

**SNA PROJECT**

**Submitted By**

Anshul Mandhana 21ucs023

Aditya Sonkusare 21ucs011

Harsh Kumar 21ucs086

Chirag Birla 21ucs051

**Submitted To**

Dr. Sakthi Balan, Associate Professor Computer Science & Engineering Department, The LNM Institute of Information Technology

**Language**

Python

**Environment**

Google Collab

# Libraries Used

* **Pandas**
  1. It is used in data manipulation and analysis .
* **Numpy**
  1. This is used while working in arrays .
* **Networkx**
  1. It is a python library for studying graphs and networks.
* **Matplotlib.pyplot**
  1. We have used the pyplot module of this library for plotting graphs, distribution charts .
* **Math**
  1. This is used for performing complex mathematical calculations with ease.

# ROUND 1

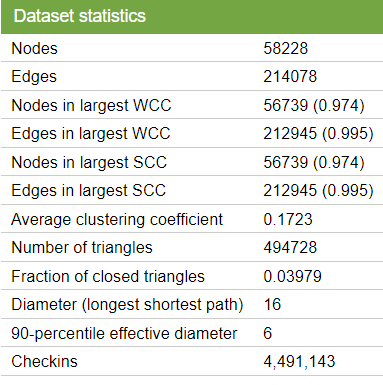
• Summarise the network statistics - Plot the degree distribution of the network, find the max and minimum degree, average degree, Standard deviation of the degree distribution.

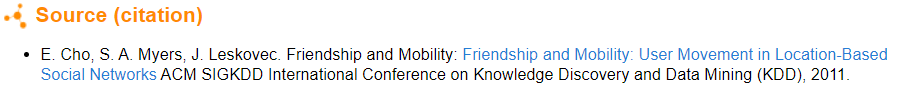
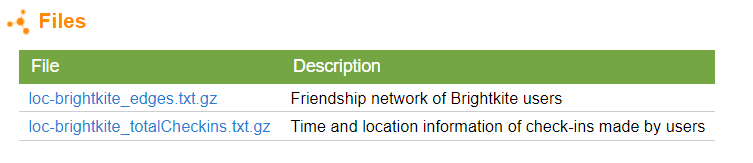
• Find all centrality measures studied in the class - Degree centrality, Eignevector centrality, Katz and PageRank centrality, clustering coefficients (both local and global), find average local clustering coefficient and compare with global clustering coefficient, find the betweenness and closeness centrality and reciprocity and transitivity that we have studied in the class using appropriate algorithms

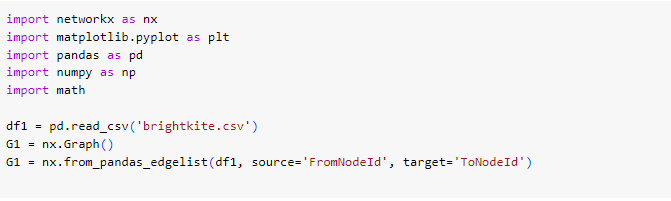
# Dataset 1 Information

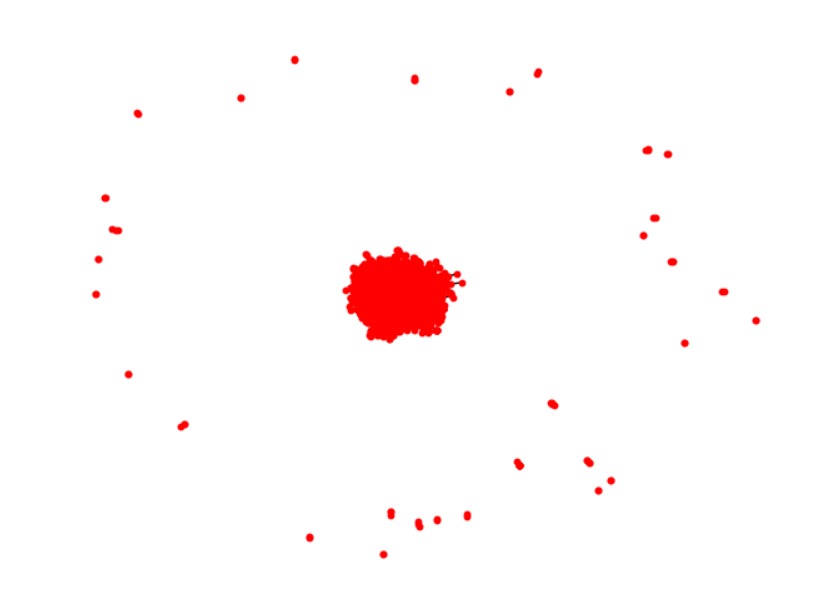
(https://snap.stanford.edu/data/loc-Brightkite.html)

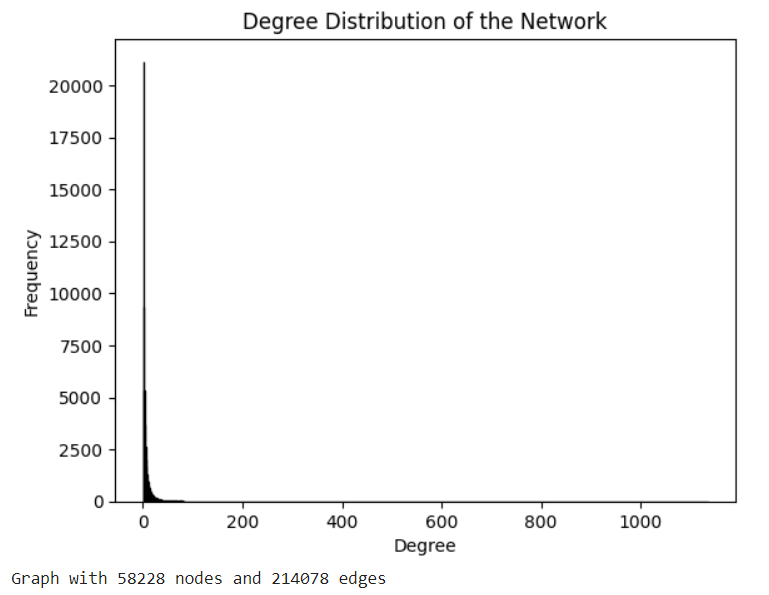
## [Brightkite](http://www.brightkite.com/) was once a location-based social networking service provider where users shared their locations by checking-in. The friendship network was collected using their public API, and consists of 58,228 nodes and 214,078 edges. The network is originally directed but we have constructed a network with undirected edges when there is a friendship in both ways. We have also collected a total of 4,491,143 checkins of these users over the period of Apr. 2008 - Oct. 2010.

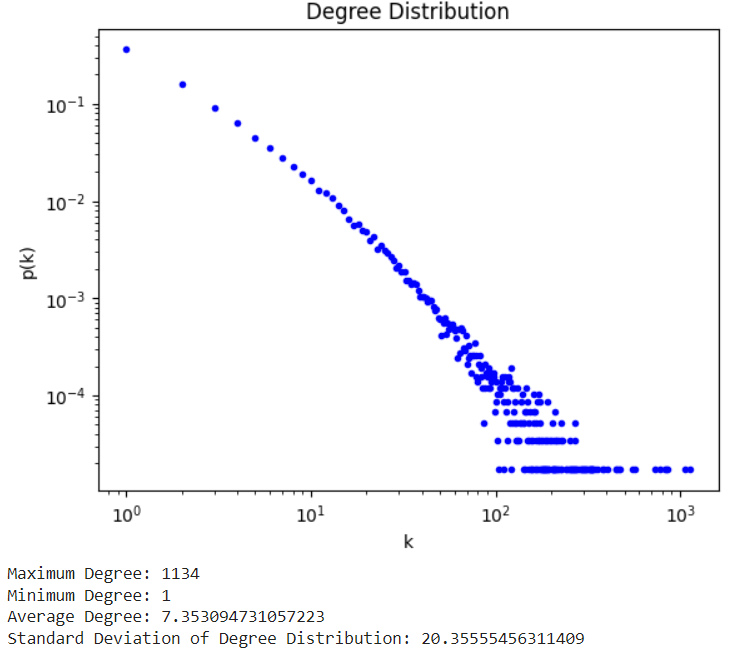
.

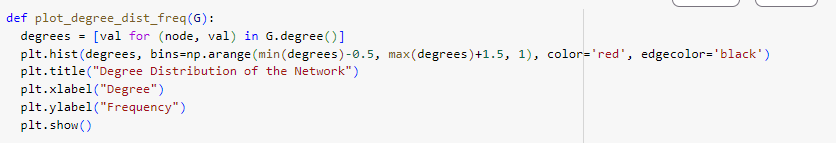


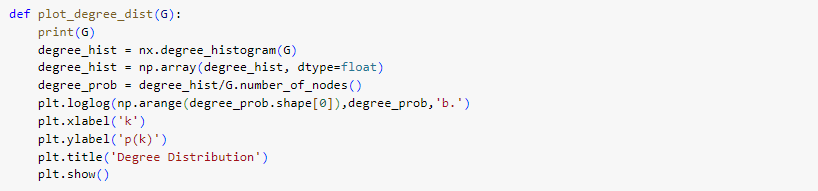
 

****

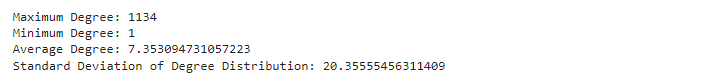
**Degree Distribution:** It is used to know about how the nodes are interrelated with other nodes, high degree means more inter-related relationship. 

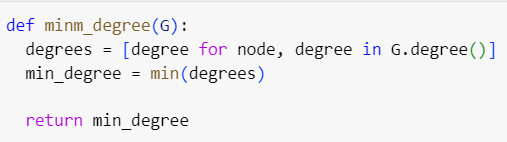


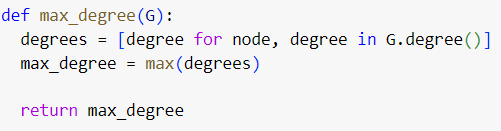




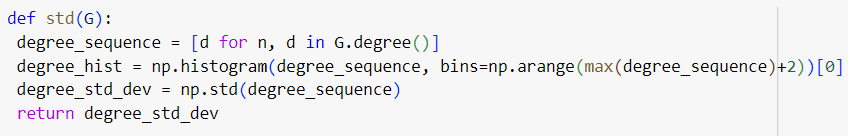


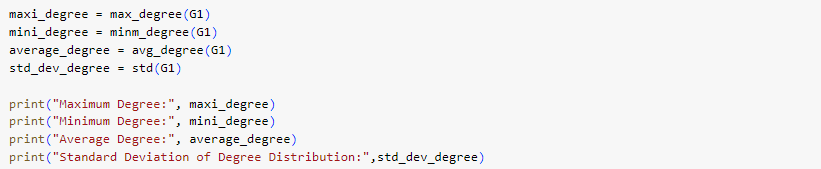
**Degree Calculation:** 







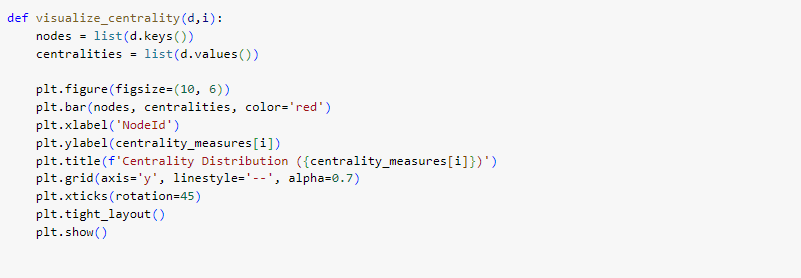


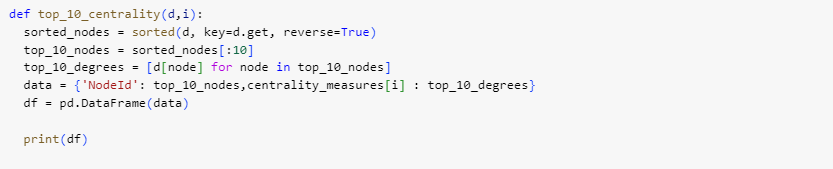
****

# Centrality and Other Measures

Centrality defines the importance of a node in the graph.





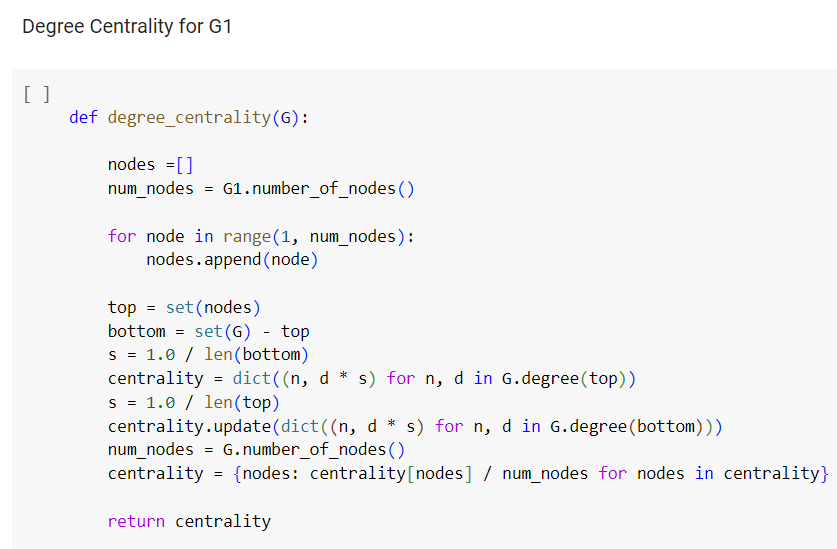


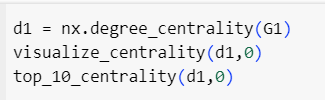
## Degree Centrality

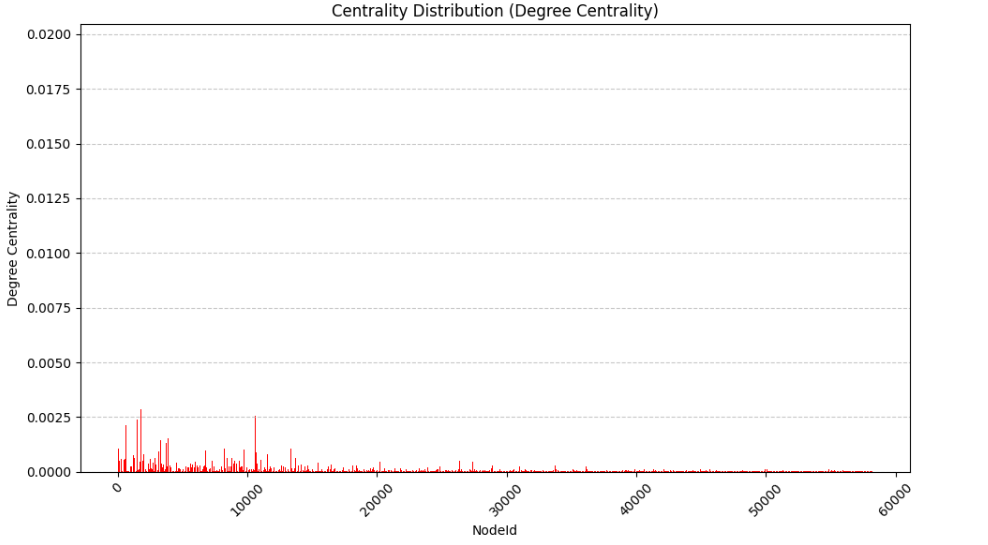
Users with lower degree centrality might be less active or have fewer connections within the network.

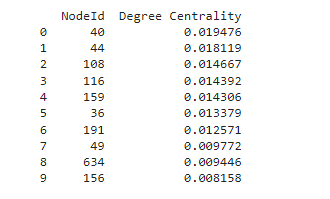
**Inference:**

1. This basically measures the node with respect to its degree. Degree of a node is the number of other nodes it is connected to . Higher degree of a node implies it has more connections, thus it is more central.
2. Users with lower degree centrality might be less active or have fewer connections within the network







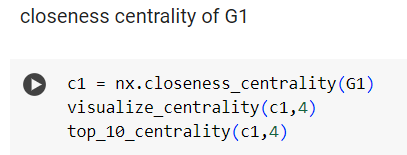


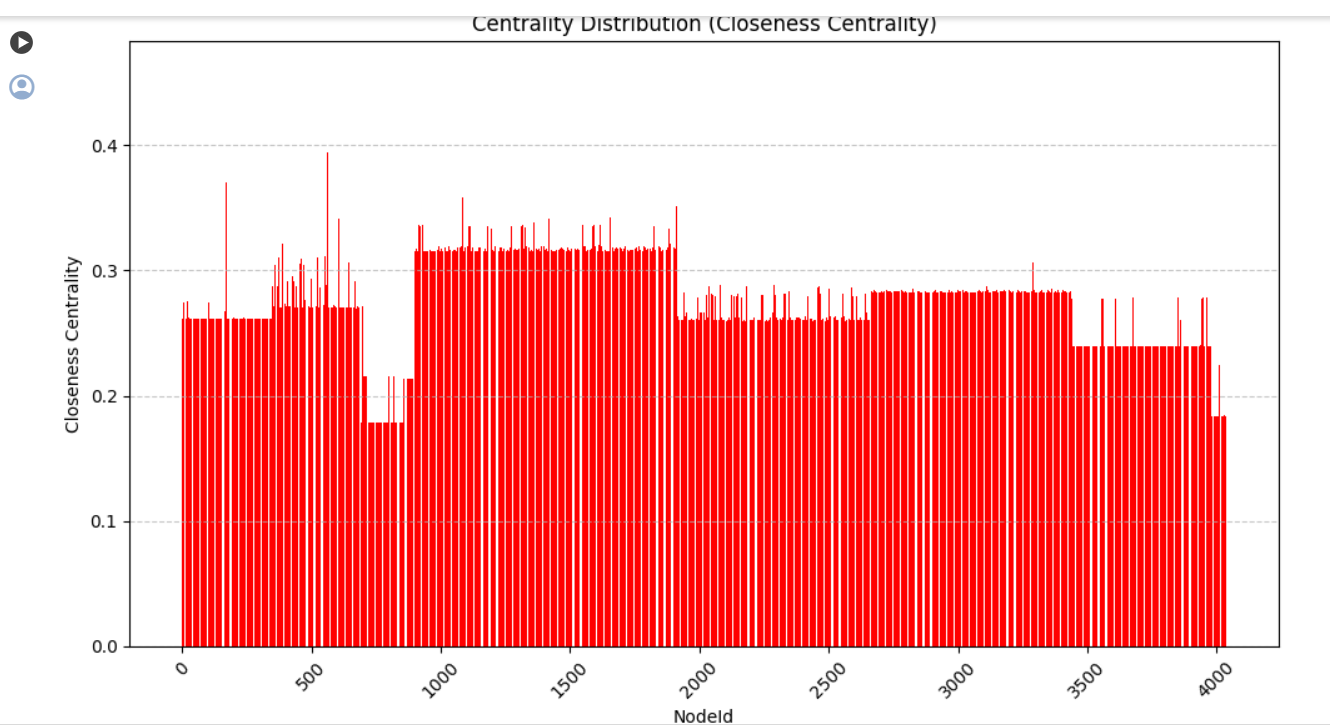
## Closeness Centrality

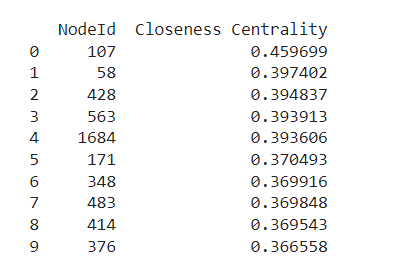
It is calculated as the reciprocal of the sum of shortest path lengths between the node *v* and all the other nodes in a graph. It suggests how efficiently a node is spreading information through the graph .

**Inference:**

1. Nodes with high closeness centrality are likely to have shorter paths to reach other nodes in the network.
2. Users with high closeness centrality might have a more significant influence on the network due to their ability to quickly disseminate information or influence to a large portion of the network.





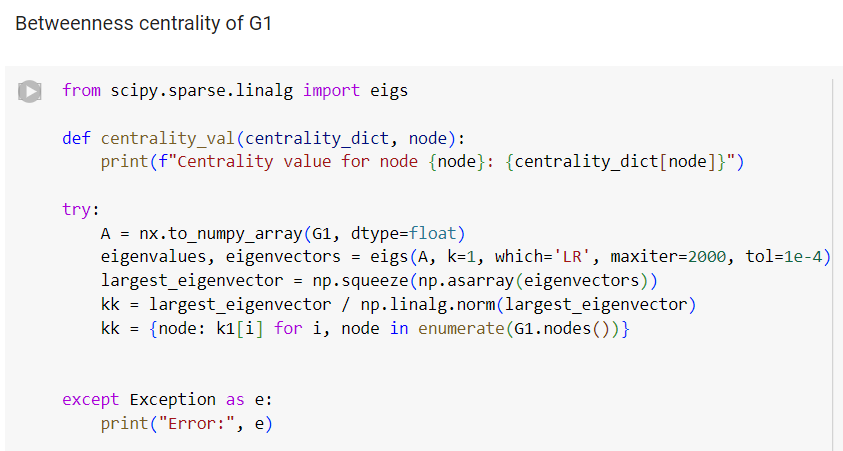


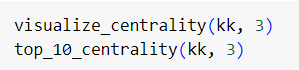
## Betweenness Centrality

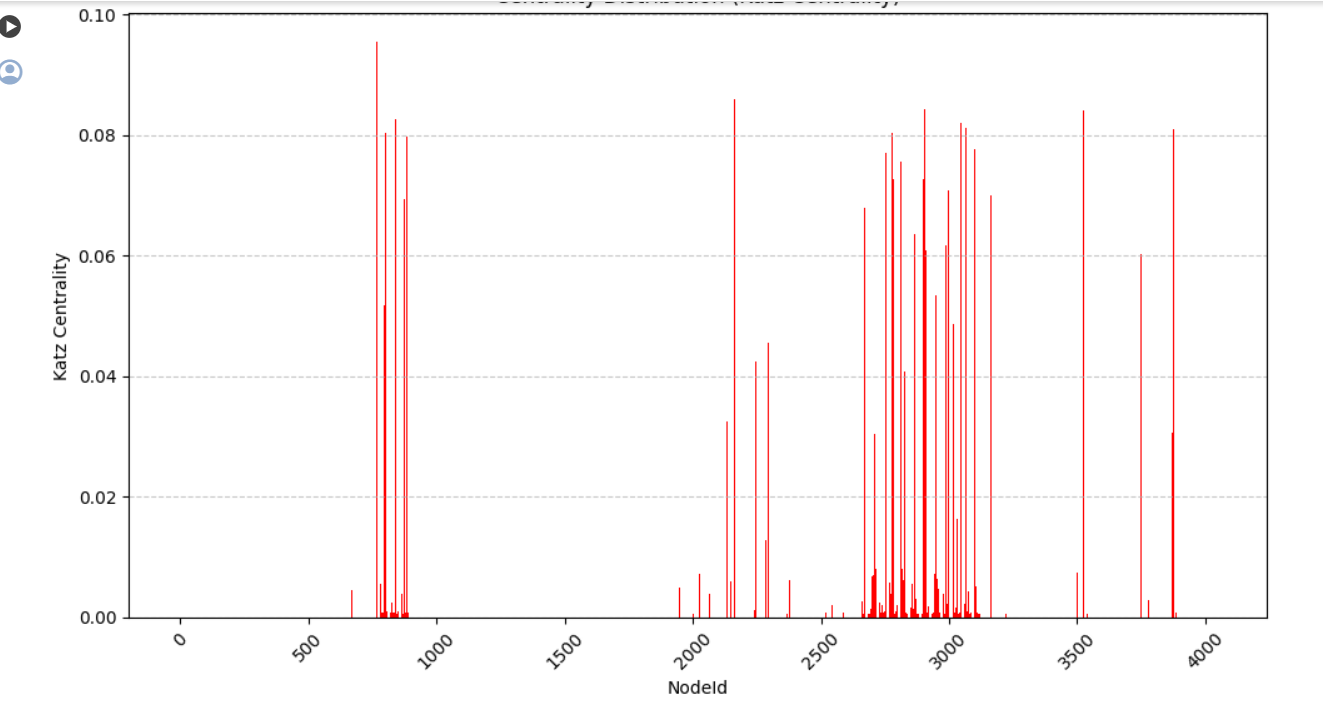
It helps to measure how important nodes are in connecting other nodes. For a vertex *vi,* we can measure the extent to which *vi* lies on shortest paths between other vertices.

### Inference

1. Nodes with high betweenness centrality act as bridges between different clusters or communities within the network.
2. These users might facilitate communication and interaction between otherwise disconnected groups of users.





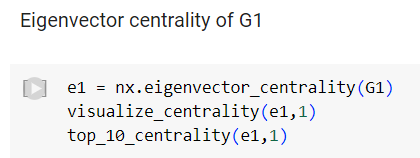


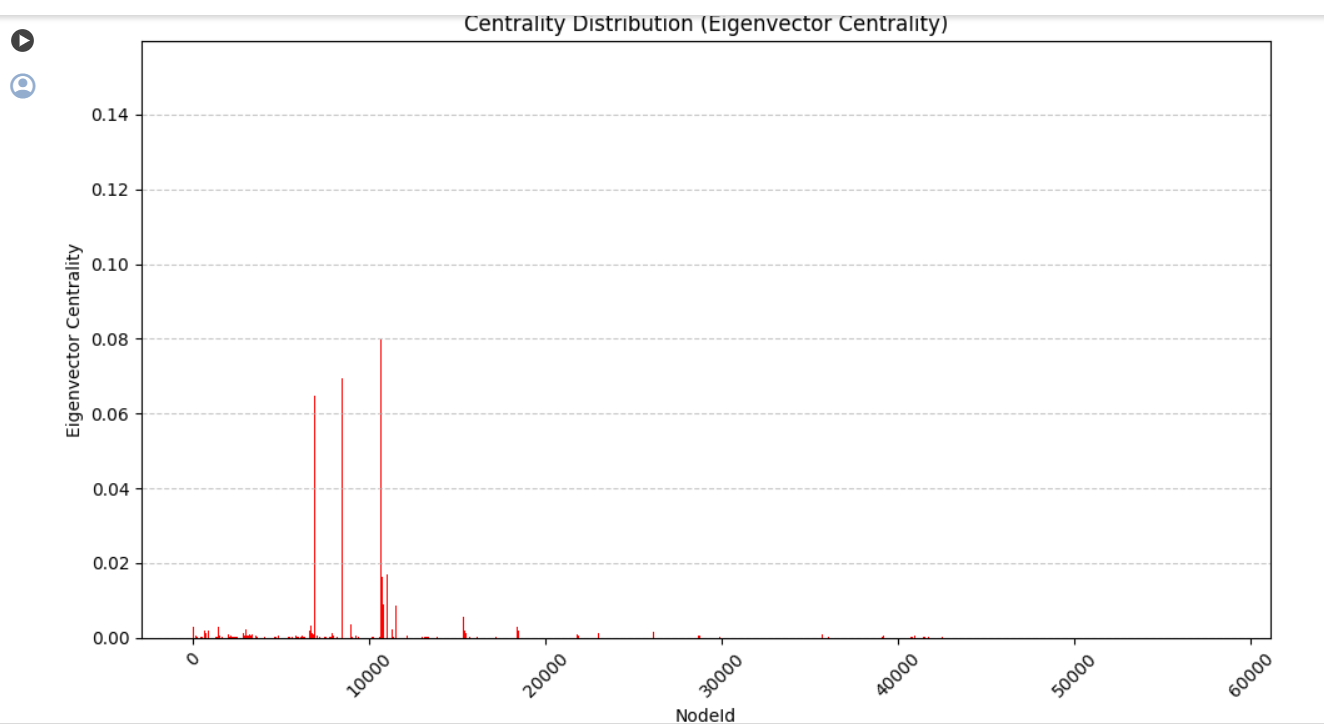
## EigenVector Centrality

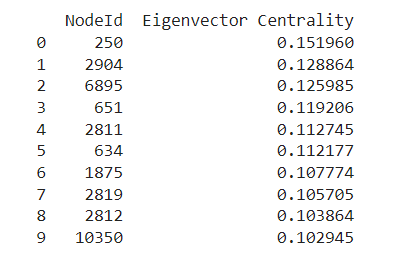
It helps us to calculate the influence of a node in the graph. More the value of eigenvector centrality, implies a node has more connections with more important nodes of the graph.

### Inference

1. Nodes with high eigenvector centrality are connected to other nodes with high centrality.
2. Users with high eigenvector centrality may not have many direct connections, but their connections are with influential users, suggesting their potential influence within the network.





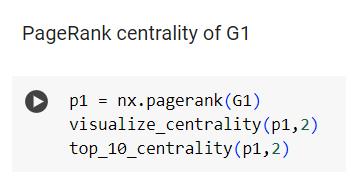


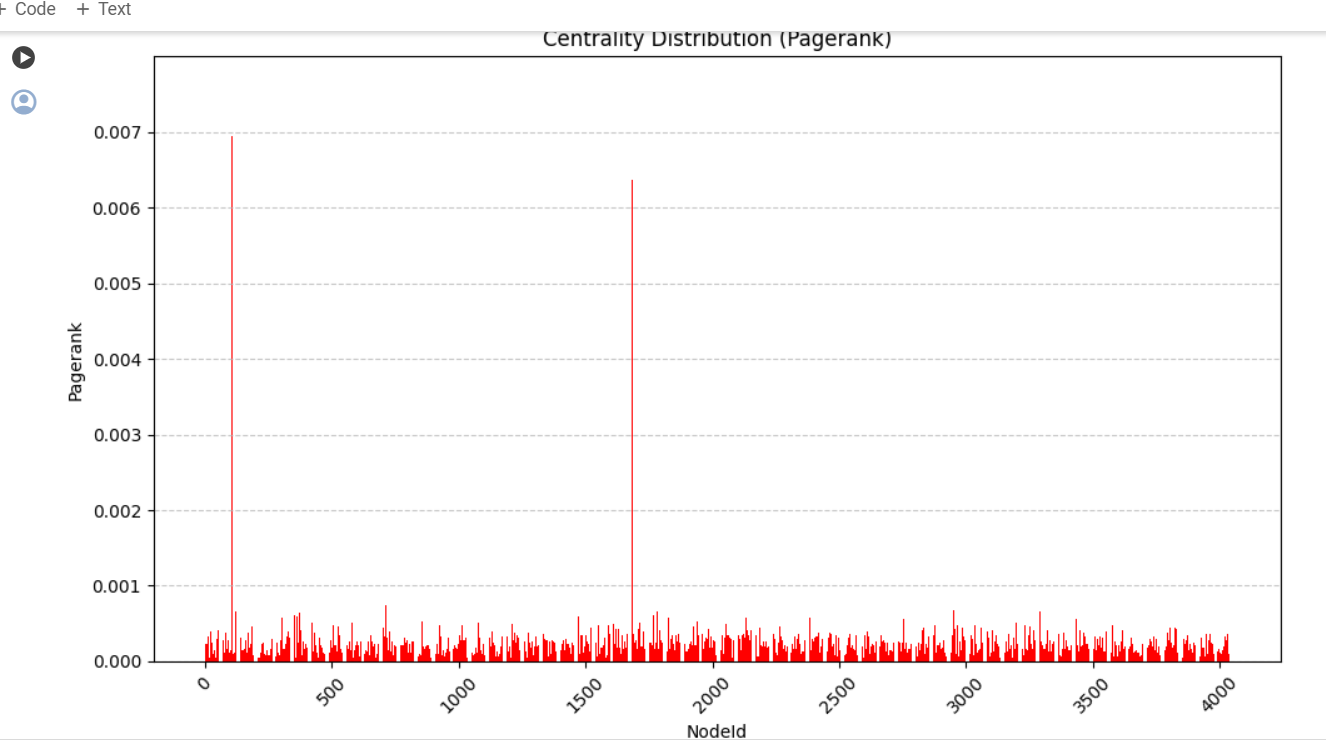
## PageRank Centrality

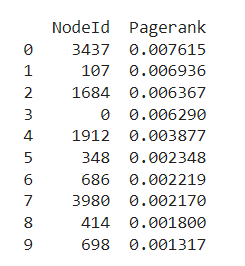
Variant of EigenVector centrality. It helps to address the problem by Katz Centrality of passing a node’s high centrality to its every connection. But in reality, a node which has connection to a higher centrality node need not be important.

### Inference

1. Nodes with high PageRank scores are important and influential within the network. They are typically well-connected and receive connections from other important nodes.
2. High PageRank users in the Brightkite network are those who are not only connected to many other users but also receive connections from other influential users. They are considered authorities or hubs in the network.





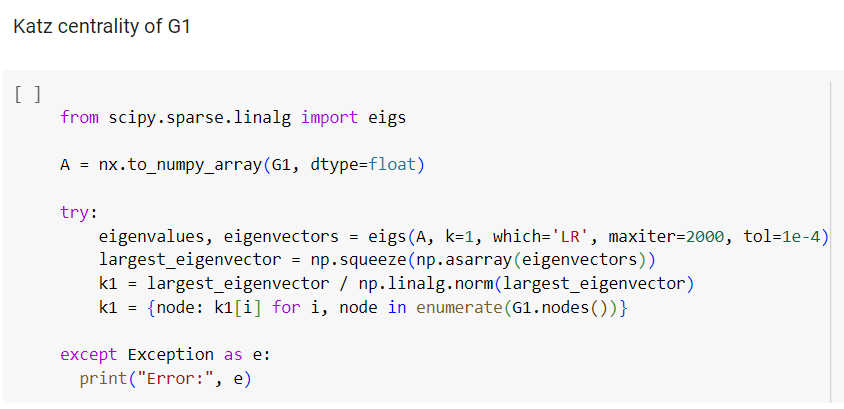


## Katz Centrality

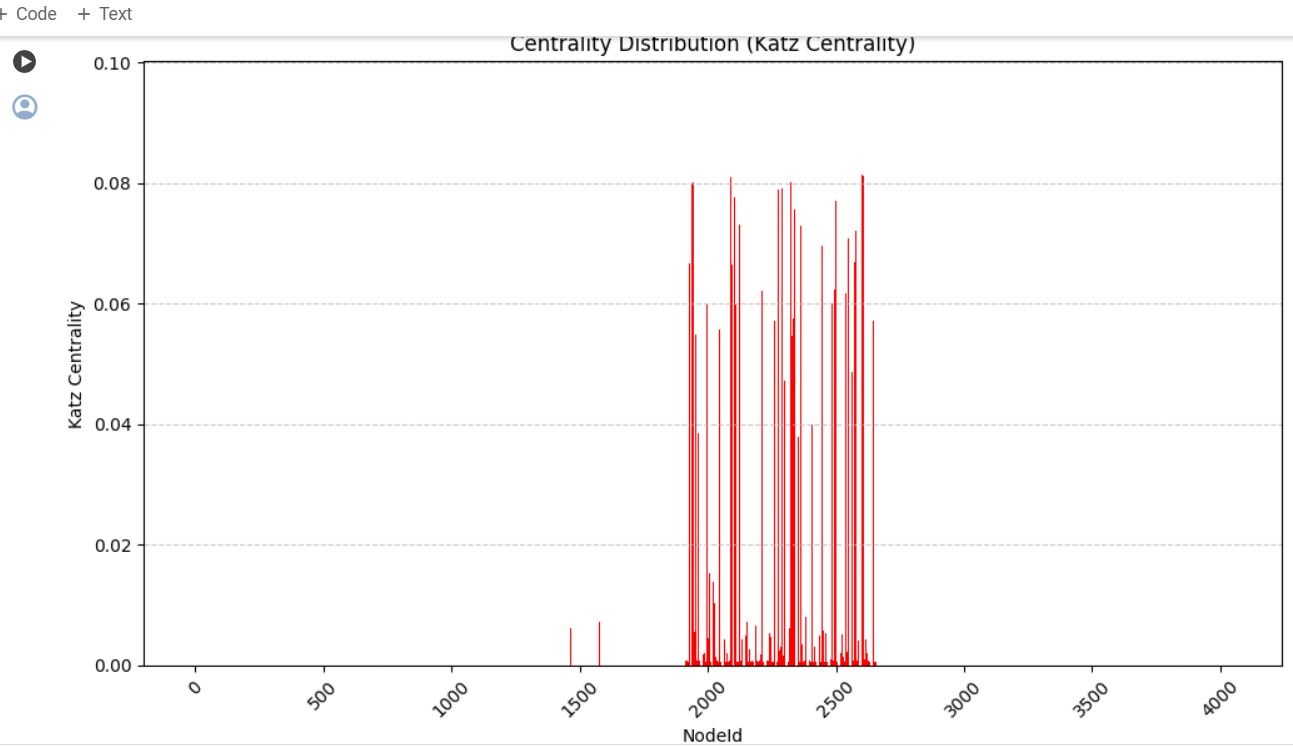
It measures the relative influence of a node within a network. It measures influence by total walks between a pair of nodes, unlike the other centrality measures which consider shortest paths.

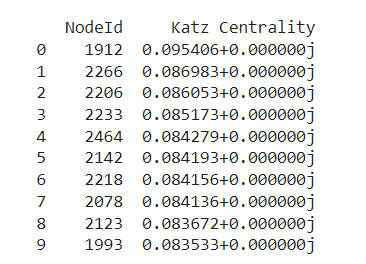
### Inference

1. High Katz centrality indicates users who not only have many direct connections but also connections to other highly central users. These users are influential due to their proximity to other influential individuals in the network.
2. It identifies users who might have indirect influence through their connections with other influential users, even if they themselves do not have a large number of direct connections.









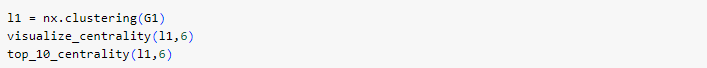
## **Clustering Coefficient**

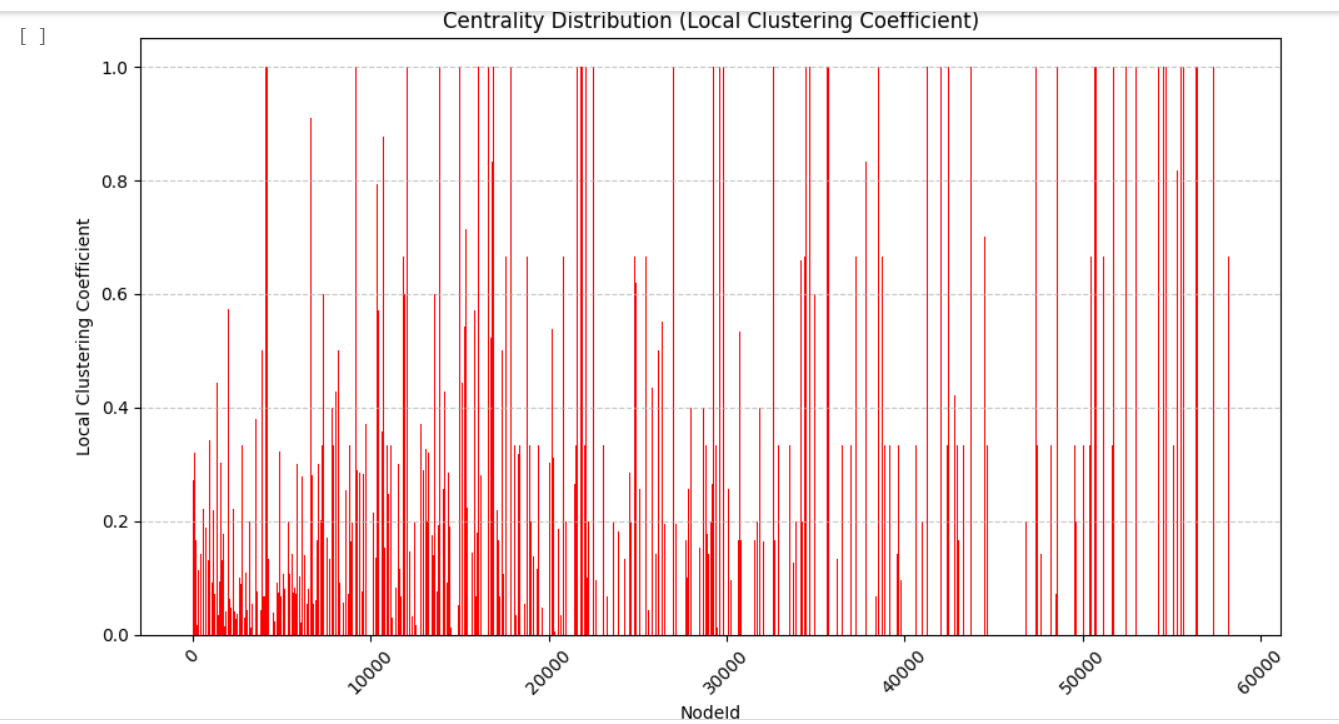
## Local Clustering Coefficient

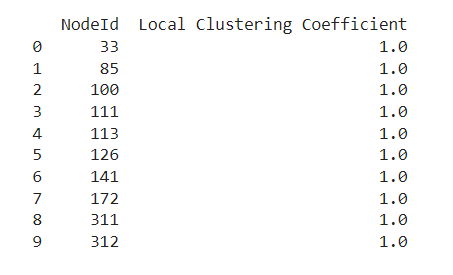
Clustering coefficients help us to measure the level of connectivity between the nodes and its neighbors.

### Inference

1. The clustering coefficient of a node measures the likelihood that its neighbors are also connected to each other.
2. In Brightkite, a high clustering coefficient for a user indicates that their friends are likely to be friends with each other as well.
3. High clustering coefficients suggest the presence of tightly knit clusters or communities within the network, where users tend to form interconnected groups.





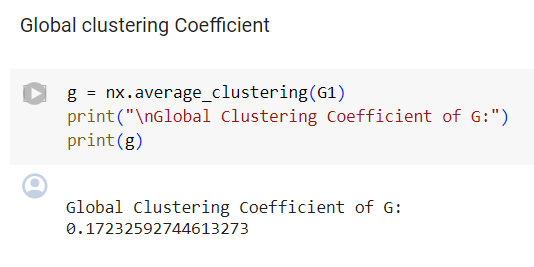


## Global Clustering Coefficient

Defines the number of closed triplets formed in our network. Helps to measure transitivity of a network.

### Inference

1. The global clustering coefficient measures the extent to which the entire network is organized into clusters or communities.
2. In Brightkite, a high global clustering coefficient indicates that the network is highly clustered overall, with many nodes forming connections within local clusters or communities.
3. It suggests a network structure where users tend to form groups where their friends are also connected to each other, leading to a globally clustered network topology.

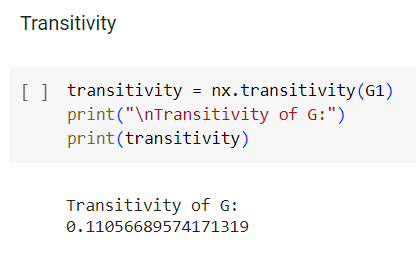


## Transitivity

It is defined as the proportion of the number of triangles forming in our network and the number of connected triples.

### Inference

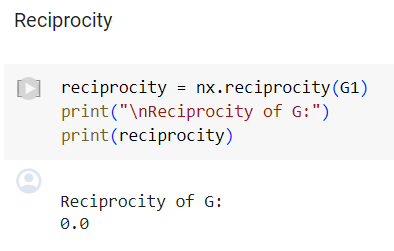
1. High transitivity suggests the presence of tightly interconnected clusters or communities within the network.
2. In the context of Brightkite, high transitivity implies that users tend to form groups where their friends are also connected to each other. This indicates the presence of cohesive social circles or communities within the network.



## Reciprocity

It is defined as the probability of 2 nodes being connected in a closed loop. It tells about the likelihood of the vertices in a directed network to be mutually linked.

Inference: The Graph is UNDIRECTED so reciprocity is 0.



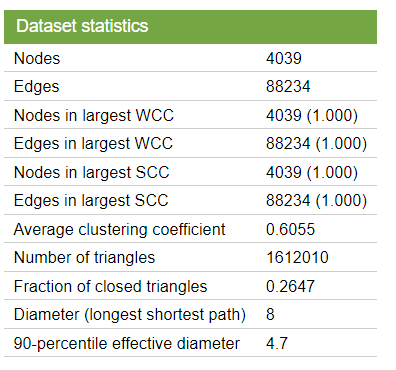
# Dataset 2 Information

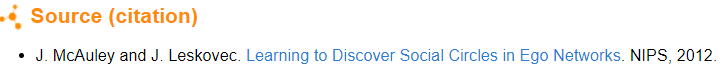
(https://snap.stanford.edu/data/ego-Facebook.html)

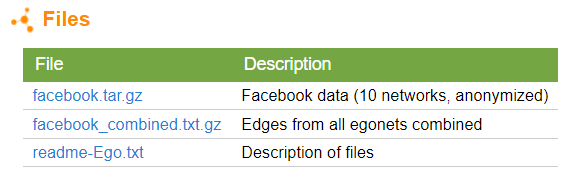
This dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook data was collected from survey participants using this [Facebook app](https://www.facebook.com/apps/application.php?id=201704403232744). The dataset includes node features (profiles), circles, and ego networks.

Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value. Also, while feature vectors from this dataset have been provided, the interpretation of those features has been obscured. For instance, where the original dataset may have contained a feature "political=Democratic Party", the new data would simply contain "political=anonymized feature 1". Thus, using the anonymized data it is possible to determine whether two users have the same political affiliations, but not what their individual political affiliations represent.

Data is also available from [Google+](https://snap.stanford.edu/data/ego-Gplus.html) and [Twitter](https://snap.stanford.edu/data/ego-Twitter.html).

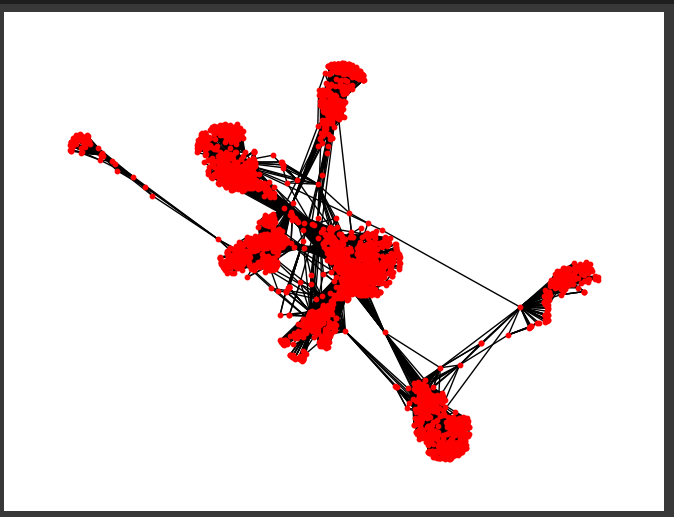




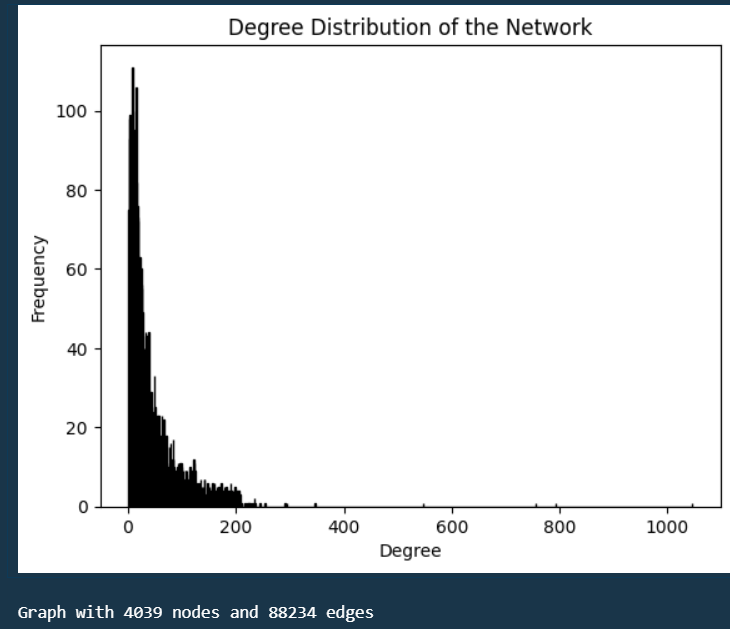


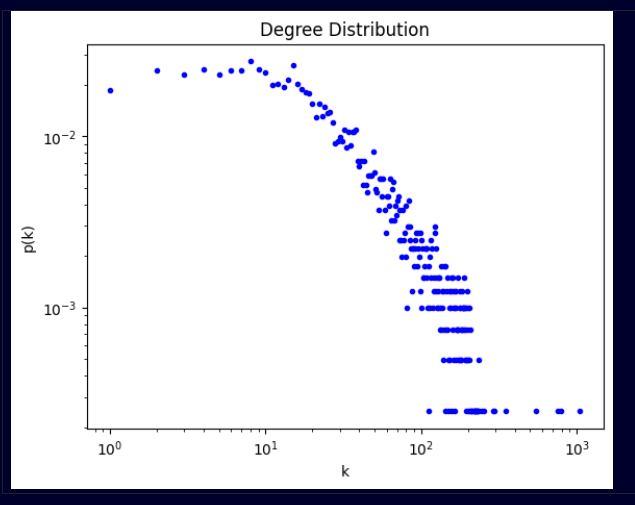


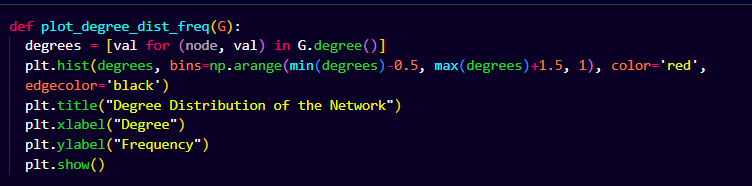


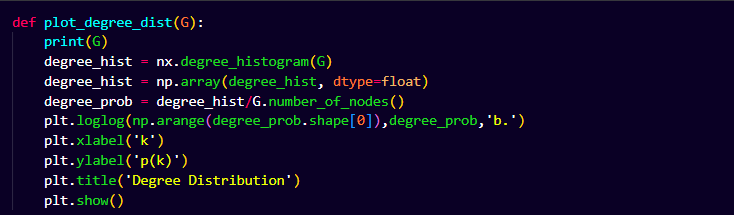
****

**Degree Distribution:**  It is used to know about how the nodes are interrelated with other nodes, high degree means more inter-related relationship.

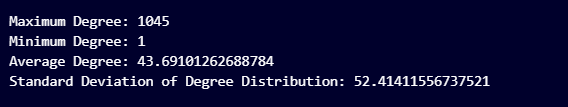


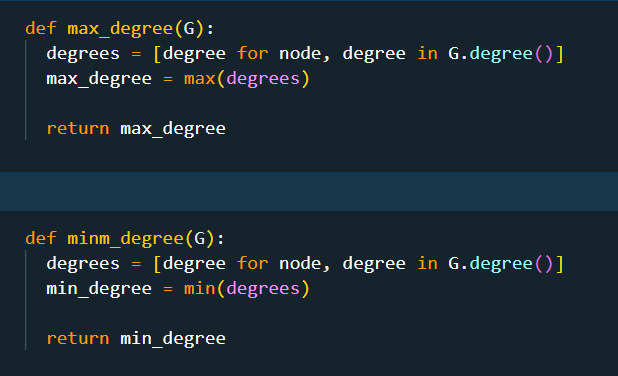


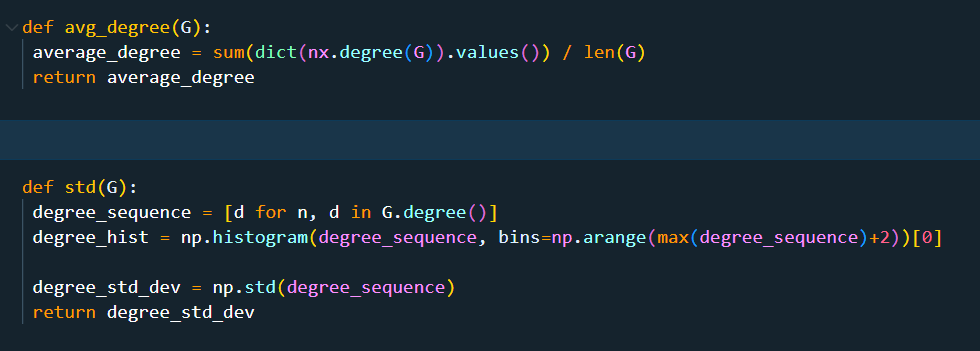


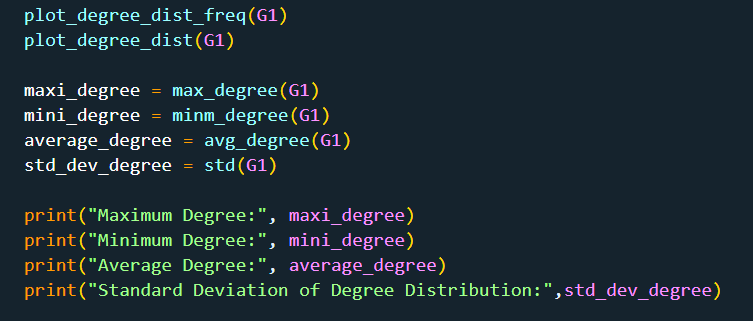




**Degree Calculation:** ****

****

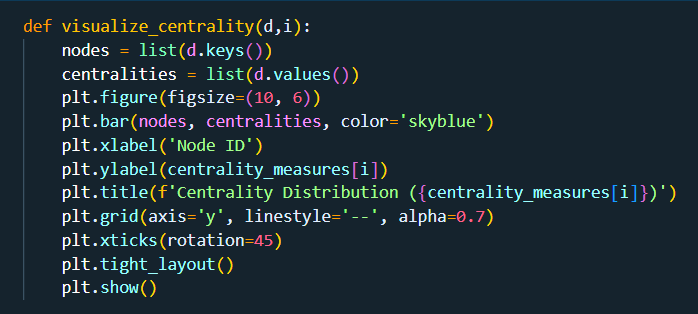
****

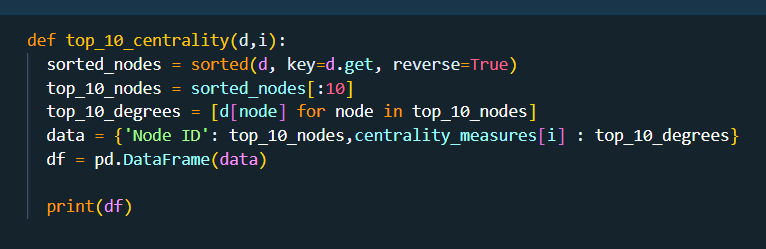
****

# Centrality and Other Measures

Centrality defines the importance of a node in the graph.







## Degree Centrality

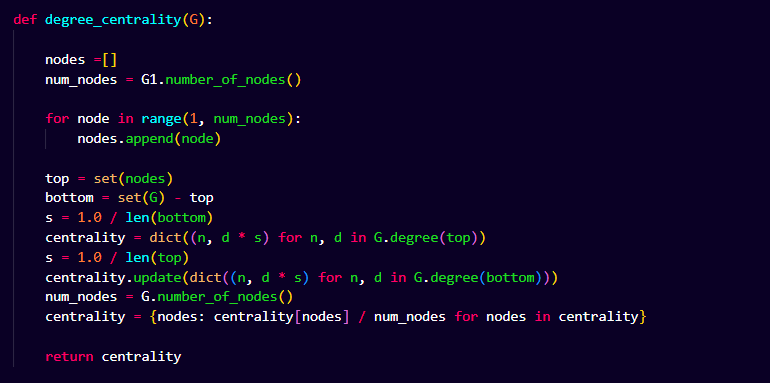
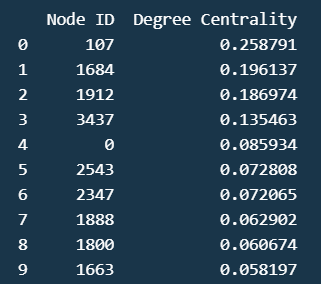
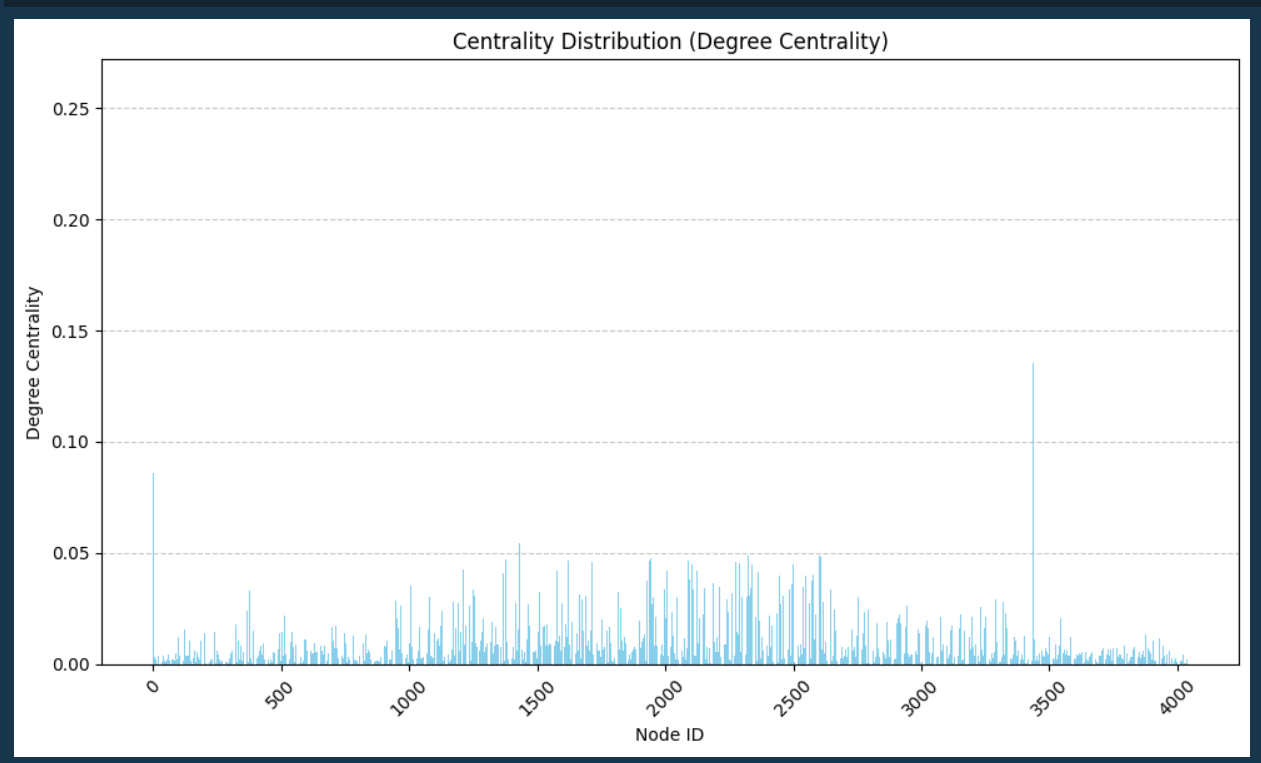
Users with lower degree centrality might be less active or have fewer connections within the network.

**Inference:**

1) Users with high degree centrality likely have numerous connections or friends within the network.

2) They might represent individuals who are actively engaged or have a wide social circle within the platform.

3) High-degree users could be influential in terms of spreading information or having a broad reach within their social circles.





## Closeness Centrality

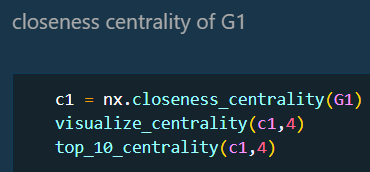
It is calculated as the reciprocal of the sum of shortest path lengths between the node *v* and all the other nodes in a graph. It suggests how efficiently a node is spreading information through the graph .

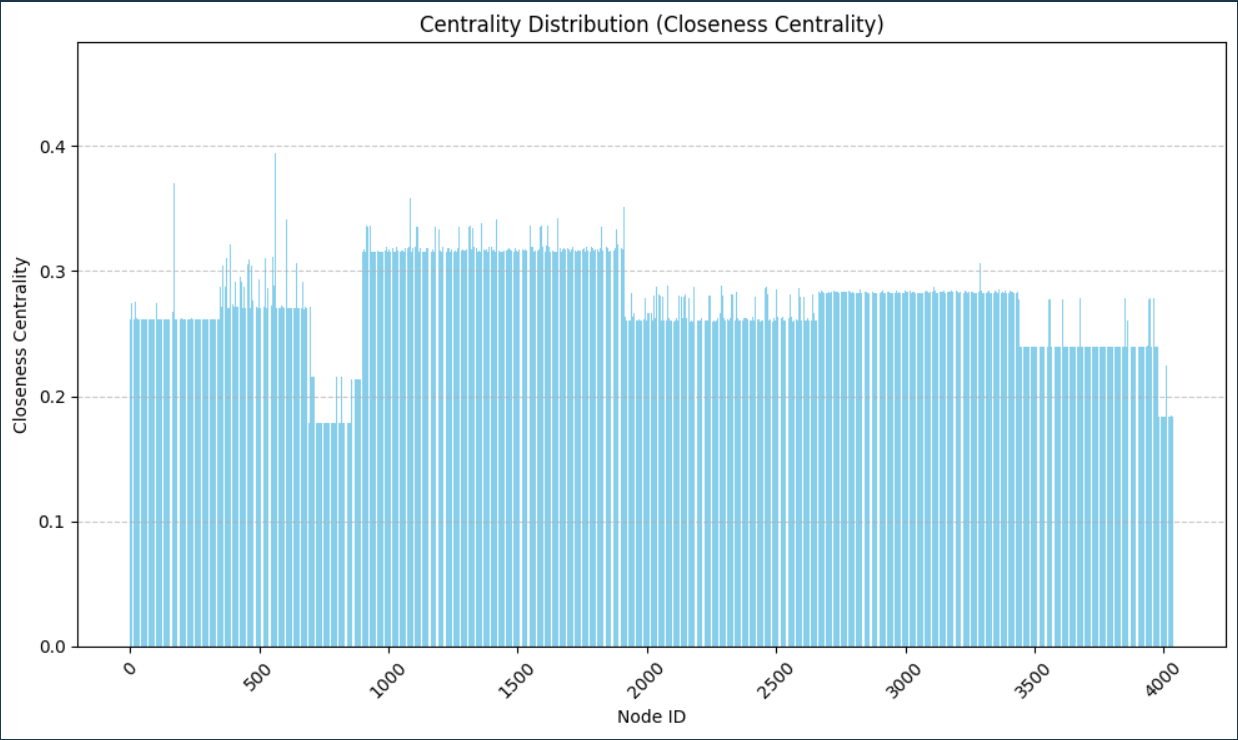
**Inference:**

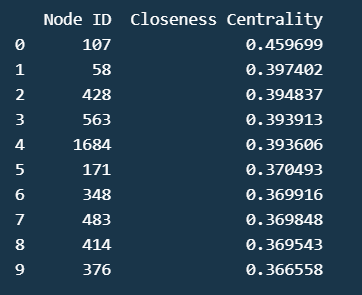
1) Users with high closeness centrality are close to many other users in terms of network distance.

2) They can quickly disseminate information or influence to a large portion of the network.

3) High-closeness users may be well-positioned to exert influence or play pivotal roles in spreading trends or messages across the platform.







## Betweenness Centrality

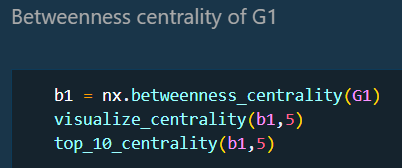
It helps to measure how important nodes are in connecting other nodes. For a vertex *vi,* we can measure the extent to which *vi* lies on shortest paths between other vertices.

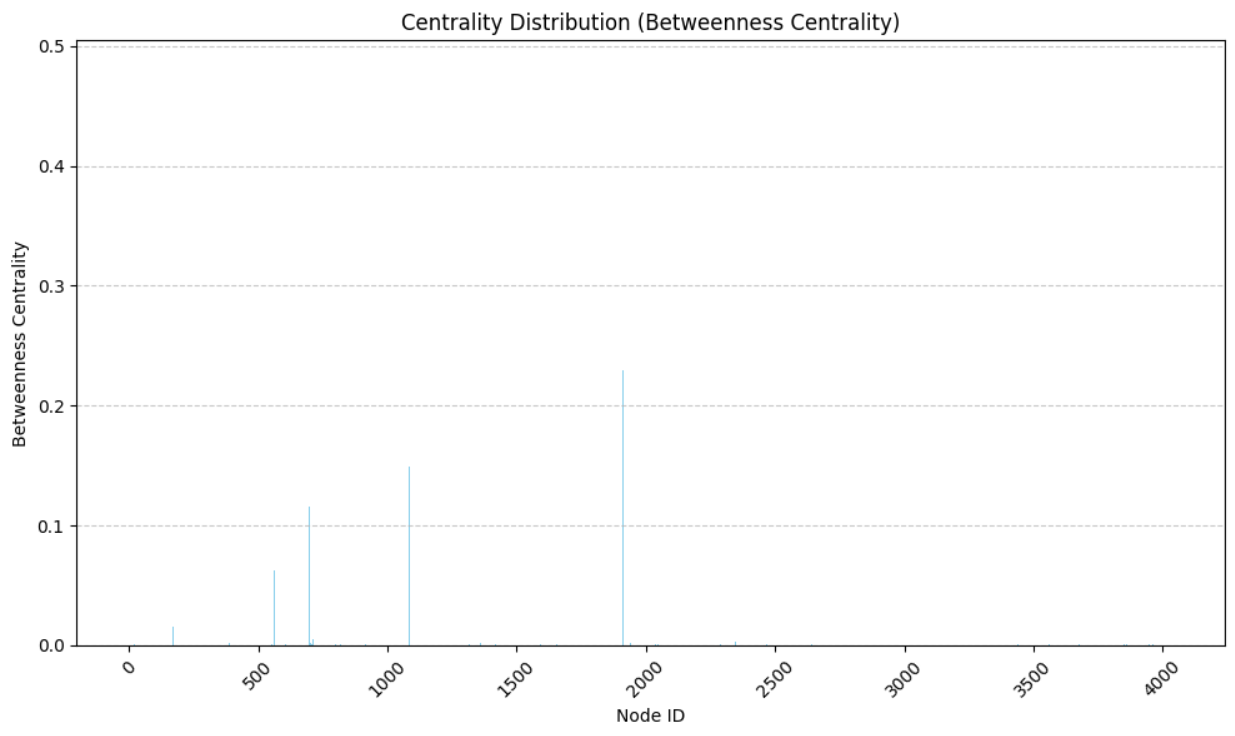
### Inference

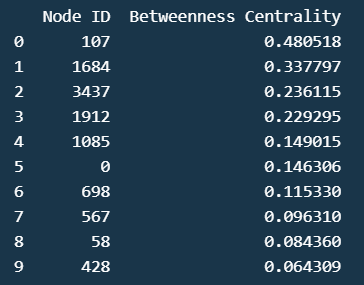
## Nodes with high betweenness centrality serve as bridges or intermediaries between different groups of users.

## They likely connect otherwise disjointed parts of the network, facilitating communication and information flow.

## Users with high betweenness centrality may act as key connectors or influencers, with the ability to bridge diverse social clusters.







## EigenVector Centrality

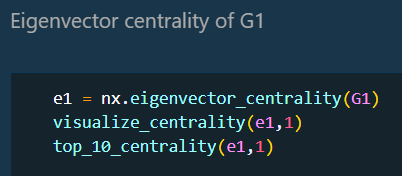
It helps us to calculate the influence of a node in the graph. More the value of eigenvector centrality, implies a node has more connections with more important nodes of the graph.

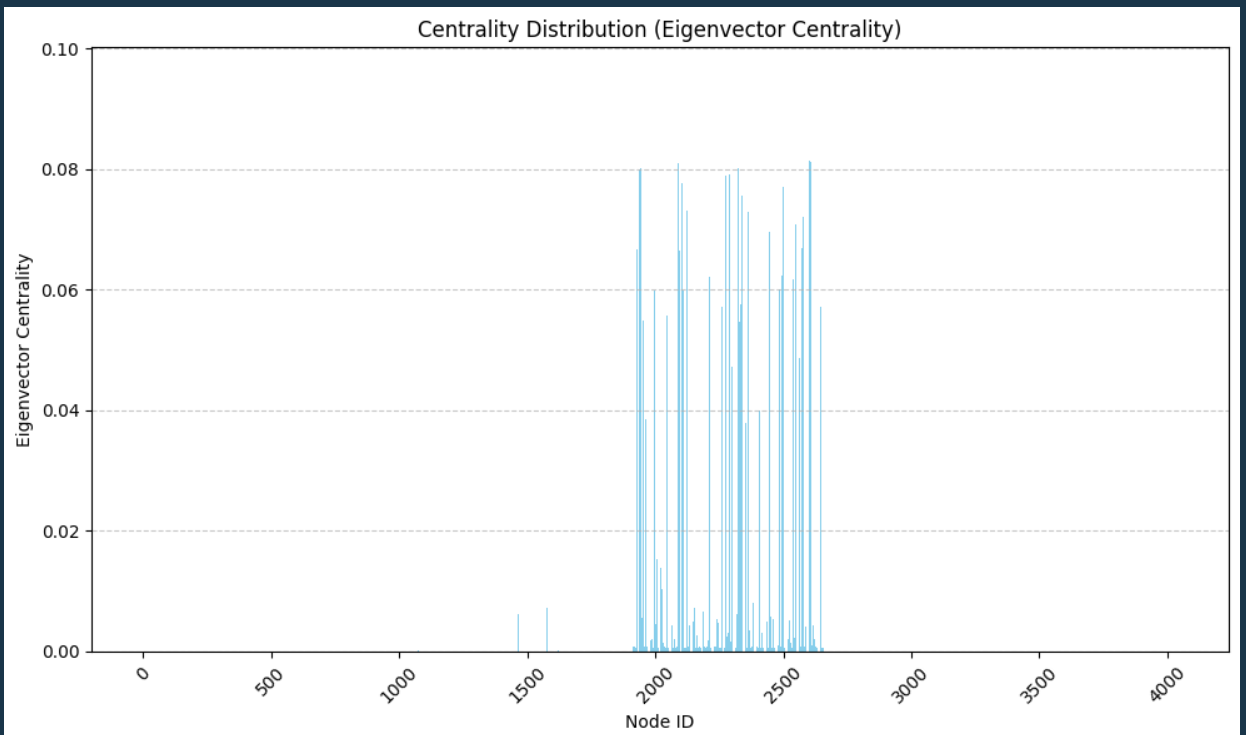
### Inference

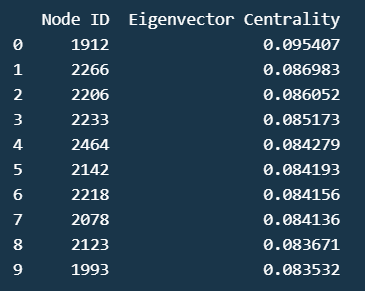
1) Nodes with high eigenvector centrality are connected to other highly central nodes.

2) They may not have many direct connections themselves but are linked to influential users.

3) Users with high eigenvector centrality likely hold significant influence within the network, as their connections are with other influential individuals.







## PageRank Centrality

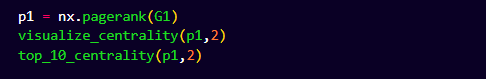
Variant of EigenVector centrality. It helps to address the problem by Katz Centrality of passing a node’s high centrality to its every connection. But in reality, a node which has connection to a higher centrality node need not be important.

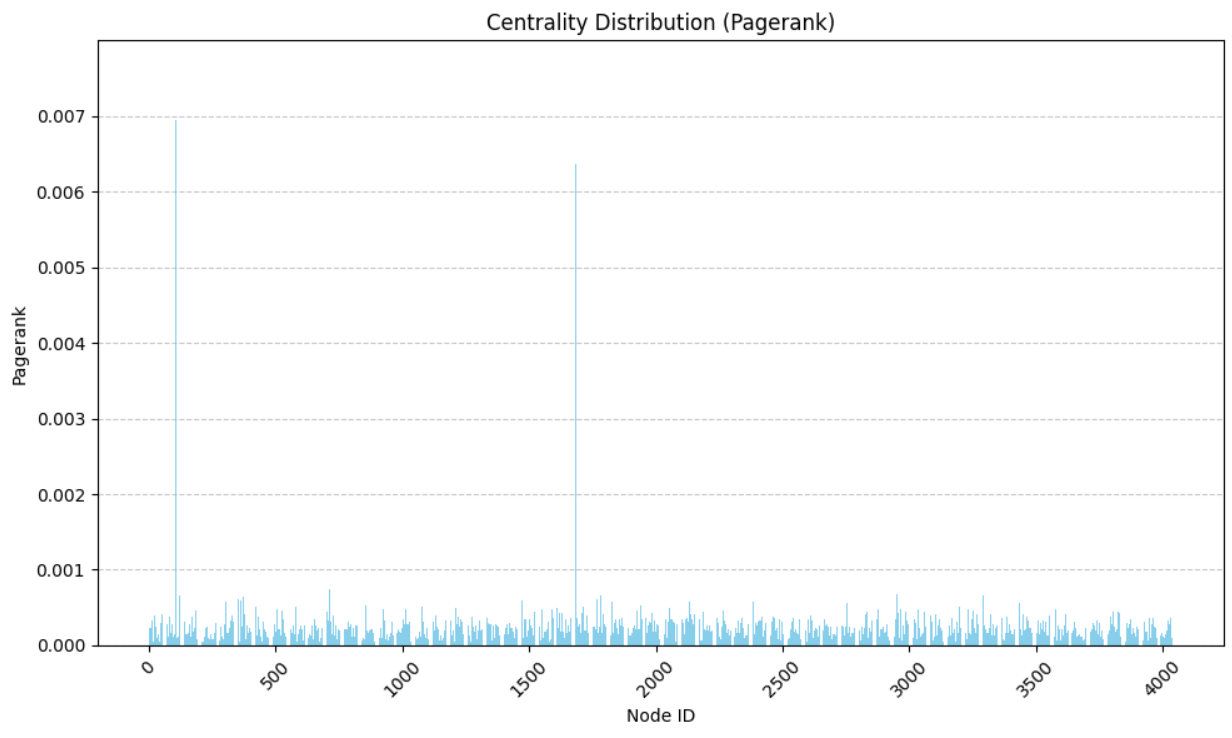
### Inference

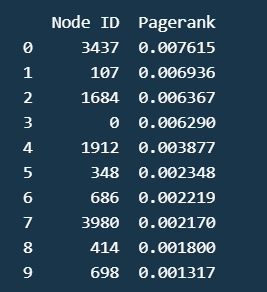
## 1) PageRank assigns a score to each node based on its incoming links and the importance of the nodes linking to it.

## 2) Nodes with high PageRank scores are considered important or authoritative within the network.

## 3) Users with high PageRank scores likely have connections from other important users, indicating their significance or influence within the network.







## Katz Centrality

It measures the relative influence of a node within a network. It measures influence by total walks between a pair of nodes, unlike the other centrality measures which consider shortest paths.

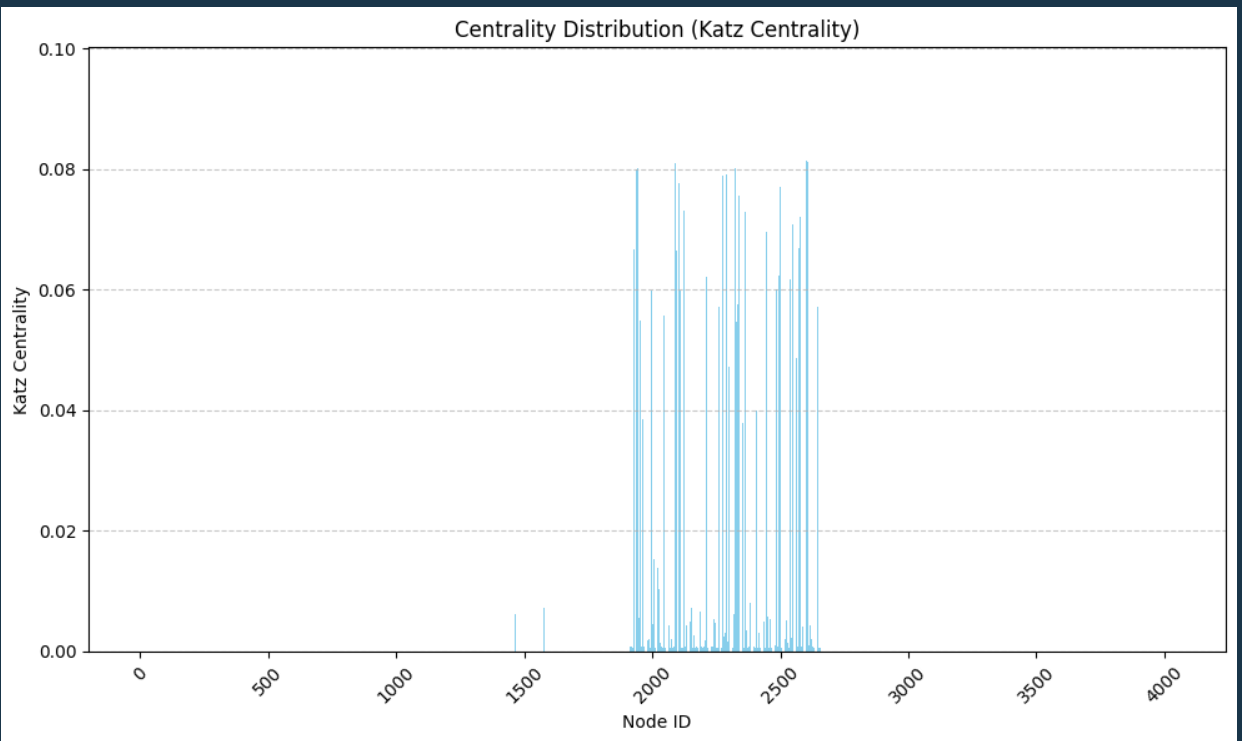
### Inference

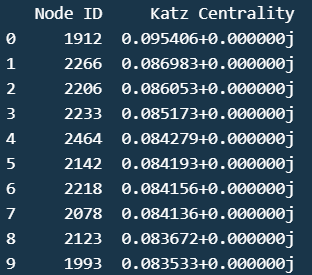
1) Katz centrality measures the influence of a node based on the contributions from its neighbors, with diminishing weights for nodes further away.

2) Users with high Katz centrality are not only directly connected to many other users but also connected to other highly central users.

3) These users likely have significant influence within the network, either through direct connections or connections to other influential individuals.





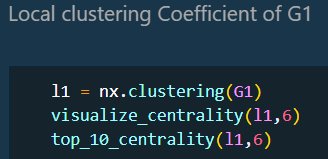
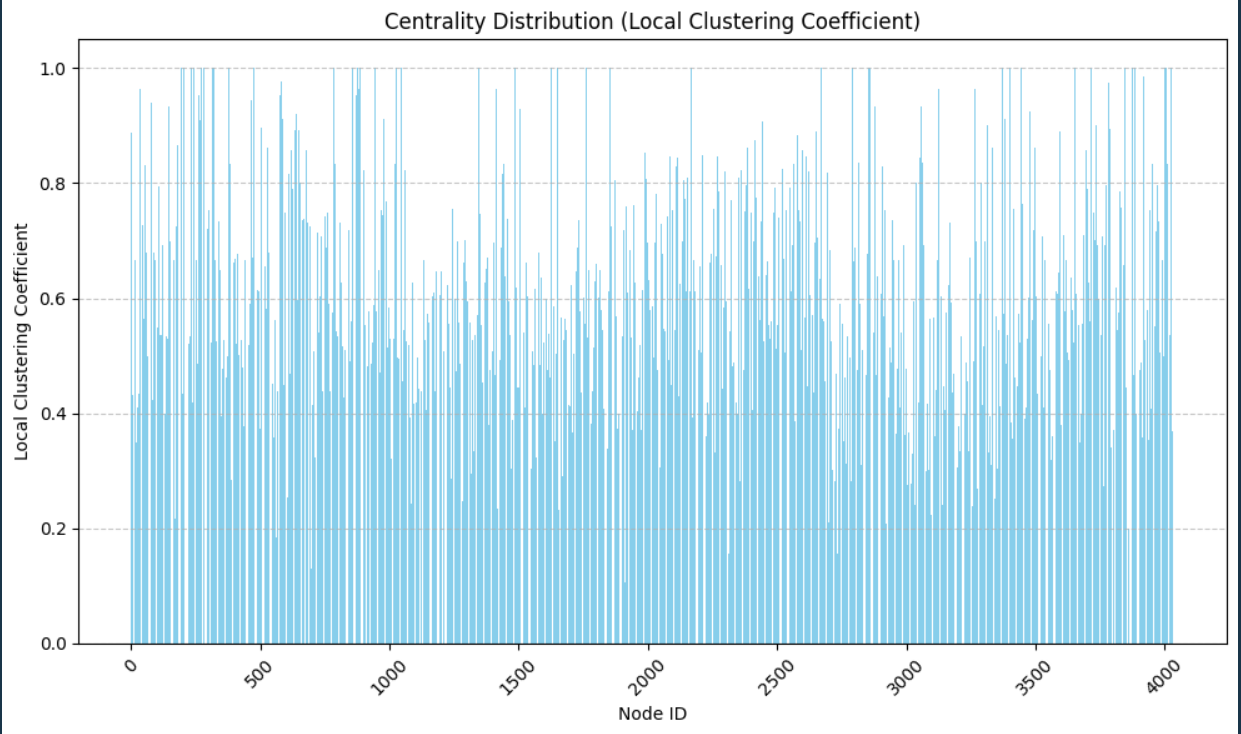


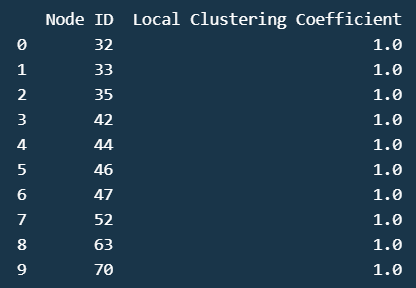
## **Clustering Coefficient:**

## Local Clustering Coefficient

Clustering coefficients help us to measure the level of connectivity between the nodes and its neighbors.

### Inference

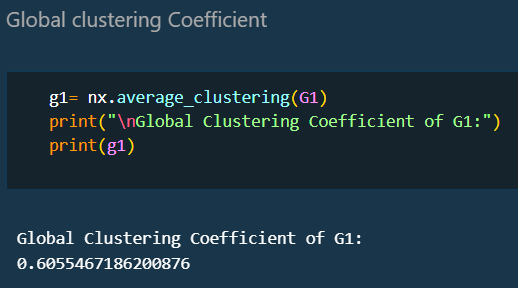
* The clustering coefficient of a node measures the likelihood that its neighbors are also connected to each other.
* In the anonymized social network dataset, a high clustering coefficient for a user suggests that their friends are likely to be friends with each other as well.
* This indicates the presence of cohesive clusters or communities within the network, where users mutually connect with each other.
* 
* 



## Global Clustering Coefficient

Defines the number of closed triplets formed in our network. Helps to measure transitivity of a network.

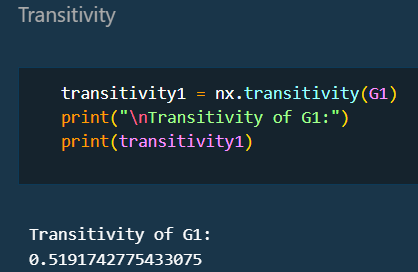
### Inference

* The global clustering coefficient measures the extent to which the entire network is organized into clusters or communities.
* In the anonymized social network dataset, a high global clustering coefficient indicates that the network is highly clustered overall.
* It suggests that users tend to form groups where their friends are also connected to each other, leading to a globally clustered network topology.
* 

## Transitivity

It is defined as the proportion of the number of triangles forming in our network and the number of connected triples.

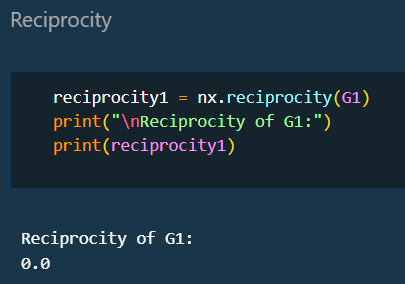
### Inference

* Transitivity measures the tendency of nodes to form triangles or clusters of interconnected nodes.
* High transitivity suggests that users' friends are likely to be friends with each other as well.
* In the anonymized social network dataset, high transitivity indicates the presence of tightly knit communities or cliques where users mutually connect with each other.
* 

## Reciprocity

It is defined as the probability of 2 nodes being connected in a closed loop. It tells about the likelihood of the vertices in a directed network to be mutually linked .

*INFRENCE:* The Graph is UNDIRECTED so reciprocity is 0.



# COMPARSIONS

***FIRST GRAPH:***

* Maximum Degree: 1134
* Minimum Degree: 1
* Average Degree: 7.3531
* Standard Deviation of Degree Distribution: 20.3556
* Global Clustering Coefficient: 0.6055
* Transitivity: 0.5192

***SECOND GRAPH:***

* Maximum Degree: 1045
* Minimum Degree: 1
* Average Degree: 43.6910
* Standard Deviation of Degree Distribution: 52.4141
* Global Clustering Coefficient: 0.1723
* Transitivity: 0.1106

**INFERENCES**:

* The 1st graph has lower avg degree(7.3531) compared to 2nd graph(43.6910), indicating the **2nd graph has nodes that are highly connected on average.**
* The Standard Deviation of degree distribution is also higher in ***second graph(52.4141) indicating a wider range of node degrees compared to first graph***(20.3556).
* The **first graph has significant higher global clustering coefficient and transitivity** values compared to second graph. This tells us that even though 1st graph has lower avg degree and narrow degree distribution, **the nodes in 1st graph tend to from more tightly knit clusters and communities.** This assertion is strongly supported by the network diagram above.
* Conversely, **the second graph**, despite having a higher average degree and wider degree distribution, exhibits lower clustering and transitivity, indicating **a more decentralized or less clustered structure*.*** This assertion is strongly supported by the network diagram above.