**DA 516**

**Social Network Analysis**

Project Report

Kunter Kunt

Mehmet Utku Acar

**Introduction**

In this project, we worked on and made network analysis on a data set which contains a novel content from Les Miserables – well known book that was written by French author Victor Hugo. The data is a basically coappearance network of all the characters. The topic and the business were quite interesting and testable for us and that why we decided to pick this one as a input data set. We wanted to reveal the network structure of the overall character interactions.

**Brief Information Regarding The Book**

In order to compare our results with true information, it is good to give short summaries about the main characters and the overall interaction.

Jean Valjean : The primary person in the book. He helps underprivileged people in the society.

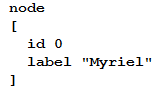
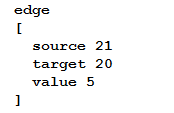
Javert: A police detective.

Cossette: A woman who hangs out and likes Jean Valjean a lot.

Fantine: A worker who labours Valjean’s factory.

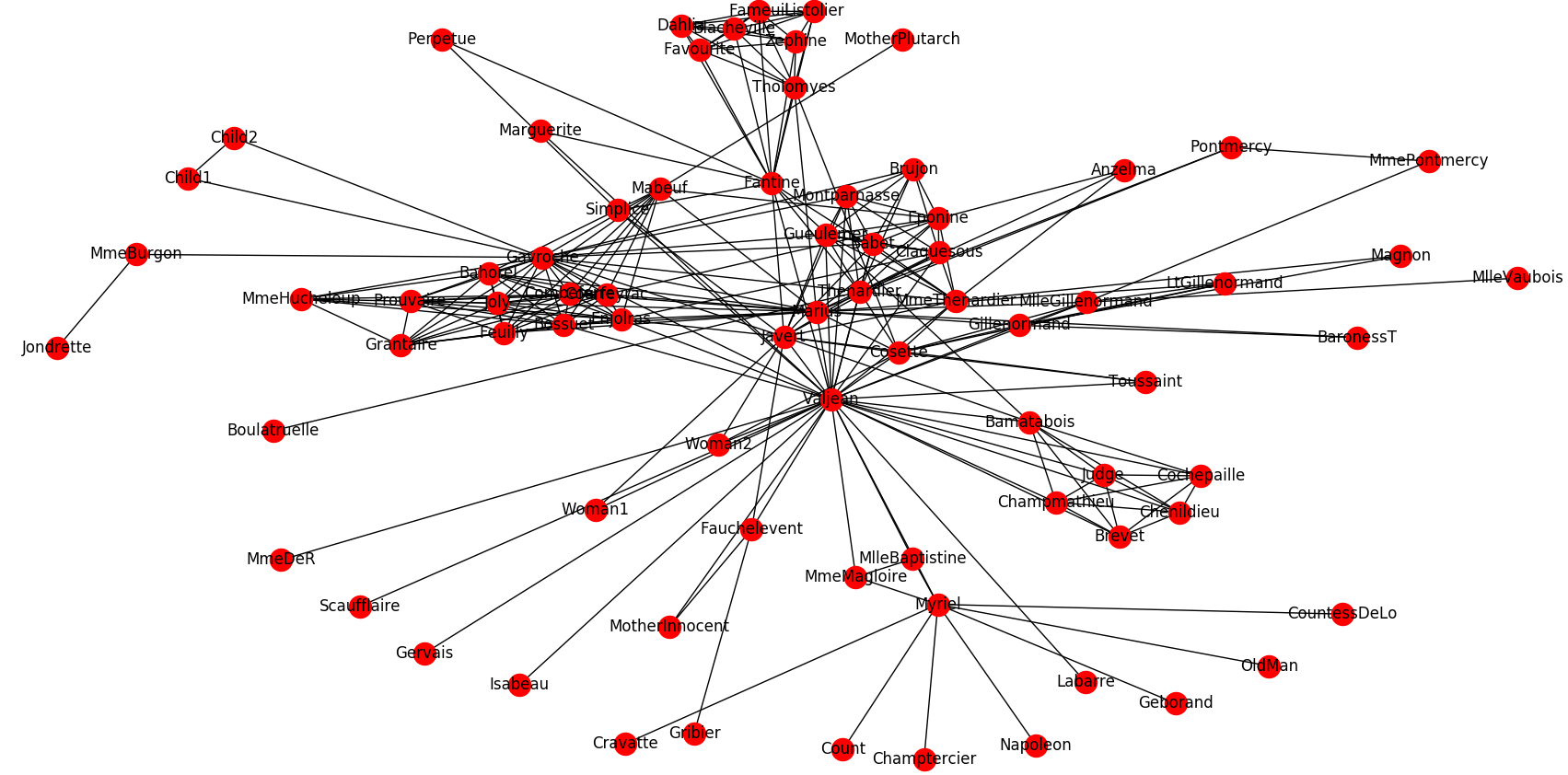
**About The Data Set**

We found and downloaded the data from the following website, <http://www-personal.umich.edu/~mejn/netdata/>. There are lots of different data sets avaliable, they differ by size, domain and other network properties. As we stated above, we selected the second one in the list. The file lesmis.gml contains the undericted & 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. In the below, you can see some samples from the original gml file.

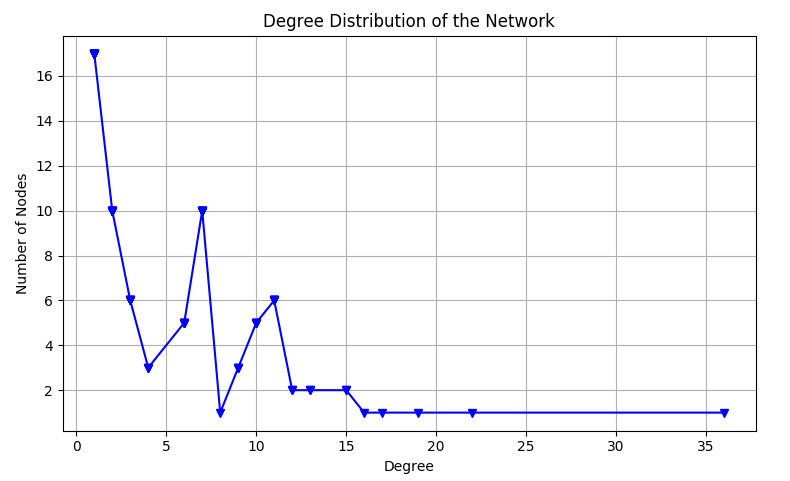
 

There was one major issue that we had to overcome. The format was problematic and therefore we couldn’t read the data. After some research, we learned that it was related with the version of the Networkx library. Either we had to downgrade or change format with our own. The second solution seemed safer than the other and we changed the gml format with a regular expression.

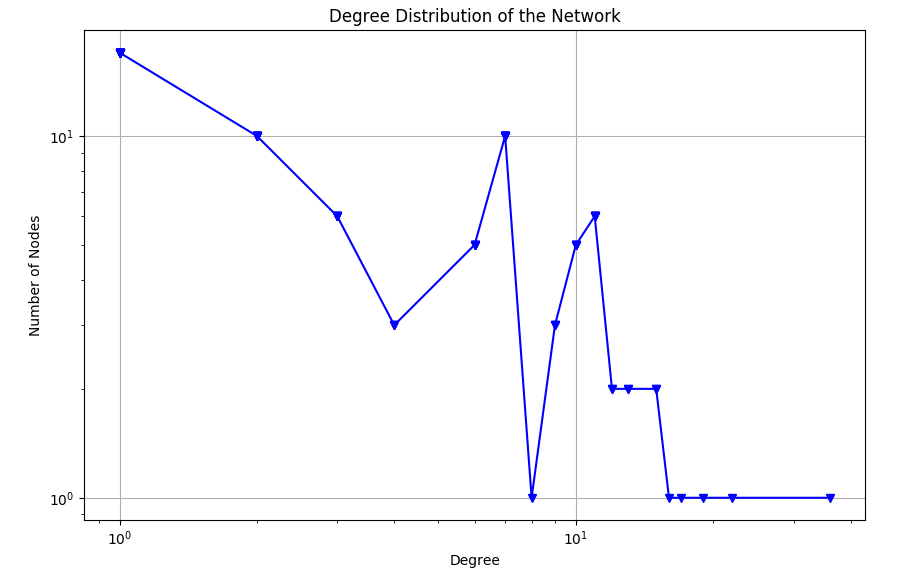
**The Whole Network**

****

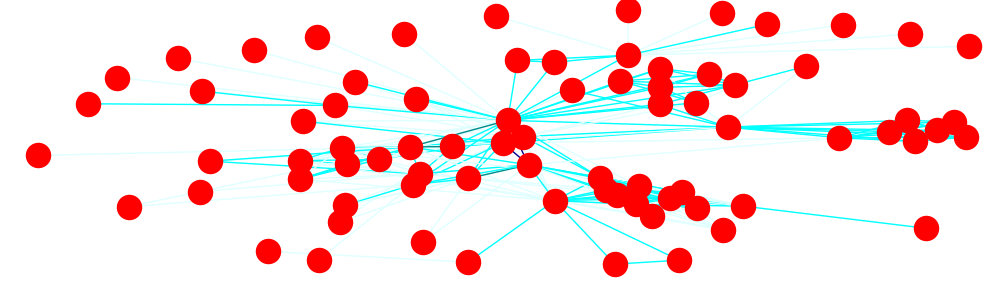
**Structure of the Network and Network Properties**

****

Like common networks, our data set also has lots of nodes whose degrees are low and a few of nodes that their degrees are above the average value.



The log-log plot appeared as expected. We can see our main characters in the right bottom corner as hubs.



The above plot displays the different edge weights. As the color goes darker, the connection between characters increases.

As based on characteristic properties of the data about the novel, it is an undirected weighted graph. Nodes correspond to characters in the novel and weights correspond to coappearances of characters. 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.

|  |  |
| --- | --- |
| Number of Nodes | 77 |
| Number of Edges | 254 |
| Average Clustering Coefficent | 0,573136 |
| Modularity | 0,55295 |
| Diameter | 5 |
| Average Shortest Path Length | 2,64114 |
| Average Node Connectivity | 2,262474 |
| Density of the Network | 0,08680 |

Our network is fully connected and so it is already one connected component so giant component (largest connected component with all 77 nodes in it ) is also itself.Because it is undirected type, no idea about strongly connected type.

As it is known that the main character of the novel is “**Jean Valjean**” and other important characters are “Cosette”, “Javert”, “Fantine”, “Marius Pontmercy”, and “Myriel” generally.Besides ,other characters take part time to time. We have calculated clustering coefficent of each node (character ) sorted and there are 26 characters that have clustering coefficent value of 1 (all neigbours communicate with each other in related chapter ) but our main character Jean Valjean has relatively small clustering coefficent value of 0.12063;from this result, we can obviously infer that characters that appear in the same chapter with Valjean do not appear too much with each other in the same chapter.Due to the result ,we can obviously infer that main character interacts with people separately as expected (most of the main characters separately communicates with people around him/her) inside chapters.

**Popularity of the Characters Based On Centrality Scores**

5 highest degree centrality values and related nodes (characters).

|  |  |
| --- | --- |
| Jean Valjean | 0.47368 |
| Gavroche | 0.28947 |
| Marius Pontmercy | 0.25000 |
| Javert | 0.22368 |
| Thenardier | 0.21052 |

We have calculated sorted betweenness centrality of the characters( nodes) and below is the highest 7 betweenness centrality values and related characters.

|  |  |
| --- | --- |
| Jean Valjean | 0.56998 |
| Myriel | 0.17684 |
| Gavroche | 0.165112 |
| Marius Pontmercy | 0.132032 |
| Fantine | 0.129644 |
| Thenardier | 0.074901 |
| Javert | 0.054331 |

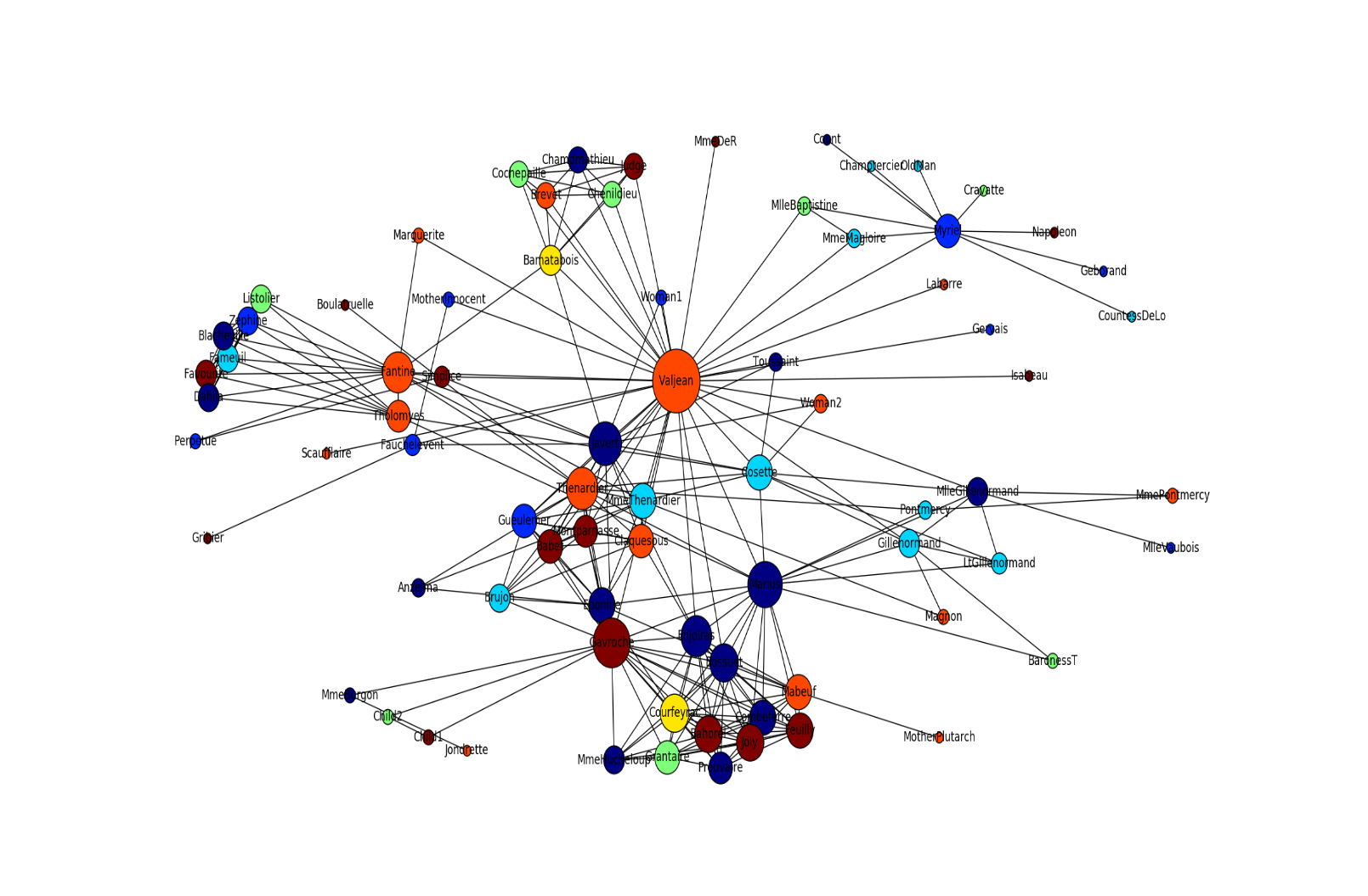
Below is the top 5 closeness centrality of the nodes

|  |  |
| --- | --- |
| Jean Valjean | 0.644067 |
| Marius Pontmercy | 0.53146 |
| Javert | 0.51700 |
| Thenardier | 0.51700 |
| Gavroche | 0.51351 |

It can be obviously seen from the results that Jean Valjean is the most popular character.He is the highest scored characted based on all centrality measures but the most important metric is betweenness centrality due to the conditions of that network. Jean Valjean’s score is much more away from other characters’ scores (it is 0.56 the the closest one to it is 0.17).Due to betweenness centrality measures, Myriel,Gvroche ,Pontmerct and Fantine have relatively important scores and they provide communication between other characters.Based on closeness centrality measures, again Jean Valjean is the highest one but we can say that Pontmercy,Javert,Thenardier and Gavroche(each score is above 0.5) are close enough to the center of the network( main reason of that is their direct connection with main character Jean Valjean).

**Communities**

Due to the results of the networkx community detection algorithm, there are 7 different communities. Below is the visualization of 7 different communities with each colored differently.



The members of each communities

**Community 1 members (16 characters):**

'Grantaire', 'Child1', 'Child2', 'Enjolras', 'Feuilly', 'Courfeyrac', 'Bossuet', 'Mabeuf', 'Gavroche', 'Bahorel', 'MmeHucheloup', 'Joly', 'Combeferre', 'Prouvaire', 'MotherPlutarch', 'Marius

**Community 2 members (10 characters):**

'MlleBaptistine', 'Champtercier', 'Geborand', 'Count', 'CountessDeLo', 'OldMan', 'MmeMagloire', 'Napoleon', 'Myriel', 'Cravatte'

**Community 3 members(11 characters ):**

'Perpetue', 'Blacheville', 'Tholomyes', 'Dahlia', 'Listolier', 'Marguerite', 'Fantine', 'Favourite', 'Zephine', 'Fameuil', 'Simplice'

**Community 4 members( 8 characters ):**

'MlleVaubois', 'Pontmercy', 'Magnon', 'BaronessT' , 'LtGillenormand', 'MmePontmercy', 'Gillenormand'

**Community 5 members(only 2 characters ):**

'Jondrette', 'MmeBurgon'

**Community 6 members(15 characters ):**

'Javert', 'Toussaint', 'Eponine', 'Anzelma', 'Cosette', 'Gueulemer', 'Babet', 'MmeThenardier', 'Woman2', 'Montparnasse', 'Brujon', 'Woman1', 'Thenardier', 'Claquesous', 'Boulatruelle'

**Community 7 members(15 characters and Valjean belong to this community ):**

'Chenildieu', 'Labarre', 'Judge', 'Brevet', 'MmeDeR', 'Valjean', Bamatabois', 'Scaufflaire', 'Gribier', 'MotherInnocent', 'Isabeau', 'Champmathieu', 'Fauchelevent', 'Gervais', 'Cochepaille'

There may be overlapping communities and we have used clique percolation method of networkx to detect overlapping communities and we have observed union of all cliques of sized 4,5,6 that can be reached through adjacent 3,4, and 5 cliques (k-1).

**Conclusion**

Each novel owns a social network that the author creates. Transforming hundreds of pages to characters network is both informative and attractive. In this project, we covered the general attributes of the network and worked on communities. As we proceeded, we encounter that the outputs are match with reality in most cases such as the main characters, the relationships between people in the novel.