

ZACHARY'S KARATE CLUB

- Vishnu Prasanth Reddy Patur

NID: vi091909

1. Summary and Purpose:

Zachary's karate club is a social network dataset of a university-based karate club that has been split into two groups after a conflict between the instructor and the president of the club. This data set has been prepared by Wayne Zachary after studying for a period of three years from 1970 to 1972 and published an article 1977. In his article, he analyzed how and why the conflict raised, and the process which led to the fission of a new club.

Zachary's karate club has a great history during his period of observation, the members ranged from 50 to 100, and the social activities including friends circle have been taken into consideration while analyzing. It all started with Mr. Hi a part-time instructor hired for karate classes, at the beginning of his study there was a petty inflammation between Mr. Hi and Mr. John, the president of the club over the price of karate classes. Mr. Hi wanted to raise the prices since he was the Instructor and Mr. John wanted to normalize the prices. Mr. Hi was fired by the officials led by Mr. John. The supporters of Mr. Hi retaliated by resigning and joining hands with their fatherly figure to form a new club.

I wanted to analyze which member of the club played a crucial role in the fission of the club by finding the member with most ties with other members of the club. While preparing the network of friendship, Zachary considered only effective relations, not like dynamic relations such as patron-client relations which gave more flexibility to the model to infer strong insights. In addition, I also wanted to find if there are any communities in the given network data.

2. Initial Graph:

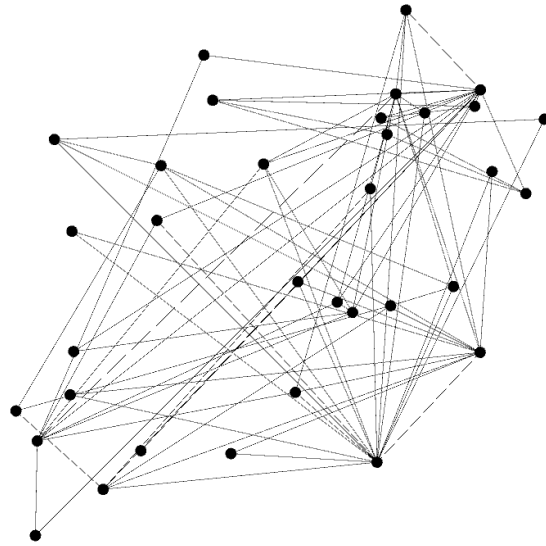


Figure 2.1 Initial Graph

The displayed network is an undirected graph and still, some basic insights from the initial graph were inferred, the nodes represent the members of the club and the links refer to the ties between 2 members. It's obvious that there are no overlapping nodes in the initial graph. Here are some statistics from the initial graph.

- Nodes: 34
- Links: 78
- Average Degree: 4.588
- Maximum Degree: 17

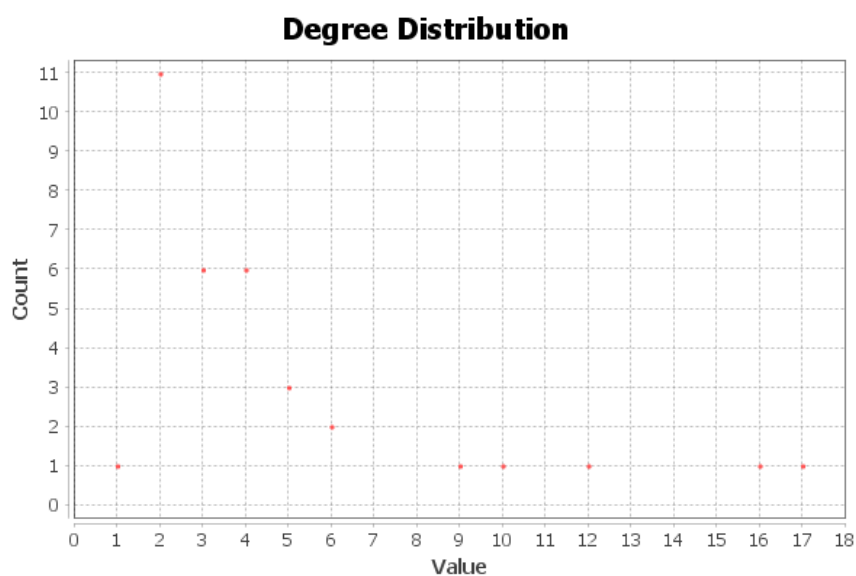
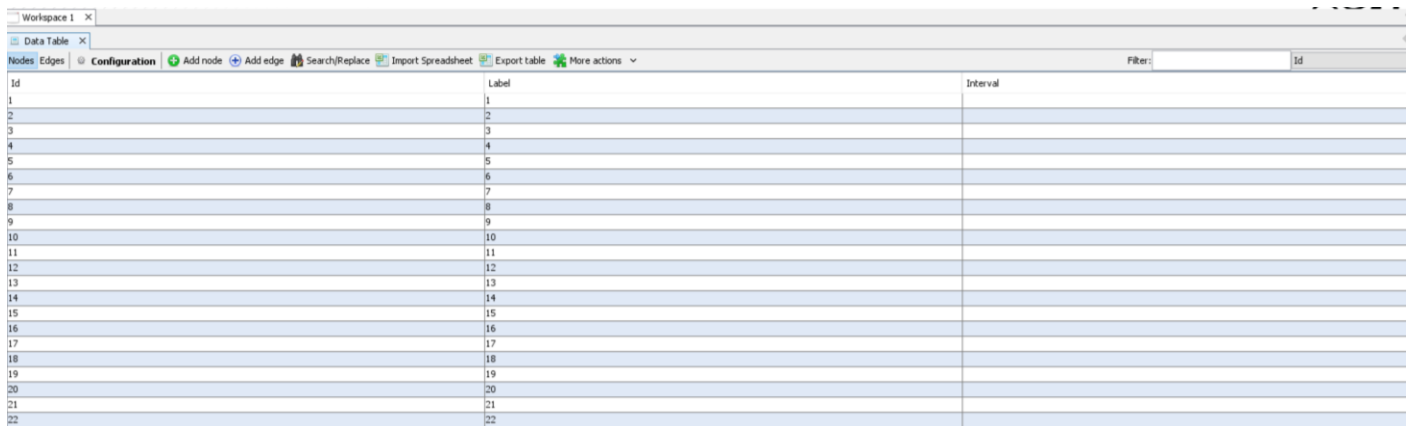


Figure 2.2 Degree Distribution

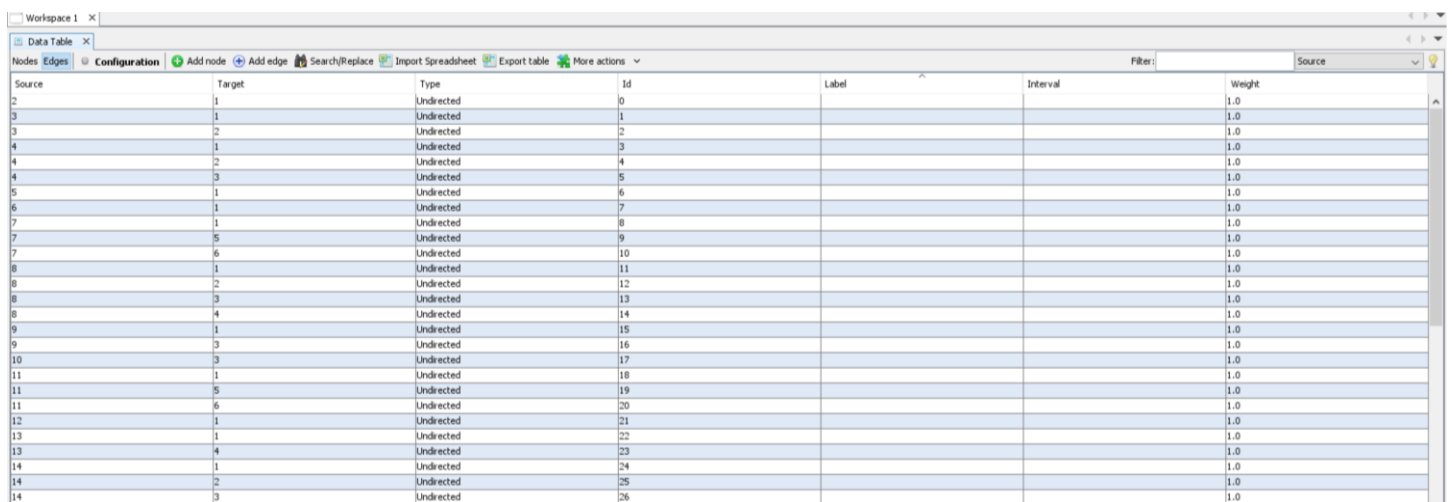
3. Data Laboratory View:



Id	Label	Interval
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	
11	11	
12	12	
13	13	
14	14	
15	15	
16	16	
17	17	
18	18	
19	19	
20	20	
21	21	
22	22	

Figure 3.1 Nodes view

The data laboratory view has two tabs on the left top corner to view some observations of nodes and edges. Initially, the node view has only 3 columns ID, Label and Interval. The ID is a unique attribute assigned to each node and the label is the name of the node, here the ID and the Label was happened to be the same for this dataset, and the Interval column seemed to have no observations.



Source	Target	Type	Id	Label	Interval	Weight
1	1	Undirected	0			1.0
1	2	Undirected	1			1.0
2	2	Undirected	2			1.0
1	3	Undirected	3			1.0
2	4	Undirected	4			1.0
3	5	Undirected	5			1.0
1	6	Undirected	6			1.0
1	7	Undirected	7			1.0
1	8	Undirected	8			1.0
5	9	Undirected	9			1.0
6	10	Undirected	10			1.0
1	11	Undirected	11			1.0
2	12	Undirected	12			1.0
3	13	Undirected	13			1.0
4	14	Undirected	14			1.0
1	15	Undirected	15			1.0
3	16	Undirected	16			1.0
3	17	Undirected	17			1.0
1	18	Undirected	18			1.0
5	19	Undirected	19			1.0
6	20	Undirected	20			1.0
1	21	Undirected	21			1.0
1	22	Undirected	22			1.0
4	23	Undirected	23			1.0
1	24	Undirected	24			1.0
2	25	Undirected	25			1.0
3	26	Undirected	26			1.0

Figure 3.2 Edges View

The Edge tab has 7 columns namely Source, Target, Type, Id, Label, Interval, and weight. The Label and Interval have no observations, Type is Undirected for all the links and the weight attribute has value 1.0 for every entry. Here the Source and Target specify the link between two nodes, we can observe that the source is ordered ascendingly, and few source nodes were repeated. This repetition of sources concludes that few nodes have ties with more than one node. It is obvious that there are no repeating ties between source and target nodes. One interesting insight from the edges tab is the edges Id attribute ranges from 0 till 77, unlike nodes ID attribute.

4. Layout Algorithms:

The algorithms used here are forced-based algorithms that give shape to the graph, each algorithm has certain parameters set to default. The layouts can be optimized by tuning the parameters to produce a precise graph that helps in analyzing the network from the obtained layout.

4.1

FORCE ATLAS

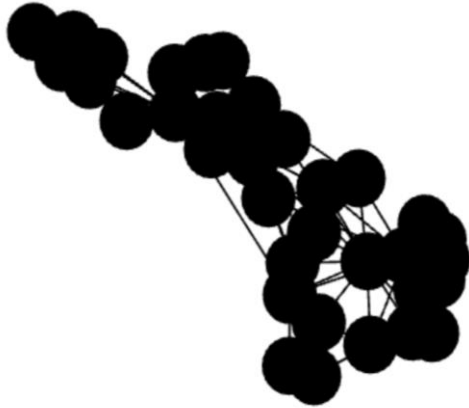


Figure 4.1 initial Graph

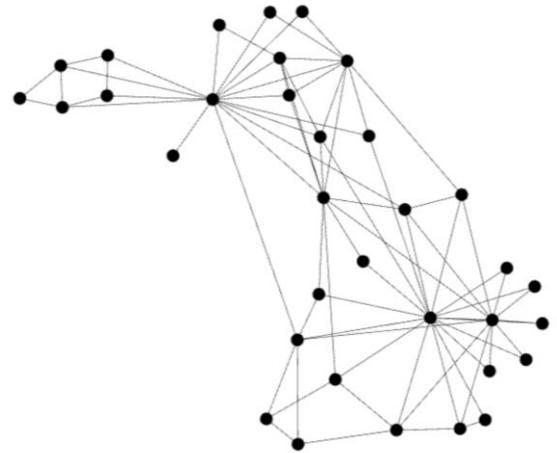


Figure 4.2 After Tuning

Force atlas allows a rigorous interpretation of the graph with the fewest biases possible, and good reliability even if it is slow. From figure 4.1, it is quite impossible to infer any insights from the initial graph with default parameters. We can observe that the nodes are overlapping, and the entire network resembles more like a cluster than a network graph. In addition, the attraction between the nodes has made it difficult to infer insights about the links since all the nodes are close to each other. After tuning a few parameters, a clear network layout is produced in figure 4.2, the resulted layout has non-overlapping nodes and less attracted to other nodes since the repulsion strength is more. Since the nodes were not closely packed, we can observe the thickness of the links and it's obvious that all the links in the network were weighted the same. The parameters tuned were mentioned below.

- Repulsion Strength: 1000.0 (node rejection strength)
- Gravity: 30 (Attract all nodes to the center)

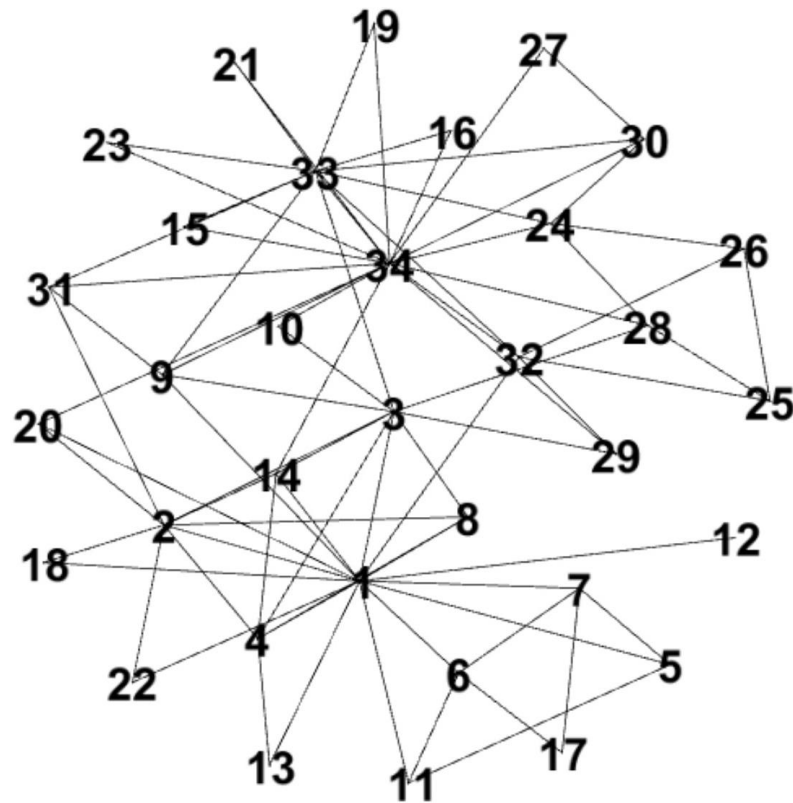


Figure 4.3

Fruchterman Reingold's layout produced a network as a system of particles in a spherical shape. The particles represent the node, here I have named the nodes with respective labels and optimized a few layout parameters to make the layout more explanatory.

- Area: 340
- Gravity: 10

The Area represents the size of the graph. In general, the size of the graph should be 10 times the number of nodes present in the network. They were 34 nodes present in the network, so the area has been set to 340 and the gravity has been set to 10 to avoid dispersion. Fruchterman Reingold tries to minimize the energy of its physical system but remains slow.

Dual Circle Layout

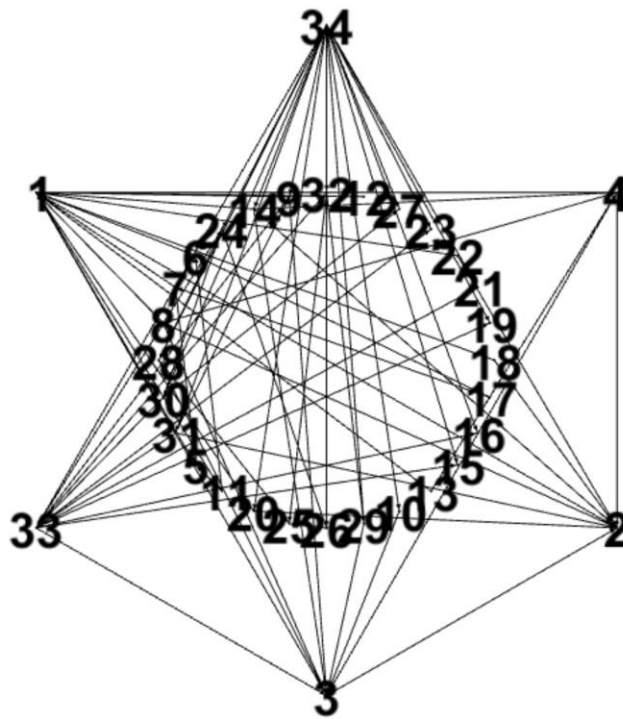


Figure 4.4

Dual circle layout assigns nodes on two concentric circles with higher degree nodes inside by default. Here I have optimized the layout to produce the higher degree nodes circle outside.

- Upper Order count: 6
- Ordered Nodes by Degree
- Layout direction: counterclockwise

Before running the layout, nodes have been ordered by degree attribute and the upper order count has been set to 6 meaning the top 6 nodes with the highest degree will be assigned on the outer circle. From figure 4.4, it can be observed that the nodes 34, 1, 33, 3, 2 and 4 are the six nodes with the highest order. This layout particularly has set a barrier in identifying the top 6 members with more ties compared to other members of the club.

Radial Axis Layout

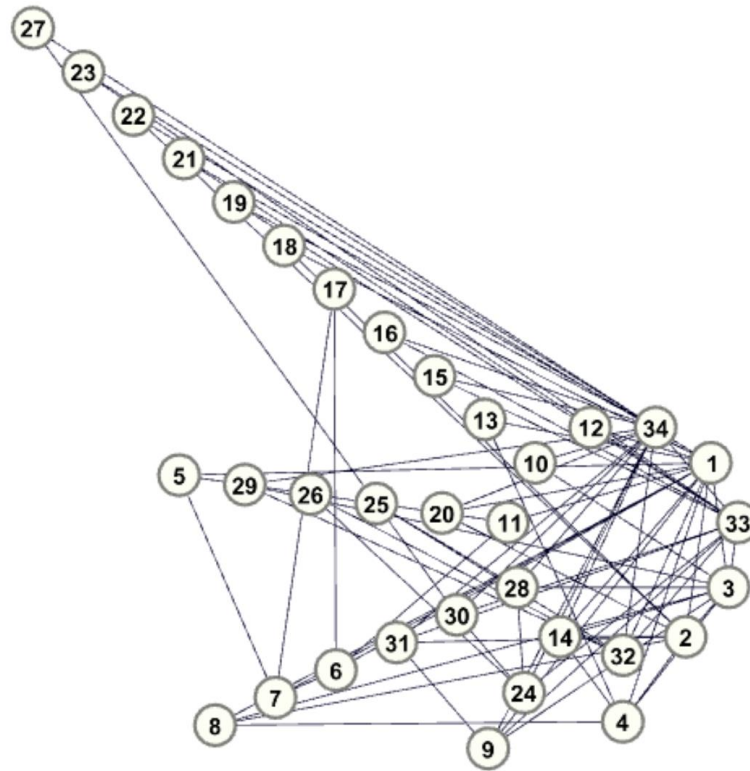


Figure 4.5

The Radial axis layout groups the nodes by degree and draws the groups in spars extending outwards from a center circle. In addition, grouping can also be generated by metrics such as directed degree and betweenness that helps in studying homophily. since the graph considered here is undirected, we are excluding grouping by directed degree.

- Grouped Nodes by Degree
- Node Layout Direction: Counterclockwise (Descending)
- Order nodes in Spar by Node ID (Ascending)

From figure 4.5, we can observe 5 spars with at least two nodes in a spar and 6 nodes with zero nodes in their Axis, the nodes in the inner circle can be referred to as base nodes for each group. The node with the highest degree is 34 and the degree of the spars decreases in a counterclockwise direction. The axis with node 10 has the highest number of nodes with the same degree compared to any other axis. Radial axis layout seems to be more useful than any other layout at this point of time followed by Dual Circle Layout.

5.

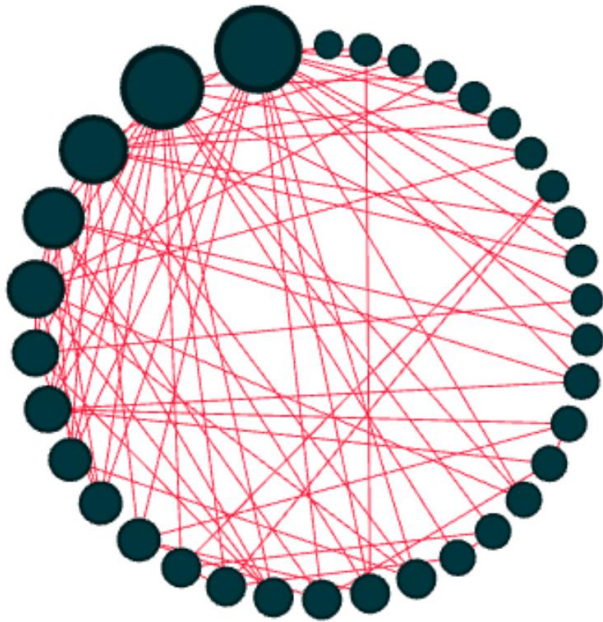


Figure 5.1

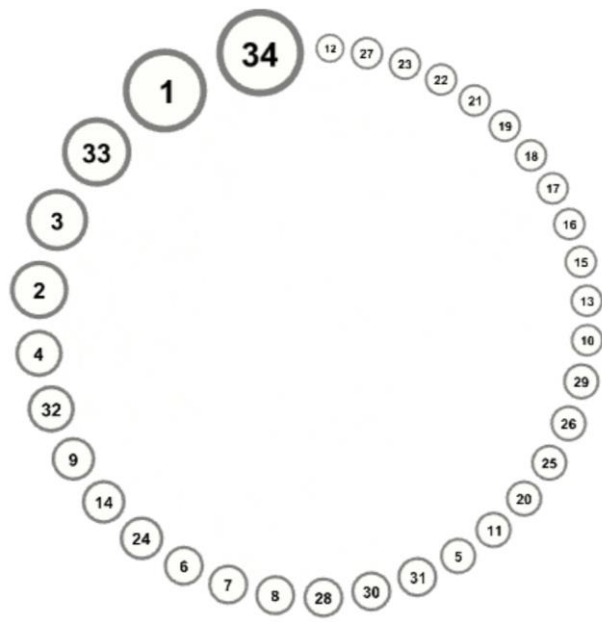


Figure 5.2

The layout used here was Circular Layout, by default the nodes will be drawn on a circle bordered by node ID in the counterclockwise direction. I have ordered the nodes by degree and ranked the size of the nodes with the degree. From figure 5.1, it is clearly noticeable that the size of the nodes increases in the clockwise direction. I have eliminated the links between the nodes and labeled the nodes to produce an explanatory graph. In addition, the label size varied with the size of the node. From figure 5.2, we can observe that the size of the node increases with degree and label size in the clockwise direction.

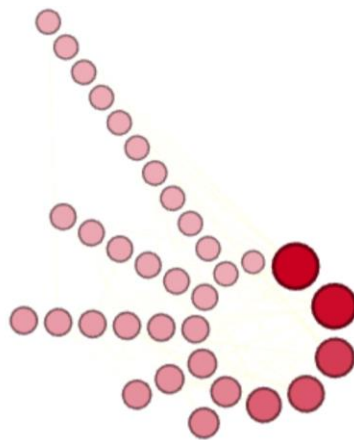


Figure 5.3

The layout used in figure 5.3 was radial axis layout, here the nodes were group by degree and I have set the partition color and size of the node to degree. We can observe that as the degree increases the size of the group with the color saturation increases in the counterclockwise direction.

6. Statistics

Here are some basic statistics of the overall network

- Average Weighted Degree: 4.588
- Diameter: 5
- Radius: 3
- Average Path length: 2.408199643493761
- Density: 0.139

Betweenness Centrality: it is a way to find the influence of a node has over other information flowing in the path. It measures the shortest weighted path between any pair of nodes.

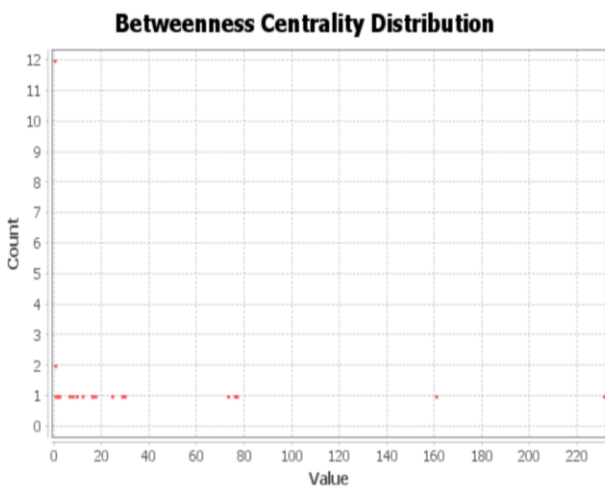


Figure 6.1

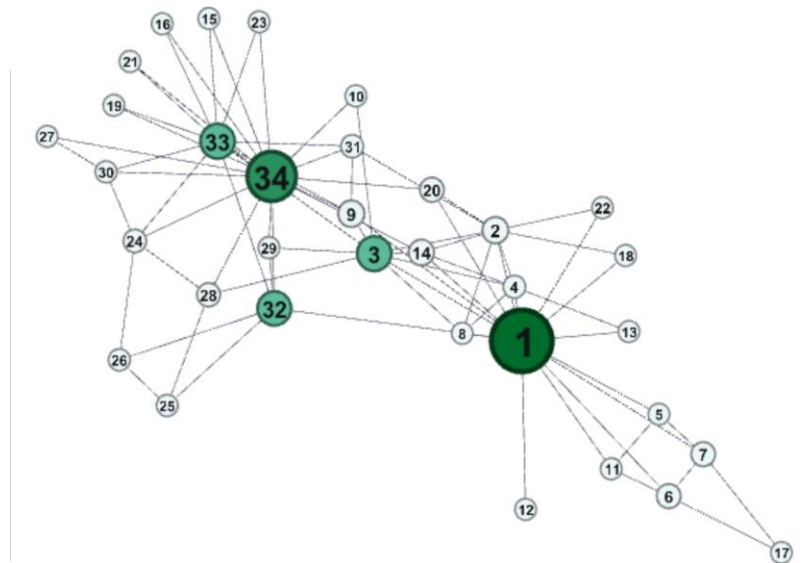


Figure6.2

From the distribution in figure 6.1, We can observe that 12 nodes have 0.0 betweenness centrality and only 2 nodes have betweenness more than 100. We can infer that these two members have more influence than any other member of the club. It is obvious that Mr. Hi and john i.e. node 1 and node 34 have more influence in the network and these two are connected through Node 32, and there are few more influencers in the network who could have been the subordinated or the officials hired by Mr. Hi.

Closeness Centrality: it is a measure that shows the closeness of a node with other nodes no matter the node lies on the shortest path between other nodes.

Eccentricity: it captures the distance between pair of nodes, where one is farthest away from the other. Low eccentricity means the node is actually close even it is farthest.

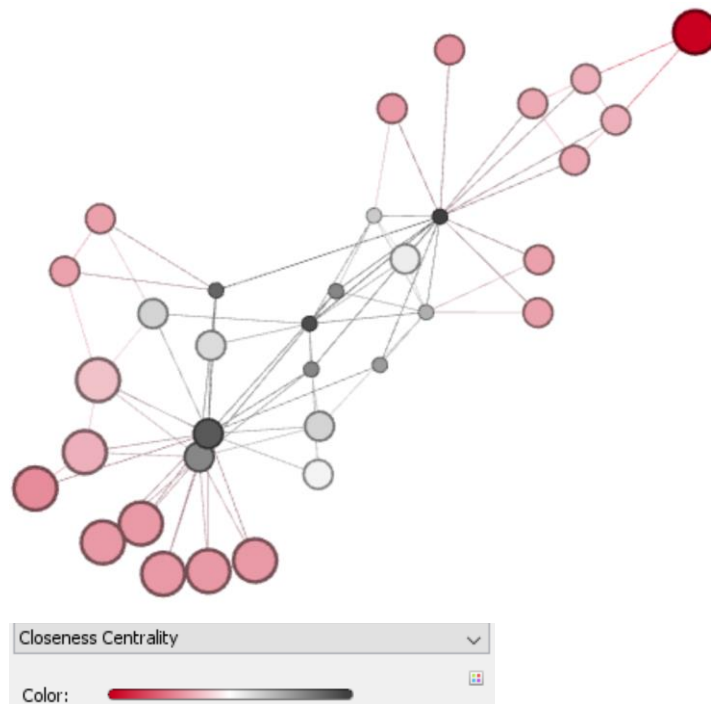
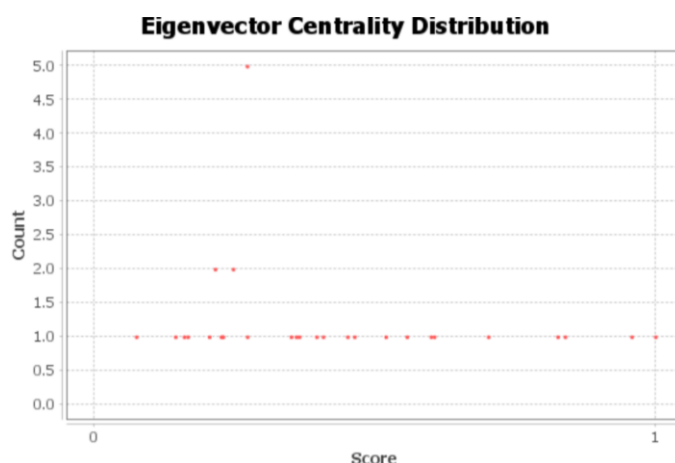


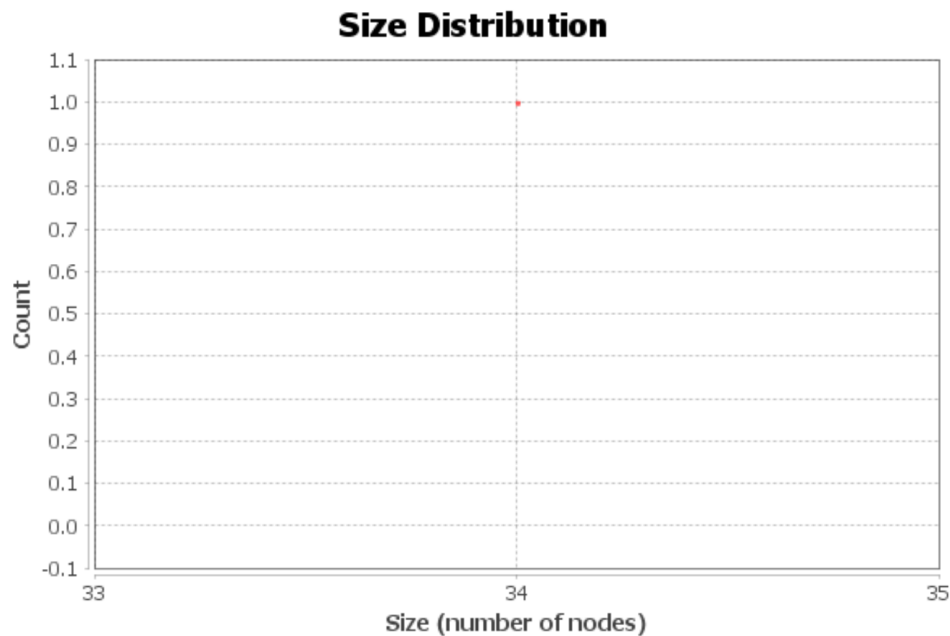
Figure 6.3

Figure 6.3 demonstrates Closeness Centrality and Eccentricity. The node color defines Closeness Centrality and the node size defines Eccentricity. It's quite easy to infer members who are well connected and who are unlikely connected in the graph.

EigenVecctor Centrality: It is also a measure of the influence of a node. Normalized relative scores are assigned to all the nodes based on connections concept. the distribution clearly shows that there is only one node that has the most relative score.



Connectivity: Determines the number of strongly or weakly connected components based on graph direction. Since our graph is undirected only weakly connected components will be generated default by gephi. Since the weight of all the links in the network is one the number of Weakly Connected Components is 34.



Clustering :

To find communities and other giant components I ran two metrics Clustering coefficient and Girvan-Newman Clustering.

Clustering Coefficient:

- Number of triangles: 45
- Number of paths (Length 2): 528
- Value of Clustering Coefficient: 0.255681812763214

Girvan-Newman Clustering:

- Number of communities: 5
- Maximum found modularity: 0.4012985

The Density of the graph is 0.139 (undirected)

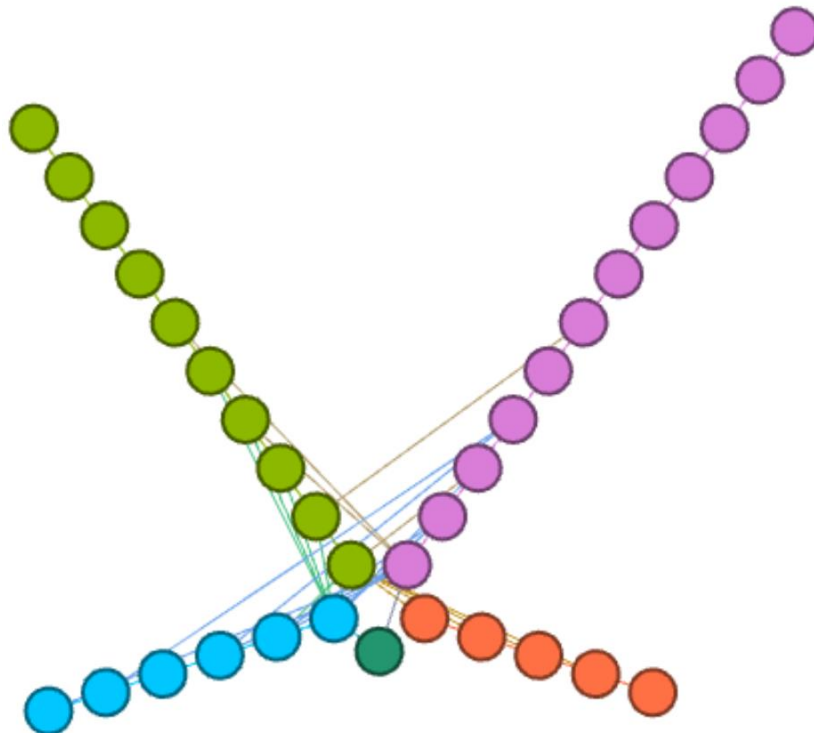
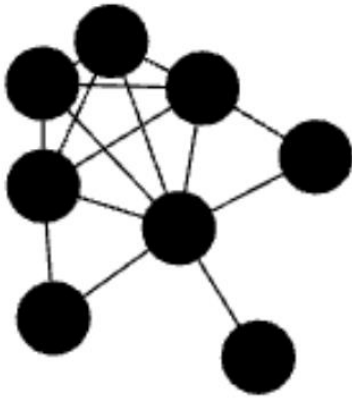


Figure 6.4

Figure 6.4 demonstrates the clustering by Grivan-Newman Clustering. It shows there are 5 components in total and one node doesn't have members in its community. We can also observe one giant component colored in purple.

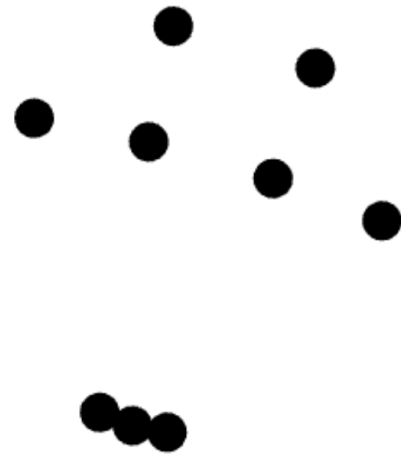
Homophily: it refers to similar nodes are likely to attach to each other in a community. Figure 6.4 shows the network homophily, the layout used here is the radial axis and the communities were grouped on the axis. So, we can observe that most nodes in each community are attached to other nodes except for a few. From the above graph we conclude, there are 4 communities with network homophily.

7. Filters:



Eccentricity partition – 3.0(less)

Figure 7.1(a)



Eccentricity Partition- 5.0 (more)

Figure 7.1(b)

Figure 7.1 demonstrates the Eccentricity of network diameter, initially I have filtered out nodes with less and more eccentricity. Figure 7.1(a) shows that nodes with less eccentricity are close to each other and from figure 7.1(b) we can infer that the nodes with high eccentricity are farthest from each other.

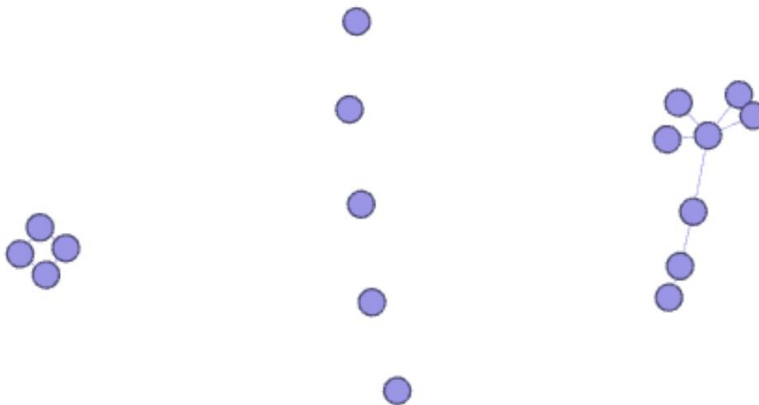


Figure 7.1(c)

Figure 7.1(c) shows that there are two communities divided by a barrier of nodes in the graph. it appears that the nodes in one community are far from the other community but they are actually close.

To drive deeply in finding giant components I have used Yifan Hu layout and Girvan-Newman Clustering.

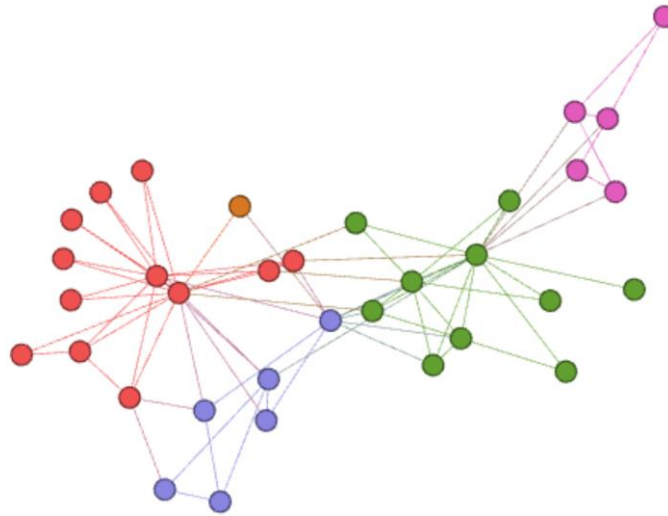


Figure 7.2

From the graph we can observe 4 large components, to find the giant component I have filtered communities by cluster-id. Figure 7.3 shows the 4 components and it is quite obvious that the component displayed in figure 7.3(d) has more nodes than any other components and can be considered as a giant component.



Figure 7.3(a)



Figure 7.3(b)

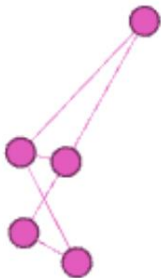


Figure 7.3(c)

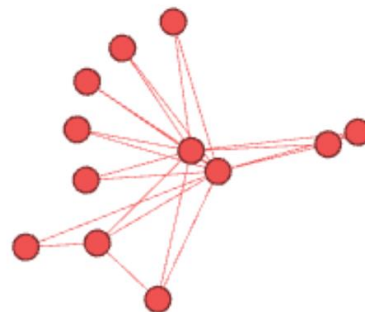


Figure 7.3(d)

8. Modularity:

Modularity is the degree to which the network's components might be separated, for the given network gephi has separated the nodes into 4 different components.

- Modularity: 0.387
- Modularity with resolution: 0.387
- Number of Communities: 4

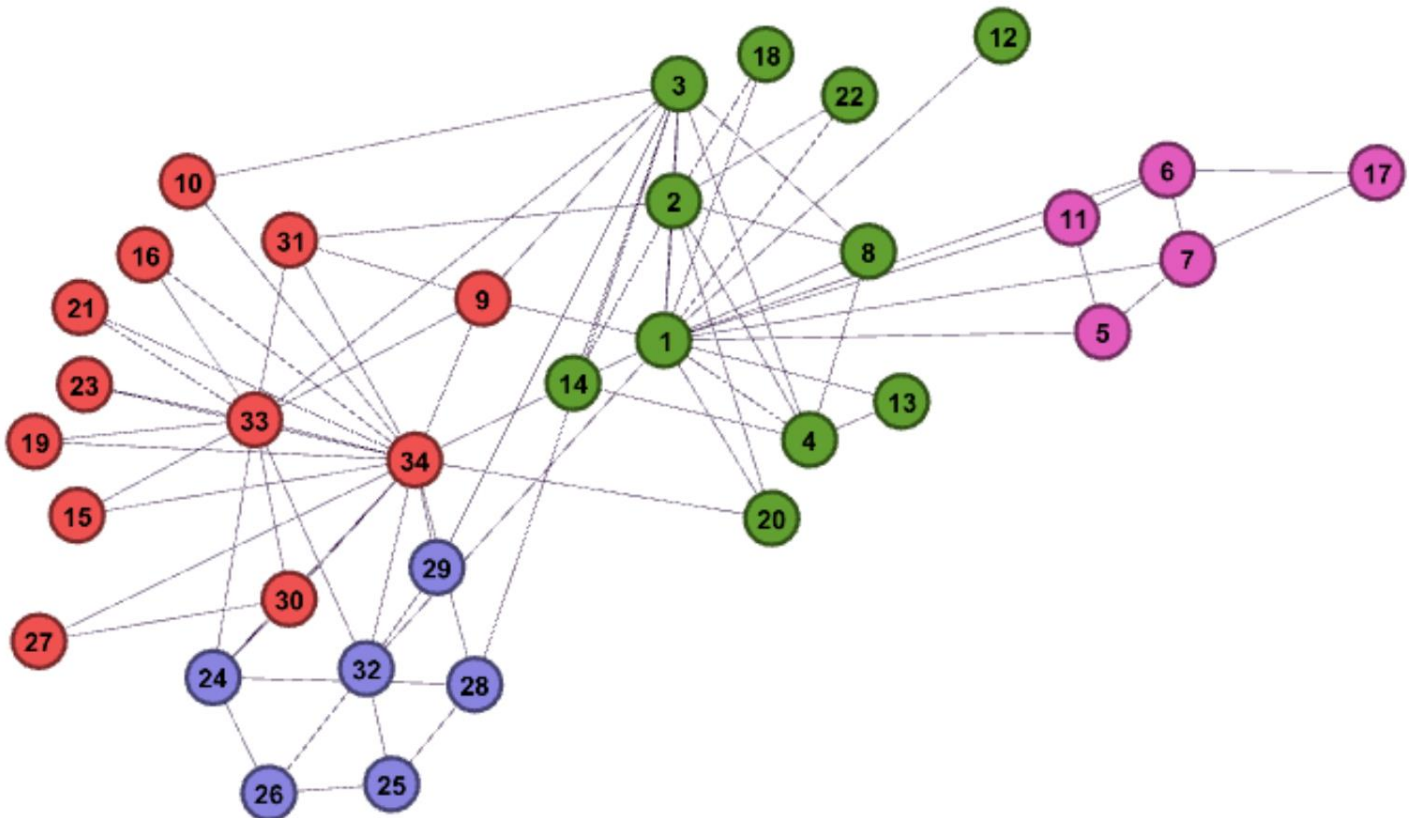


Figure 8.1(a)

The figure demonstrates the communities based on the modularity concept, here I have used muffin hu layout and colored with modularity. we can observe two giant communities (Orange and Green) and two minor communities(Purple and violet). One interesting insight was the two small communities are close to at least one of the giant community. By filtering, we can obtain a final network graph with only two communities. To make the network graph more explanatory I have labeled the nodes, ranked the nodes with the degree and scaled the labels with node size in the final graph (Figure 8.2(b)).

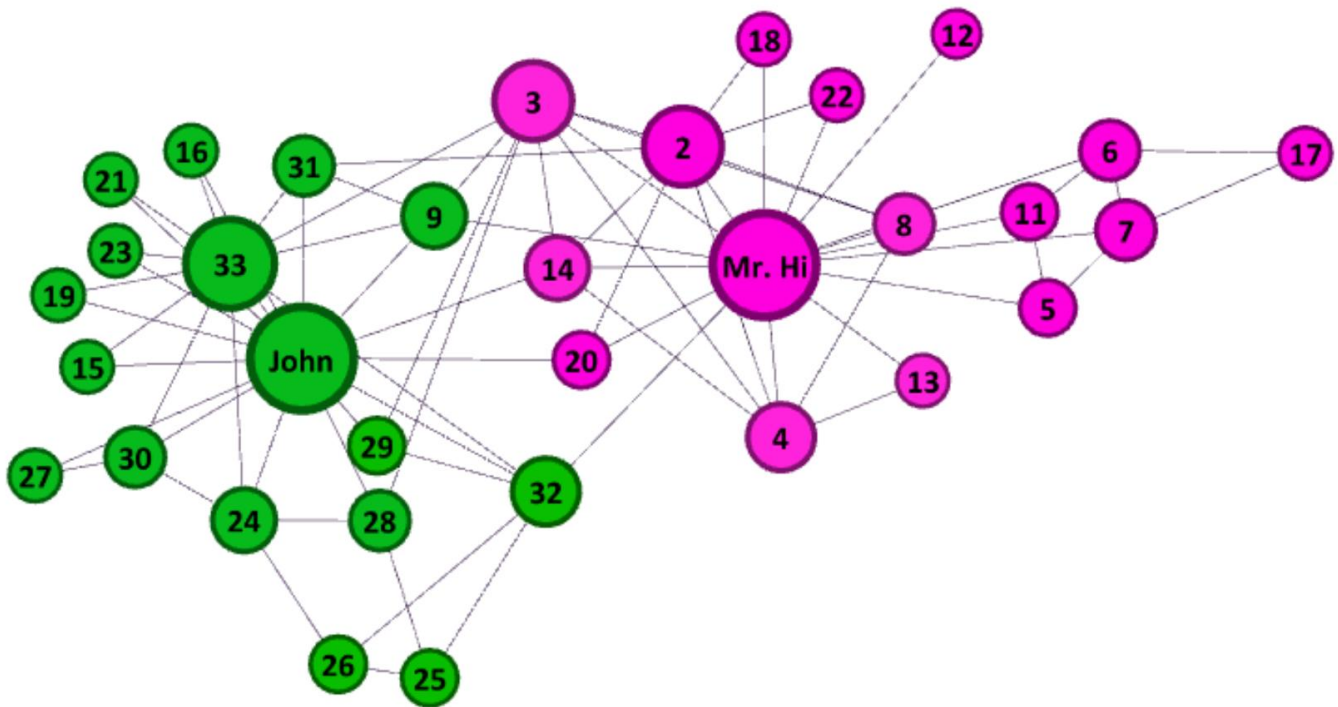


Figure 8.2(b) Final Graph

Figure 8.2(b) demonstrates the final network graph of Zachry's karate club, it shows the two communities formed after separation. It is noticeable that Node 34 and Node 1 are the base nodes for the Green community and Purple community i.e John and Mr.Hi's group. Here I have labeled the nodes 34 and 1 with actual names for better visualization. We can also observe that Mr. Hi and John are connected through nodes 3, 9, 14, 20 and 32. But due to the clustering coefficient, these are clustered with the respective Groups. Without these nodes, there won't be a connection between Mr. Hi and John.

9. Contemplation:

While studying this network I got a better understanding of which algorithm or layout to consider for a particular task. I found that the radial axis is a better option for grouping based on a given attribute, the circular layout is the best available option for arranging the nodes by parameters like degree, etc.

Since the weight of the links is 1, I would like to extend my analysis with a weighted graph to identify the thickness between two nodes i.e. to identify how strongly the members in the group are connected. It's been predicted that Mr.9 allied with john's group due to the clustering coefficient and closeness but in real-life he allied with Mr. Hi, I would further want to investigate this particular event.