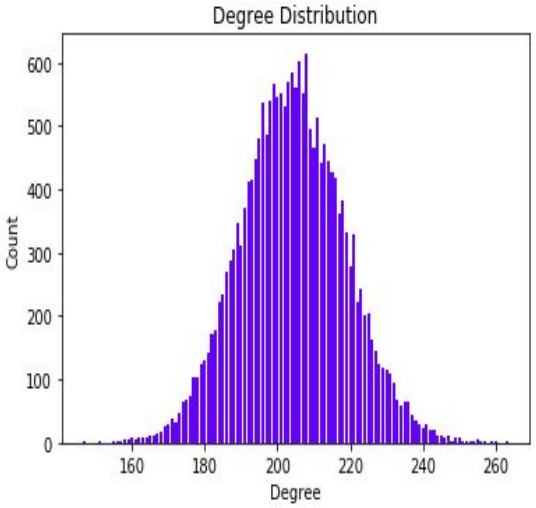
For Network Analysis part with degree distribution and coefficient clustering and other techniques were used.

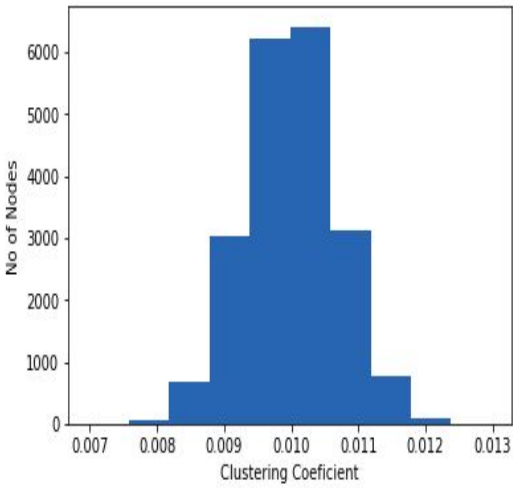
- The graph is an undirected graph with 20419 nodes and 30213 edges and average degree of nodes was 2.96.
- Average clustering coefficient is found to be 0.39 .
- Average shortest path averaged over all connected components is 1.1.
- Giant component by size had 8407 nodes and has diameter equals 15.

Here is Degree Distribution plot generated by SLAP Dataset.

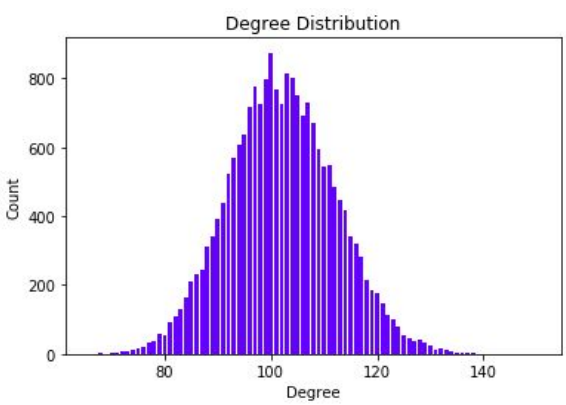


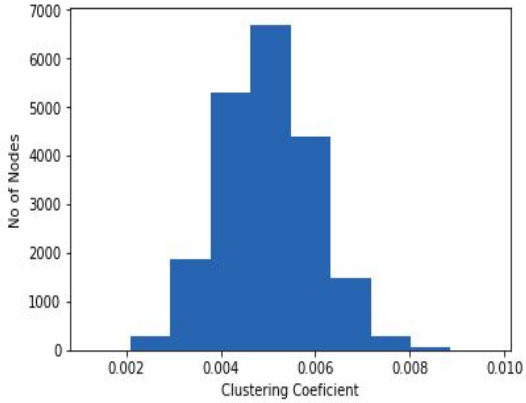
Here are few plot generated by random graph:
With probability 0.01-





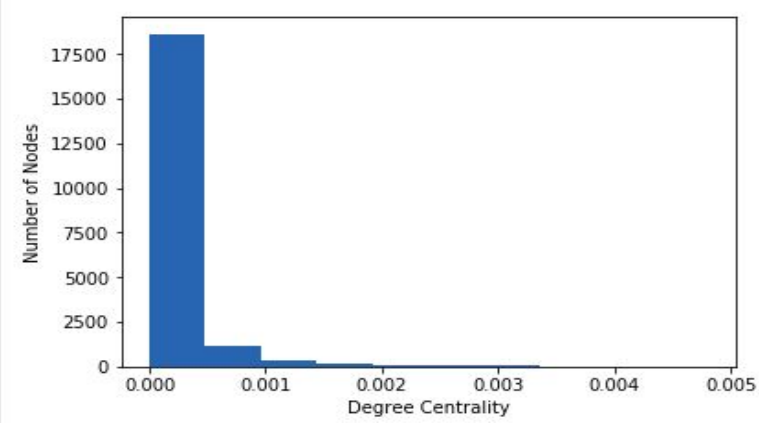
With probability 0.005-



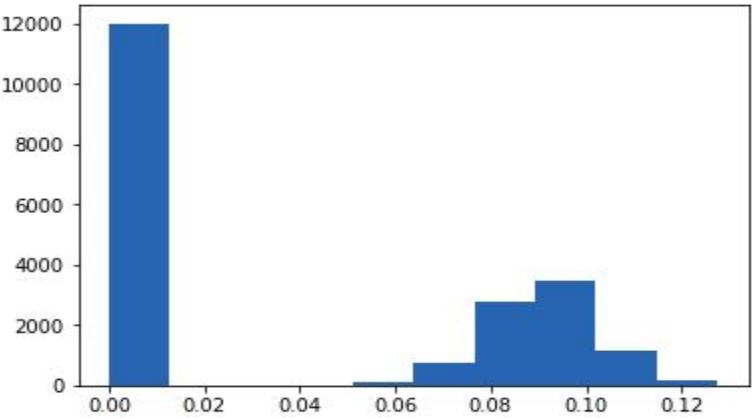


There were many centrality measures which were calculated from this network.

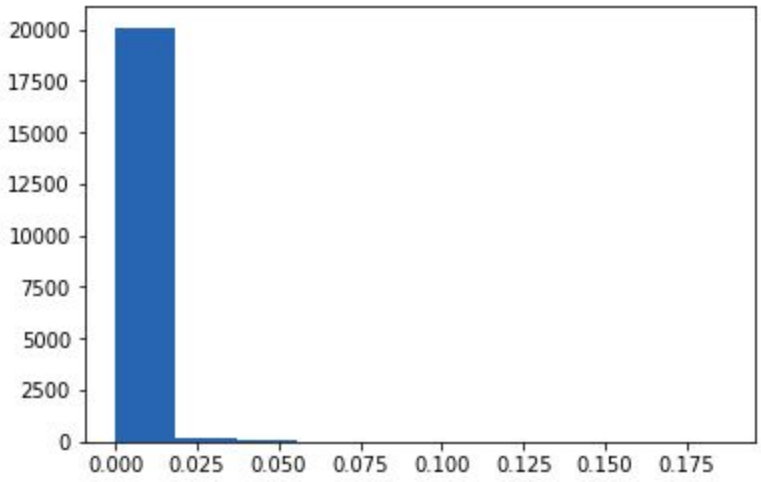
Degree Centrality-



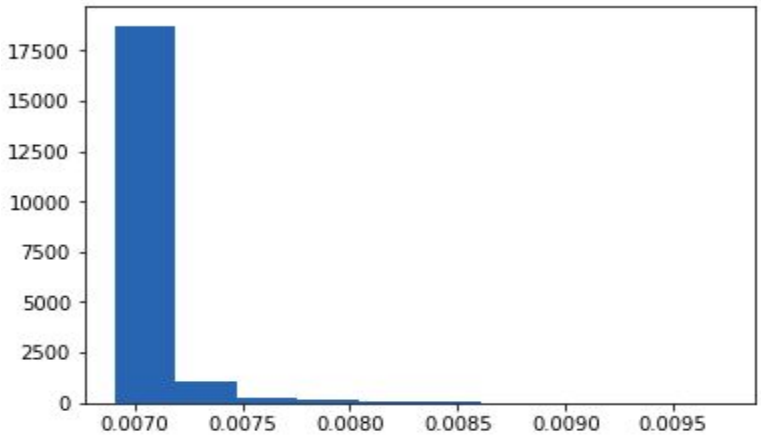
Closeness Centrality -



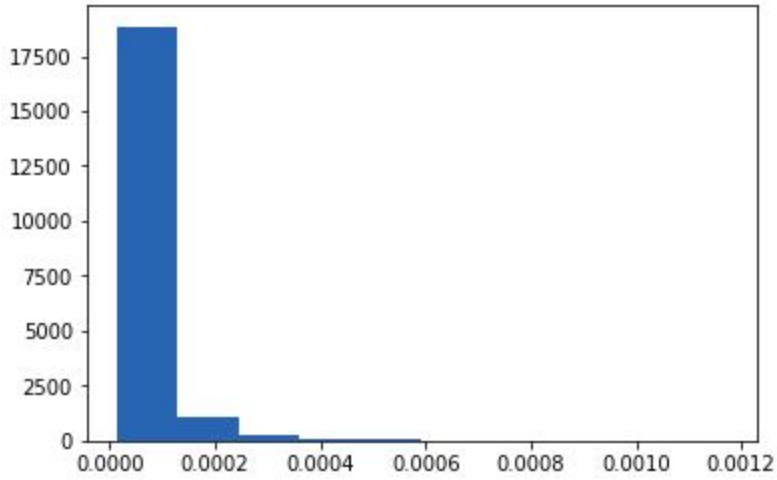
EigenVector Centrality -



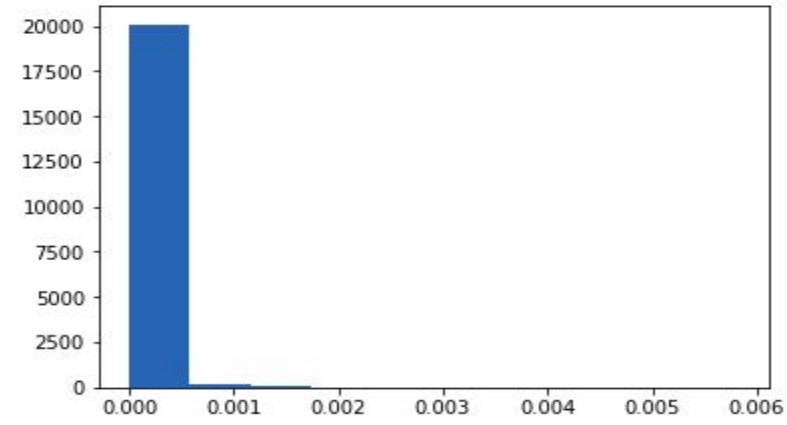
Katz Centrality -



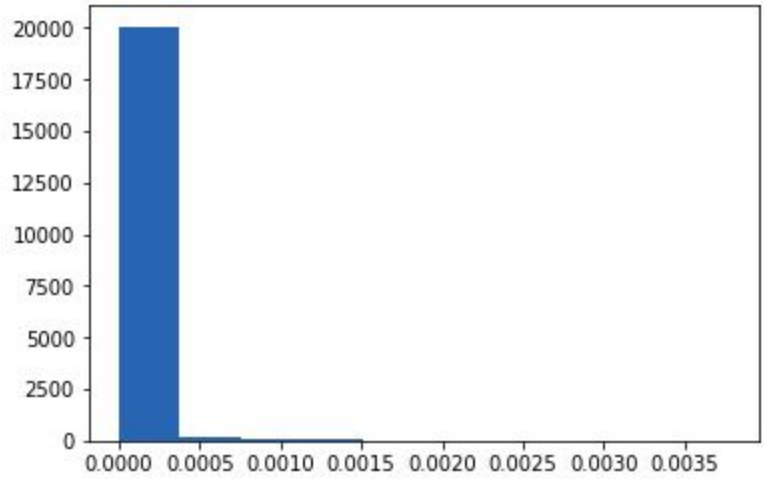
PageRank Centrality -



Hits Centrality -



Betweenness Centrality -

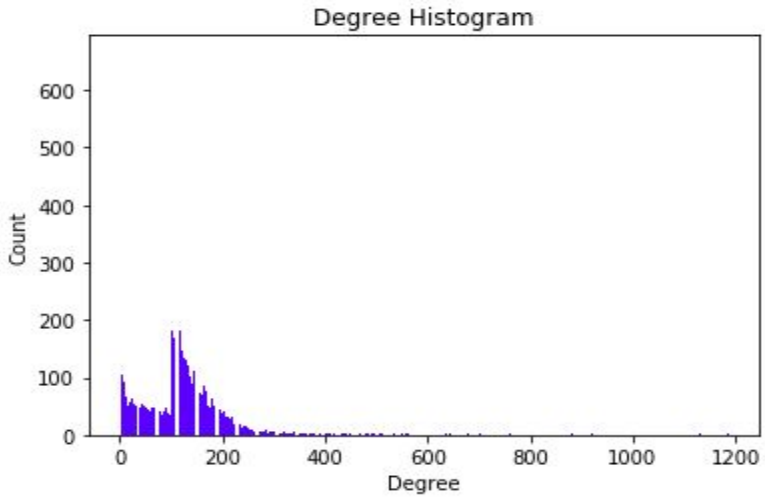


2. IMDB Movie Dataset

For Network Analysis part with degree distribution and coefficient clustering and other techniques were used.

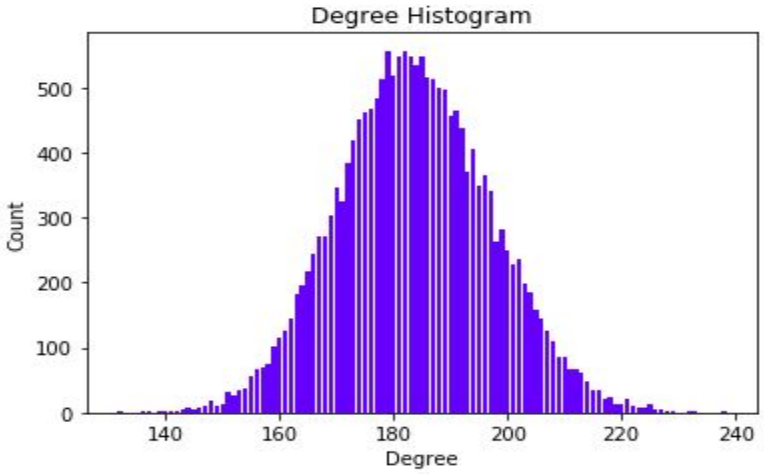
- The graph is an undirected graph with 18352 nodes and 1085810 edges and average degree of nodes was 118.33.
- Average clustering coefficient is found to be 0.19 .
- Giant component by size had 17667 nodes.

Here is the Degree Distribution plot generated by IMDB Dataset.

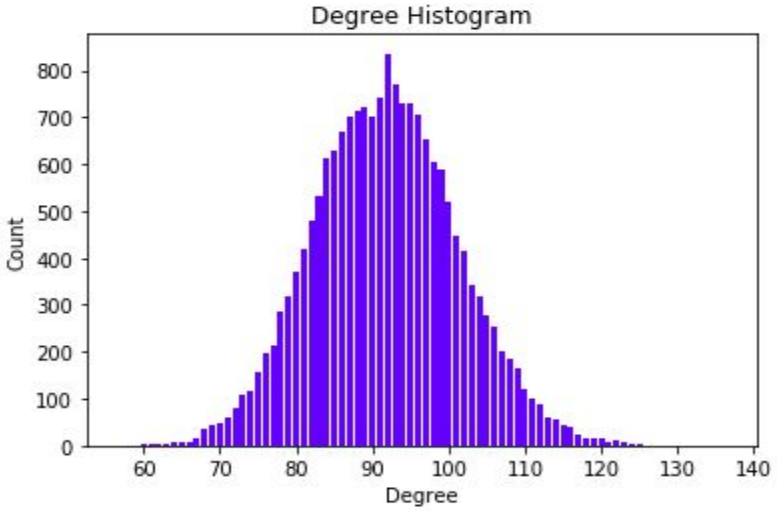


Here are few plot generated by the random graph:

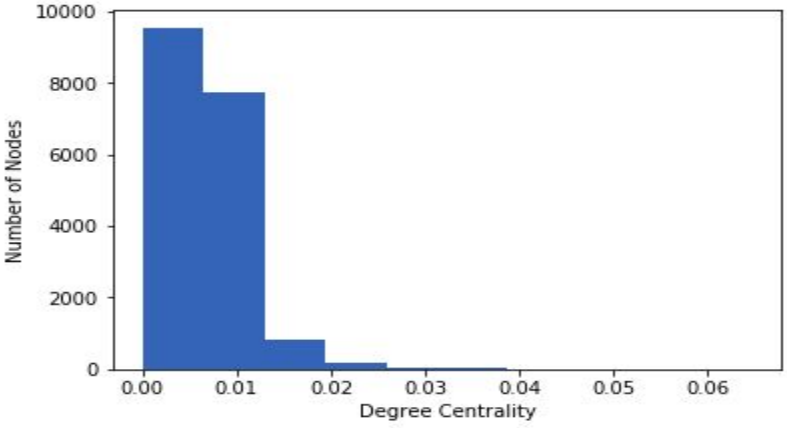
With probability 0.01-



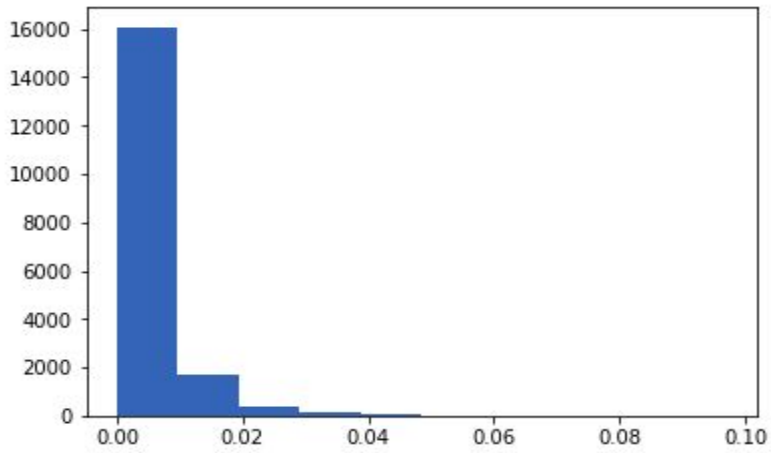
With probability 0.005-



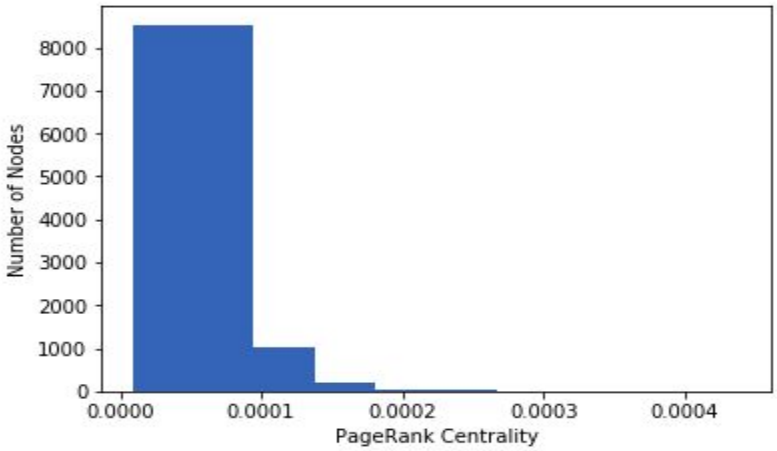
There were many centrality measures which were calculated from this network.
Degree Centrality-



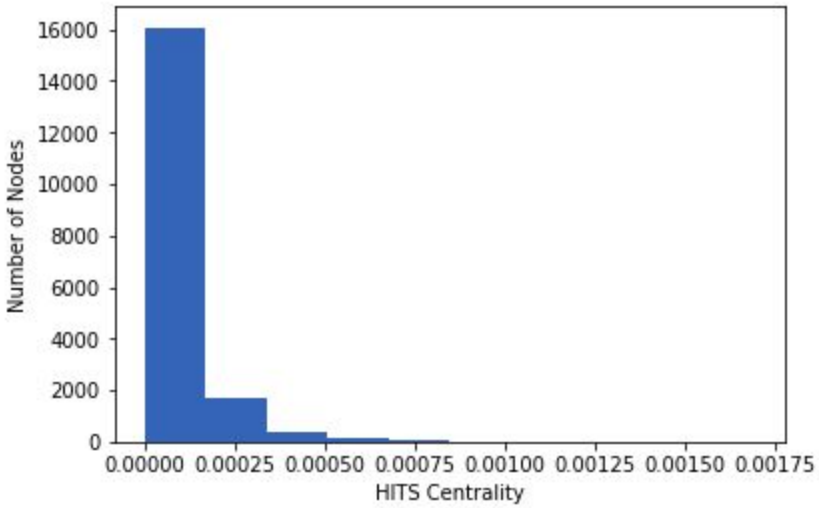
Eigen Vector Centrality -



PageRank Centrality -



Hits Centrality -

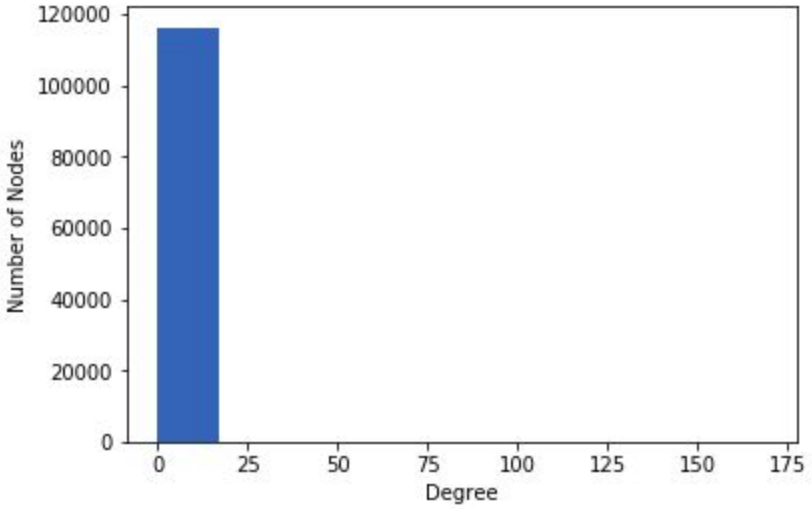


3. Higgs Twitter Dataset

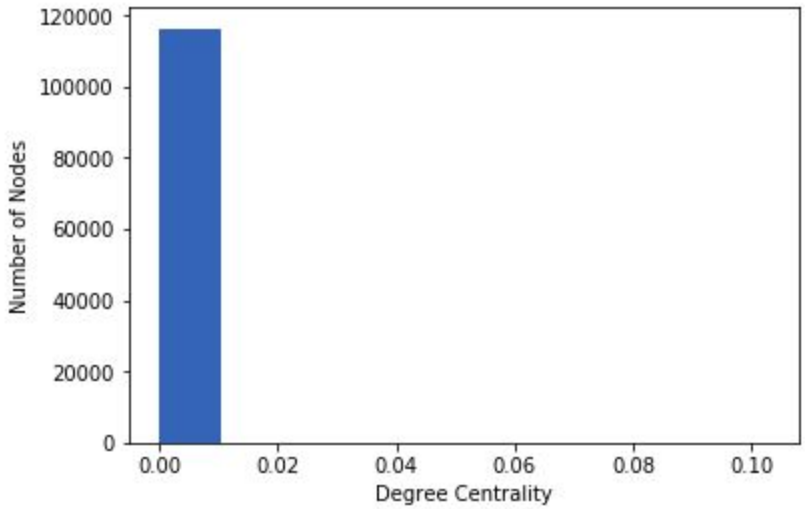
For Network Analysis part with degree distribution and coefficient clustering and other techniques were used.

● The graph is an undirected graph with 116408 nodes and 150818 edges and average degree of nodes was 2.96.

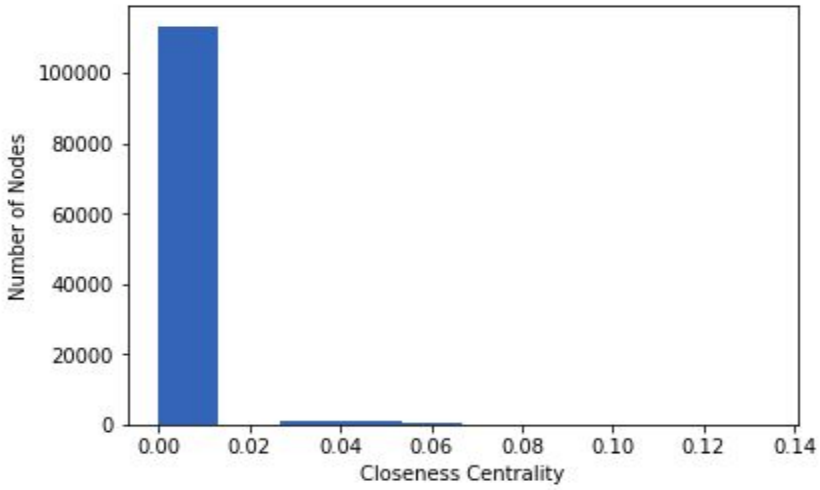
Here is Degree Distribution plot generated by Higgs Dataset.



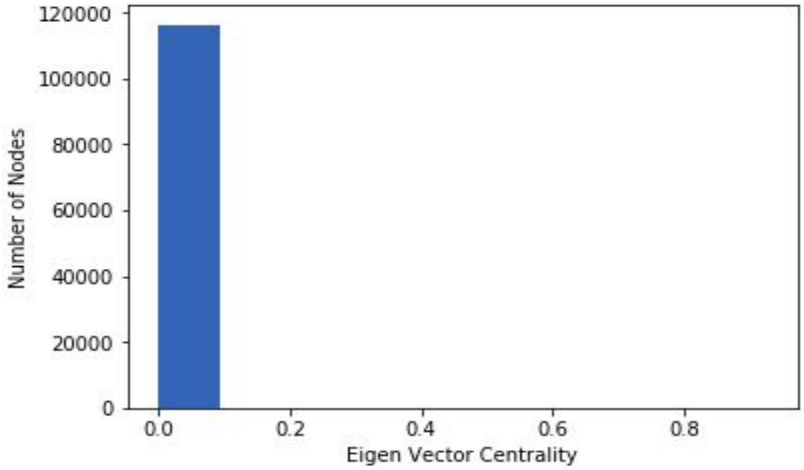
There were many centrality measures which were calculated from this network.
Degree Centrality-



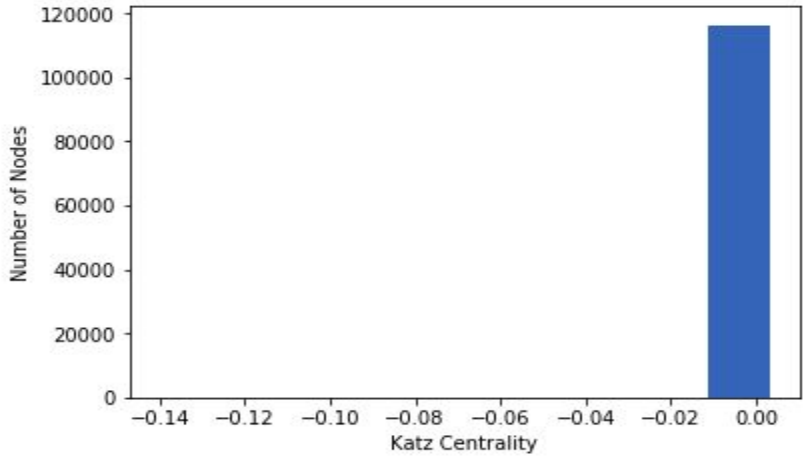
Closeness Centrality -



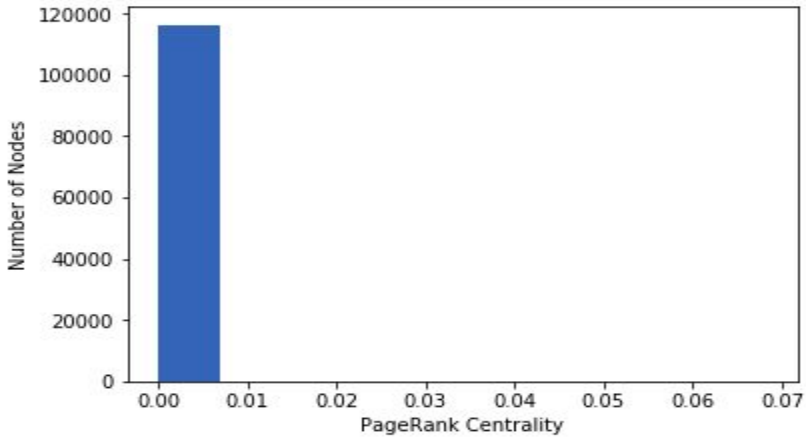
Eigen Vector Centrality -



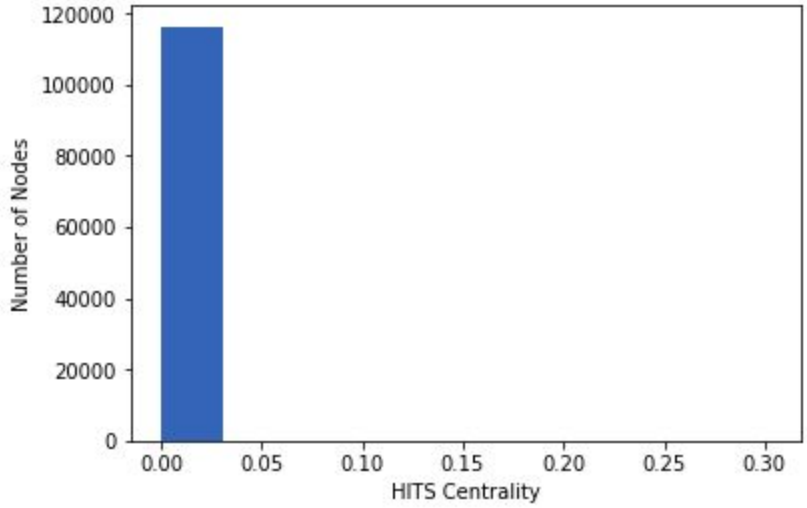
Katz Centrality -



PageRank Centrality -



Hits Centrality -

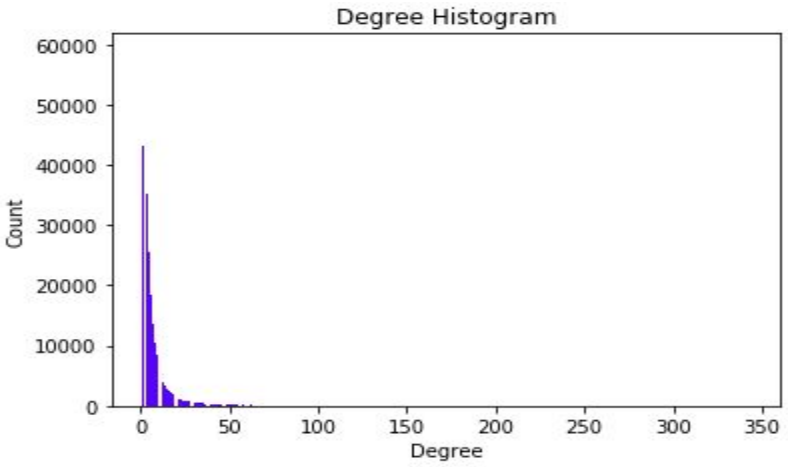


4. DBLP Co-authorship network

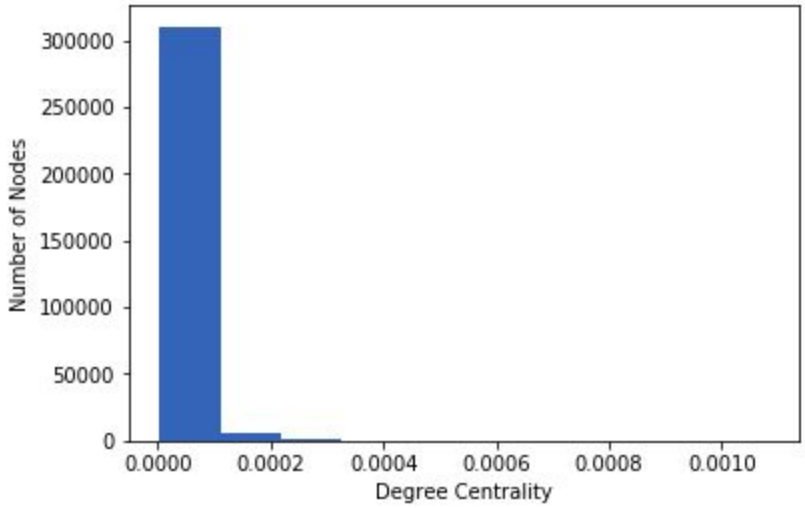
For Network Analysis part with degree distribution and coefficient clustering and other techniques were used.

- The graph is an undirected graph with 317080 nodes and 1049866 edges and average degree of nodes was 6.62.
- Average clustering coefficient is found to be 0.63 .

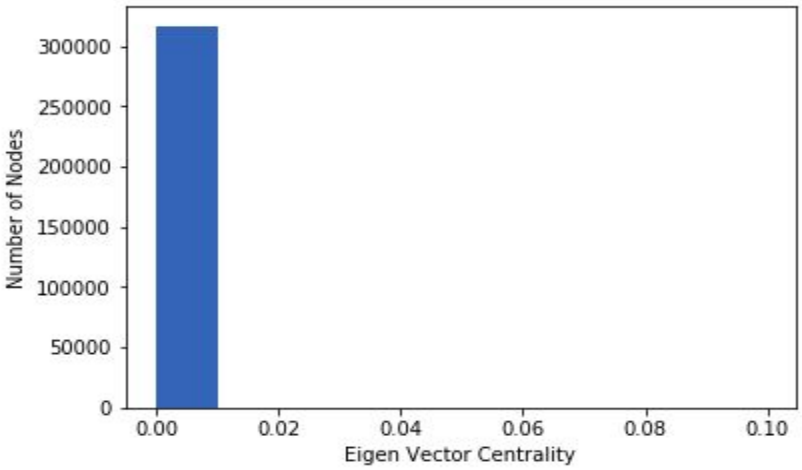
Here is Degree Distribution plot generated by DBLP Dataset.



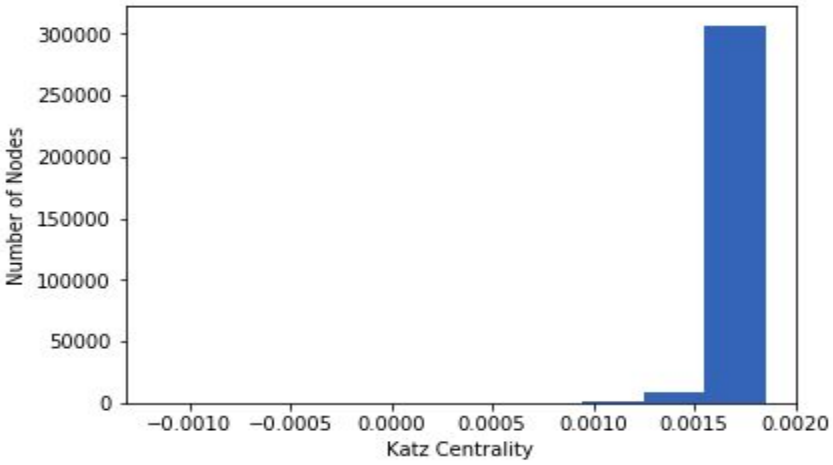
There were many centrality measures which were calculated from this network.
Degree Centrality-



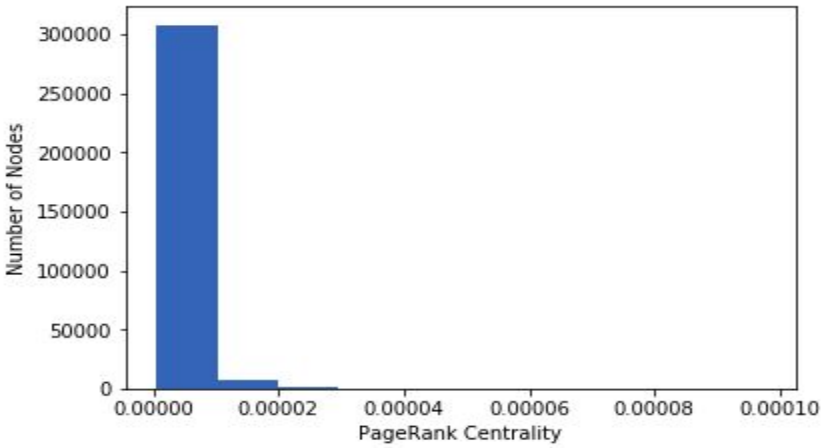
Eigen Vector Centrality -



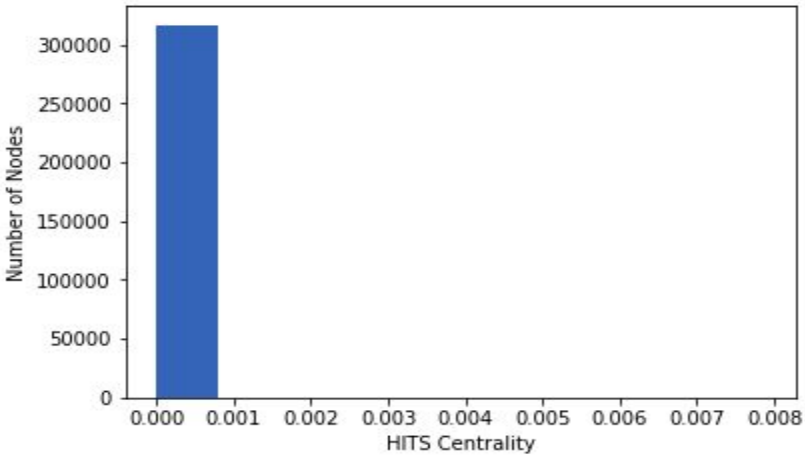
Katz Centrality -



PageRank Centrality -



Hits Centrality -

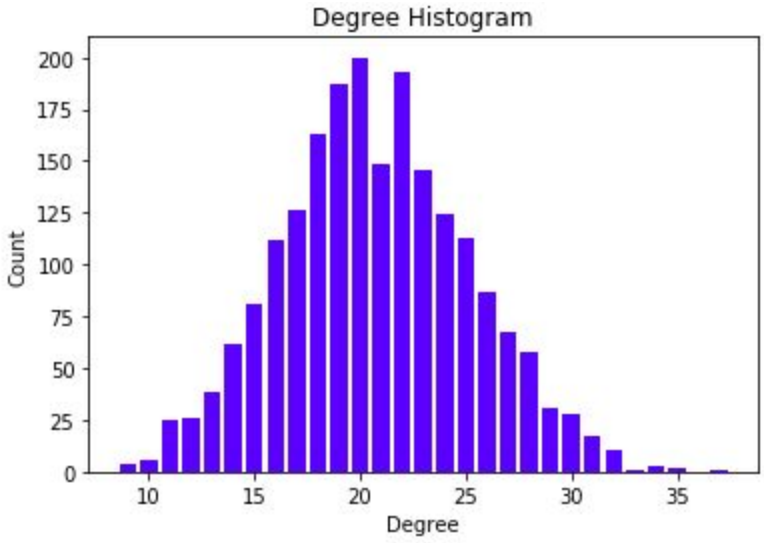


5. Foursquare Restaurant Review Dataset

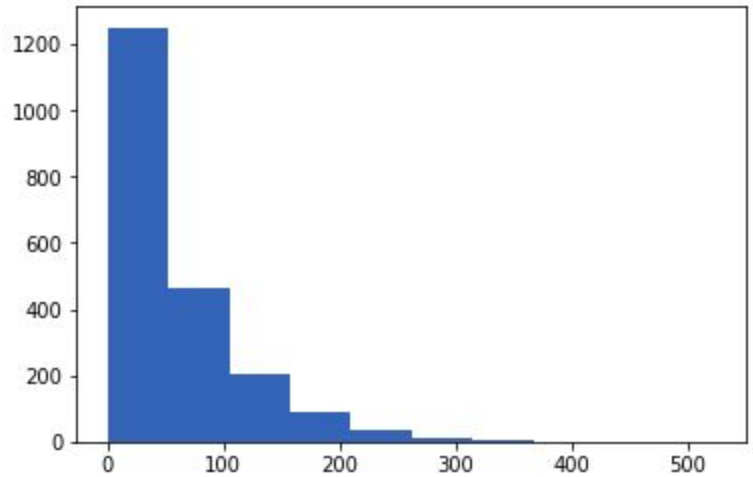
For Network Analysis part with degree distribution and coefficient clustering and other techniques were used.

- The graph is an undirected graph with 2060 nodes and 58810 edges and average degree of nodes was 57.09.
- Average clustering coefficient is found to be 0.45 .
- Giant component by size had 2045 nodes and has diameter 6.

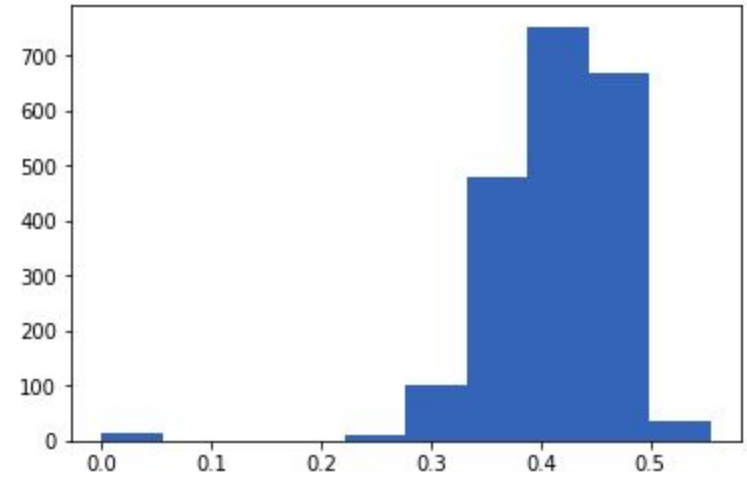
Here is Degree Distribution plot generated by Foursquare Restaurant Review Dataset.



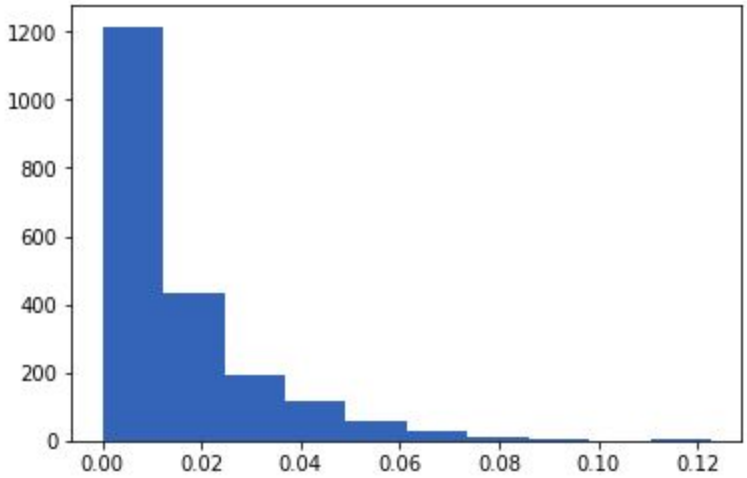
There were many centrality measures which were calculated from this network.
Degree Centrality-



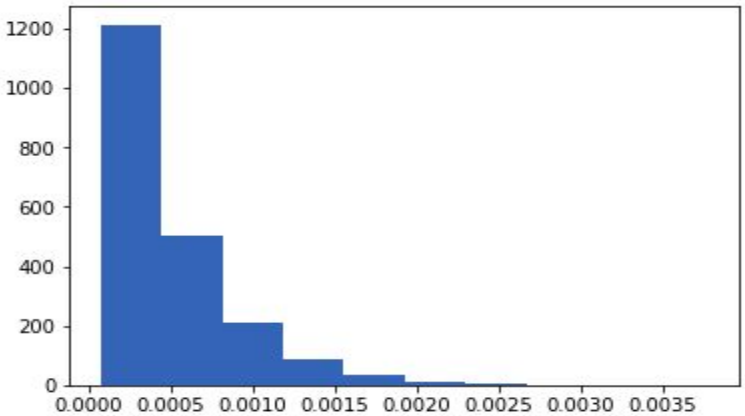
Closeness Centrality -



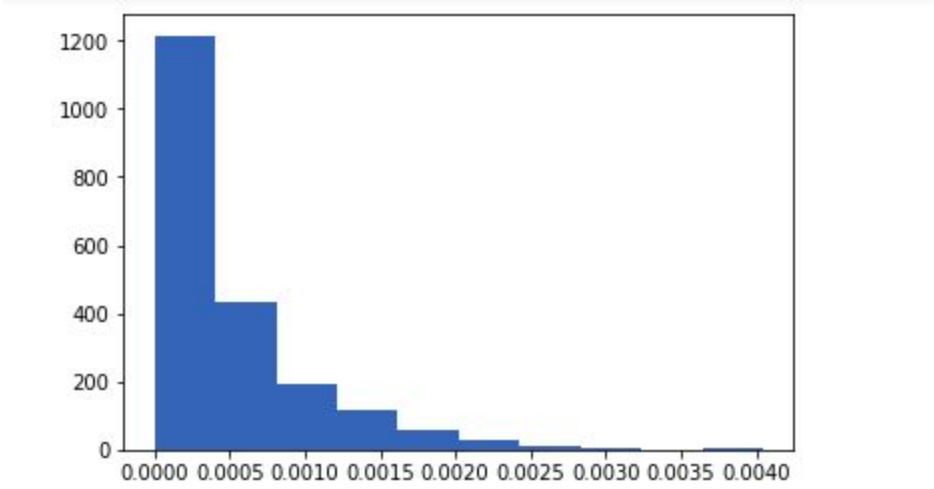
EigenVector Centrality -



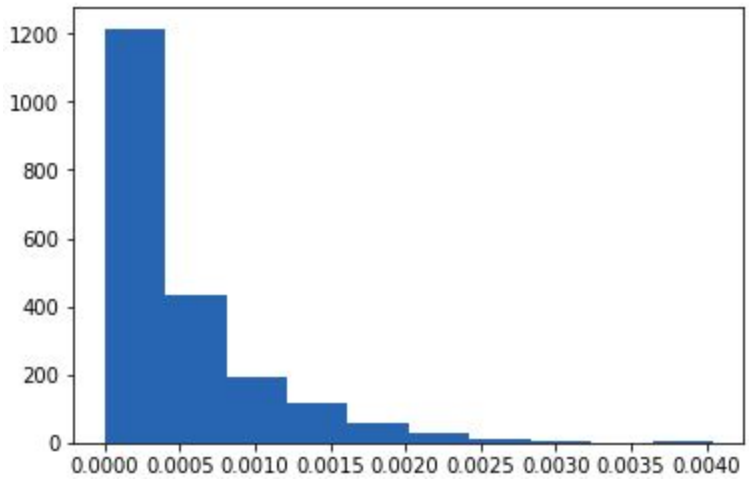
PageRank Centrality -



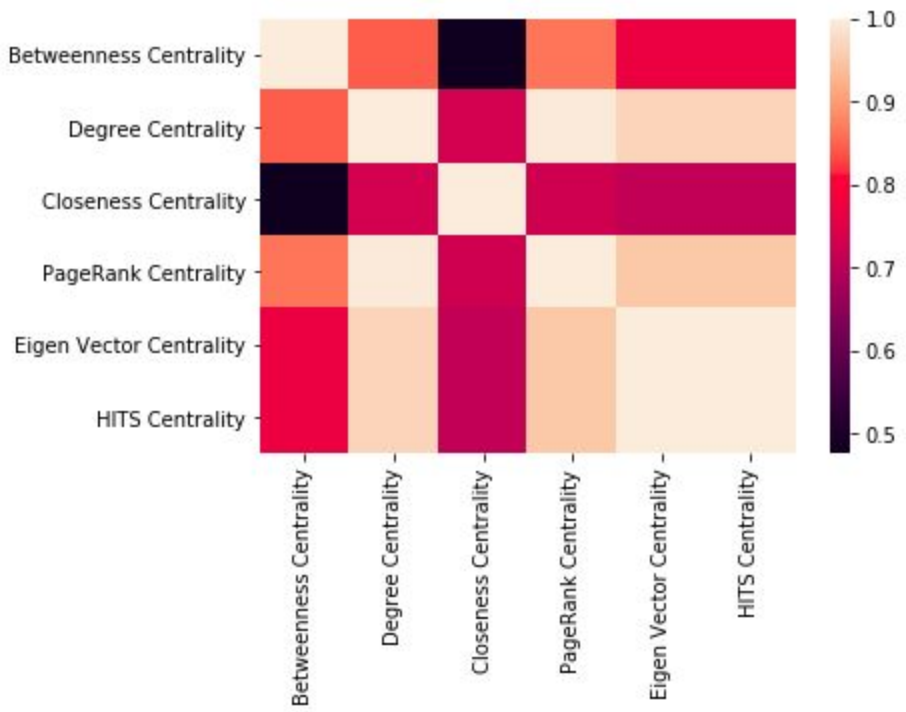
Hits Centrality -



Betweenness Centrality -



Correlation Between Centralities



The above correlation between centrality is found on Foursquare Restaurant Review Dataset. Here we can see that PageRank Centrality is closely associated with Degree Centrality i.e., if there is an increase in value of one centrality others will show the same effect. Similarly we can see Eigenvector Centrality and Hits Centrality are also closely related. We can see that Betweenness And Closeness are too different i.e., change in one value will almost show negative change in another value.