Task 1

a)

Set a suitable threshold for each measure (look at 5–8 most central nodes):

- For the following a threshold of 8 most central nodes for visualization is used

Identify most central and influential nodes with different measures:

- node degree:

Label	Degree
1477	43
1443	43
1457	42
1502	42
1563	41
1452	41
1428	41
1458	40

-

- weighed degree:

Label	Weighted Degree
1437	221.000003
1563	216.800002
1457	186.800001
1458	183.799999
1452	171.800002
1477	165.2
1498	164.599999
1480	161.8

-

- closeness centrality:

Id	Closeness Centrality
1443	0.957447
1477	0.957447
1457	0.9375
1502	0.9375
1428	0.918367
1452	0.918367
1563	0.918367
1426	0.9

-

- betweenness centrality:

Id	Betweenness Centrality
1443	10.267852
1477	9.288583
1502	8.774473
1457	8.27553
1563	8.193991
1480	7.875956
1522	7.730772
1585	7.57164

-

What do these measures tell about nodes?

- node degree:
  - The degree of a node tells us the number of edges connected to the node.

- weighed degree:
  - The weighted node degree is the sum of the edge weights for edges incident to that node.

- closeness centrality:
  - $$\textbf{Closeness centrality: } C_C(v) = \frac{1}{avg_{u \in V}\{Dist(v,u)\}}$$
  - The closeness centrality of a node measures its average farness (inverse distance) to all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes. Therefore, the more central a node is, the closer it is to all other nodes.

- betweenness centrality:
  - $$\textbf{Betweenness centrality: } C_B(v) = \frac{\sum_{u,w \in V, u \neq w} \frac{\#\{\text{shortest-paths(u,w) through v}\}}{\#\{\text{shortest-paths(u,w)}\}}}{\binom{n}{2}}$$
  - Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Betweenness centrality measures the extent to which a vertex lies on paths between other vertices. Vertices with high betweenness may have considerable influence within a network. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

b)

Definition of community measures:

- Modularity:
  - Modularity measures the strength of division of a network into clusters. Networks with high modularity have dense connections between the nodes

within the same cluster but sparse connections between nodes in different clusters. Modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules.

- Girvan-Newman clustering:
  - Divisive hierarchical clustering based on edge betweenness. Number of shortest paths passing through the edge. The algorithm removes the “most valuable” edge, traditionally the edge with the highest betweenness centrality, at each step.

Identify communities:

Processed Graph Data

Nodes: 46

Edges 809

-

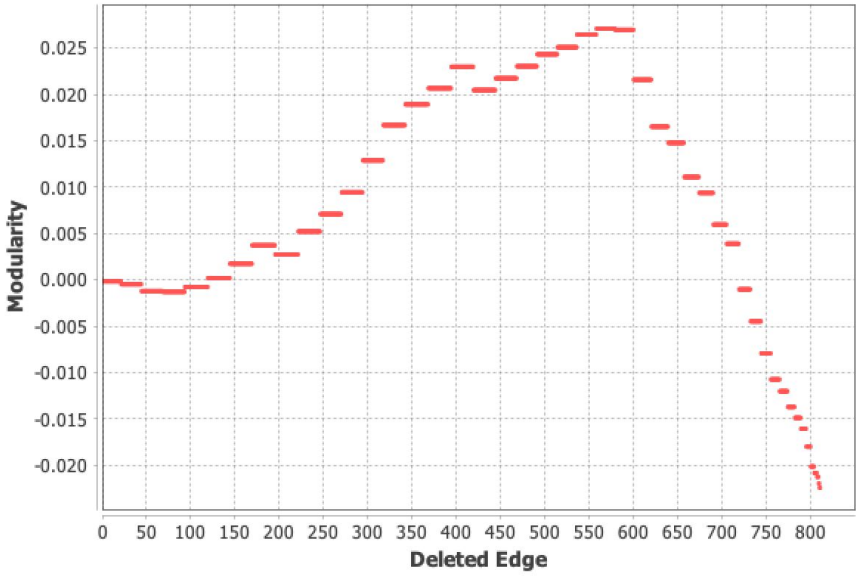
- Girvan-Newman clustering:

Communities

Number of communities: 24

Maximum found modularity: 0.027242184

-



-

- We can see that the Girvan-Newman algorithm is not that well suited for the given network. The network is very dense and therefore well connected. The algorithm needs to delete a lot of edges, here 600 out 800, to get the maximum found modularity of 0.027. That score is still very low. Furthermore, 24 communities are detecting while having only 46 nodes in the graph. That would mean around two nodes per cluster which is a bad community detection.

- Modularity:
    - For this algorithm we can fine tune the resolution by running the algorithm with different resolutions.

Resolution	Number of communities	Modularity	Modularity with resolution
0	46	-0.026	-0.026
0.2	17	0.207	-0.010
0.4	10	0.309	0.056
0.6	7	0.341	0.144
0.8	5	0.371	0.251
1.0	4	0.379	0.379

    - For resolution = 0.8 we can see that modularity and modularity with resolution converged already very closely. We can assume that for modularity 4-5 clusters are reasonable results.

Compare results (similarities and differences):

- Done above.

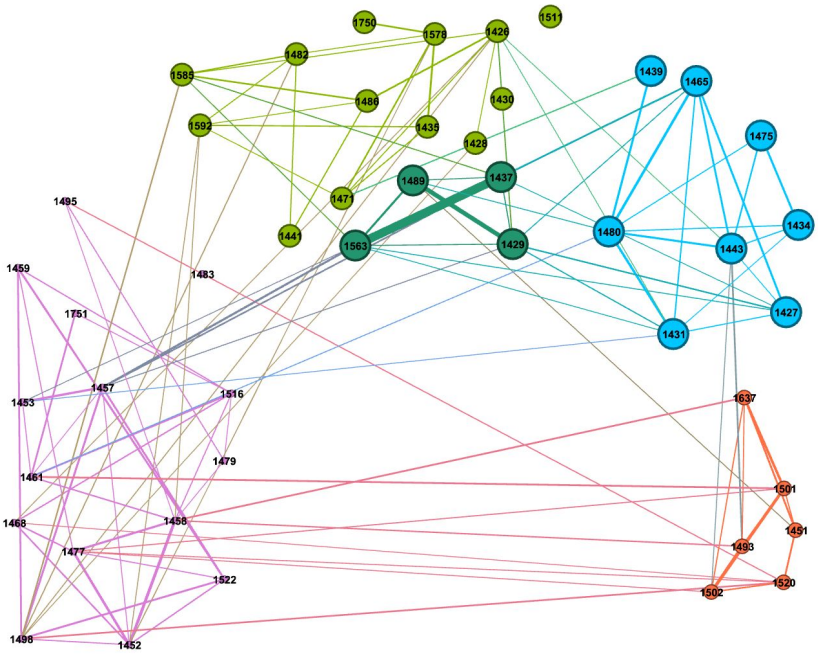
Does the background information (class and gender) explain communities?

- Yes, the background information explains communities. Since the community results for Girvan-Newmann are not really good (approx. 24 communities), I analyzed the background information for modularity. We have about 3-4 communities. So I picked out the nodes of each cluster separately and double-checked with metadata from “t1_schoolclass5meta.txt”. And turned out that the background information explain communities. For example, in one cluster mainly female, and then also divided by class. However, I didn’t figure out how to import the metadata into gephi. Because importing another txt/csv result in creating a new graph. Therefore, I needed to it separately manually on paper

c)

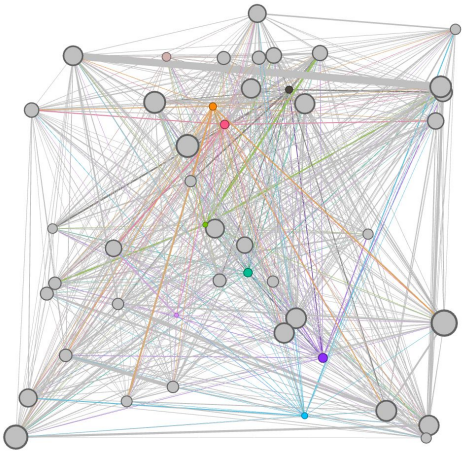
Present the communities visually

- Task:
    - Use distinct colours to show nodes, add labels, hide low weight edges to simplify the graph, move nodes so that communities become separated.
    - Take snapshots of both Modularity and Girvan-Newman results.
- Modularity:



-
- each class has its own color
- nodes were dragged to separate from other cluster and closer to own cluster
- size of the node represents the modularity
- weights < 5.8 are filtered out

- Girvan-Newman Algorithm

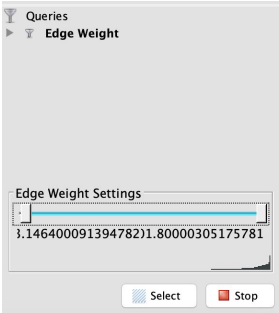


-
- I could give every cluster a color. But that's unnecessary work because we can't conclude much out of it. Since anyways each cluster only contains 2 nodes. Maybe more interesting is to differentiate the size of nodes by clustering coefficient. However, also doesn't tell us much.

d)
Modularity function for community detection:

Hiding low weight edges with different thresholds and analyze links inside and between communities:

- Gephi has the nice option to filter by edge weights:



- By moving the bar slowly to the right and increasing the threshold you can observe the following. First of all, the intra edges - edges between different clusters - tend to disappear more frequent than the inter edges - edges within the cluster.

Which communities have strongest interconnections?

- The light green cluster is definitely the cluster with the weakest interconnections. These disappear the first.

What are bridge nodes that combine two communities (end points of strong links between communities)?

- Deletion of bridge nodes increases the graph's number of connected components. It is not contained in any cycle. For a connected graph, a bridge can uniquely determine a cut. Followed by blue, then purple, afterwards orange. See the colors in the screenshot above. The strongest links are of the dark green cluster with node IDs: 1429, 1437, 1489, 1563

Are these the same as central nodes? Or what is the role of central nodes in communities?

- Centrality identifies the most important vertices within a graph. Which node are the most influential nodes to other nodes in the graph. Deleting that node doesn’t necessarily result in more clusters or disconnects the graph.