# **Community Detection - Empirical Study**

**Assignment 1 Report – November 2021**

*Submitted in partial fulfillment of the requirements of*

*CS G519 Social Media Analytics*

*By*

*Shahid Nazir*

*BITS ID Number: 2020H1030120P*

*Vishal John Varghese*

*BITS ID Number: 2020H1030134P*

*Under the supervision of:*

Prof. Poonam Goyal



BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS

November 2021

**Statistics of graphs**

In this section the statistics describing the datasets have been computed and have been tabulated in table1 below. The statistics that have been calculated are number of nodes in the graph, number of edges of the graph, average path length of the graph and the average clustering coefficient of the graph.

| **Dataset** | **No. of nodes** | **No. of edges** | **Avg. path length** | **Avg. clustering coefficient** |
| --- | --- | --- | --- | --- |
| Karate Club Network | 34 | 78 | 2.4082 | 0.570638 |
| Dolphins social Network | 68 | 159 | 3.35695 | 0.258958 |
| Jazz musicians network | 198 | 2742 | 2.23504 | 0.617451 |

**Table 1 : Statistics of the three graphs**

**Comparison**

In this section, the comparison of three datasets has been shown based on the number of clusters found, modularity score and run time of the algorithm. The comparison has been done and results have been tabulated for each of the clustering algorithms.

**1. Girvan Newman Clustering**

| **Dataset** | **Number of clusters found** | **Modularity score** | **Run time of the algorithm**  **(seconds)** |
| --- | --- | --- | --- |
| **Karate Club Network** | 4 | 0.3632478632478 | 0.070544481277 |
| **Dolphins social Network** | 4 | 0.4580712788259 | 0.216091156005 |
| **Jazz musicians network** | 5 | 0.0036263573735 | 5.782495975494 |

**Table 2 : No of clusters, modularity score and run time of three datasets using girvan-newman algorithm**

**2. Modularity based clustering**

| **Dataset** | **Number of clusters found** | **Modularity score** | **Run time of the algorithm**  **(seconds)** |
| --- | --- | --- | --- |
| **Karate Club Network** | 3 | 0.3806706114398422 | 0.015624523162841797 |
| **Dolphins social Network** | 4 | 0.4954906847039 | 0.015621662139 |
| **Jazz musicians network** | 4 | 0.4389078153753 | 0.200640678405 |

**Table 3 : No of clusters, modularity score and run time of three datasets using modularity based clustering**

**3. Spectral Clustering**

| **Dataset** | **Number of clusters found** | **Modularity score** | **Run time of the algorithm (seconds)** |
| --- | --- | --- | --- |
| **Karate Club Network** | 4 | 0.4101742274819 | 0.23172116279 |
| **Dolphins social Network** | 4 | 0.4833076223250 | 0.25385951995 |
| **Jazz musicians network** | 3 | 0.0143895164023 | 0.21608448028 |

**Table 4 : No of clusters, modularity score and run time of three datasets using spectral clustering**

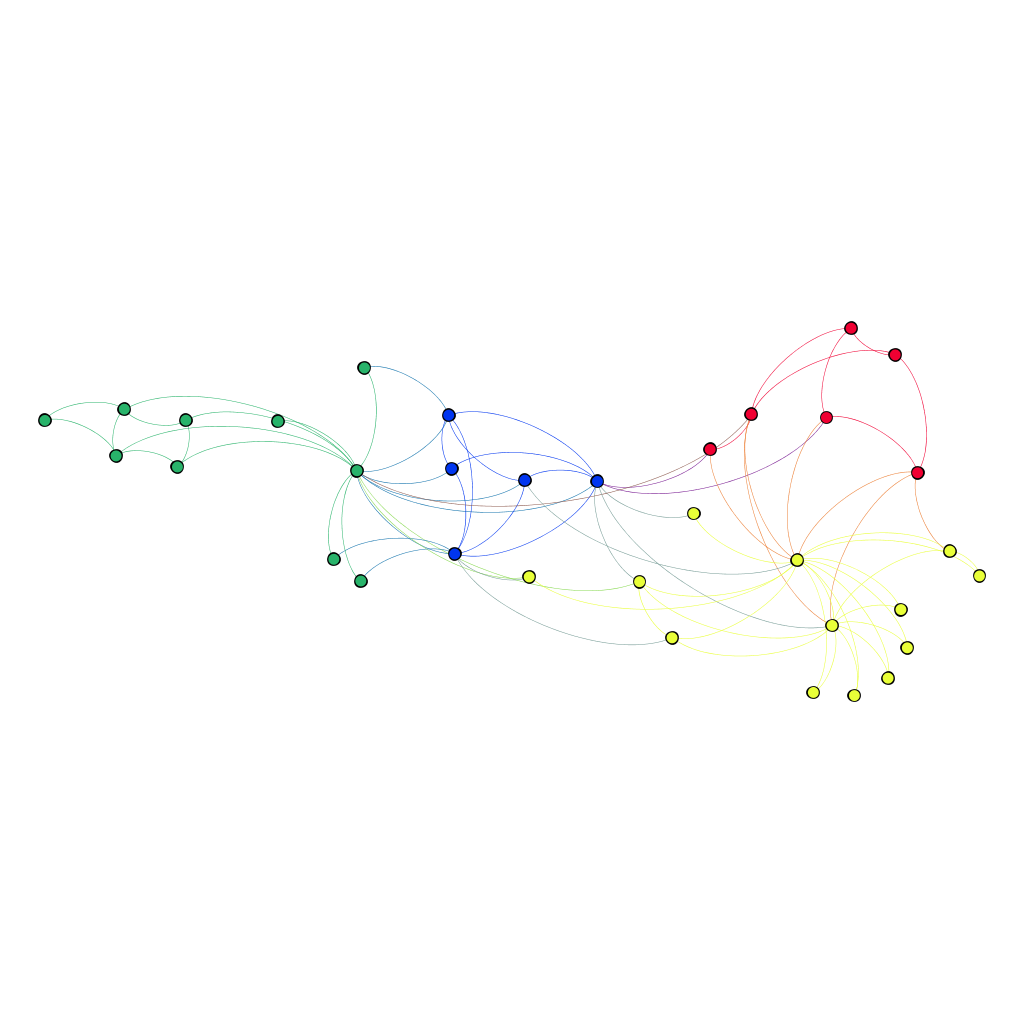
**Observations**

After running all the algorithms on all the datasets we found that the run time is least in modularity-based clustering which uses modularity maximization.

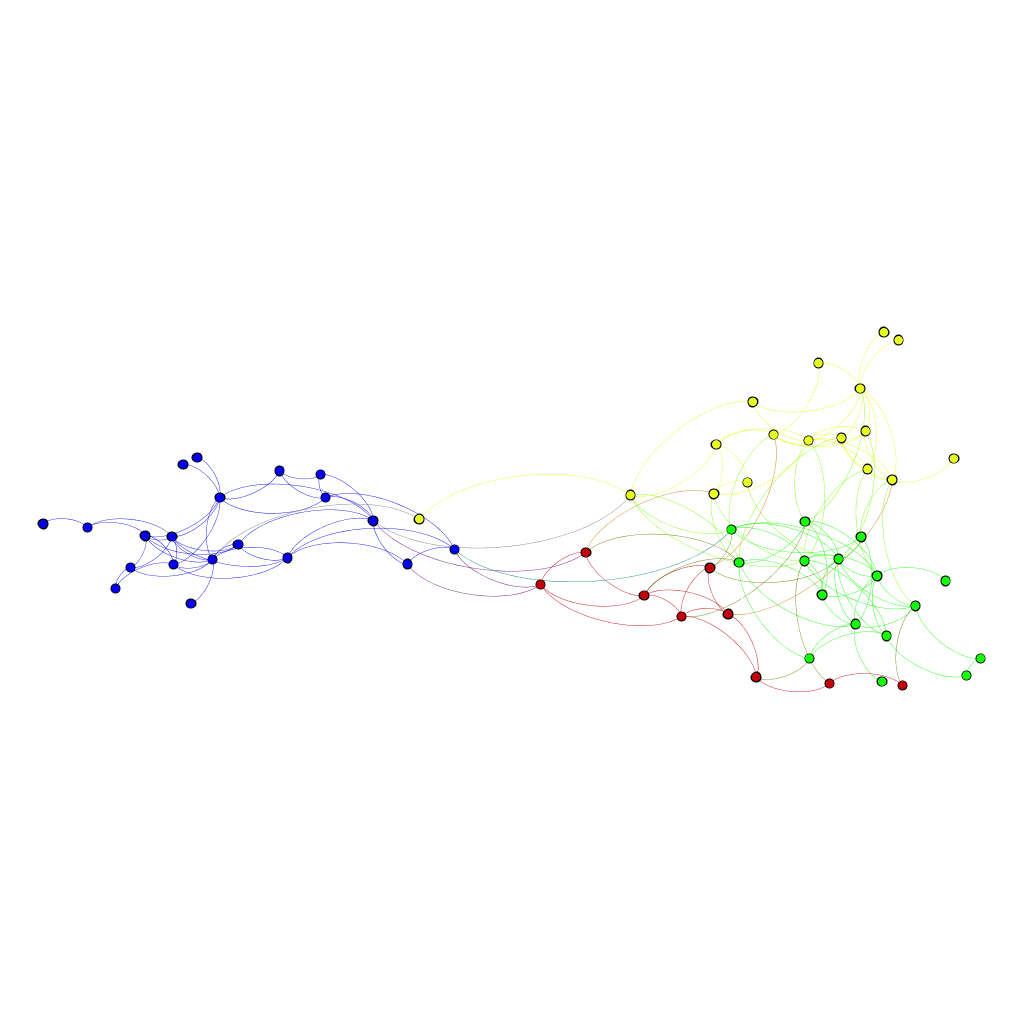
Also the modularity score for all the datasets is approximately the same for all the algorithms except for the third dataset whose modularity score is way higher in modularity-based clustering than girvan newman and spectral clustering. Therefore, the modularity-based algorithm is performing faster and better than the other two algorithms for the given datasets.

**Visualization of the clusters found**

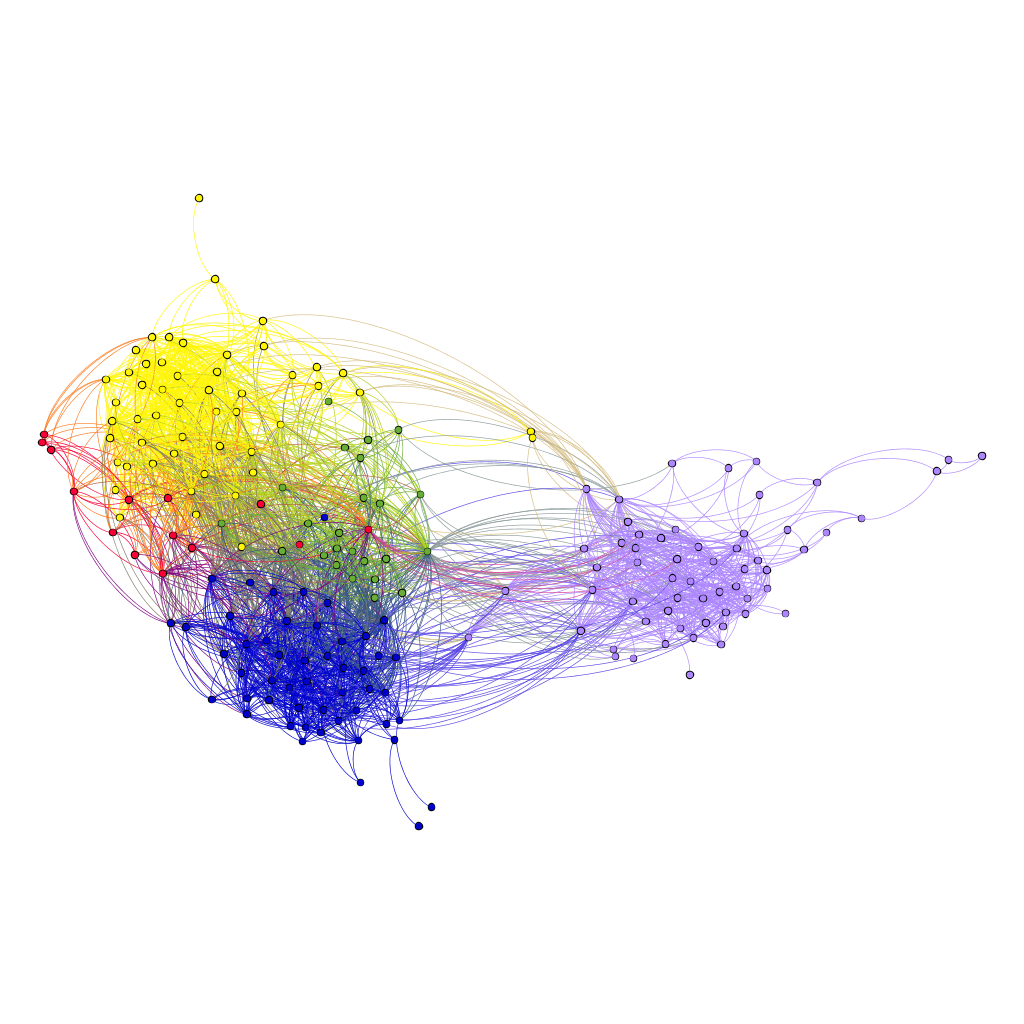
The visualization of clusters has been done using Gephi which is an open source graph visualization and analysis tool. The gephi finds clusters based on the modularity maximization. The graphs for three datasets consisting of communities formed during clustering have been shown below.

****

**Figure 1 : Visualization of the clusters of karate dataset found by modularity based clustering in gephi**

****

**Figure 2 : Visualization of the clusters of dolphins dataset found by modularity based clustering in gephi**

****

**Figure 3 : Visualization of the clusters of jazz dataset found by modularity based clustering in gephi**

**Observations**

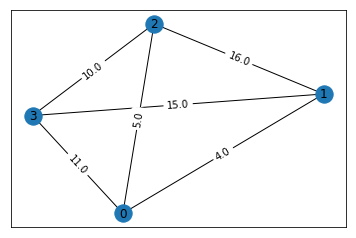
The graphs shown above have been taken from the gephi software for the three given datasets. The clusters have been shown in different colors. The clusters have been formed using modularity based clustering which uses modularity maximization method. For the first dataset, four clusters were formed. Also for the second dataset, four clusters were formed. And for the third dataset, five clusters were formed. From the above we can observe that our results are similar to girvan newman algorithm based clustering and the number of clusters which we have got are same in gephi and girvan newman algorithm.

**Representative Networks**

In this section, the representative networks of the given three datasets are shown.

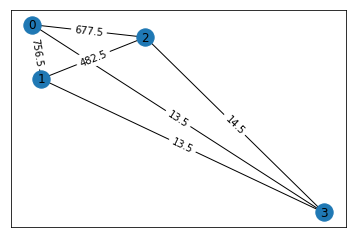
For the first graph, the betweenness-based clustering using the Girvan-Newman

algorithm has been used to find clusters. Then its representative network has been created which is shown below in figure 4.

****

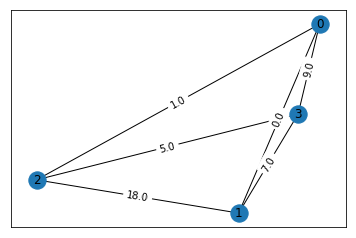
**Figure 4 : Representative network of the karate dataset**

To create the representative network of the second graph, the clusters have been found using modularity-based clustering. Then its representative network has been created which is shown below in figure 5.

****

**Figure 5 : Representative network of the dolphins dataset**

To create the representative network of the third graph, the clusters have been found using spectral clustering. Then its representative network has been created which is shown below in figure 6.

****

**Figure 6 : Representative network of the jazz dataset**

**Github Link :**

<https://github.com/Shahid-Nazir-Shah/Community-Detection---Empirical-Study>