PROJECT
ON
FUNDAMENTALS OF DATABASE MANAGEMENT SYSTEM (FDBMS)

Submitted to: Prof. Ashok Harnal

Submitted by: Group 15
Yog Raj Singh (025040)
Yogesh Sachdeva (025041)
Manyata Manocha (025053)

This project contains an analysis of the weighted network of coappearances of characters in Victor Hugo's novel "Les Miserables". Nodes represent characters as indicated by the labels and edges connect any pair of characters that appear in the same chapter of the book. The values on the edges are the number of such coappearances. The data on coappearances were taken from D. E. Knuth, The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley, Reading, MA (1993).
The file contained a network representing “who appears next to whom” in the 19th century novel *Les Misérables* by Victor Hugo. It displays a link between characters **A** and **B** means they appeared on the same page or paragraph in the novel.

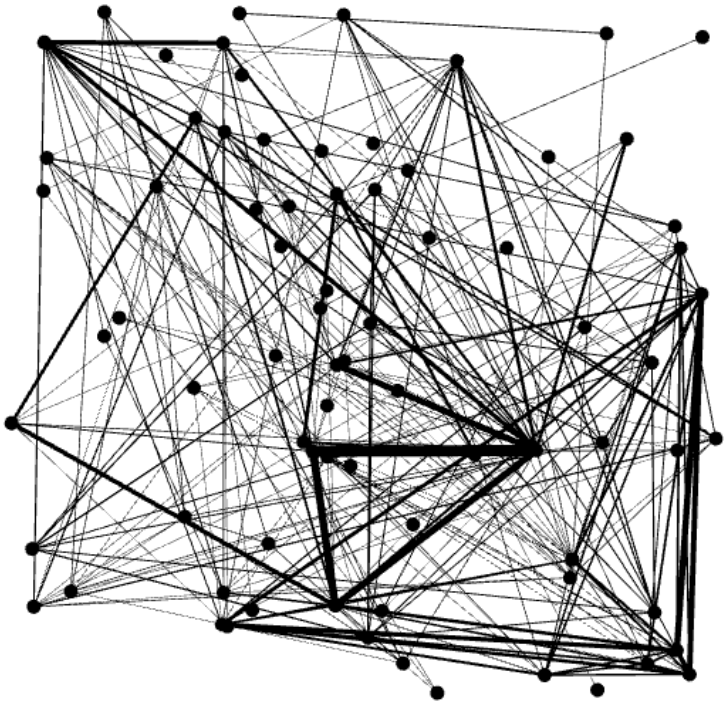
Initially, when we loaded the file, we analysed *a few things:*

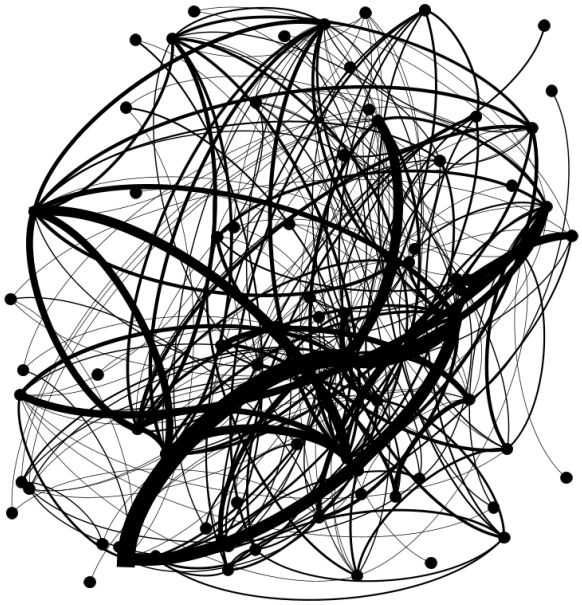
- The network comprises **74** characters, and they’re all connected by **254** links.
- Links are **undirected**, meaning that if A is connected to B, then it is the same as B connected to A.
- The report also tells us the graph is **not dynamic**: it means there is no evolution or chronology, it won’t “move in time”.

The “Les Misérables” dataset was downloaded as a GML file. As the dataset was clean and did not require any adjustments, we were able to easily upload it into Gephi. We then changed the appearance of my network’s nodes and edges by altering the colors and size. Within the preview settings, we also made adjustments to the nodes’ label size, the thickness and curvature of the edges, as well as the background color. we continued to make such adjustments until satisfied with my visualization.

1. Initial Plot

Below is the initial rap we received when we loaded the GML file.

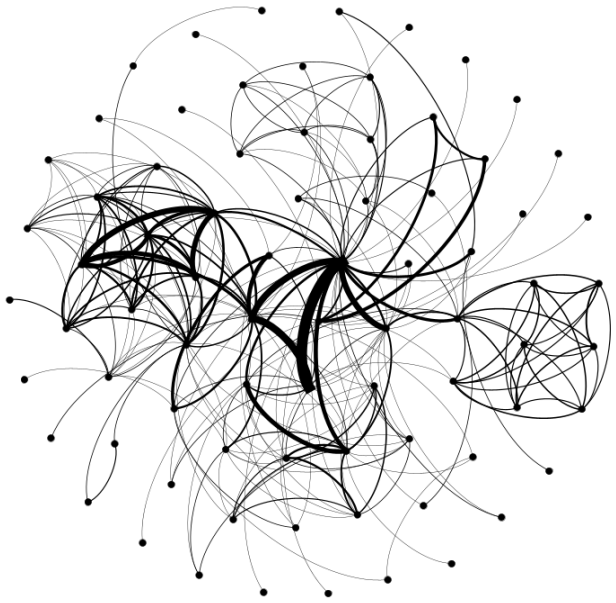




This is how the network initially appeared in Gephi. It doesn’t look very useful.

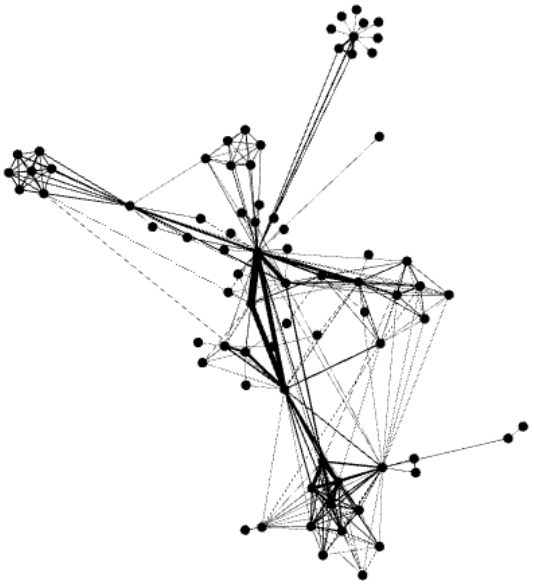
2. Gephi: Fruchterman Reingold Layout

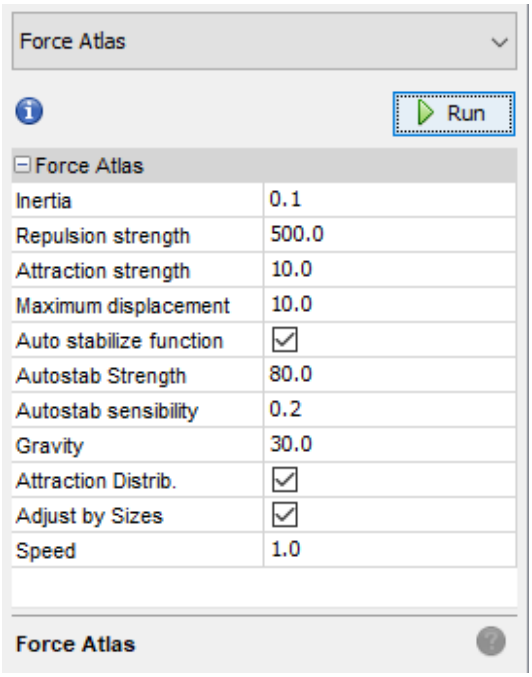
Next, we applied the Fruchterman Reingold Layout to give network an arrangement and shape. The Fruchterman-Reingold benefits of working well on small to medium-sized graphs and converging relatively quickly.



3. Gephi: Force Atlas layout

Next, we applied the Force Atlas layout and made some adjustments to the tuning, behaviour, and performance settings of the layout to give it the shape that it currently has. ForceAtlas2 simulates a system in order to spatialize a network. Nodes repulse each other like magnets, while edges attract their nodes, like springs. These forces create a movement that converges into a balanced state. This forms clusters of characters who are related to each other.

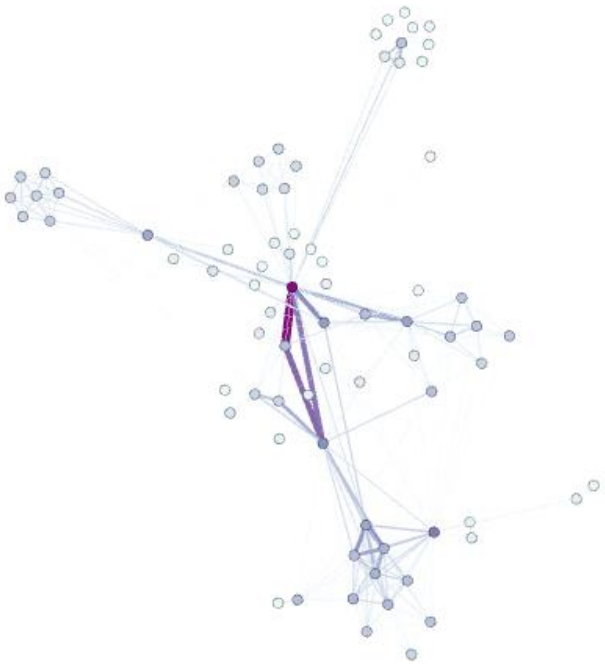


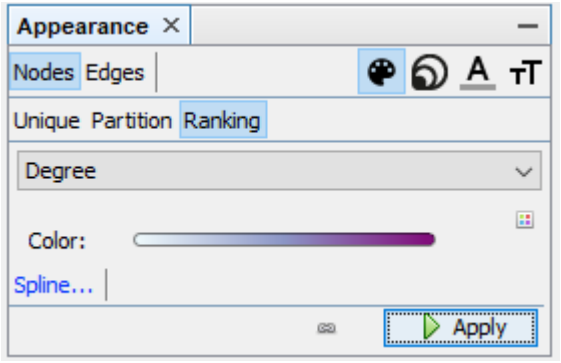


A network consists of entities and their relations. This is what we just visualized. Yet, the properties of these entities remain invisible.
The characters in the novel “Les Misérables” are male or female. Are males more likely to be connected to males, or females? Just looking at the network in Gephi, we can’t tell.

4. Gephi: Ranking (Degree)

Degree = number of connections

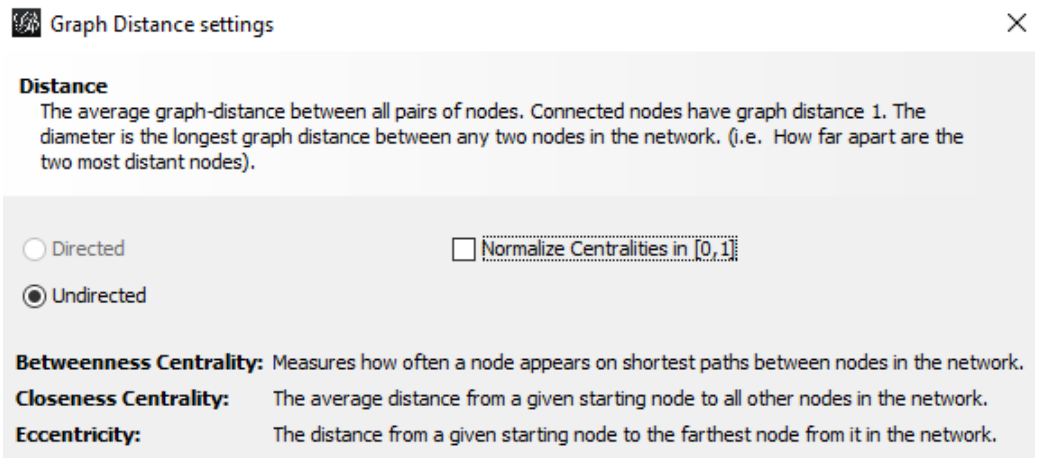




We then changed the appearance of the network's nodes and edges by altering the color and size. The darker color shows the degree of coappearances between any two particular nodes/characters is higher.

5. Gephi: Statistics (Betweenness)

We can calculate the average path length, or length of the edges using the **Statistics** tab.



Graph Distance Report

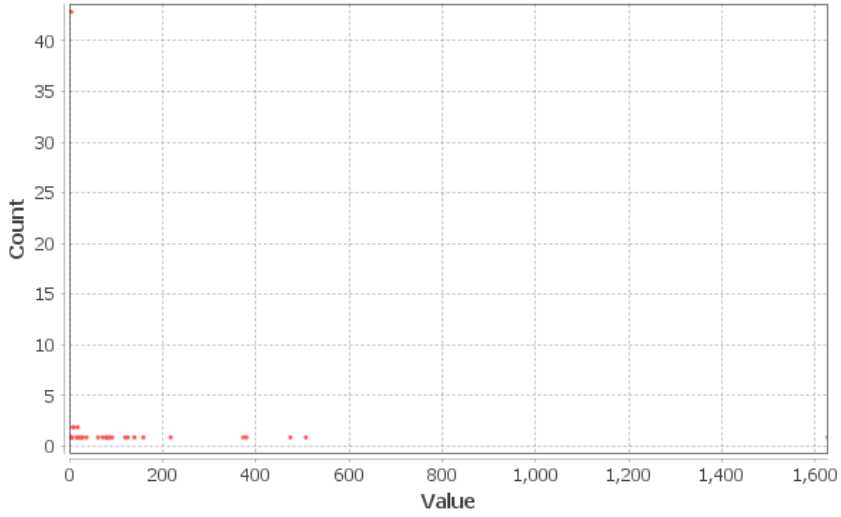
Parameters:

Network Interpretation: undirected

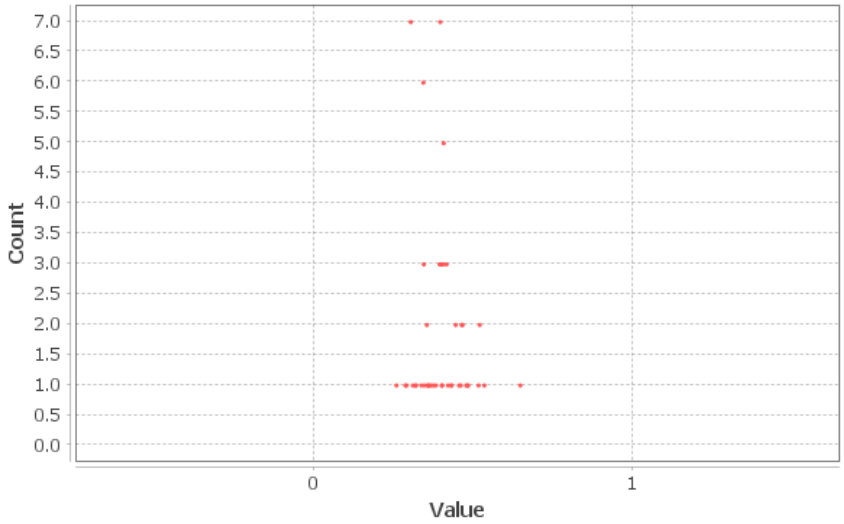
Results:

Diameter: 5
Radius: 3
Average Path length: 2.6411483253588517

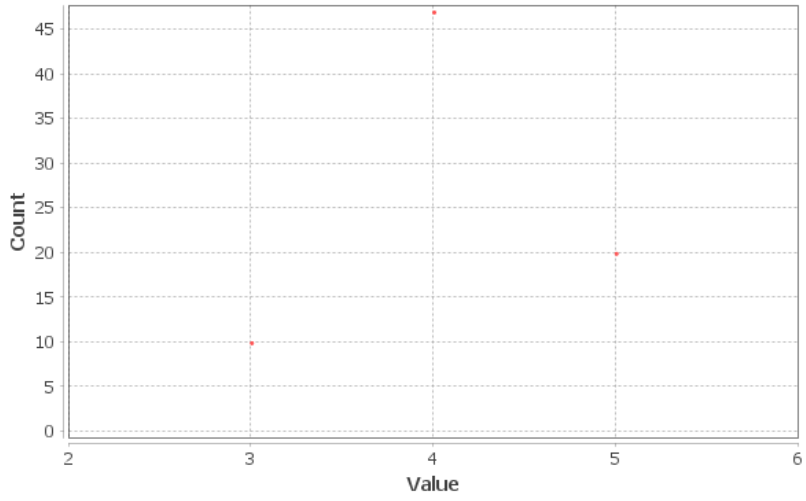
Betweenness Centrality Distribution

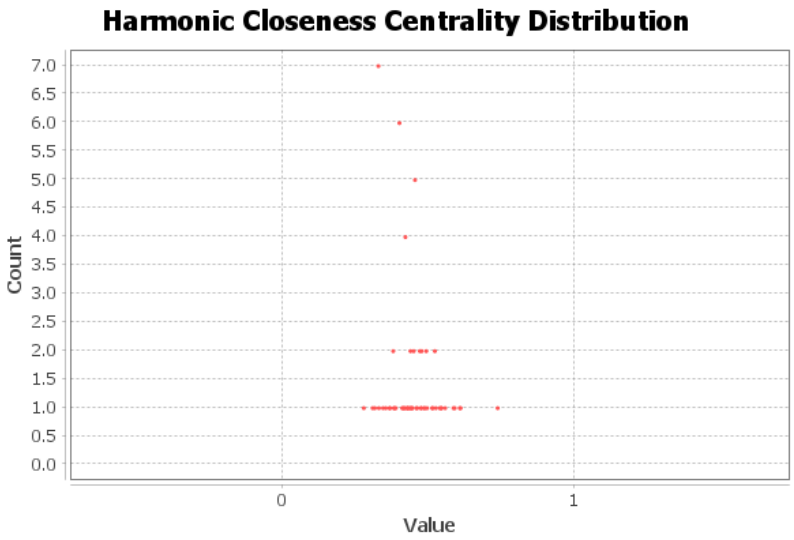


Closeness Centrality Distribution



Eccentricity Distribution

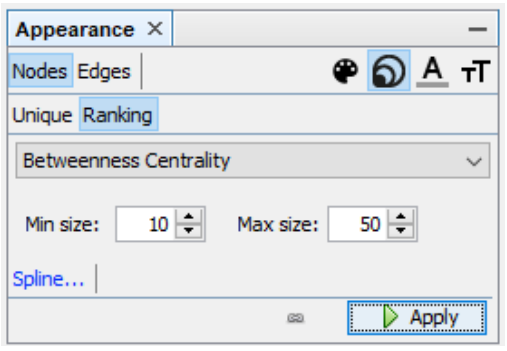


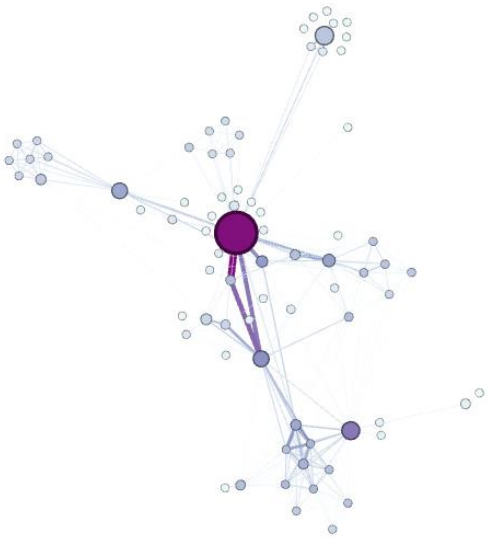


6. Gephi: Rank (Betweenness)

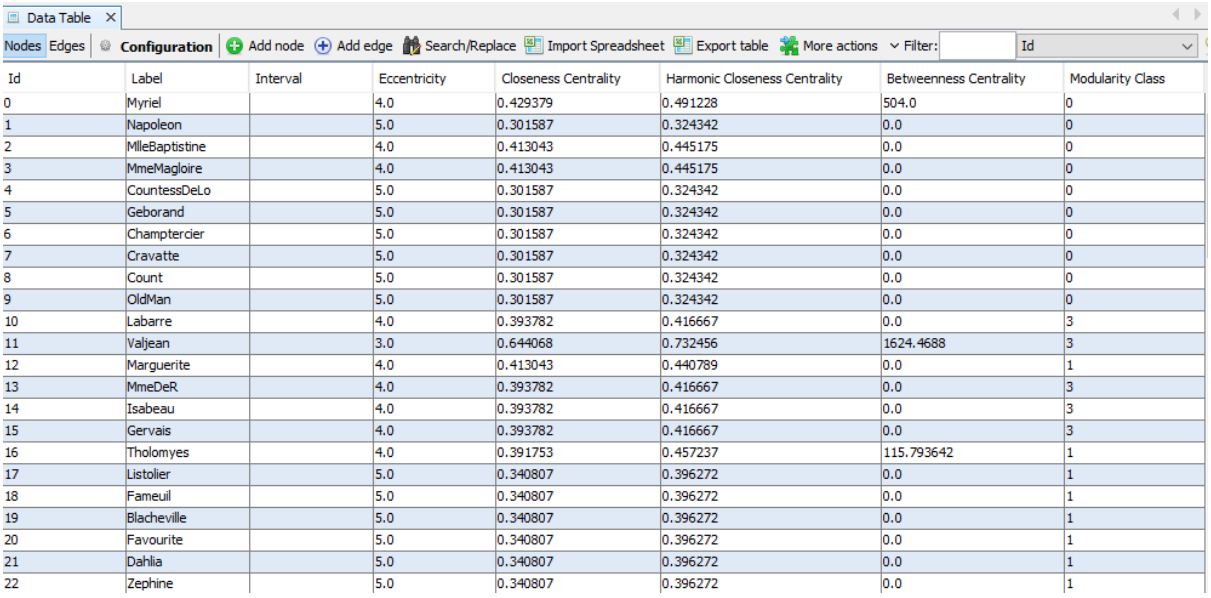
Metrics generates general reports but also results for each node. Thus, three new values have been created by the “Average Path Length” algorithm we ran.

- Betweenness Centrality
- Closeness Centrality
- Eccentricity





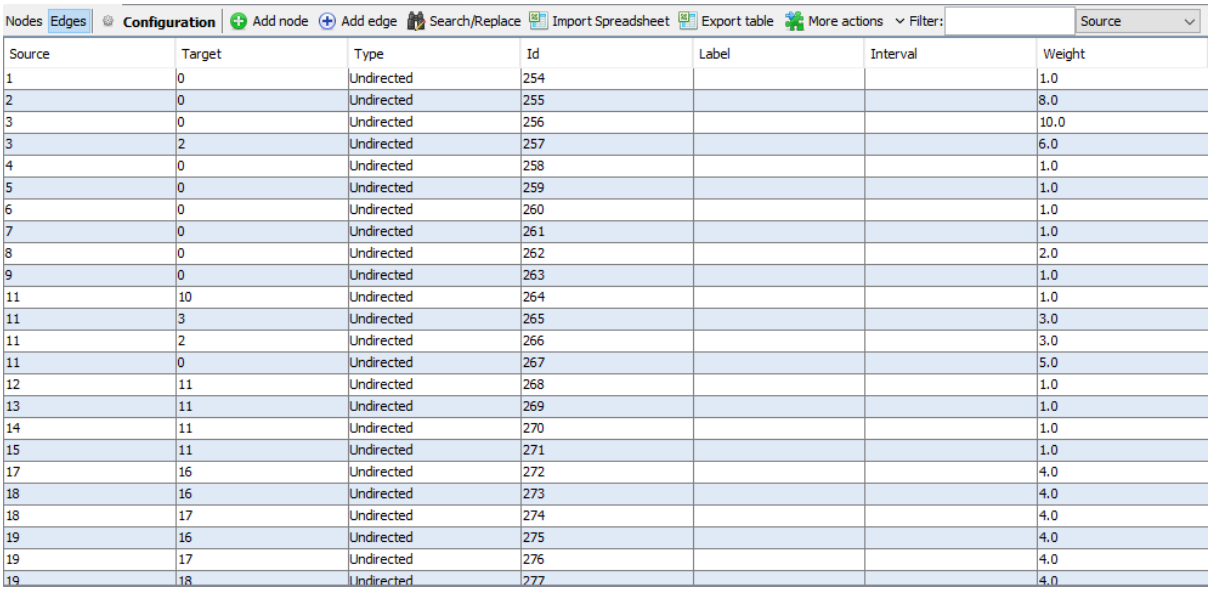
Valjean has the highest betweenness centrality, whereas Myriel comes in second. As a result, while not knowing many characters personally, he is frequently on the shortest path between strangers.



Data Table

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

Id	Label	Interval	Eccentricity	Closeness Centrality	Harmonic Closeness Centrality	Betweenness Centrality	Modularity Class
0	Myriel		4.0	0.429379	0.491228	504.0	0
1	Napoleon		5.0	0.301587	0.324342	0.0	0
2	MlleBaptistine		4.0	0.413043	0.445175	0.0	0
3	MmeMagloire		4.0	0.413043	0.445175	0.0	0
4	CountessDeLo		5.0	0.301587	0.324342	0.0	0
5	Geborand		5.0	0.301587	0.324342	0.0	0
6	Champtercier		5.0	0.301587	0.324342	0.0	0
7	Cravatte		5.0	0.301587	0.324342	0.0	0
8	Count		5.0	0.301587	0.324342	0.0	0
9	OldMan		5.0	0.301587	0.324342	0.0	0
10	Labarre		4.0	0.393782	0.416667	0.0	3
11	Valjean		3.0	0.644068	0.732456	1624.4688	3
12	Marguerite		4.0	0.413043	0.440789	0.0	1
13	MmeDeR		4.0	0.393782	0.416667	0.0	3
14	Isabeau		4.0	0.393782	0.416667	0.0	3
15	Gervais		4.0	0.393782	0.416667	0.0	3
16	Tholomyes		4.0	0.391753	0.457237	115.793642	1
17	Listolier		5.0	0.340807	0.396272	0.0	1
18	Fameuil		5.0	0.340807	0.396272	0.0	1
19	Blacheville		5.0	0.340807	0.396272	0.0	1
20	Favourite		5.0	0.340807	0.396272	0.0	1
21	Dahlia		5.0	0.340807	0.396272	0.0	1
22	Zephine		5.0	0.340807	0.396272	0.0	1

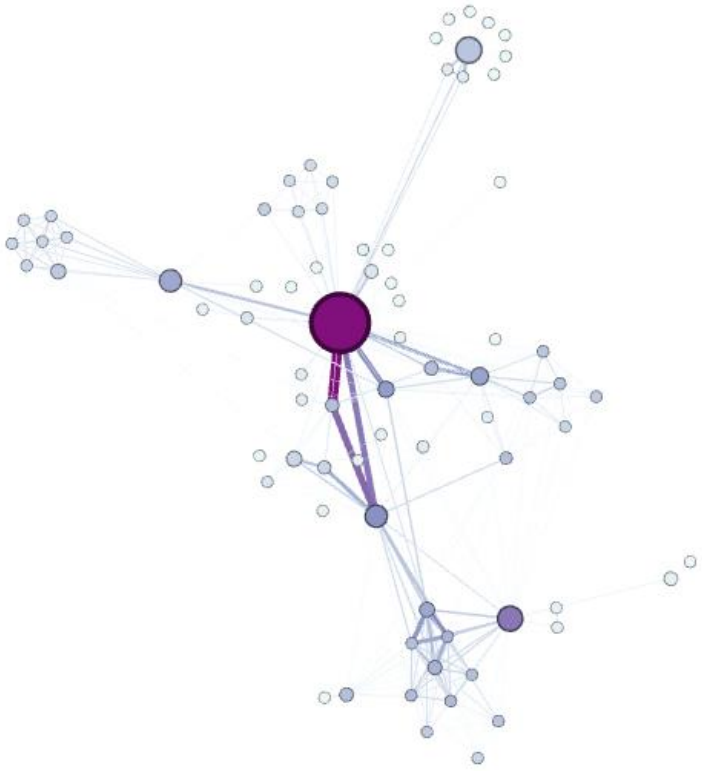


Nodes Edges Configuration Add node Add edge Search/Replace Import Spreadsheet Export table More actions Filter: Source

Source	Target	Type	Id	Label	Interval	Weight
1	0	Undirected	254			1.0
2	0	Undirected	255			8.0
3	0	Undirected	256			10.0
3	2	Undirected	257			6.0
4	0	Undirected	258			1.0
5	0	Undirected	259			1.0
6	0	Undirected	260			1.0
7	0	Undirected	261			1.0
8	0	Undirected	262			2.0
9	0	Undirected	263			1.0
11	10	Undirected	264			1.0
11	3	Undirected	265			3.0
11	2	Undirected	266			3.0
11	0	Undirected	267			5.0
12	11	Undirected	268			1.0
13	11	Undirected	269			1.0
14	11	Undirected	270			1.0
15	11	Undirected	271			1.0
17	16	Undirected	272			4.0
18	16	Undirected	273			4.0
18	17	Undirected	274			4.0
19	16	Undirected	275			4.0
19	17	Undirected	276			4.0
19	18	Undirected	277			4.0

7. Gephi: Layout (Betweenness)

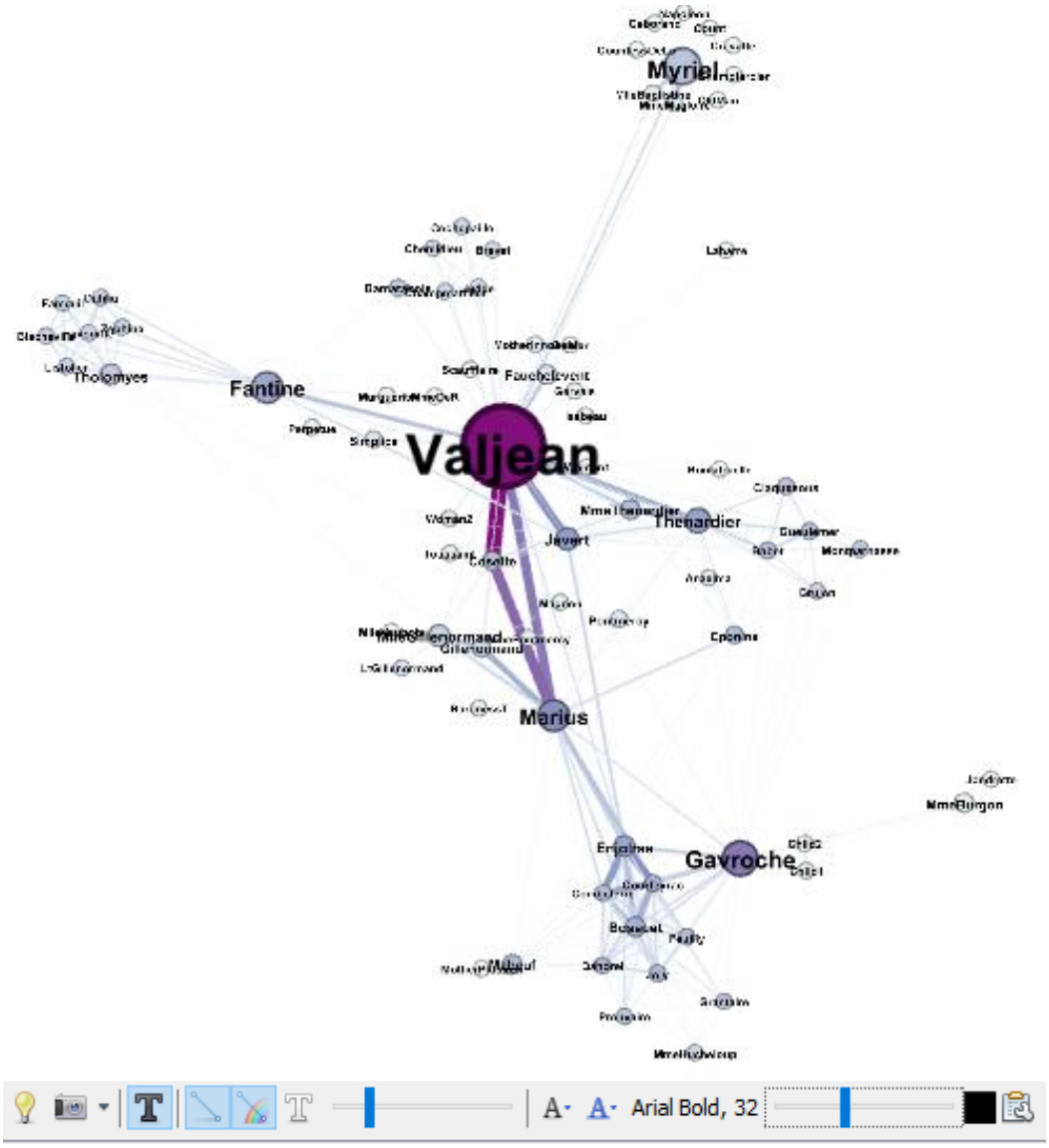
This metric indicates influential nodes for the highest value.



This change was made with this layout to eliminate node overlapping and provide a better look. The nodes are now clearly visible and may be identified. Because the dataset was not much larger, there were only a few changes to notice.

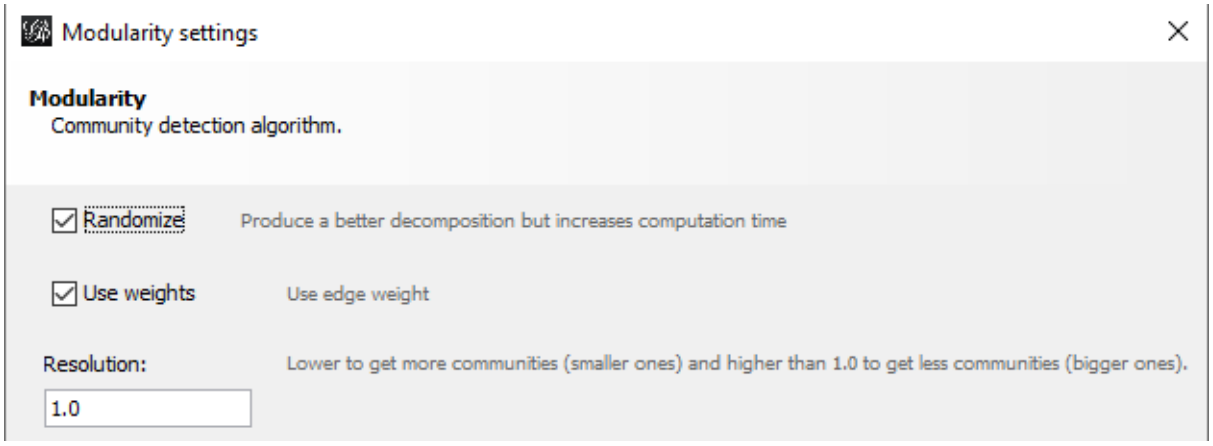
8. Gephi: Labels

To label all the characters with names and using below toolbar the font style and size has been adjusted.



9. Gephi: Community Detection

The ability to detect and study communities is central to network analysis. We would like to colorize clusters in our example. This forms colored clusters to efficiently analyze the communities.



Modularity Report

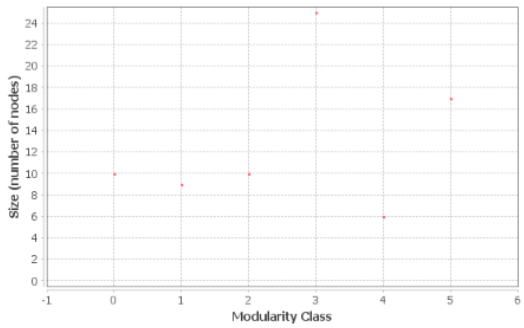
Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

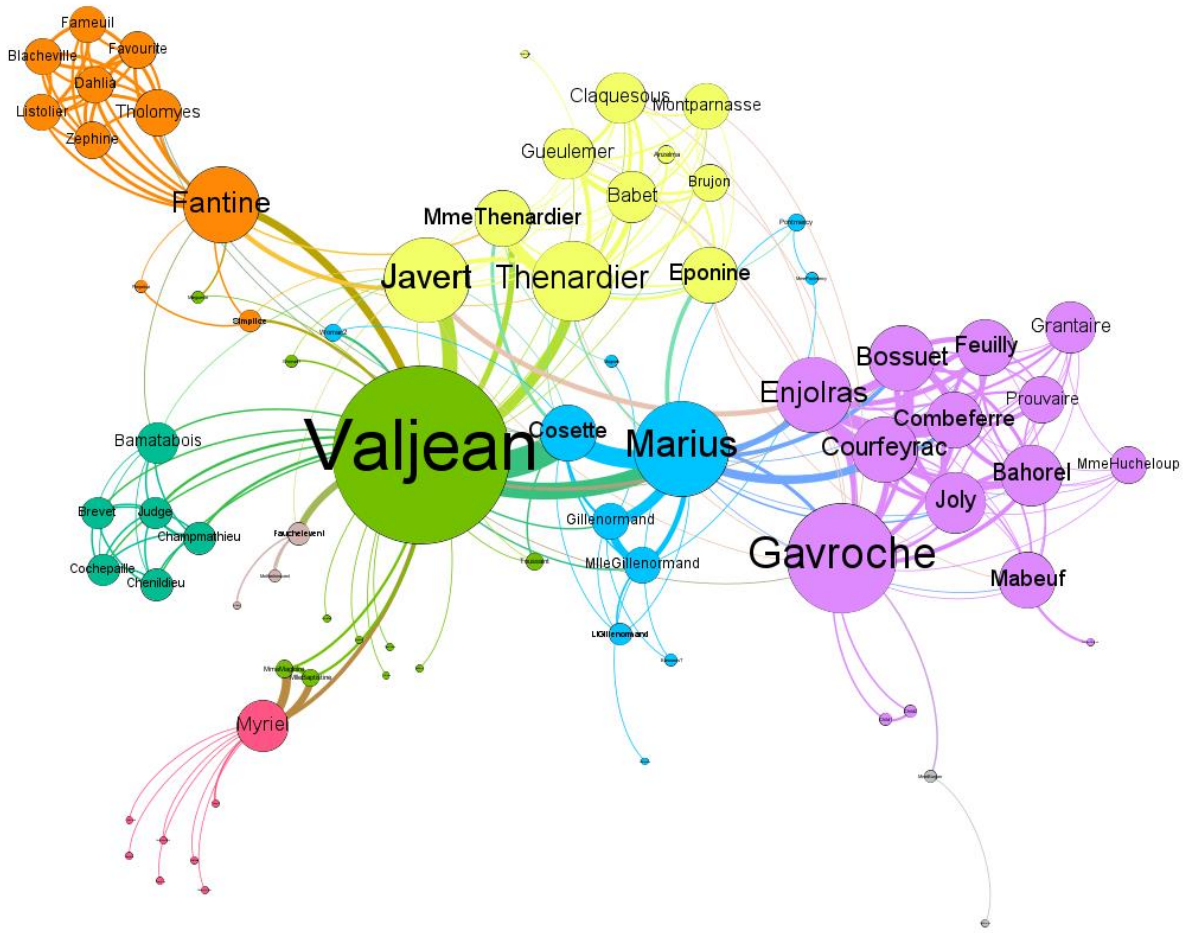
Modularity: 0.565
Modularity with resolution: 0.565
Number of Communities: 6

Size Distribution



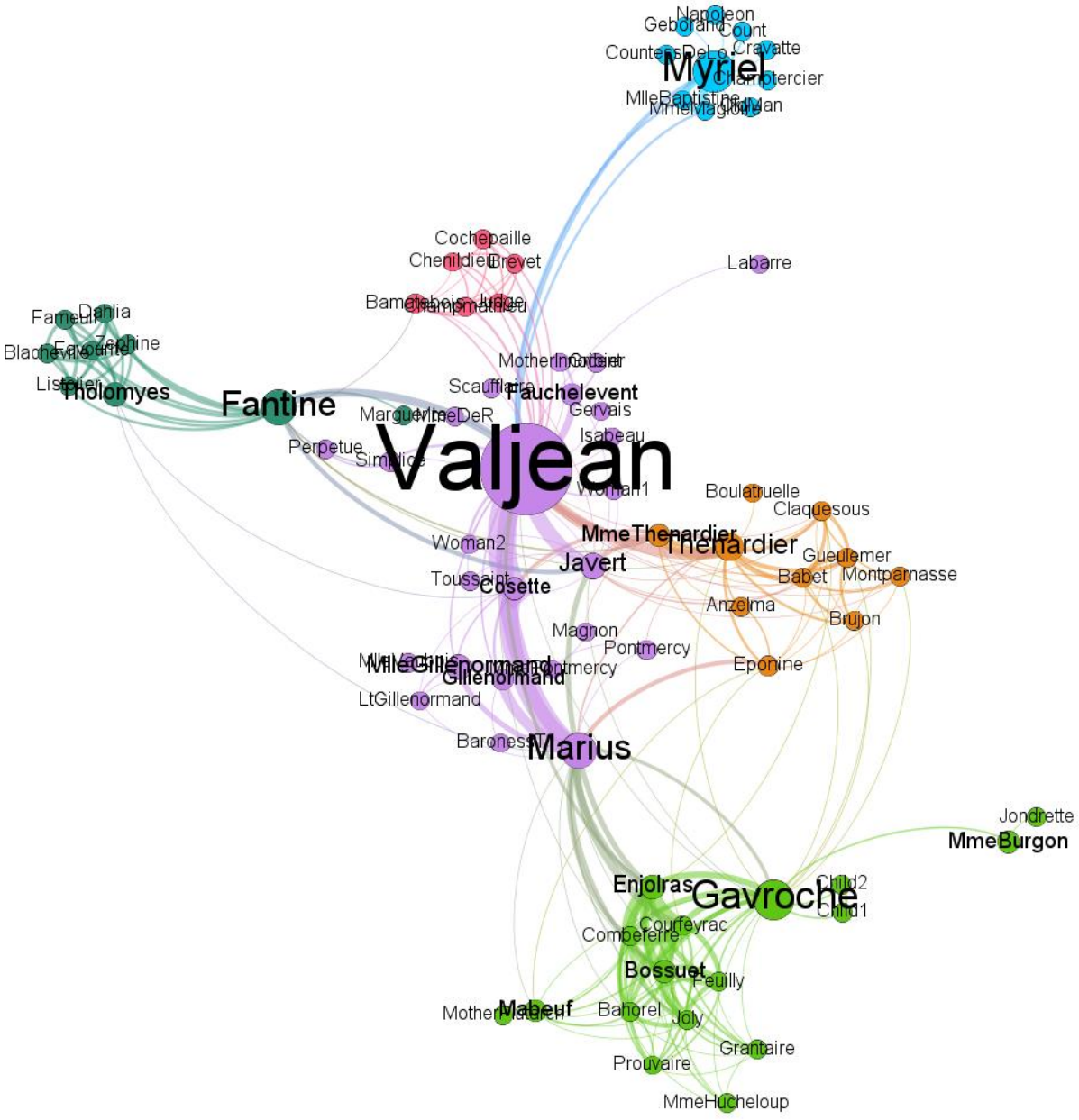
Algorithm:

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, *Fast unfolding of communities in large networks*, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000


Fameuil
Favourite
Blacheville
Dahlia
Listolier
Tholomyes
Zephine
Fantine
Claquesous
Montparnasse
Gueulemer
Babet
Brujon
MmeThenardier
Javert
Thenardier
Eponine
Simplice
Grantaire
Bossuet
Feuilly
Enjolras
Prouvaire
Combeferre
Courfeyrac
Bahorel
MmeHucheloup
Joly
Valjean
Cosette
Marius
Bamatabois
Brevet
Judge
Champmathieu
Cochepaille
Chenildieu
Fauchelevent
Gillenormand
MlleGillenormand
Gavroche
Mabeuf
Myriel

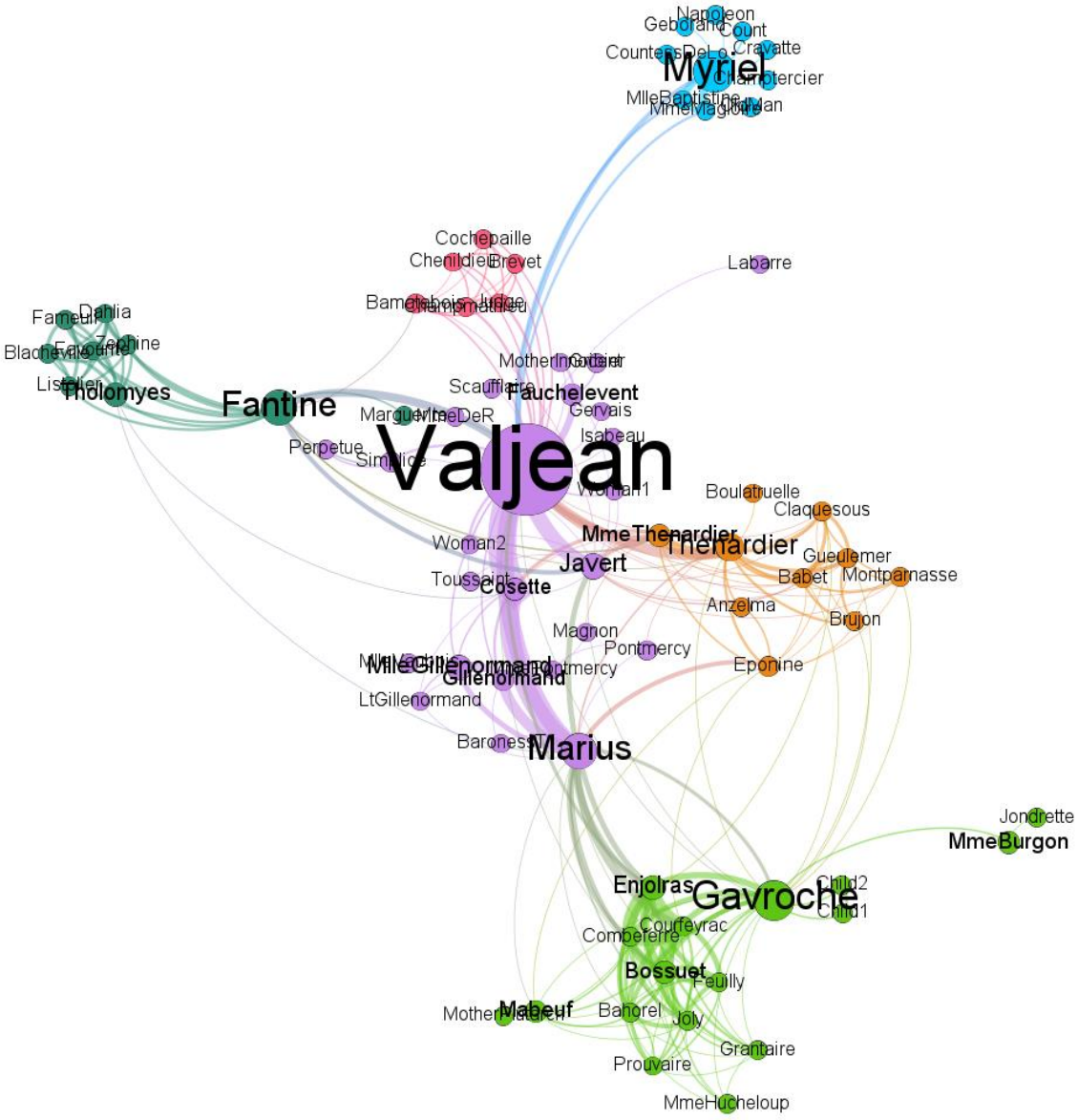
10. Gephi: Filter

To make better visualizations, we use the filter of degree range. This option basically removes the “leaves” in the network that are not connected to many other nodes. As the dataset was not large, we have kept the lower range to 1 only.



11. Gephi: Preview

For the final preview of the graph



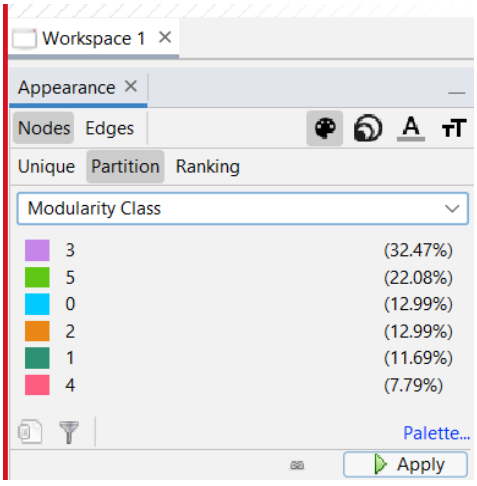
Analysis:

a) Who are important entities from different points of view?

Ans. As seen below, it is clear that Valjean is the key character of *Les Misérables*. Gavroche, Marius, Javert, Thenardier, and Fantine are clearly also central characters overall and within their own modular communities. The first thing we can see is that **Valjean** is the central character; not only because he has the largest degree but because he also has the largest betweenness centrality by a long shot. Valjean also has a comparatively low clustering coefficient, meaning that many of his acquaintances don't know one another Gavroche is the character who has the most influential connections in the novel, even if he does not have the most connections.

b) How many communities exist within the network? Examine the characteristics of each community. Why a community is different from other communities. For this purpose, examine community attributes in the Data Laboratory of gephi.

Ans. It was clear from the clusters and colors of the network to identify 6 groups of characters. In addition, the node size, which was determined according to the node betweenness centrality, allowed for the characters that have the most connectivity to be easily identifiable in comparison to lesser-connected ones.
The community in purple having the key character Valjean is the most spread one. The community in blue has a key character, Myrial is tightly coupled. The community in brown is also loosely coupled. We can observe the distance between the nodes is wide in the green community. The nodes in the community in dark green are coupled but are distant from the main character, Fantine.
The modularity percentage of the purple community that is class 3 is the highest which is 32.47%. The least is of class 4 which is 7.79%.
The darkness of the color of nodes also distinguishes.



c) Examine relationship of nodes within and outside communities

The relationship of the nodes of the blue community shows that they have fewer coappearances outside the community. However, the brown, green and purple community have more coappearances outside their communities, with other communities.

d) Any further insights that you may draw by analyzing the network

The weight (thickness) of the edge between each node also indicates how often co-appearances between characters occur throughout the novel. Therefore, a thicker edge informs us that those characters appear together more often than those with a thinner edge. Several lesser characters also have very high clustering coefficients, so their networks are comparatively smaller and denser than Valjean's.

We were not able to find any attribute like gender on the basis of which we could distinguish.