Methods of data mining

Assignment 4

Task 1

Miro-Markus Nikula

712686

1. a)

The following statistics were calculated with Gephi, and the values presented are the two highest values among all nodes. The value of the metric is inside the parenthesis.

i) **Degree**: 1477 (43) & 1443 (43). The value describes the number of edges connecting the node.

ii) **Weighted degree:** 2437 (221) & 1563 (217). Weighted degree is like degree but also considers the weights of the edges. The value is the sum of the weights of the edges connecting a node.

iii) **Closeness centrality:** 1477 (0.96) & 1443 (0.96). Closeness centrality is the average of all the shortest paths from a node to each node. A small value would indicate that the node on average is close to its connected nodes. These two nodes however have the largest values, which means that on average they're not very close to other nodes, but this is due to the high degree of them. Node 1750 has the smallest closeness centrality (0.64), which is mostly because it has the smallest degree (20). That student has a "smaller and closer" network.

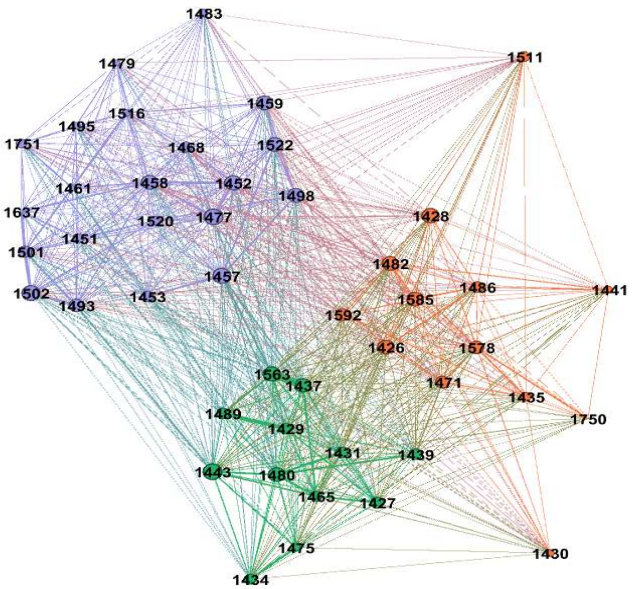
iv) **Betweenness centrality:** 1443 (10.3) & 1477 (9.3). Betweenness centrality describes the most central nodes. Each node has a shortest path to each node. If a node has a high betweenness centrality, it means that it is included in many of these shortest paths.

v) The two highest degree nodes are not the same as highest weighted degree nodes, because the edges they're connected by don't have very large values.

vi) I'd say that a high betweenness centrality is crucial for information flow. The most information in general flows through the most central nodes and for that reason the most important nodes for information flow are 1477 and 1443.

b)

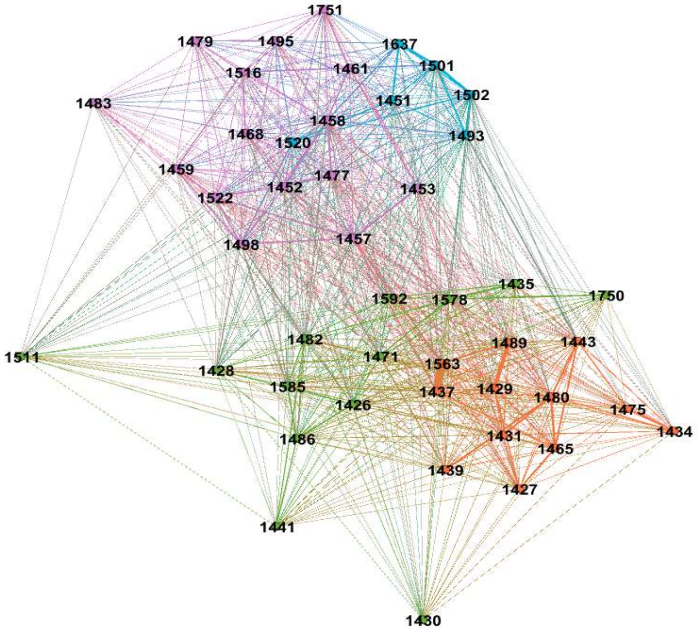
I found 3 communities using Gephi's modularity function:



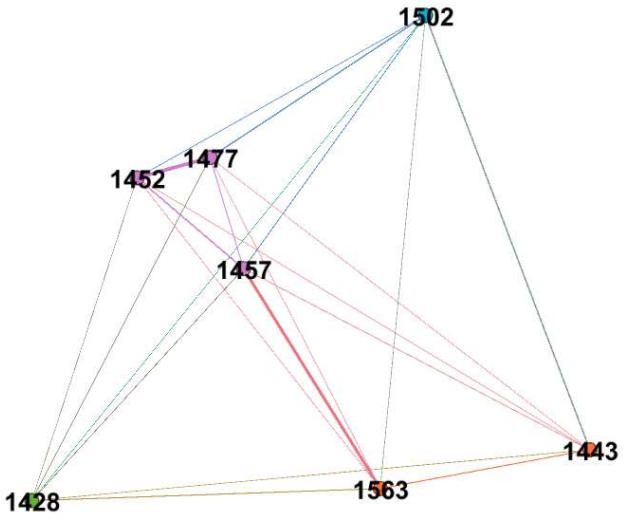
The different communities can be seen in the above graph as different colors. Blue is about 46%, orange 28% and green 26% of the nodes. The distinction between the three communities is quite clear: in blue are the 5A-class, in green the females from class 5B and in orange the males from 5B. At least one exception is that in green there's at least one male from 5B (node 1439). Even though the 5B-class is separated in two in here, it still can be seen from the graph that there is not as clear distinction between 5B's different genders as there is between 5B and 5A. Remembering my time as a 5th grader, this distinction between classes and genders makes a lot of sense.

c)

- I played around with the functions and tried the modularity function again and unlike in part b, this time ended up with 4 communities. The distinction is otherwise the same as in part b, but this time there's a small community consisting of 7 females in class 5A.



- As said before, the number of edges per node (connections per student) is called degree. The degree values in this graph range from 20 to 43 meaning that there is a lot of variation between the number of connections a student has. Some could also call the more connected students as more social. What I'm curios in is how are the most social students divided among the four communities. Ie. do the social people all belong to one class or are they all boys or girls or do social people have their own community. For this, I filtered the students by their degree. I filtered out everyone with a degree lower than 41. This left us 7 students who in this context I consider as the most social ones and as can be seen from the image below, they divide between all the 4 groups. Among these students we have males from 5A and 5B as well as females from 5A and 5B.



- I also tried running the clustering coefficient function. Clustering coefficient (<=1) is supposed to tell how connected a node's neighbors are to one another. I would translate this in this context as "how withdrawn a student's group of friends are". As expected, the smallest degreed student 1750 has the highest clustering coefficient with 1.0 and the highest values are with the most connected students 1477 and 1443 with 0.79.