STUDY OF GRAPH MEASURES AND METRICS WITH SOCIAL NETWORKING DATASET

A PREPRINT

Kamal Aghayev
Computer Science 2
kamal.aghayev@ufaz.az

Amin Azimov
Computer Science 2
amin.azimov@ufaz.az

Hikmat Pirmammadov
Computer Science 1
hikmat.pirmammadov@ufaz.az

January 3, 2020

1 Introduction

The goal of this project is to study different graph measures and metrics (e.g. centrality measures). To do this we decided to used social networking dataset [1]. The process of study is as follows: we get the case of the graph in the original state as the baseline, then we remove some nodes from the graph and see the changes that happen in different measurements. This way, if we remove some nodes that influence some measurements the most, we can also see the relationship between different measurements and metrics.

2 Dataset

Huawei Social Network Data [1] is the dataset containing data about interaction between people by posts and comments in Facebook, Instagram and Twitter. We used the data collected from Facebook; however, the code of the project has a good modularity and may easily be used on any dataset. The dataset contains 1000 nodes and 250315 edges. Graph is directed since each edge represent a post or a comment. For the graph representation of the dataset see Fig. 1. The measurements of the dataset are described in the section 3 Measures and Metrics.

3 Measures and Metrics

The following measures and metrics were studied during the project:

- Degree Centrality [3.1]
- Betweenness Centrality [3.2]
- Closeness Centrality [3.3]
- Katz Centrality [3.4]
- Network Density [3.5]
- Network Diameter [3.6]
- Network Average Shortest Path Length [3.7]

3.1 Degree Centrality

Degree Centrality of a node is defined as the number of edges incident in the node. Since we use NetworkX [2] library for computations, we need to note that "The degree centrality values are normalized by dividing by the maximum possible degree in a simple graph $n-1$ where n is the number of nodes in G." [3]
We can also define the Degree Centrality of the network as a sum of values of degree centrality for all nodes. Degree Centrality of the baseline network is 200.8128.

On the Fig. 2 you can see the representation of the graph with colors of a node associated to the degree centrality value of the node.

3.2 Betweenness Centrality

Betweenness Centrality is the number of the shortest paths that pass through the node in a connected graph. It can be described by the following formula:

$$BC(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of the shortest paths from s to t and $\sigma(s,t|v)$ is the number of the shortest paths from s to t passing through the node v. The values for the Betweenness Centrality are normalized [4], since this is the standard for the Networkx [2] library.
We can also define the Betweenness Centrality of the network as a sum of values of betweenness centrality for all nodes. Betweenness Centrality of the baseline network is 0.9014
On the Fig. 3 you can see the representation of the graph with colors of a node associated to the betweenness centrality value of the node.

3.3 Closeness Centrality

Closeness Centrality is a reciprocal of the sum of the shortest path distances from a node to all other $n-1$ nodes of a graph. The values is usually normalized by multiplying the value by $n-1$, since Closeness Centrality depends on the number of nodes. It can be represented by the following formula:

$$CC(v) = \frac{n-1}{\sum_{u \in V} d(u,v)}$$

where V is the set of nodes, $d(u,v)$ is the distance between u and v and n is the number of nodes.
We can also define the Closeness Centrality of the network as a sum of values of closeness centrality for all nodes. Closeness Centrality of the baseline network is 526.4307
On the Fig. 4 you can see the representation of the graph with colors of a node associated to the closeness centrality value of the node.

3.4 Katz Centrality

Katz Centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors [5]. It can be represented by the following formula:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

where A is the adjacency matrix of the graph and α and β are parameters that control the initial centrality.
We can also define the Katz Centrality of the network as a sum of values of Katz centrality for all nodes. Katz Centrality of the baseline netowk is 0.3359
On the Fig. 5 you can see the representation of the graph with colors of a node associated to the Katz centrality value of the node.

3.5 Network Density

The density of a graph is a fraction of the number of connections of a node by the number of possible connections of a node. It can be represented by the following formula for an undirected graph:

$$d = \frac{2m}{n(n-1)}$$

and by the following formula for a directed graph:

$$d = \frac{m}{n(n-1)}$$

where m is the number of edges in the network and n is the number of the nodes in the network. For the baseline network, the value of the density is 0.1004.

3.6 Network Diameter

Network Diameter is defined as the length of the longest path between all the shortest paths from a node to another node. According to 6 handshakes rule [6], any 2 people are six, or fewer, social connections away from each other. It was interesting for us to check this on the real-life example. On out dataset, any 2 people are three, of fewer, social connections away from each other, which means that the diameter of the network is 3.

3.7 Network Average Shortest Path Length (NASPL)

Network Average Shortest Path Length is defined as the mean value of the shortest paths from all nodes to all other nodes. It can be represented by the following formula:

$$NASPL = \frac{\sum_{u \in V} \sum_{v \in V} d(u, v)}{n(n-1)}$$

where V is the set of nodes, $d(u, v)$ is the distance from u to v and n is the number of nodes.
For the baseline network, the value of the average shortest path length is 1.8996, which means that the length of the shortest path from a node to another node is around 1.8996, i.e. about 2 people.

4 Metrics Relationship

The purpose of the project was to study different metrics and find relationship between them. We decided to remove the nodes that are the most important for different metrics and see what happens to the network by means of the measures and metrics. For each metric, we extracted the nodes that impacted the metric for the 90%, e.g. if the sum of the values of some metric for the nodes is 1, then we extracted the nodes with the highest value for the given metric such that the sum of the metric values of the nodes is 0.9. This way we can compare the results of different metrics when nodes impacting other metrics are removed from the graph. Note that for the analysis of the metrics, after removing the nodes we did not recalculate the values for the metrics related to nodes, but used the ones computed before removing nodes to see the relationship between the metrics in a better way.

4.1 Degree Centrality

For the Degree Centrality, we removed 882 nodes to decrease its value by 90% from 200.8128 to 20.0280. On the Fig. 6 you may see the network with extracted nodes and on the Table 1

Metrics	Value
Degree Centrality	20.0280
Betweenness Centrality	0.0750
Closeness Centrality	61.6109
Katz Centrality	-0.4024
Network Density	0.0708
Network Diameter	4
NASPL	2.4650

Table 1: Metrics for the graph with high Degree Centrality nodes extracted

4.2 Betweenness Centrality

For the Betweenness Centrality, we removed 861 nodes to decrease its value by 90% from 0.9014 to 0.0899. On the Fig. 7 you may see the network with extracted nodes and on the Table 2

Metrics	Value
Degree Centrality	23.7977
Betweenness Centrality	0.0899
Closeness Centrality	72.6040
Katz Centrality	-0.2534
Network Density	0.0721
Network Diameter	4
NASPL	2.3854

Table 2: Metrics for the graph with high Betweenness Centrality nodes extracted

4.3 Closeness Centrality

For the Closeness Centrality, we removed 882 nodes to decrease its value by 90% from 526.4307 to 52.1937. On the Fig. 8 you may see the network with extracted nodes and on the Table 3

Metrics	Value
Degree Centrality	16.8368
Betweenness Centrality	0.0624
Closeness Centrality	52.1937
Katz Centrality	-0.3831
Network Density	0.0680
Network Diameter	4
NASPL	2.5826

Table 3: Metrics for the graph with high Closeness Centrality nodes extracted

4.4 Katz Centrality

For the Katz Centrality, we removed 80 nodes to decrease its value by 90% from 0.3359 to 0.0198. On the Fig. 9 you may see the network with extracted nodes and on the Table 4

Metrics	Value
Degree Centrality	181.9459
Betweenness Centrality	0.8016
Closeness Centrality	483.9244
Katz Centrality	0.0198
Network Density	0.0972
Network Diameter	3
NASPL	1.9028

Table 4: Metrics for the graph with high Katz Centrality nodes extracted

4.5 Relationship

Below you may find the Table 5 that combines all 4 table above for the better view of the situation. In the left most row you may see the name of the metrics measured and on the top most column you may see the name of the metrics nodes of which were removed from the graph.

Metrics / Centrality	Baseline	Degree	Betweenness	Closeness	Katz
Degree Centrality	200.8128	20.0280	23.7977	16.8368	181.9459
Betweenness Centrality	0.9014	0.0750	0.0899	0.0624	0.8016
Closeness Centrality	526.4307	61.6109	72.6040	52.1937	483.9244
Katz Centrality	0.3359	-0.4024	-0.2534	-0.3831	0.0198
Network Density	0.1004	0.0708	0.0721	0.0680	0.0972
Network Diameter	3	4	4	4	3
NASPL	1.8996	2.4650	2.3854	2.5826	1.9028

Table 5: Comparison matrix

As it easily may be seen from the table, Degree Centrality is highly related to Closeness Reality. When we decrease the value for the Closeness Centrality by 90% we strongly decrease the value of the Degree Centrality. This is strongly logical, since the nodes with higher Closeness Reality are closer to the center of the graph, such that they may reach other nodes in a smaller, and therefore, they usually have a big number of connections.

Similarly, Betweenness Centrality is closely related to Closeness Centrality, since higher Betweenness Centrality means that big number of shortest paths lay from the given node. Therefore, if the node has a big value for Closeness Centrality, it usually lays on the shortest path between other nodes.

As we may see from the table, decreasing Degree Centrality influence the value of the Katz Centrality a lot. This happens because Katz Centrality shows "importance" of the node and if the node has a lot of connections it usually means that it is important.

As well as Katz Centrality, Closeness Centrality as well relates to Degree Centrality, since number of connections in social networks mean that a person is close to the society.

Network diameter value for the Baseline network was 3 and did not change when we removed the nodes with the highest Katz values, but the value has increased to 4 when we removed the nodes with the highest values for the other metrics. This happens because the Katz value may be negative and 90% of the Katz Centrality is collected in only 80 nodes from 1000 nodes. This mean, that decreasing 90% of the nodes responsible for Katz Centrality we are left with the network with 920 nodes. Thus, although we removed the most important nodes, the metrics of the graph do not change a lot. Compared to other metrics, other metrics require about 800 nodes to be removed, which means that we are left with about 200 nodes, thus we see huge difference in the values of the metrics after our removing process.

Network Average Shortest Path Length is highly related to Closeness Centrality value, since the latter is the reciprocal of the sum of the shortest path distances and the former is the mean value of the shortest path distances. Thus, removing the nodes with the high value of Closeeness Centrality measure, we highly increase the values of the Network Average Shortest Path Length.

5 Further Work

For the further work, we may study more metrics as Eigenvector Centrality, Google's PageRank Centrality, etc. and compare them to the metrics described in the current project. The other interesting experiment would be checking other datasets with less and higher network density values to see the difference between the behaviour of the metrics on different types of datasets.

References

[1] Huawei Social Network Data.

[2] NetworkX library.

[3] Definition of `degree_centrality` function from NetworkX library.

[4] Definition of `degree_centrality` function from NetworkX library.

[5] Definition of `katz_centrality_numpy` function from NetworkX library.

[6] Six degrees of separation. Six Handshakes rule.

[7] graph.png.

[8] Images for the Baseline network

[9] Images for the Degree Centrality network

[10] Images for the Betweenness Centrality network

[11] Images for the Closeness Centrality network

[12] Images for the Katz Centrality network

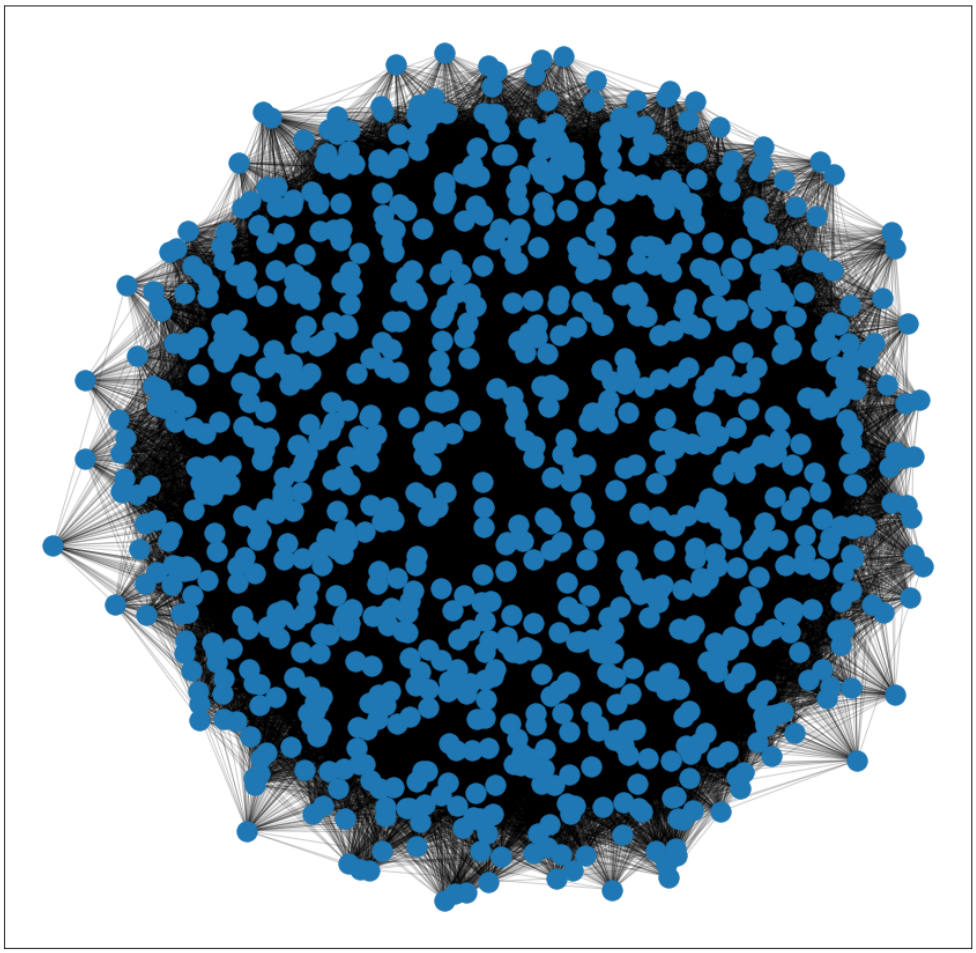


Figure 1: Graph representation of the dataset [7]

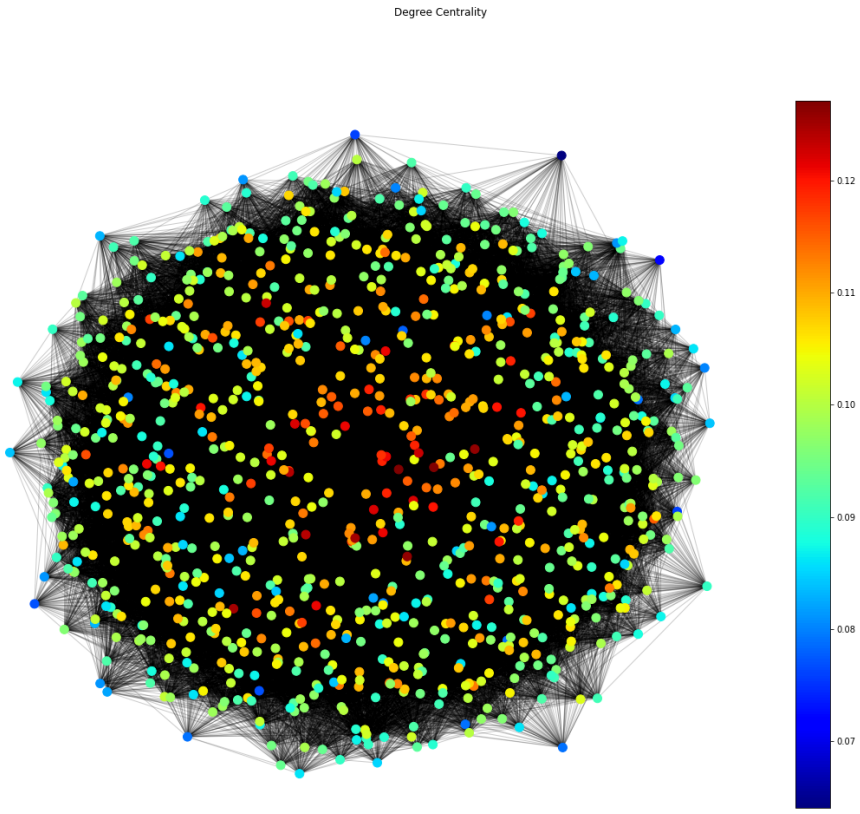


Figure 2: Degree Centrality of the baseline graph [8]

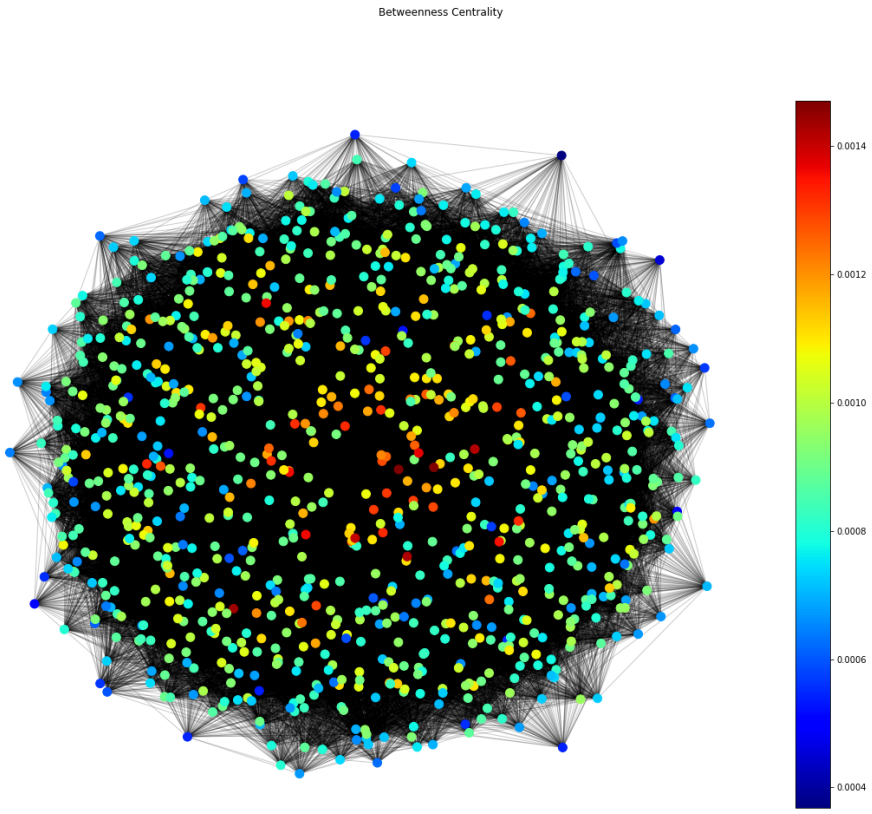


Figure 3: Betweenness Centrality of the baseline graph [8]

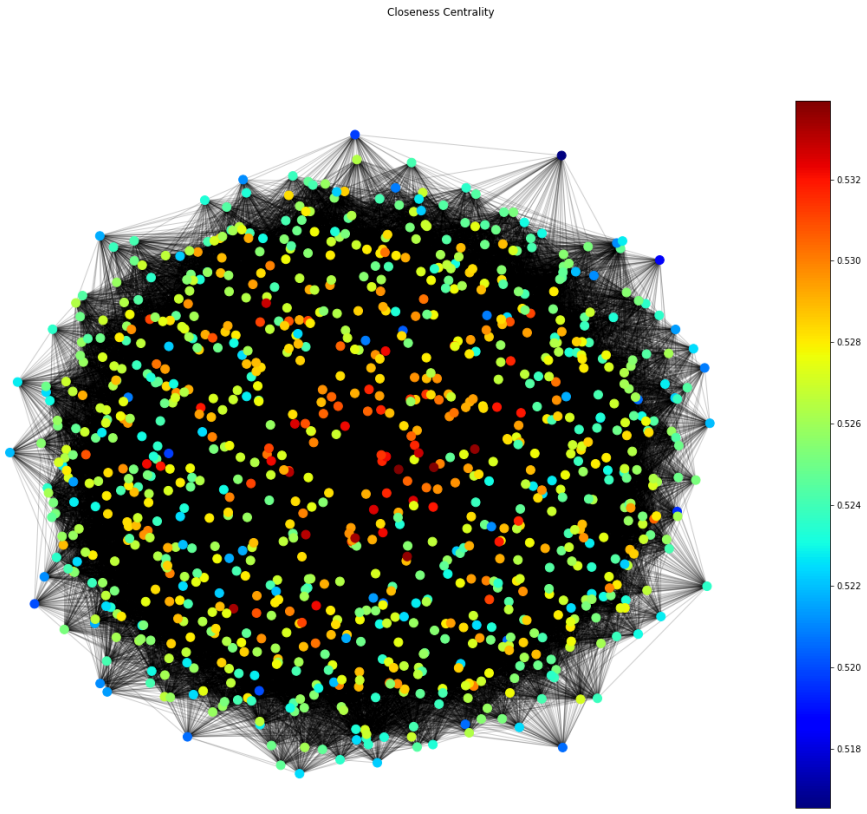


Figure 4: Closeness Centrality of the baseline graph [8]

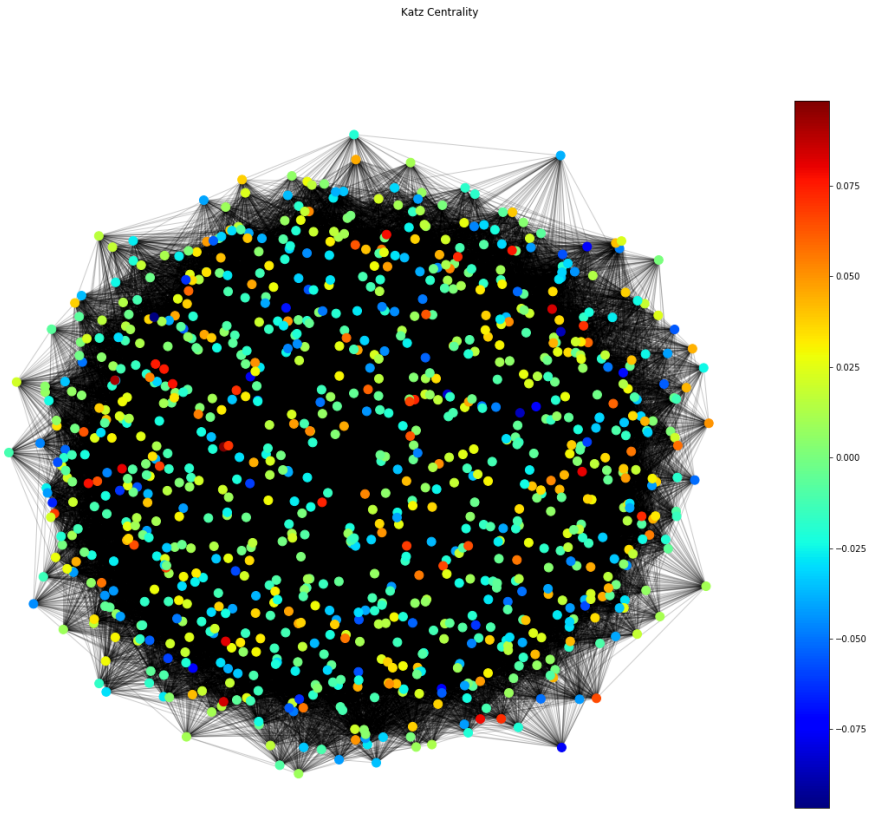
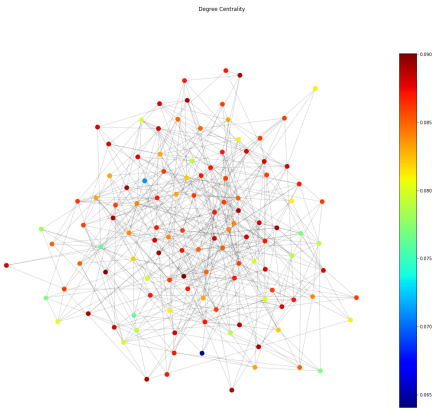
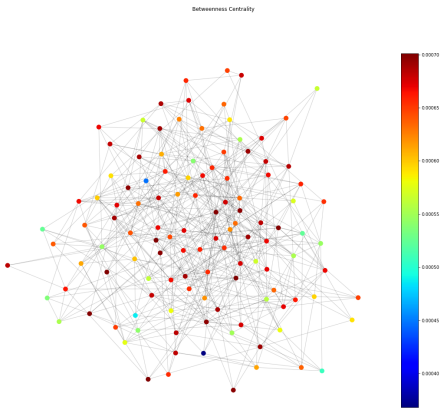


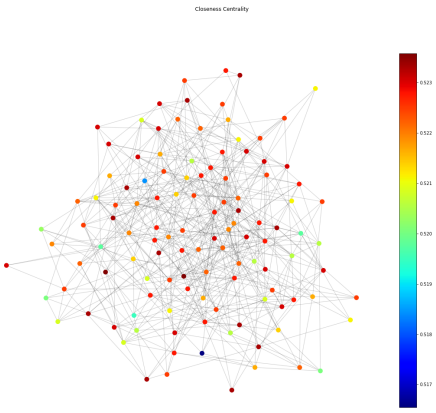
Figure 5: Katz Centrality of the baseline graph [8]



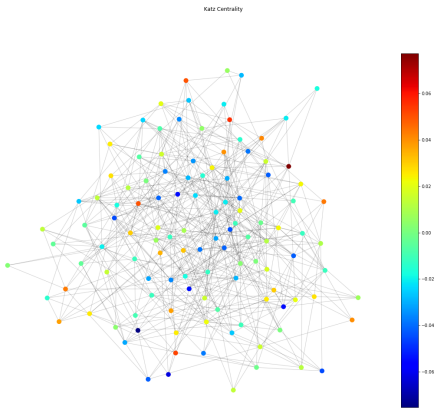
(a) Degree Centrality



(b) Betweenness Centrality

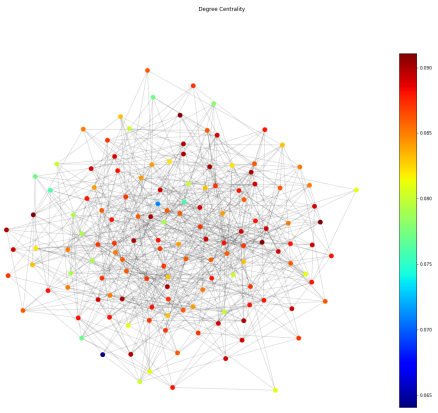


(c) Closeness Centrality

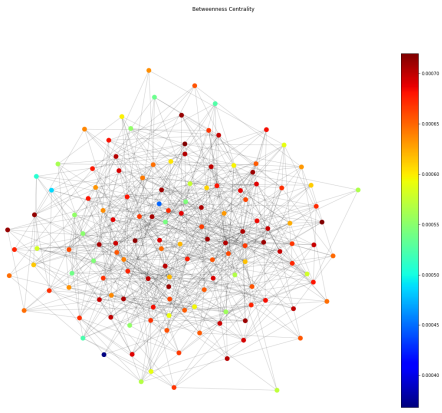


(d) Katz Centrality

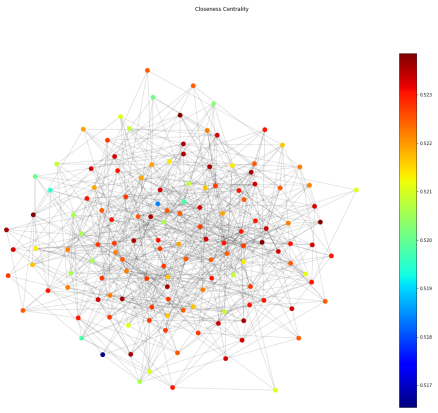
Figure 6: Graph with high nodes Degree Centrality nodes extracted [9]



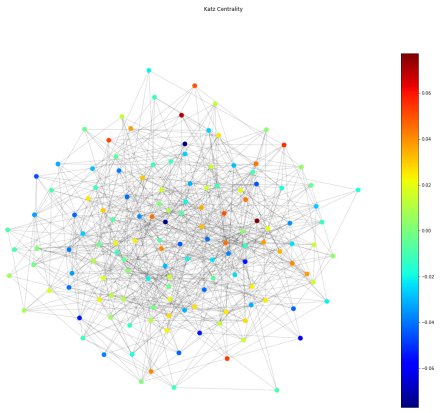
(a) Degree Centrality



(b) Betweenness Centrality

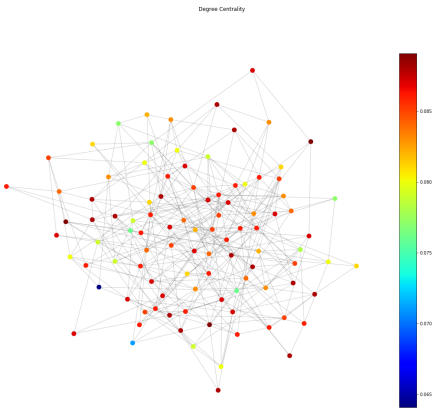


(c) Closeness Centrality

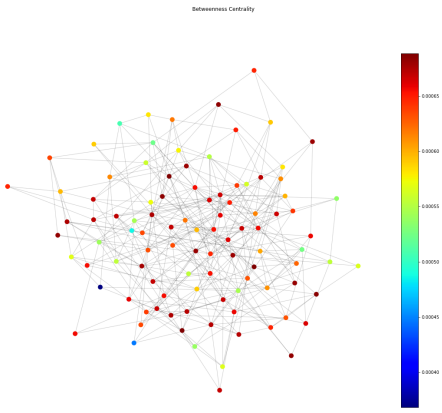


(d) Katz Centrality

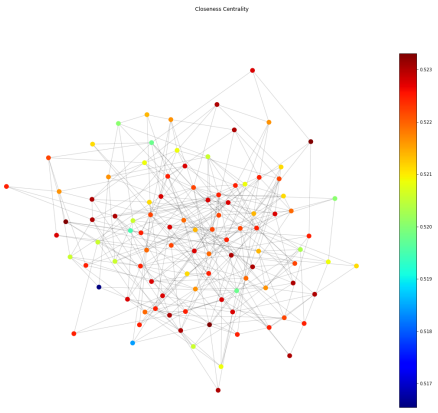
Figure 7: Graph with high nodes Betweenness Centrality nodes extracted [10]



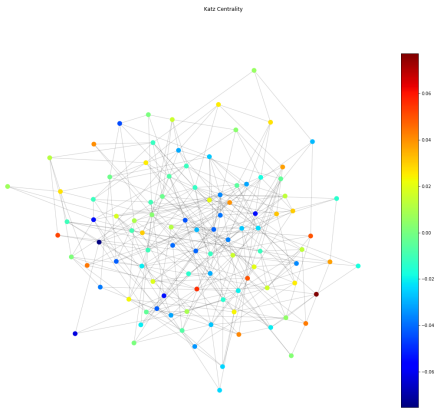
(a) Degree Centrality



(b) Betweenness Centrality

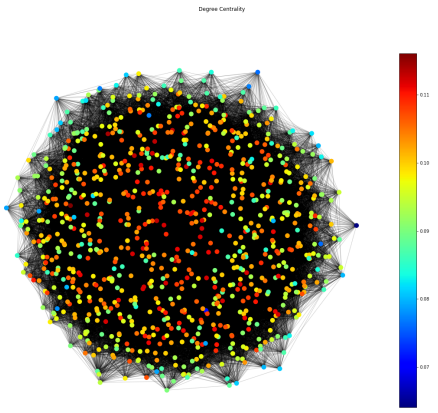


(c) Closeness Centrality

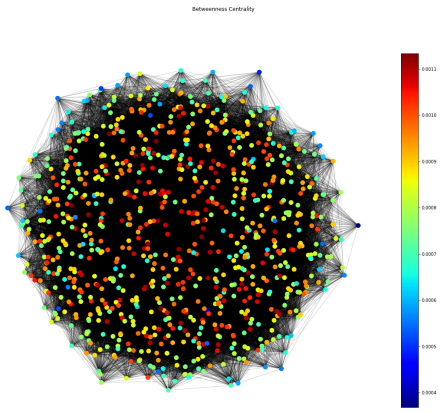


(d) Katz Centrality

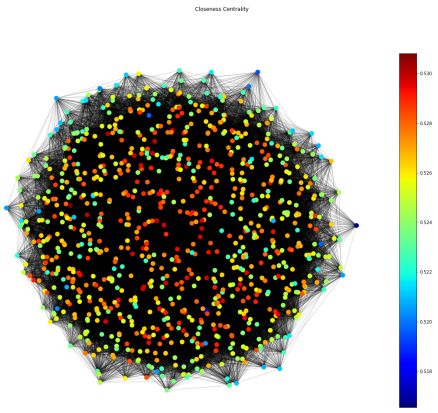
Figure 8: Graph with high nodes Closeness Centrality nodes extracted [11]



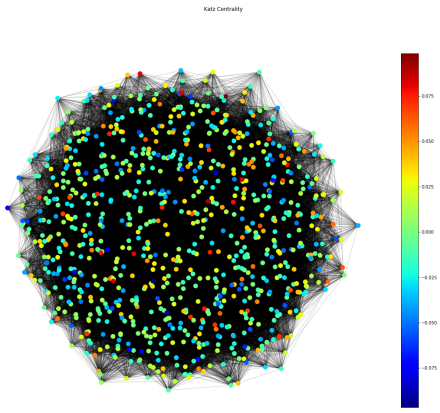
(a) Degree Centrality



(b) Betweenness Centrality



(c) Closeness Centrality



(d) Katz Centrality

Figure 9: Graph with high nodes Katz Centrality nodes extracted [12]