*3 January 2021*



**Network and Algorithms  
Final Project***STUDY OF MEASURES ON  
SOCIAL MEDIA DATA*

*Groupmates*Isa Baghirov

Nadir Abdullayev  
Kanan Mikayilov

Shoykyat Sharafyabi

# ***Introduction***

*This project* consists of analyzing measures such as centrality, betweenness, and much more on data collected from a popular social media app called Instagram. These measures help us identify the most influential members of the social network. For each metric, we will display the graphs and elaborate more further on each metric.

*The dataset* used in this project is [***Huawei Social Network Data***](https://www.kaggle.com/andrewlucci/huawei-social-network-data?select=Instagram_Data.xlsx)collected three years ago from which we used only Instagram data. This data was collected by crawling social networks, in our case, Instagram pages. Huawei’s web crawlers extracted the posts, comments, followers, and likes from social media. This data is directed and labeled, it has 1000 nodes and 4933 edges.

*Why is this data important?*It helps to grow businesses, understand how much impact marketing campaigns have on this social network, increase engagement and responsiveness, measure and improve brand awareness, learn from competitors, and so much more.

*Before we start* working with our data, we drop the first column and the first row of our data table since it consists of the names of the Instagram members that were included in this dataset. Additionally, we converted the data into *int64* format since before it was in *object* format whereas our data is composed of numbers.  
  
To make the data more visual, among different heat maps such as *jet, hsv, viridis, etc.* we chose the *rainbow heatmap* since we think it is the best way to make data more descriptive. Basically, the hotter heatmap gets, the more a value of a specific metric is for each node.

*Measures and metrics* we used in this project:

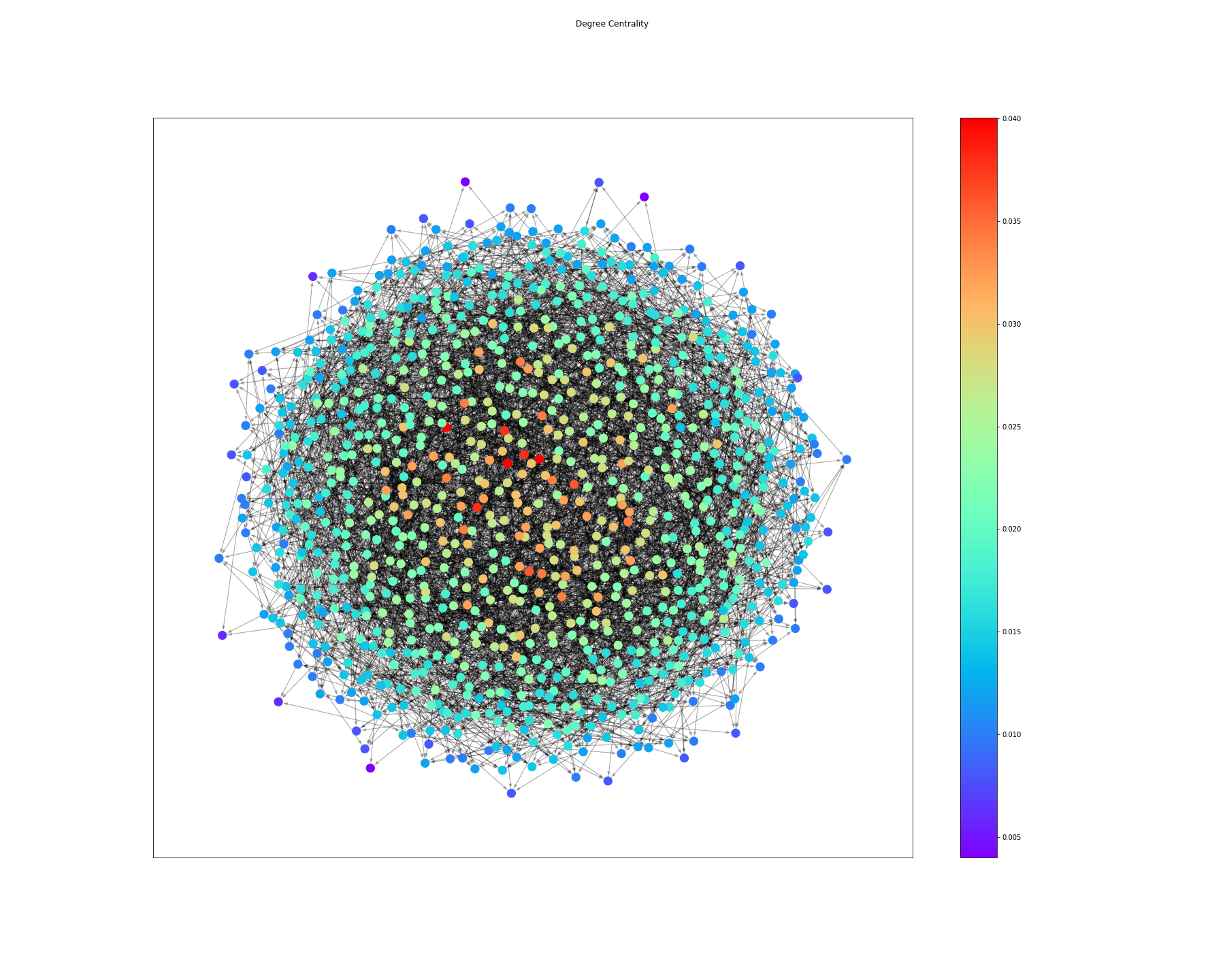
1. [Degree centrality](#5aljd3a2yaq)
2. [Betweenness centrality](#n0gv0zbixu4d)
3. [Katz centrality](#hvt096pgqd4n)
4. [Eigen centrality](#7cavzh434eyq)
5. [Closeness centrality](#1xozltsxnv3z)
6. [Network diameter](#ubongt4ja8jh)
7. [Network density](#9xss807bk6t9)
8. [Network average shortest path length](#b7qe6mggmxt0)

*Further* in this document, we will get more into detail about what these measures mean and what their values are for each analysis we make.

# **1. Degree Centrality**

Degree Centrality is defined as the number of links a node has. As required, we used the NetworkX library for computation inside the program. In a directed network, there are have two separate measures of degree centrality: *indegree* and *outdegree*.  
*Indegree* means the number of ties that are directed to the node.  
*Outdegree* means the number of ties that are directed from the node to other nodes.

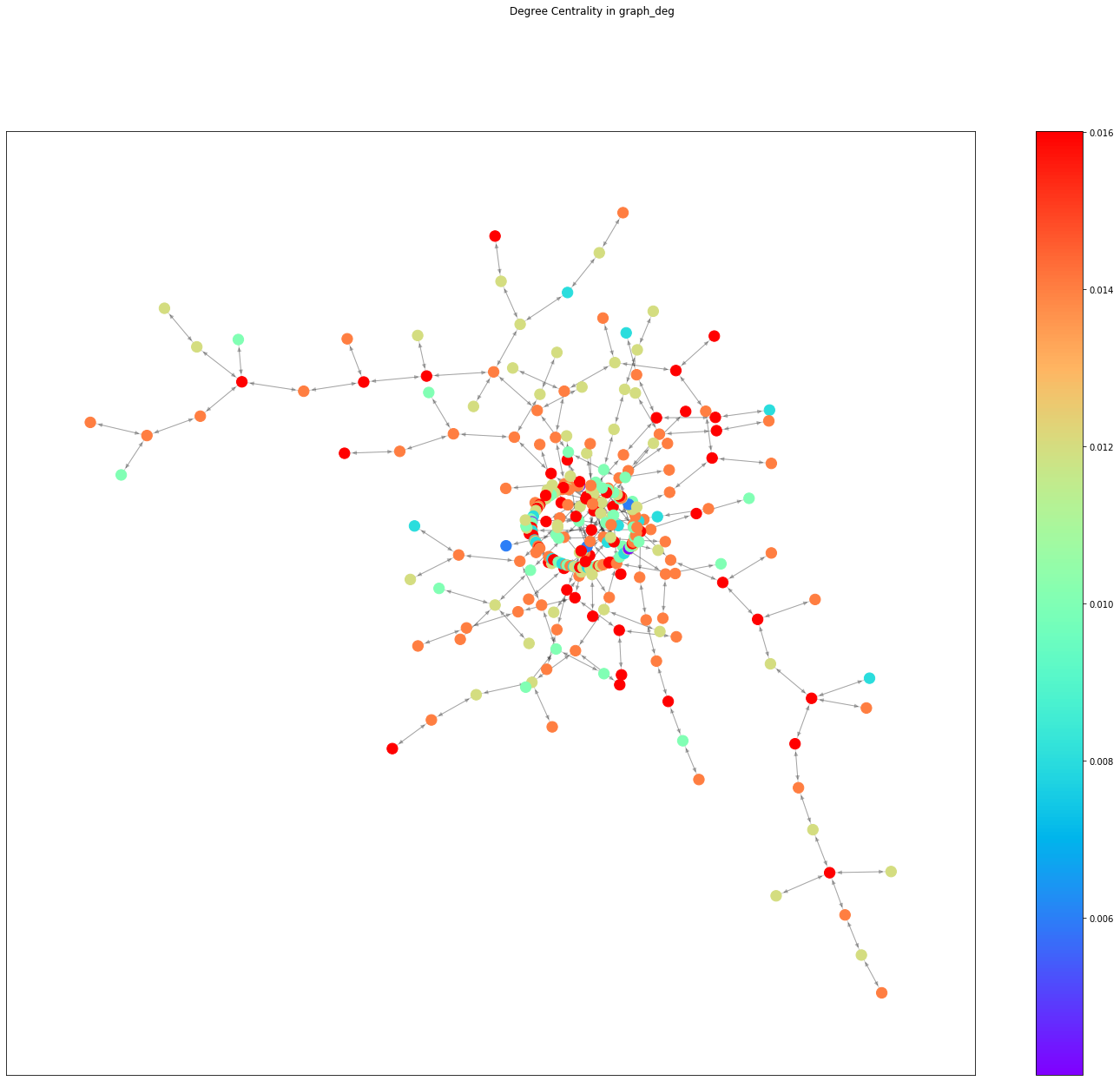
In our case, since our network is directed, degree centrality is the sum of values of indegree and outdegree, so the degree centrality of the baseline network is 19.75175.



## 

## **Degree Centrality Analysis**

For Degree Centrality Analysis we selected nodes that contribute to 80% of the cumulative sum which resulted in the removal of 307 nodes from the graph.



|  |  |
| --- | --- |
| The graph diameter | 5 |
| Density | 0.004087628536756722 |
| Shortest path length | 3.273137137137137 |
| Degree Centrality | 3.9439439439439425 |
| Closeness centrality | 88.90439689916666 |
| Betweenness Centrality | 0.26766295188296063 |
| Katz Centrality | 5.11576088075858 |
| EigenVector centrality | 5.6597264133737895 |

# 

# **2. Betweenness centrality**

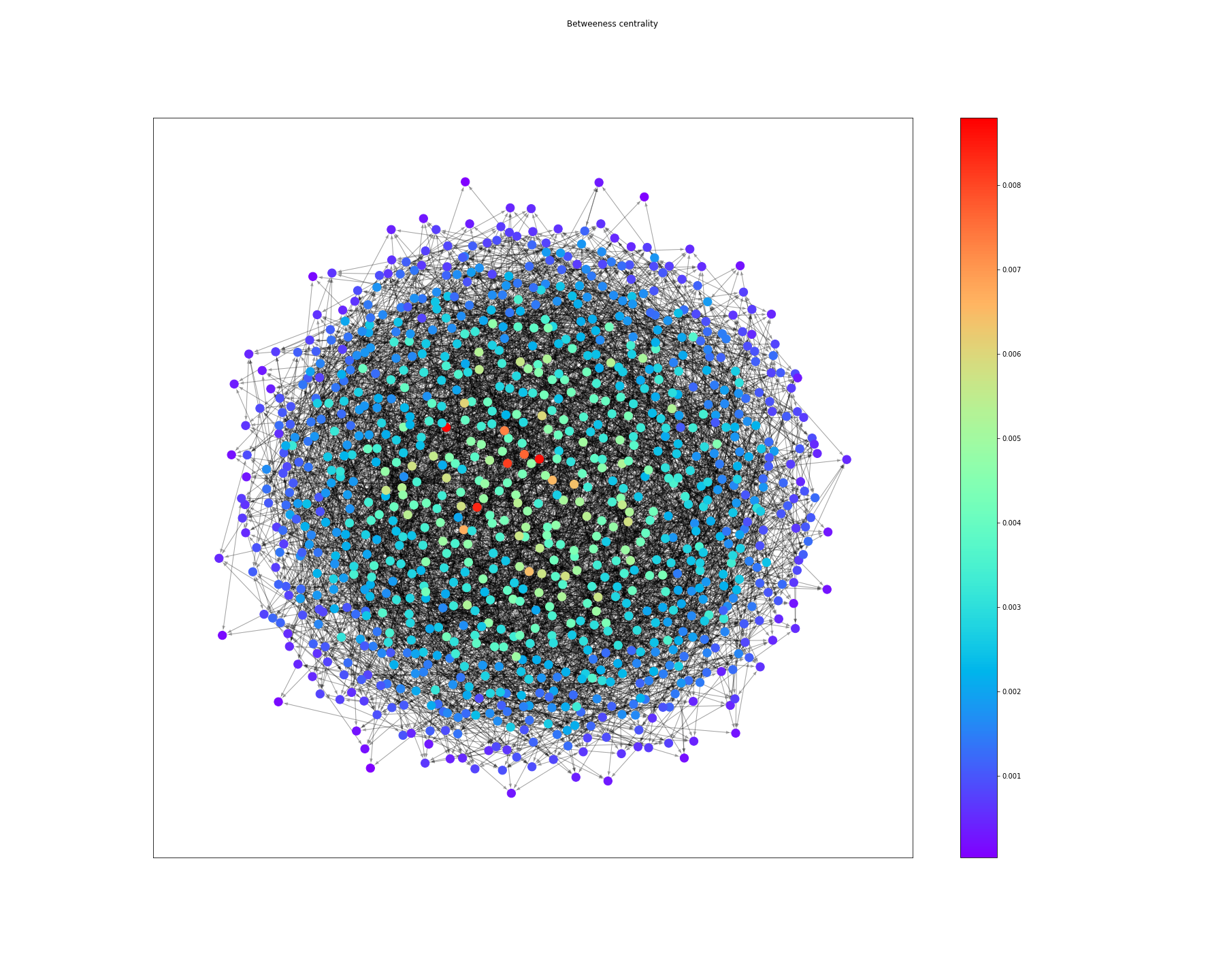
Betweenness centrality is, first of all, a graph, which is based on the shortest paths. For each of the pairs of vertices in a connected graph, there exists at least one shortest path between the vertices.

Betweenness centrality can be defined as a sum of values of betweenness centrality for all nodes. For our network, the Betweenness Centrality is 2.27769.

Betweenness centrality is given by the following expression:



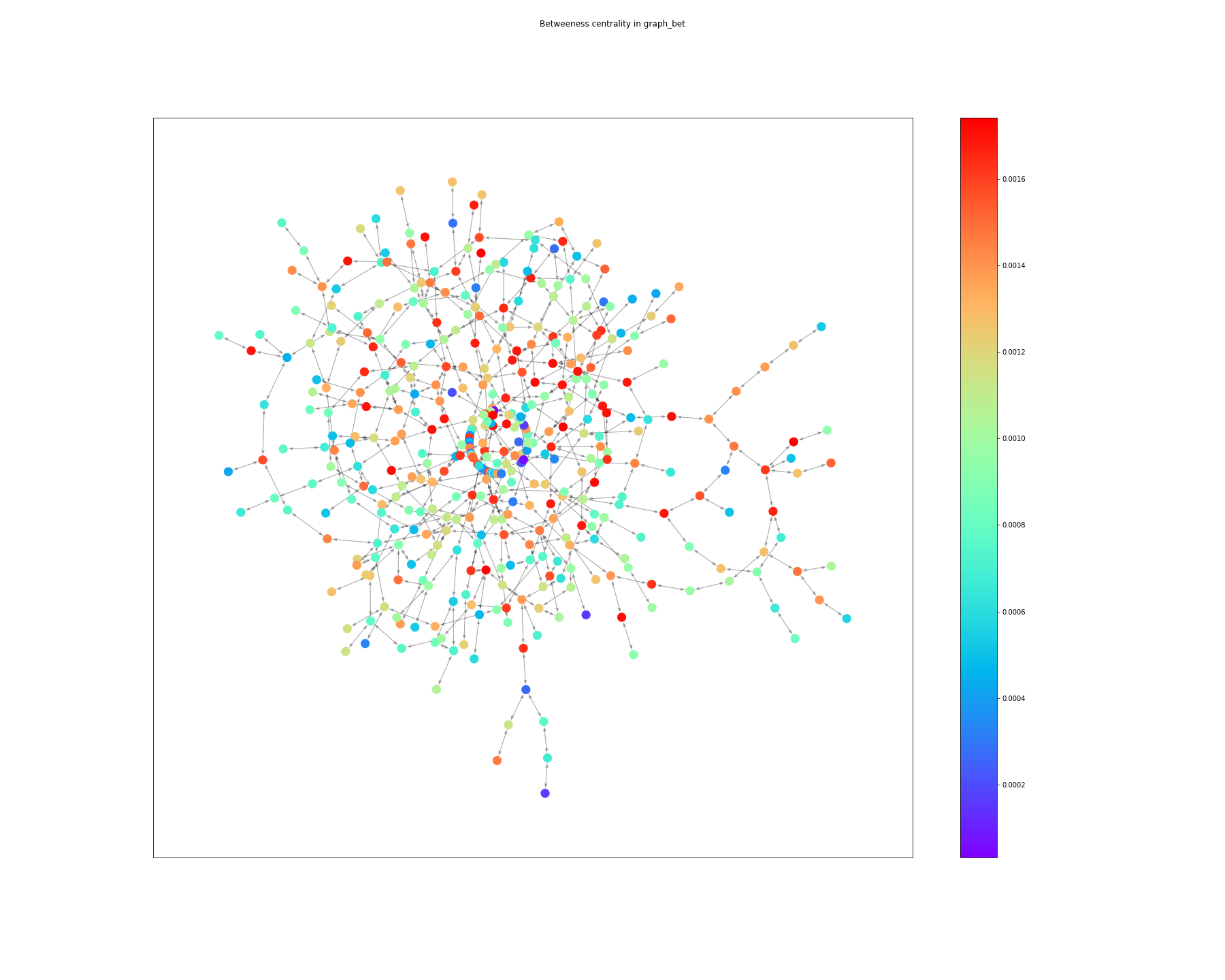
where V is the set of nodes, σ(s, t) is the number of the shortest paths from s to t, and σ(s, t|v) is the number of the shortest paths from s to t passing through the node v.



## 

## **Betweenness Centrality Analysis**

For Betweenness Centrality Analysis we selected nodes that contribute to 80% of the cumulative sum which resulted in the removal of 427 nodes from the graph.

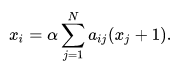


|  |  |
| --- | --- |
| The graph diameter | 5 |
| Density | 0.005013688689514134 |
| Shortest path length | 3.273137137137137 |
| Degree Centrality | 5.999999999999999 |
| Closeness centrality | 125.1961919017603 |
| Betweenness Centrality | 0.45509005629514293 |
| Katz Centrality | 8.075678554348308 |
| EigenVector centrality | 8.737919359406446 |

# 

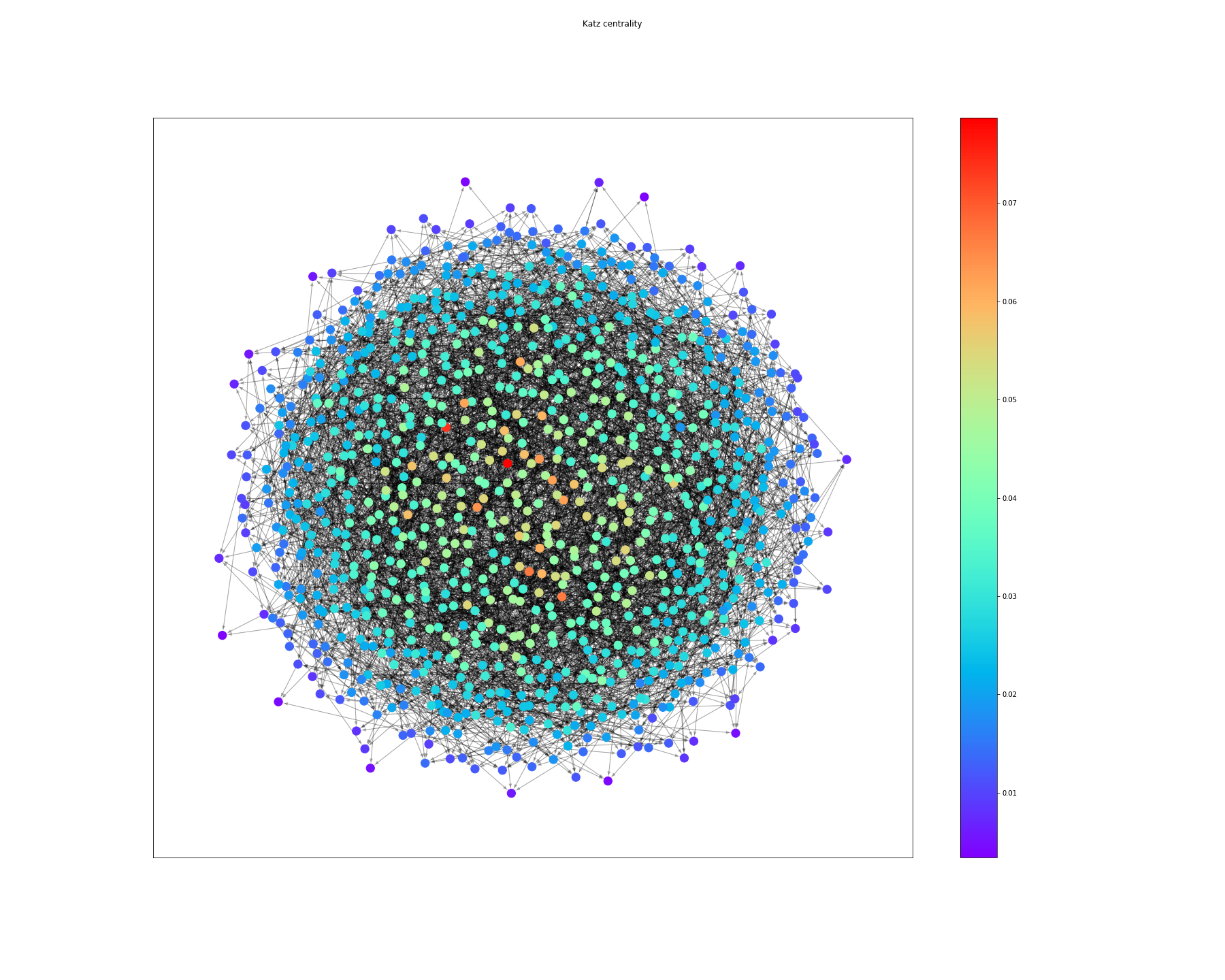
# **3. Katz centrality**

Katz centrality is the relative influence a node has within a network by measuring the number of the neighbors of the node. These neighbors are also called first-degree nodes which means that they are immediate neighbors. The Katz centrality for node *i* can be defined by this formula:



α approaches from below and compared to Eigenvector centrality, is replaced by

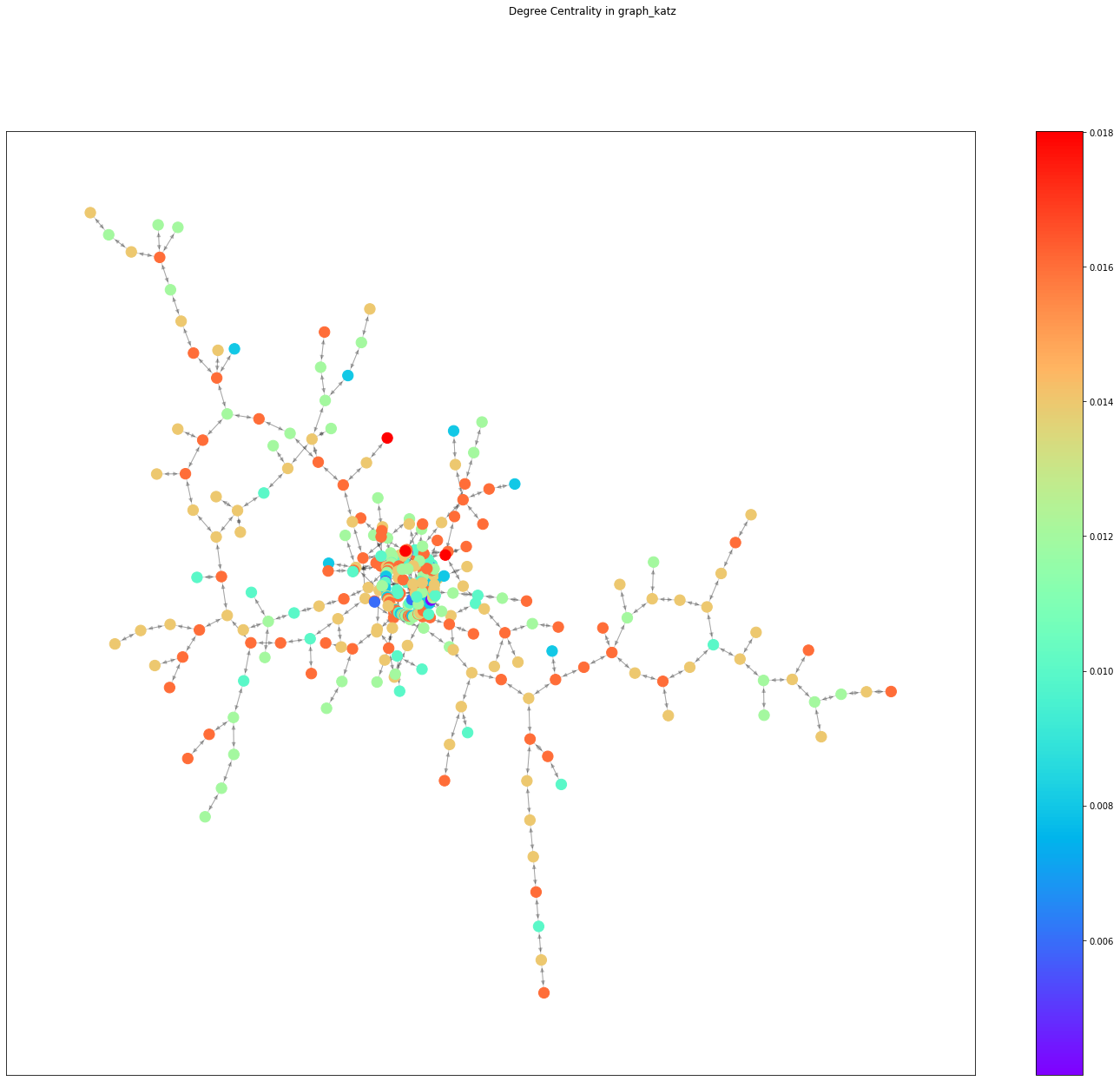
For our network, the Katz Centrality is 29.13566.



## 

## **Katz Centrality Analysis**

For Katz Centrality Analysis we selected nodes that contribute to 80% of the cumulative sum which resulted in the removal of 337 nodes from the graph.

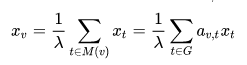


|  |  |
| --- | --- |
| The graph diameter | 5 |
| Density | 0.0034989813092390822 |
| Shortest path length | 3.273137137137137 |
| Degree Centrality | 2.502502502502499 |
| Closeness centrality | 60.96937660878829 |
| Betweenness Centrality | 0.15272709754981095 |
| Katz Centrality | 3.1381732414557066 |
| EigenVector centrality | 3.5518385691107923 |

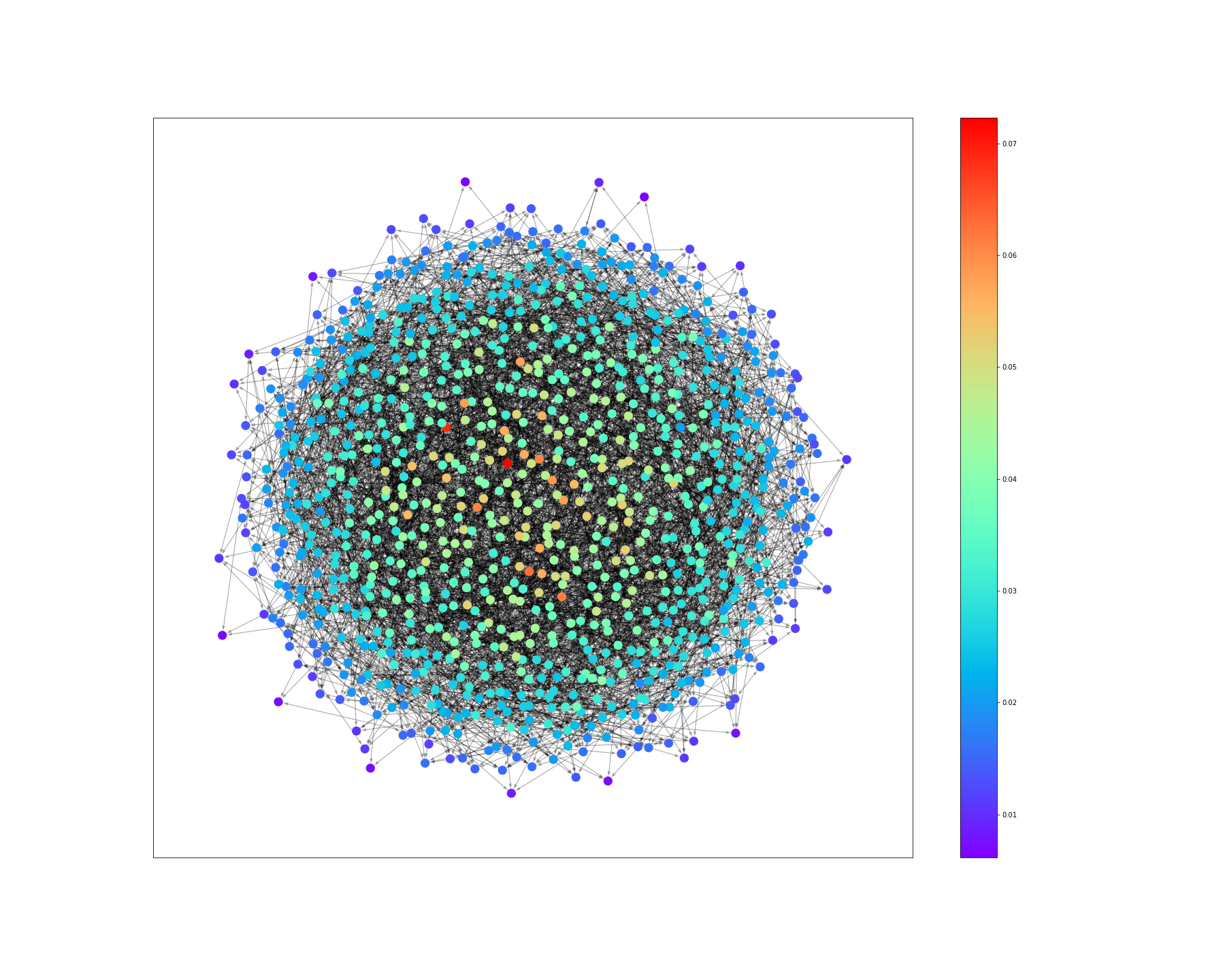
# 

# **4. Eigenvector centrality**

Eigenvector centrality is an influence of a node in the network. So, the relative scores are set to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. An eigenvector with a high score means that a node is connected to many nodes who themselves have high scores.

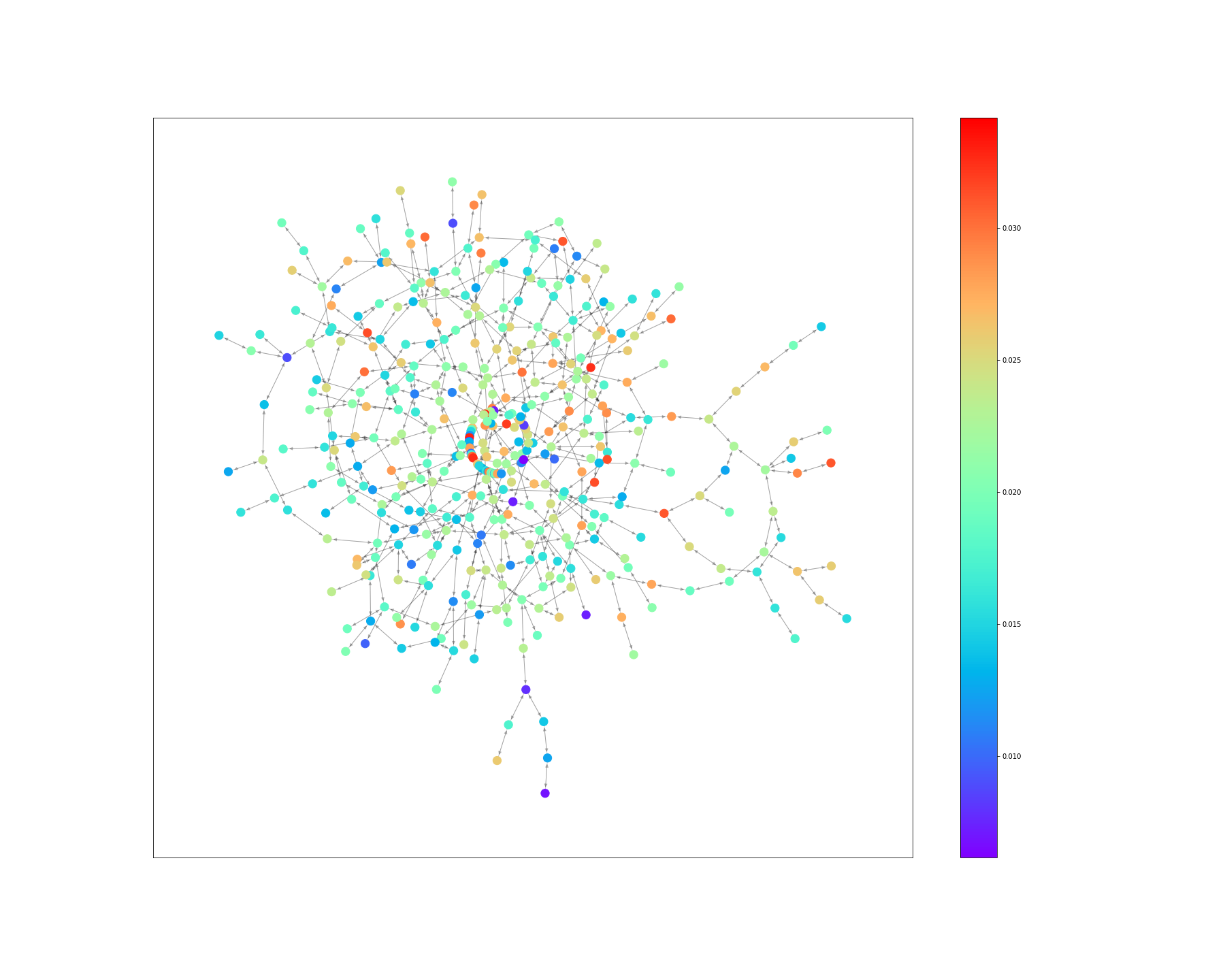


, where M(v) is a set of neighbors of v and **λ** is an eigenvector, which is constant.

For our network, the Eigenvector Centrality is 29.69391.  


## **Eigenvector centrality Analysis**

For Katz Centrality Analysis we selected nodes that contribute to 80% of the cumulative sum which resulted in the removal of 318 nodes from the graph.



|  |  |
| --- | --- |
| The graph diameter | 5 |
| Density | 0.0041267384877884256 |
| Shortest path length | 3.273137137137137 |
| Degree Centrality | 4.122122122122123 |
| Closeness centrality | 92.2082119978657 |
| Betweenness Centrality | 0.28281274021136943 |
| Katz Centrality | 5.376479430681673 |
| EigenVector centrality | 5.93140458049938 |

# 

# **5. Closeness centrality**

Closeness centrality is the measure of

centrality in the network, which is calculated by summing the length of the shortest paths between the node and all other nodes in the graph.

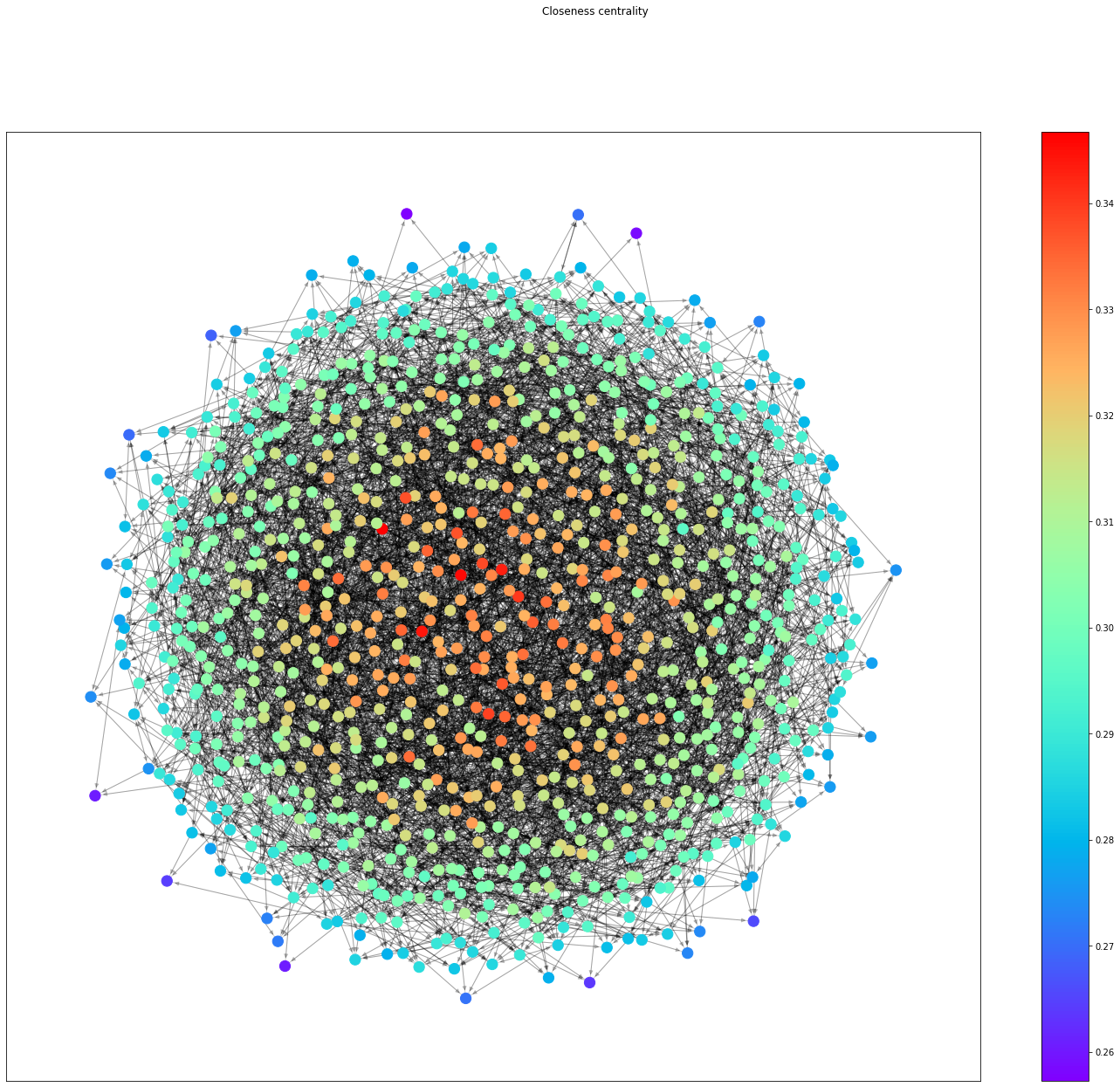
The Closeness centrality is calculated by the following formula:



, where d(y, x) is a distance between the vertices “x” and “y”.

The Closeness Centrality of the network can be defined as a sum of values of closeness centrality for all nodes.

For our network, the Closeness Centrality is 306.24685.

****

## 

## **Closeness Centrality Analysis**

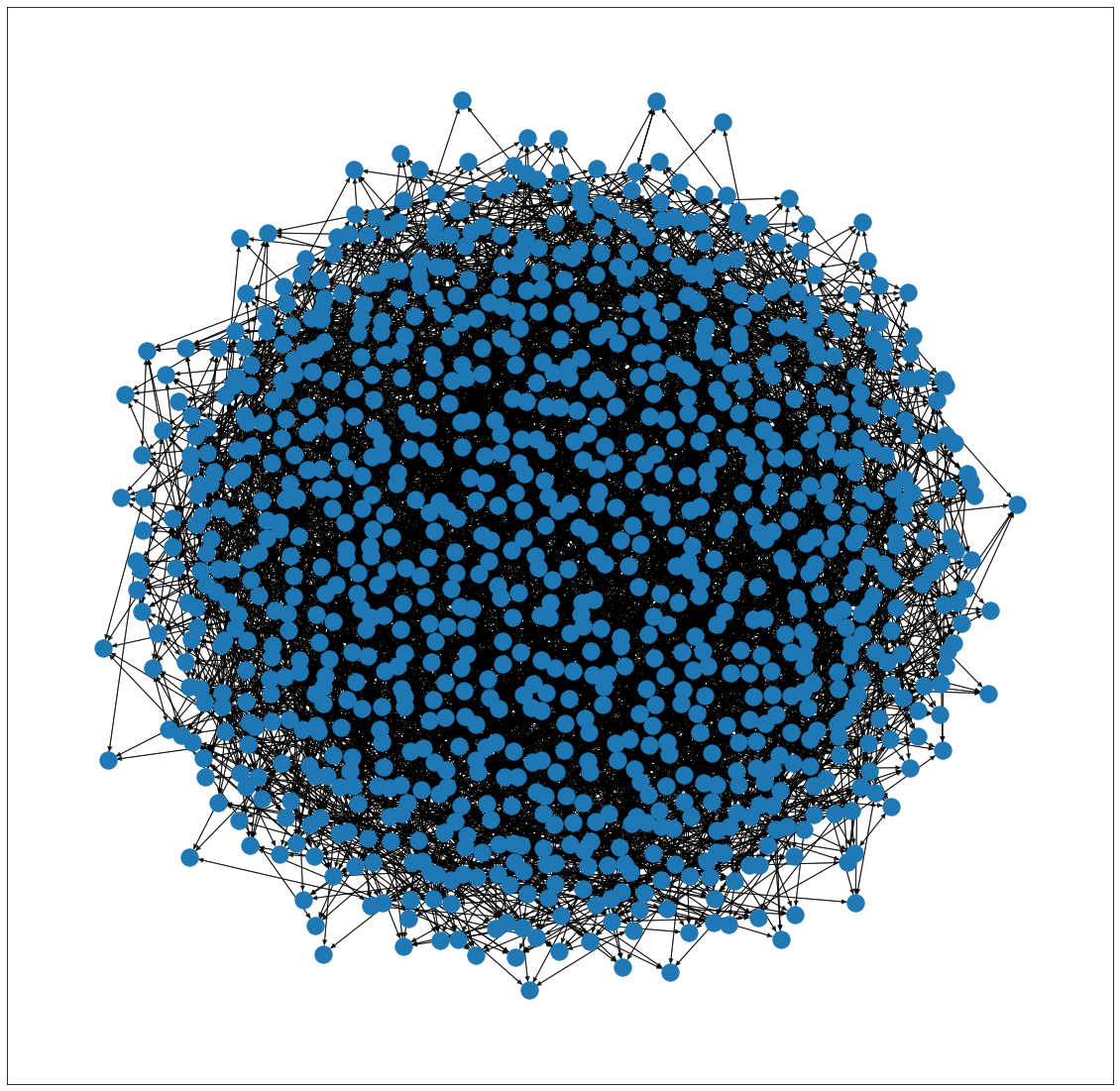
For Closeness Centrality Analysis we selected nodes that contribute to 80% of the cumulative sum which resulted in the removal of 318 nodes from the graph.

|  |  |
| --- | --- |
| The graph diameter | 5 |
| Density | 0.0034989813092390822 |
| Shortest path length | 3.273137137137137 |
| Degree Centrality | 2.502502502502499 |
| Closeness centrality | 60.96937660878829 |
| Betweenness Centrality | 0.15272709754981095 |
| Katz Centrality | 3.1381732414557066 |
| EigenVector centrality | 3.5518385691107923 |

**6. Network diameter**

The diameter of a network is the longest of all its geodesics, which means that it is the length of the longest path from all the shortest paths from one node to another one.

For our network, the Network diameter is 5.



**7. Network density**

Network density is the division of the actual number of edges into the possible number of edges in the graph. The formula for the network density can be changed depending on the type of graph.

**For undirected graph network density:**

**For directed graph network density:**

**where:**

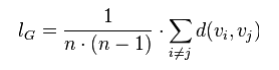
E → number of edges.

V → number of vertices.

Our network represents a directed graph, the Network density is 0.009875.

# **8. Network average shortest path length**

The average path length is the average of all geodesics, which means it is the average of all shortest paths from all nodes to other nodes. The formula for these metrics is the following:



**where:**

**n**: the number of vertices

**v**i: ith vertice in the graph

**vj**: jth vertice in the graph

For our network, the Network average shortest path length is 3.27313.

**Conclusion**

We have just analyzed the Huawei Social Network Data and displayed different metrics and their own analysis. However, this is one of the examples of analyzing a network. Further work might be analyzing the same type of data given by other sources such as Google or Instagram itself, where the data might be slightly different. For example, they could have higher density, or much more nodes and edges, and we could use that data to see and understand the difference between these metrics even deeper.