Kyle Scott
NID: ky016987

Project 2 – Zachary Karate Club

1. **WW Zachary Article Summary:**

Wayne W. Zachary wanted to create a study that would help explain how and why fission takes place within small social groups. He did so by studying a karate club from 1970-1972 for a total of three years. In this club, there was the club president, John A., as well as the club mentor and karate instructor, Mr. Hi. The existence of two power positions within the club eventually led to a power struggle between John A. and Mr. Hi. John A. wanted to keep Mr. Hi's salary at the same level it had been at, while Mr. Hi wished for a raise in his wage. Both parties saw themselves as the authority figure with the right to make such a decision, so they reached a stalemate, and the club began to gravitate towards the two possible sides. As the club reached its eventual fission, club meetings included members from both sides fighting over issues where the simple majority would win. The meetings turned into a competition of which side could bring in more support so that rules and decisions could be created and repealed based on which side had the current majority. At this point, there was no concrete organizational structure to either of the two divisions, it was merely born from the already in place friendships within the club. For the most part, both sides would still interact with one another. The only times there would be a true divide were during times of conflict where both sides would talk only with the side they agreed with. Over time, this strengthened the division of the groups by pulling those with ideological similarities closer together during the stressful crisis times. Once these divisions had strengthened enough, the club split up for good and the fission had fully completed.

The model created for the karate club is that of a network of friendship. Any type of relationship could have factored into the model, giving it a large degree of freedom on predictability. The karate club created edges between two people if they were seen to interact with one another on a regular basis outside of the club. This formed an undirected graph as there are no "one-way relationships" in the karate club study. The data can be visualized, as shown in the paper, through a matrix. In this matrix, there is a simple binary code to show whether two people form an edge between their nodes. If two people do have such a relationship, it is shown with a 1, and if not, it is shown with a 0. The matrix forms N rows and N columns, where N is the number of nodes in the study, and the data can be filled out row by row. In row 1, column 1 is automatically filled with a 0 (Can't be connected to itself) and then column y is filled with a 1 or 0 depending on if node 1 and node y form an edge or not. This, however, was not enough to show the strength of a relationship because all relationships here are quantified as equals. This matrix is to be known as the *existence matrix*, whereas the must also be another called the *capacitated matrix* which measures relationship strength. For the karate club, this included a list of eight possible activities that two people may share outside of class, and the number out of eight of shared activities would become the value in the capacitated matrix. Now, the capacitated matrix can take on any values between 0 and 8, representing weights for the edges instead of the binary existence from before.

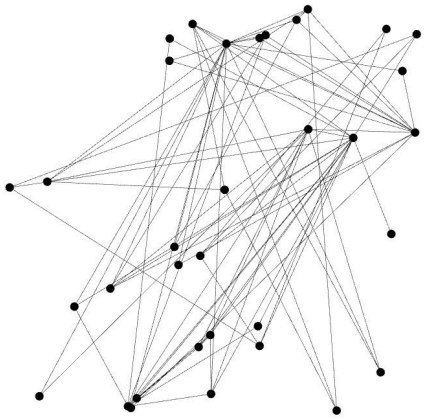
Finally, the model can be used to analyze network flows within the club. This represents the start and end points of the diffusion of information. Those that hear the information in the

first half can be labeled as the “source” and those in the back half the “sink”. Zachary found the label of source vs. sink to be equivalent (34/34 correctly predicted) to the factions within the club (John A’s side or Mr. Hi’s side). Upon testing the side after fission, the same model predicted the correct side in 33/34 cases. Zachary concluded through these results that the basis for the split in the club was rooted in the members friendships, whether it was conscious or unconscious.

Purpose for Analyzing this Network:

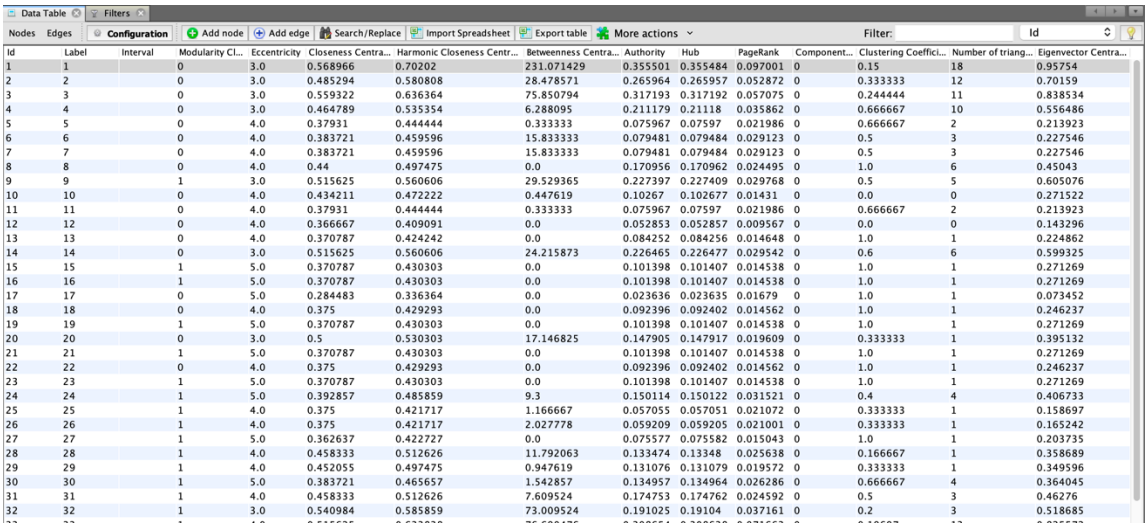
The purpose for analyzing this network is to first understand how relationships between individuals impact the diffusion of information and the likelihood of those individuals to take specific sides during conflict. With the data we have been given, the goal is clearly to predict whether a specific member of the karate club would end up on the side of John A. or Mr. Hi based on relationships and network flow, as shown by Zachary in his paper.

2. **Initial GEPHI Report**



The initial graph doesn’t allow us to ascertain too much – We can’t see the weights of the edges or the size of any of the nodes. In addition, the edges are so scattered that it’s hard to tell how many of the nodes have many connections. From the appearance of the outside of the graph, we can at least see that there are a significant number of nodes (roughly 12-15) with three connections or less.

3. **Data Laboratory**



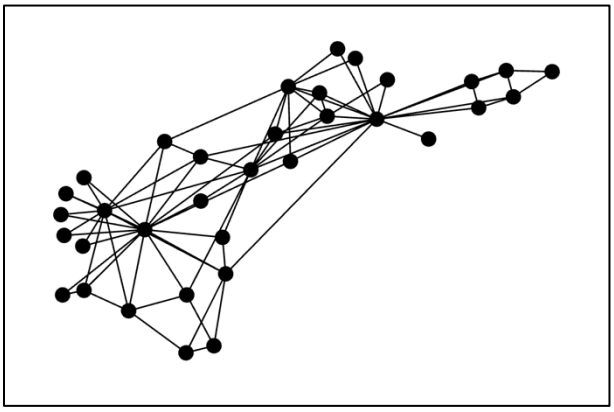
Data Table | Filters

Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions | Filter: | Id

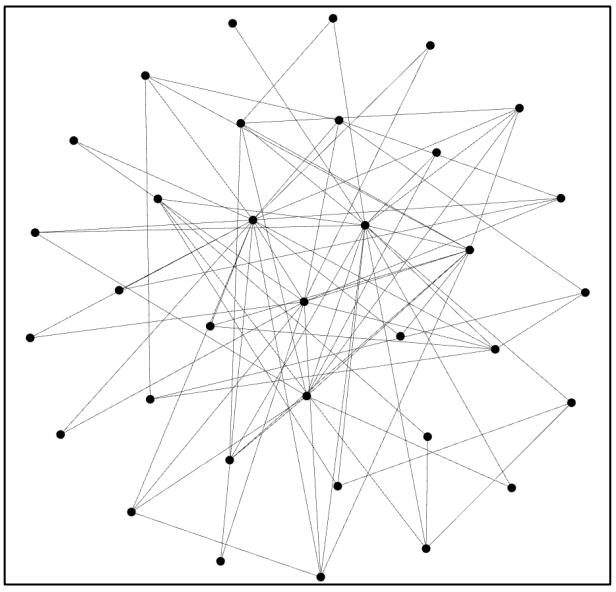
Id	Label	Interval	Modularity Cl...	Eccentricity	Closeness Centra...	Harmonic Closeness Centr...	Betweenness Centra...	Authority	Hub	PageRank	Component...	Clustering Coeffici...	Number of triang...	Eigenvector Centra...
1	1		0	3.0	0.568966	0.70202	231.071429	0.355501	0.355484	0.097001	0	0.15	18	0.95754
2	2		0	3.0	0.485294	0.580808	28.478571	0.265964	0.265957	0.052872	0	0.333333	12	0.70159
3	3		0	3.0	0.559322	0.636364	75.850794	0.317193	0.317192	0.057075	0	0.244444	11	0.838534
4	4		0	3.0	0.464789	0.535354	6.288095	0.211179	0.21118	0.035862	0	0.666667	10	0.556486
5	5		0	4.0	0.37931	0.444444	0.333333	0.075967	0.07597	0.021986	0	0.666667	2	0.213923
6	6		0	4.0	0.383721	0.459596	15.833333	0.079481	0.079484	0.029123	0	0.5	3	0.227546
7	7		0	4.0	0.383721	0.459596	15.833333	0.079481	0.079484	0.029123	0	0.5	3	0.227546
8	8		0	4.0	0.44	0.497475	0.0	0.170956	0.170962	0.024495	0	1.0	6	0.45043
9	9		1	3.0	0.515625	0.560606	29.529365	0.227397	0.227409	0.029768	0	0.5	5	0.605076
10	10		0	4.0	0.434211	0.472222	0.447619	0.10267	0.102677	0.01431	0	0.0	0	0.271522
11	11		0	4.0	0.37931	0.444444	0.333333	0.075967	0.07597	0.021986	0	0.666667	2	0.213923
12	12		0	4.0	0.366667	0.409091	0.0	0.052853	0.052857	0.009567	0	0.0	0	0.143296
13	13		0	4.0	0.370787	0.424242	0.0	0.084252	0.084256	0.014648	0	1.0	1	0.224862
14	14		0	3.0	0.515625	0.560606	24.215873	0.226465	0.226477	0.029542	0	0.6	6	0.599325
15	15		1	5.0	0.370787	0.430303	0.0	0.101398	0.101407	0.014538	0	1.0	1	0.271269
16	16		1	5.0	0.370787	0.430303	0.0	0.101398	0.101407	0.014538	0	1.0	1	0.271269
17	17		0	5.0	0.284483	0.336364	0.0	0.023636	0.023635	0.01679	0	1.0	1	0.073452
18	18		0	4.0	0.375	0.429293	0.0	0.092396	0.092402	0.014562	0	1.0	1	0.246237
19	19		1	5.0	0.370787	0.430303	0.0	0.101398	0.101407	0.014538	0	1.0	1	0.271269
20	20		0	3.0	0.5	0.530303	17.146825	0.147905	0.147917	0.019609	0	0.333333	1	0.395132
21	21		1	5.0	0.370787	0.430303	0.0	0.101398	0.101407	0.014538	0	1.0	1	0.271269
22	22		0	4.0	0.375	0.429293	0.0	0.092396	0.092402	0.014562	0	1.0	1	0.246237
23	23		1	5.0	0.370787	0.430303	0.0	0.101398	0.101407	0.014538	0	1.0	1	0.271269
24	24		1	5.0	0.392857	0.485859	9.3	0.150114	0.150122	0.031521	0	0.4	4	0.406733
25	25		1	4.0	0.375	0.421717	1.166667	0.057055	0.057051	0.021072	0	0.333333	1	0.158697
26	26		1	4.0	0.375	0.421717	2.027778	0.059209	0.059205	0.021001	0	0.333333	1	0.165242
27	27		1	5.0	0.362637	0.422727	0.0	0.075577	0.075582	0.015043	0	1.0	1	0.203735
28	28		1	4.0	0.458333	0.512626	11.792063	0.133474	0.13348	0.025638	0	0.166667	1	0.358689
29	29		1	4.0	0.452055	0.497475	0.947619	0.131076	0.131079	0.019572	0	0.333333	1	0.349596
30	30		1	5.0	0.383721	0.465657	1.542857	0.134957	0.134964	0.026286	0	0.666667	4	0.364045
31	31		1	4.0	0.458333	0.512626	7.609524	0.174753	0.174762	0.024592	0	0.5	3	0.46276
32	32		1	3.0	0.540984	0.585859	73.009524	0.191025	0.19104	0.037161	0	0.2	3	0.518685

In the Laboratory, we can see the values for many important calculations for each node. The first of which is the Modularity Class. The original Modularity calculation will spit out 4 different communities when run with a Resolution of 1.0 (Lower number creates more communities). The problem is we know our data includes two different classes: Mr. Hi's side and John A's side. To rectify this problem, I switched the Resolution to 1.4 (Lowest number that yielded 2 classes), so we now have the breakdown for our prediction which we will discuss later. Another interesting column is the column for Betweenness Centrality. In this column, we can see Node 1 has a staggering value of 231 and Node 34 (Not in screenshot) has a similarly staggering value of 160. Not so coincidentally, these two nodes represent Mr. Hi (1) and John A. (34). Similarly, the authority/hub sections are headed by the same two nodes representing Mr. Hi and John A., as one should expect to see. The final interesting thing I noticed in the data is that Node 3 has the second highest value for Closeness Centrality. As we will see in our analysis, this Node is the most difficult node to place as it has many connections, and those connections are to both sides of the graph with no majorly discernible pattern.

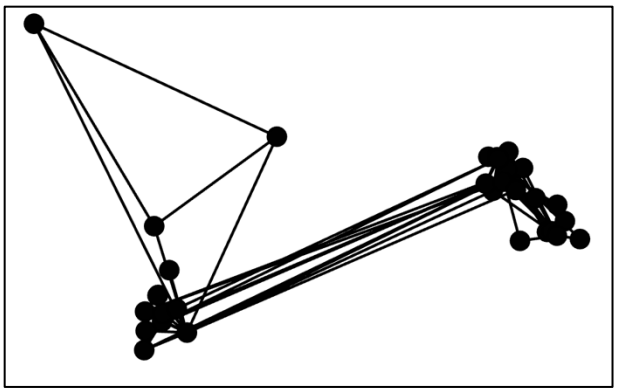
4. **Layout Algorithms**



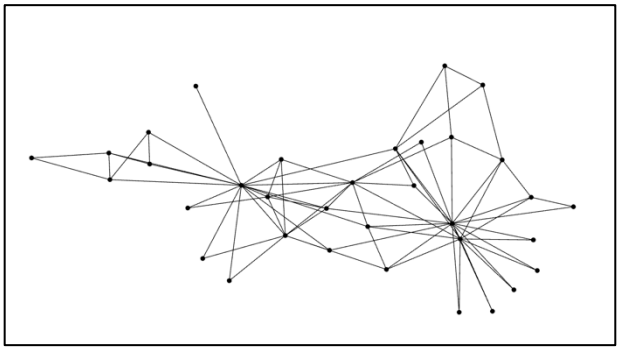
FORCE ATLAS 2:
This algorithm has created two clear communities in the graph – The hubs of Mr. Hi and John A. are also easily distinguished here. Another interesting thing to note is the bridge (Node 3) right in the middle discussed in the last section. Overall, this algorithm produced a clean graph that shows off exactly what we need for this analysis.



FRUCHTERMAN REINGOLD:
This algorithm creates a graph with roughly uniform edge length. This is nice because we can see exactly how many relationships each node has (no edges are supposed to intersect), making it easier to conclude which nodes are hubs in the graph. The problem here is we would like to study the two sides of the karate club, so this graph is not the best choice for us as it does not split up nicely for communities.



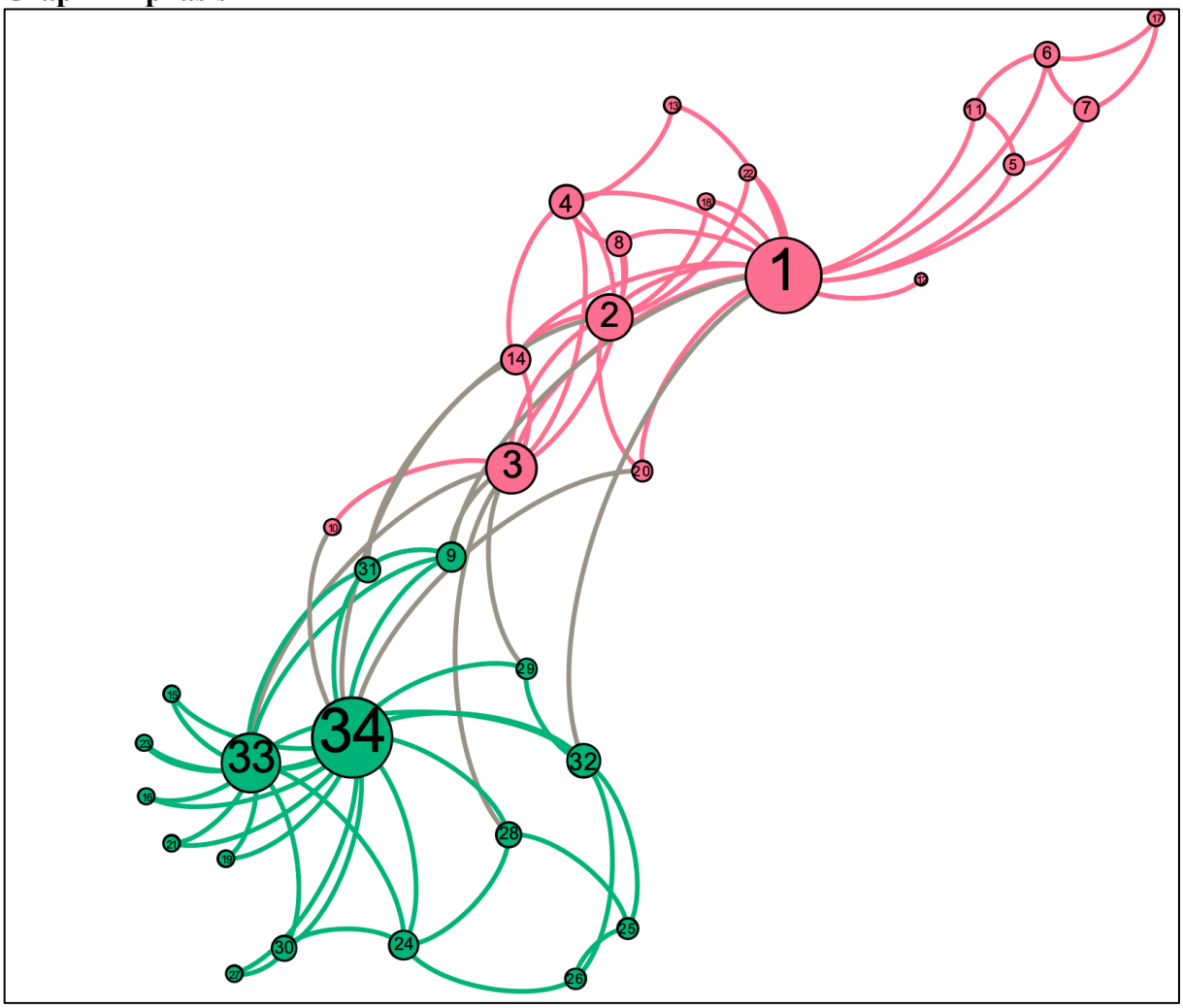
OPENORD:
This layout clearly doesn't work well for this data. It is designed to show clusters which is partially what we want to see (Mr. Hi vs John A.), but the algorithm will only create a nice-looking graph when there are thousands of nodes, whereas we only have 34. There is also no case for distinguishing the hubs in either of these clusters and overall doesn't look very appealing. It would be interesting to run some modularity analysis to see whether the clusters are accurate for predicting the data, despite being a difficult graph to read.



YIFAN HU:
This is another nice-looking layout. Here there is a similar split to the Force Atlas model, however there is a little more clutter in this graph between the two communities. There isn't quite as much separation even though the hubs are still easy to distinguish. This is definitely a good option, even if it looks a little less clean when compared to Force Atlas.

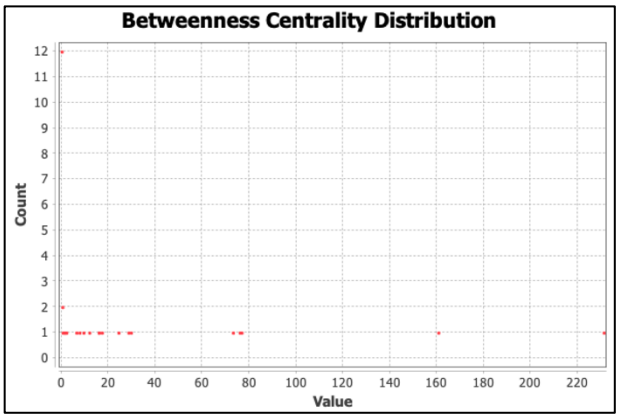
I will be choosing the Force Atlas 2 model to run my analysis due to the clear split into communities and easily identifiable hubs. I want a graph that will show off how Zachary came to his predictive analysis in his paper, and this algorithm will be able to best show how the friendships between club members shaped the two communities.

5. **Graph Emphasis**

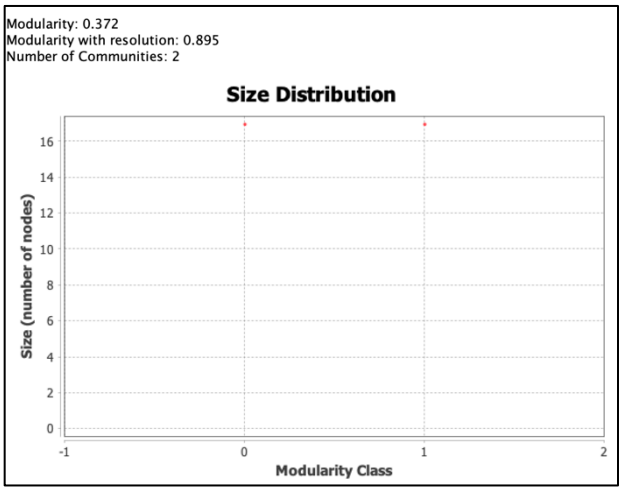


This new diagram shows the two colored communities, shown in Pink and Green. On the green side, Node 34 represents John A. and on the pink side, Node 1 represents Mr. Hi. Not surprisingly, these two nodes are the largest in the whole graph. The data did not come with weights for the edges as shown in Zachary's paper, so the coloring for the edges was done based upon the modularity, just like the nodes. In this case, any pink edges are a relationship between two pink nodes, green edges for green nodes. However, the brown edges show a relationship between one pink node and one green node. The more brown edges that a node has, the more split its relationships are. The interesting thing here is that the only node (Aside from John and Hi) to have more than two brown edges is Node 3 – The "bridge" from earlier. Node 9, which was the only node incorrectly predicted in Zachary's study, has two brown edges. This tells us that in general, the two groups are friends with those in their group with very little crossover to the other group. This makes predicting sides quite a lot easier in our modeling. The phenomena described here is that of homophily, and we can conclude that the karate club does show signs of it. In general, each community associates with itself and not with the other.

6/7. **Statistics and Filters**



CENTRALITY:
The graph to the left shows the distribution of betweenness centrality values for each node. This value is a measure of how many shortest paths each node is involved in between nodes on the graph. In our study, this isn't the *most* important measure, but it does show the power that Mr. Hi and John A. have over the whole network (They are the two highest values here).



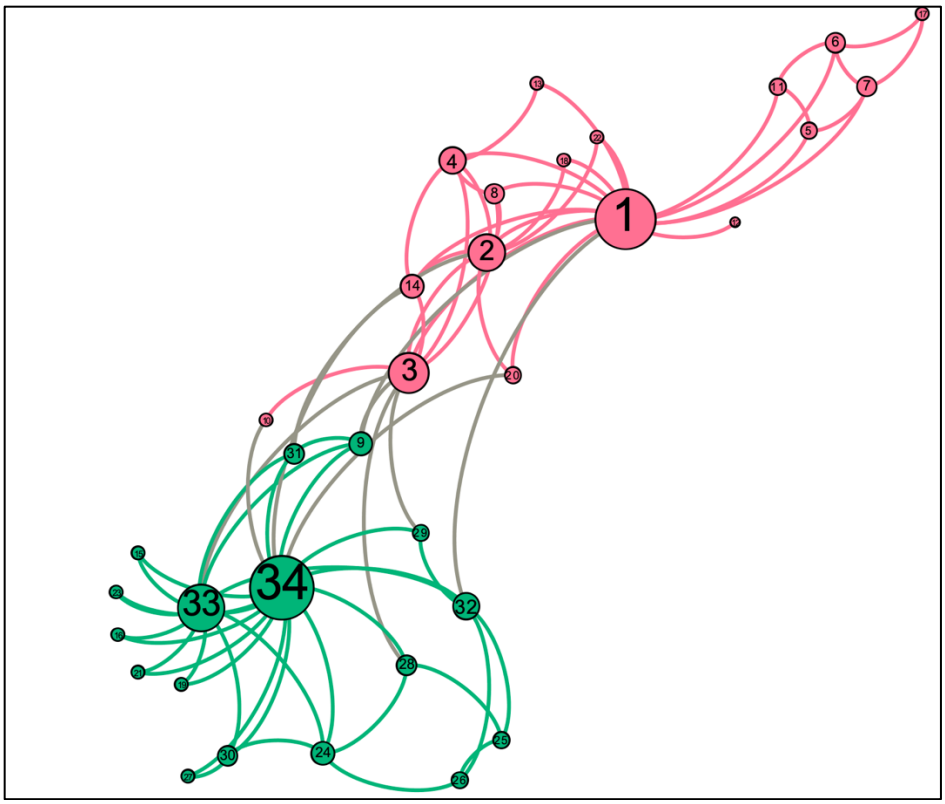
MODULARITY:
Here we get a good idea of whether our graph shows indications of strong community structure or not. With a range of [-1,1], the modularity value will tell us that it is likely to find the existence of community structure for values above 0. We have a value of .372 and a value of .895 with resolution. This is an indication that community structure is likely, but we cannot be sure without considering the context of our data.

Luckily, we do know within the context of our data that there *should* be two communities and we have shown the existence of exactly two communities in our graph previously – One being the group of John A. and the other being the group of Mr. Hi. This structure proves to be quite accurate as 32/34 of the people belong to the community that they would ultimately choose in the karate club split.

The graph has a density of .139 which shows that most nodes are not connected to one another. With the exception of one or two nodes aside from John A. and Mr. Hi, most people are connected with only a couple others in the dataset. (Homophily was discussed earlier in the analysis)

8. **Modularity Analysis**

(Same graph as before, but I will paste it again)



The final graph after running modularity analysis shows that the karate club was split into two main groups (Mr. Hi and John A.). The initial report of the karate club analysis written up by Zachary showed the following breakdown of sides for each node:

INDIVIDUAL NUMBER IN MATRIX *C*	FACTION MEMBERSHIP FROM DATA	FACTION MEMBERSHIP AS MODELED	HIT/ MISS	CLUB AFTER SPLIT FROM DATA	CLUB AFTER SPLIT AS MODELED	HIT/ MISS
1	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
2	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
3	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
4	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
5	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
6	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
7	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
8	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
9	John	John	Hit	Mr. Hi's	Officers'	Miss
10	John	John	Hit	Officers'	Officers'	Hit
11	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
12	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
13	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
14	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
15	John	John	Hit	Officers'	Officers'	Hit
16	John	John	Hit	Officers'	Officers'	Hit
17	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
18	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
19	John	John	Hit	Officers'	Officers'	Hit
20	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
21	John	John	Hit	Officers'	Officers'	Hit
22	Mr. Hi	Mr. Hi	Hit	Mr. Hi's	Mr. Hi's	Hit
23	John	John	Hit	Officers'	Officers'	Hit
24	John	John	Hit	Officers'	Officers'	Hit
25	John	John	Hit	Officers'	Officers'	Hit
26	John	John	Hit	Officers'	Officers'	Hit
27	John	John	Hit	Officers'	Officers'	Hit
28	John	John	Hit	Officers'	Officers'	Hit
29	John	John	Hit	Officers'	Officers'	Hit
30	John	John	Hit	Officers'	Officers'	Hit
31	John	John	Hit	Officers'	Officers'	Hit
32	John	John	Hit	Officers'	Officers'	Hit
33	John	John	Hit	Officers'	Officers'	Hit
34	John	John	Hit	Officers'	Officers'	Hit
TOTALS		34 hits, 0 misses 100% hits, 0% misses			33 hits, 1 miss 97% hits, 3% misses	

The column labeled as "CLUB AFTER SPLIT FROM DATA" is the important column we need to look at. This column will show us exactly how accurate our modularity analysis is. By looking through the two communities in our graph, we can determine if the node belongs to the proper community based upon the data from this table. Upon looking through our graph, the conclusion is that Node 9 (Predicted as Officers', actual Mr. Hi's) and Node 10 (Predicted as Mr. Hi's, actual Officers') were the only misplaced nodes on our graph. This shows that we have reached similar results to Zachary, as we have 32 hits to only 2 misses for a success rate of 94%, a sufficient number to accept the accuracy of our model. It is also important to note that the model predicted 33/34 correctly for the question of faction membership. The only node incorrectly placed was Node 10 which appears in Mr. Hi's faction instead of John A. Once again, the 97% success rate is enough to accept the validity of our model for that prediction.

9. **Final Contemplation**

I noticed when loading the data into GEPHI that it did not include any information on the weights of the edges in the graph. I think it would've added a lot to the analysis to be able to see how strong specific relationships were between people like it showed in Zachary's paper. I would've been interested to see if the misplaced nodes were nodes with much smaller weights in their edges, implying harder prediction for the faction split on that node. One could assume that nodes with higher edge weight are better friends and therefore more firmly a part of the same faction together.

I think that the most interesting part of the project was the introduction to prediction within the Network Science world. In the past, I have done different things with Machine Learning, working with large datasets in python and R with the goal of predicting some sort of outcome using linear modeling and statistics. I found it entertaining to get to take some of that experience and start applying it to network graphs, and I'm interested to see what more complex things can be done using GEPHI and other network science tools.