Clustering Analysis on Les Misérables Social Network

DUNDA, Gerry Windiarto Mohamad 20491372

Abstract

This paper employs a comprehensive clustering analysis approach, utilizing Fiedler, normalized Ncut, transition path theory, modified Ncut, and Louvain community detector to delve into the Les Misérables social network. This network captures the co-occurrence relationships between characters in Victor Hugo's renowned novel, offering a captivating backdrop to explore the intricate web of interactions. By leveraging these advanced clustering techniques, we partition the Les Misérables network into distinct groups, unveiling the applicability of each algorithm and its interpretability.

1 Introduction

Victor Hugo's masterpiece, "Les Misérables," has captivated readers for generations with its intricate narrative and richly woven web of characters. The interactions and relationships between these characters form a complex social network that underlies the novel's compelling storyline. Exploring this social network through clustering analysis provides a unique lens through which to unravel the underlying patterns and dynamics within the narrative.

In recent years, network analysis techniques have gained significant traction in various fields, including social sciences, biology, and information technology. By applying clustering algorithms, we can identify cohesive groups of characters within the Les Misérables social network, revealing insights into the intricate social dynamics and thematic connections that shape the story.

In this paper, we employ a range of clustering methods to analyze the Les Misérables social network. Our methodology encompasses Fiedler, normalized ncut, transition path theory, modified Ncut, and Louvain community detector to identify two or more communities based on their co-appearance. Through the application of these clustering techniques, we aim to uncover hidden structural patterns and thematic dynamics within Les Misérables. By identifying distinct groups of characters, we can gain a deeper understanding of their roles, interactions in the narrative without reading the novel.

2 Dataset and Methodology

Dataset. The Les Misérables dataset is available at *https://github.com/yao-lab/yao-lab.github.io/blob/master/data/lesmis.mat*. Alternatively, the dataset can be accessed from *networkx* python package via les_miserables_graph() API. The network is a weighted directed graph where the weight counts how many co-appearances between two characters. The short description of the novel is stated as follows. The ex-convict, Jean Valjean, is the main character of the novel. He spends a great deal of time in the novel running away from Inspector Javert. He is also closely tied to his adopted daughter, Cosette, and her future husband, Marius. If one were interested in influencing Jean Valjean, once could target Cosette or Marius if Valjean cannot be directly targeted.

Bipartition technique based on Fiedler. Given a weighted undirected graph $G = (V, E, W)$, we represent all connections in vertices V as E encoded in the adjacency matrix A, where $A_{i,j} = w_{i,j}$ when there is a connection between i and j with weight $w_{i,j}$ and $A_{i,j} = 0$ otherwise. Define $D = \mathbf{diag}(d_i)$ where d_i is the degree of node $i \in V$. We can write the (unnormalized) Laplacian

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

matrix $L = D - A$. The bipartitioning is realized thorugh Fiedler theorem. Specifically, the produced subgraphs are $\mathcal{N}_+ := \{i : v_1(i) > 0\}$ and $\mathcal{N}_- := \{i : v_1(i) < 0\}$, where v_1 is the eigenvector of L with the second smallest eigenvalue.

Normalized Ncut. We describe the method from [3] to bipartition a graph. We consider the normalized Laplacian matrix as $\mathcal{L} = D^{-1/2}LD^{1/2}$. After that, we consider the eigenvector of $\mathcal{L}$ with the second smallest eigenvalue. In other words, we do the similar technique as Fiedler except we consider the normalized version.

Transition path theory. Firstly, we define the Markov transition matrix $P = D^{-1}A$. Secondly, we find the stationary distribution π such that $\pi P = P$. We need to define the source and sink set as V_0 and V_1 respectively. We define the committor function as $q_i = \mathbb{P}[\text{trajectory starting from } i \in V \text{ hit } V_1 \text{ before } V_0]$. Then, we aim to find $q = [q_1, \ldots, q_{|V|}]^T$ subjected to Dirichlet boundary condition, that is $Lq_i = 0, \forall i \in V - V_0 - V_1$, $q_j = 0, \forall i \in V_0$, and $q_j = 1, \forall i \in V_1$. After we obtain q, we can cluster the nodes based on the threshold which is 0.5. We can compute the effective flux on each edge as $J^+(x, y) = \max(J(x, y) - J(y, x), 0)$, where $J(x, y) = \pi(x)(1 - q(x))P(x, y)q(y)$ when $x \neq y$ and $J(x, y) = 0$ when $x = y$. This quantity is meaningful if one is interested in the current of trajectories making their way from a set of states V_0 to set of states of V_1.

Modified Ncut [2]. This is a multiple clustering method. Firstly, select k largest eigenvalues with their corresponding eigenvectors of P ($\forall i, Px_i = \lambda_i x_i$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{|V|}$). After that, we can perform k-means clustering by considering $k - 1$-dimensional vector space by the row vector of the matrix $[x_2, \ldots, x_k]$.

Louvain community detection [1]. It tries to find communities such that it can maximize the density of the links inside each community compared to links between communities. Specifically, it considers the modularity function Q_c of a community c, defined as

$$Q_c = \frac{\Sigma_{in}}{2m} - (\frac{\Sigma_{tot}}{2m})^2$$

where Σ_{in} is the sum of edge weights between nodes within the community c, Σ_{tot} is the sum of all edge weights for nodes within the community, and m is is the sum of all of the edge weights in the graph. The basic idea of the algorithm is to perform greedy algorithm to increase the objective. The initial condition is all nodes are considered as separate community. The algorithm then performs heuristic iteration by removing node i and pick its neighbor j, resulting in the change of modularity ΔQ, written as

$$\Delta Q = [\frac{\Sigma_{in} + 2k_{i,in}}{2m} - (\frac{\Sigma_{tot}^2 + k_i}{2m})^2] - [\frac{\Sigma_{in}}{2m} - (\frac{\Sigma_{tot}}{2m})^2 - (\frac{k_i}{2m})^2],$$

where w_i is weighted degree of i and $k_{i,in}$ is the sum of the weights of the links between i and other nodes in the community that i is moving into. This process is repeated sequentially for each node until no further increase in ΔQ is possible.

3 Results

We present the result of using bipartition technique based on Fiedler, as shown in Figure 1. The Fiedler value of this dataset is 0.554. It shows that the cluster is somehow centered around the main characters, Valjean and Javert. Even though Marius and Cosette should be closely related to Valjean, the method assigns them to Javert's cluster.

The bipartitioning result using normalized Ncut algorithm is shown in Figure 2. The algorithm gives less interpretable result compared to Fiedler as Javert and Valjean are located on the same cluster.

The bipartitioning result using transition path theory is illustrated in Figure 3. The edge is directed as we build it from the effective flux, and the thickness quantifies the magnitude of the flux. Moreover, the size of the node quantifies the transition flux. We set Valjean as a sink node and Javert as a source node.

The result seems more intuitive compared to previous two methods. Firstly, we discover a large flux from Cosette and Marius to Valjean node. This corresponds to them being close relatives to Valjean. Secondly, both Valjean and Javert point towards Enjolras and Thénardier, suggesting intermediary role in the story. Indeed, Javert and Valjean appears with Enjolras in the barricade scene. Also,

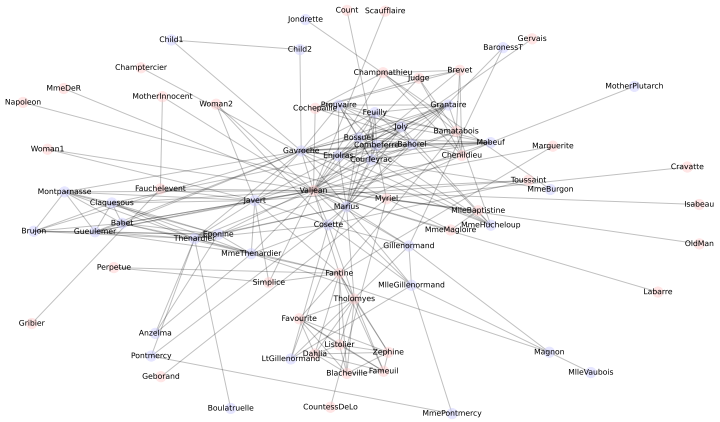

Figure 1: Bipartition result based on Fiedler.

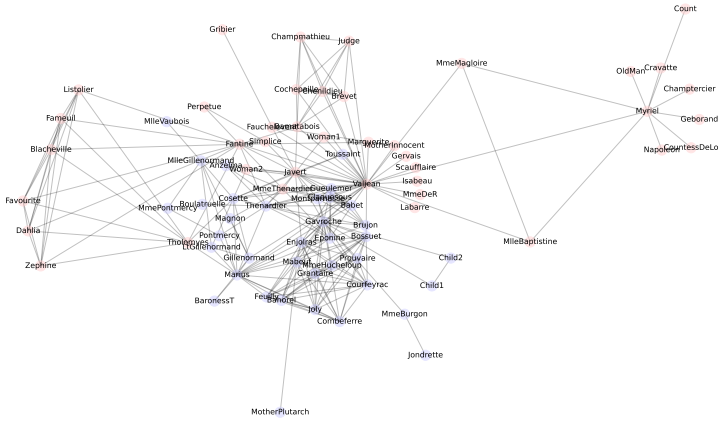

Figure 2: Bipartition result using normalized Ncut algorithm.

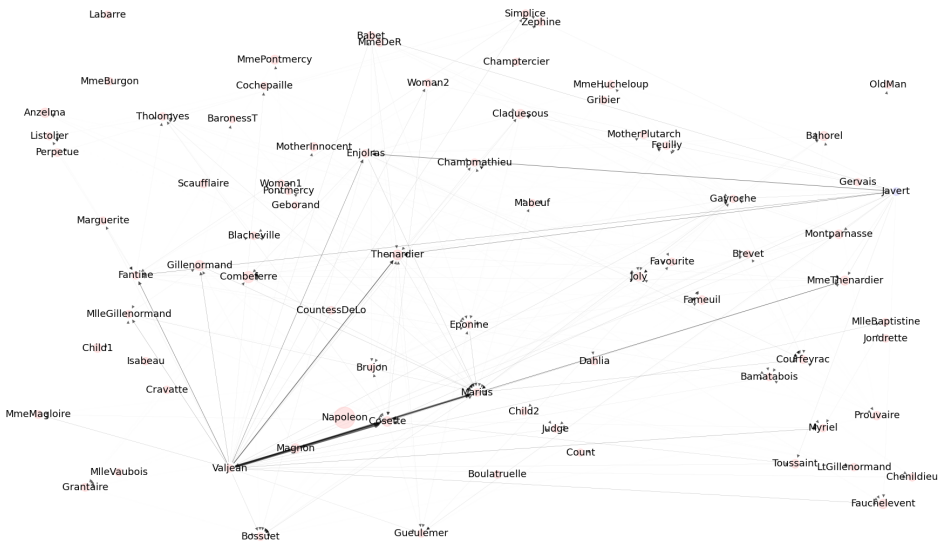

Figure 3: Bipartition result using transition path theory.

Thénardier often makes deals with Javert and Valjean. Thirdly, Napoleon has the largest transition flux to bridge many characters in the novel. This observation is reflected on him being an emperor of France. These observations alone may highlight several important characters, which lead us to their interesting roles in the story.

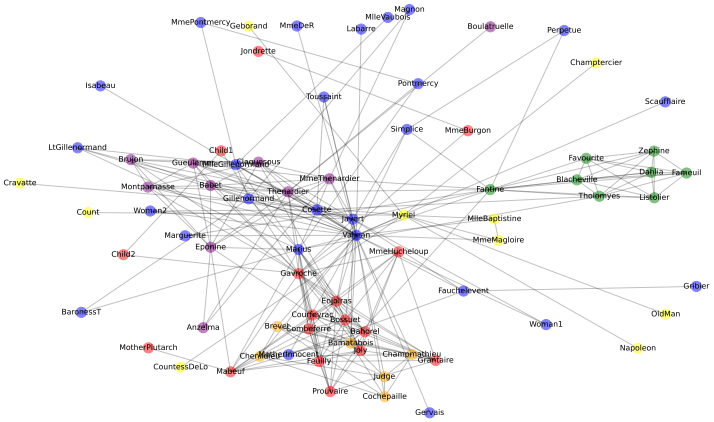

Figure 4: Bipartition result using modified Ncut algorithm.

The result for multiple clustering method using modified Ncut is shown in Figure 4. We choose $k = 6$. The blue cluster captures the community formed by Valjean and Javert. We can see that some relatives of Valjean are included in the community. We also observe that the red community is relatively large. Because Gavroche good friends are Enjolras, Grantaire, Joly, Combeferre, and Courfeyrac, they are

classified in the same community by the algorithm. We can see other community such as green where Zephine, Dahlia, and Favourite are categorized in the same community. This corresponds to the story where they are Grisette and members of Fantine's group. Some side characters are clustered into the same community as observed in yellow. The orange community is the smallest, and it captures the relationship of prisoners between Champmathieu, Brevet, Cochepaille and Chenildieu.

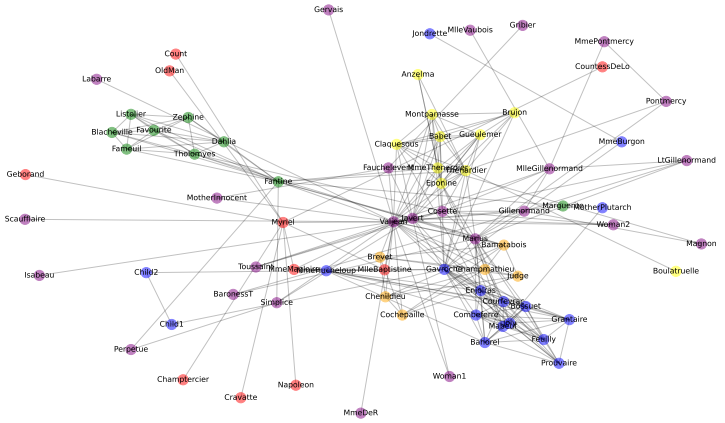

Figure 5: Bipartition result using modified Ncut algorithm.

The result for multiple clustering method using Louvain community detection is shown in Figure 5. Interestingly, this clustering method yields similar result as modified Ncut except one person is assigned differently, Marguerite. The person is assigned to Valjean group in the modified Ncut result, whereas in the Louvain result she is assigned to Fantine group. In fact, Marguerite is closer to Fantine as Marguerite gave Fantine lessons on poverty.

4 Conclusion

We explored Les Misérables dataset using some graph clustering techniques, namely Fiedler, normalized Ncut, transition path theory, modified Ncut, and Louvain community detection. The results of the exploration of different algorithms have different appeals. Using bipartition method did not give significant insights. In contrast, multi clustering method gives better analysis with their own strength. The result using transition path theory unvravels some important intermediary characters in the story. The multi-clustering methods give a meaningful clustering result.

References

[1] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: https://doi.org/10.1088/1742-5468/2008/10/p10008.

[2] Marina Meilă and Jianbo Shi. "A Random Walks View of Spectral Segmentation". In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*. Ed. by Thomas S. Richardson and Tommi S. Jaakkola. Vol. R3. Proceedings of Machine Learning Research. Reissued by PMLR on 31 March 2021. PMLR, Apr. 2001, pp. 203–208. URL: https://proceedings.mlr.press/r3/meila01a.html.

[3] Jianbo Shi and J. Malik. “Normalized cuts and image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905. DOI: 10.1109/34.868688. URL: https://doi.org/10.1109/34.868688.