CSIC5011 Project2 Report
Spectral Clustering and Transition Paths Analysis of Karate Club Network

Wei Cheng
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology
wchengad@connect.ust.hk

Abstract

In this projects, spectral clustering and transition path analysis are studied on a typical Zachery's Karate Club Network. The Cheeger vector is applied to bipartite the network into two components with spectral clustering, which is alternatively achieved by solving Laplacian equation with Dirichelet boundary conditions on a *committor* function in transition path analysis. Then relative importance of the nodes in bridging communities are shown by effective and transition flux. Finally, one social network and protein binding transition network are investigated following the same clue.

1 Introduction

Networks have widespread its application for describing structures in a variety of fields in the past fifteen years, including social sciences, biology, economics and engineering. One of the simplest mathematical structure to study a network is the graph structure, which introduces the notations such as degrees, paths, connectivities, etc. Researchers can also endow other network structures, for example geometry structures in manifold learning theory, and topological structures in persistent homology.

Given a graph with a starting point, a random walk is a finite Markov chain that is time-reversible. In fact, there is not much difference between the theory of random walks on graphs and the theory of finite Markov chains; every Markov chain can be viewed as random walk on a directed graph, if we allow weighted edges. Similarly, time-reversible Markov chains can be viewed as random walks on undirected graphs, and symmetric Markov chains, as random walks on regular symmetric graphs. In this paper we'll formulate the results in terms of random walks, and mostly restrict our attention to the undirected case.

General theory of random walks of finite Markov chains on graphs related to data analysis is well studied. Perron-Frobenius theroy for non-negative matrices leads to the characterization of non-negative primary eigenvectors, such as stationary distributions of Markov chains; application examples include Google's PageRank. Fiedler theroy and Cheeger's Inequality give the characteristic which is about the second eigenvector in a Markov chain, mostly reduced from graph Laplacians, which is the basis for spectral clustering. The transition path theory which was originally introduced in the context of continuous time Markov process on continuous state space and discrete state space, is adapted to the setting of discrete time Markov chain with transition probability matrix.

In this projects, following the project instructions, spectral clustering and transition paths analysis of random walk on a network graph structure are investigated. The following parts is organized as, methodology of spectral clustering and transition paths analysis are described in Sec. 2, detailed information analysis on Karate Club network is discussed in Sec. 3, and finally similar evaluation on social network of Les Misérables and LAO protein binding transition network are performed using the same clue in Sec. 4.

2 Methodology

In this section the methodology used in this project are discussed. We first review the Fiedler theory and Cheeger vector's application in spectral clustering to bipartite the network. Then we consider the row Markov matrix and calculate its Perron eigenvector, which represents the stationary distribution of the network. Finally, the effective and transition flux are computed depending on a defined *committor* function to analysis the relative importance of the node bridging the communities.

2.1 Fiedler Theory and Spectral Clustering

Let $G = (V, E)$ be an undirected, unweighted simple 1 graph. Although the edges here are unweighted, the theory below still holds when weight is added. We can get a similar conclusion with the weighted adjacency matrix. However the extension to directed graphs will lead to different pictures. We use $i \sim j$ to denote that node $i \in V$ is a neighbor of node $j \in V$. So the adjacency matrix and the diagonal Matrix $D = diag(d_i)$ are defined as

$$\begin{aligned} A_{ij} &= \begin{cases} 1, & i \sim j, \\ 0, & \text{otherwise} \end{cases} \\ d_i &= \sum_{j=1}^{n} A_{ij} \end{aligned} \tag{1}$$

Then the graph Laplacian L and normalized graph Laplacian $\mathcal{L}$ can be defined as

$$\begin{aligned} L &= D - A \\ \mathcal{L} &= D^{-1/2}(D - A)D^{-1/2} \end{aligned} \tag{2}$$

So the Fiedler theory tells that for graph Laplacian L or normalized graph Laplacian $\mathcal{L}$, here we take $\mathcal{L}$ as an example. Let $\mathcal{L}$ has n eigenvectors

$$\mathcal{L}v_i = \lambda_i v_i,\ v_i \neq 0,\ i = 0, \dots, n-1 \tag{3}$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. For the second smallest eigenvector v_1 , define

$$\begin{aligned} V_0 &= \{i : v_1(i) < 0\}, \\ V_1 &= \{i : v_1(i) > 0\}, \\ V_u &= V - V_0 - V_1. \end{aligned} \tag{4}$$

We have the following results.

- $\#\{i, \lambda_i = 0\} = \#\{connected\ components\ of\ G\}$;
- If G is connected, then both V_0 and V_1 are connected. $V_0 \cup V_u$ and $V_1 \cup V_u$ might be disconnected if $V_u \neq \varnothing$.

where the second smallest eigenvector can be used to bipartite the graph into two connected components by taking V_0 and V_1 when $V_u = \varnothing$. For this reason, we often call the second smallest eigenvalue λ_1 as the *algebraic* connectivity, the corresponding eigenvector *Cheeger* vector.

2.2 Transition Path Theory

We noticed that normalized graph Laplacian $\mathcal{L}$ keeps the same connectivity measure of unnormalized graph Laplacian L. Furthermore, L is more related with random walks on graph, through which eigenvectors of $P = D^{-1}A$ are easy to check and calculate. That's why we choose this row Markov matrix to analysis the graph. So in this section, we first introduce the row Markov matrix and its stationary distribution, then the *committor* function which is the solution of a Laplacian equation with Dirichlet boundary conditions is deducted, and finally the effective and transition current is investigated.

2.2.1 Stationary Distribution of Markov Chain

Given a graph $G = (V, E)$, consider a random walk on G with transition probability $P_{ij} = P(x_t{+}1 = j|x_t = i) \geq 0$. Thus P is a row-Markov matrix, and the stationary distribution π^T is

$$\pi^T P = \pi^T \tag{5}$$

such π is invariant/equilibrium distribution.

If P is primitive, then the largest eigenvalue λ with $|\lambda| = 1$ is unique w.r.t

$$\lim_{t\to\infty} \pi_0^T P^k = \pi^T,\ \forall \pi_0 \geq 0,\ 1^T \pi_0 = 1 \tag{6}$$

This means when we take powers of P , i.e.P^k, all rows of P^k will converge to the stationary distribution π^T . Such a convergence only holds when P is primitive.

If P is irreducible, then π is unique.

2.2.2 Committor Function

Given two sets V_0 and V_1 in the state space V , the transition path theory tells how these transitions between the two sets happen (mechanism, rates, etc.). If we view V_0 as a reactant state and V_1 as a product state, then one transition from V_0 to V_1 is a reaction event. The reactive trajectories are those part of the equilibrium trajectory that the system is going from V_0 to V_1 . Let the hitting time of V_l be

$$\tau_i^k = inf\{t \geq 0 : x(0) = i, x(t) \in V_k\}, k = 0, 1. \tag{7}$$

The central object in transition path theory is the *committor* function. Its value at $i \in V_u$ gives the probability that a trajectory starting from i will hit the set V_1 first than V_0 , i.e., the success rate of the transition at i. The defined *committor* function satisfies the following Laplacian equation with Dirichlet boundary conditions

$$\begin{gathered}(Lq)(i) = [(IP)q](i) = 0, i \in V_u \\ q_{i\in V_0} = 0,\ q_{i\in V_1} = 1.\end{gathered} \tag{8}$$

which is in the following formulation

$$q_i = Prob(\tau_i^1 < \tau_i^0) = \begin{cases} 1, & x_i \in V_1 \\ 0, & x_i \in V_0 \\ \sum_{j\in V} P_{ij} q_j, & i \in V_u \end{cases} \tag{9}$$

The *committor* function provides natural decomposition of the graph. If $q(x)$ is less than 0.5, i is more likely to reach V_0 first than V_1 ; so that $\{i|q(x) < 0.5\}$ gives the set of points that are more attached to set V_0 . Once the *committor* function is given, the statistical properties of the reaction trajectories between V_0 and V_1 can be quantified.

2.2.3 Reactive Current

For any stationary trajectory, we call each proportion of the trajectory from A to B a $AB - reactive$ trajectory. We also define the reactive current from A to B of a directed edge ij as

$$J(ij) = \begin{cases} \pi(i)[1 - q(i)]P_{ij} q(j), & i \neq j \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

One can see that it shows the probability of the directed edge ij being a proportion of a $AB{-}reactive$ trajectory. Note that for all $i \in B, q(i) = 1$ and then $J(ij) = 0$ for all j connected to i, which can be explained as the movement leaving $i \in B$ for the other nodes cannot be a part of any $AB - reactive$ trajectory.

2.2.4 Effective Current and Transition Current

The reactive current gives a clue for us to measure the importance of an edge and a node among all $AB - reactive$ trajectories. In particular, we define the effective current of an edge ij as

$$J^{+}(ij) = \max\{J(ij) - J(ji), 0\}. \tag{11}$$

Note that for every directed edge ij, (i) $J^{+}(ij) \geq 0$ and (ii) $J^{+}(ij) > 0$ implies $J^{+}(ji) = 0$. Therefore, we can think of the effective current as the "direction preference" and its "strength" of an edge in terms of the successful transition from A to B. We also observe that $J^{+}(ij) = 0 (i.e. J^{+}(ji) \geq 0)$ for all $i \in B$. For a node $i \in V$, we define the transition current through i as

$$T(i) = \begin{cases} \sum_{j \in V} J^{+}ij, & x_i \in A \\ \sum_{j \in V} J^{+}ji, & x_i \in B \\ \sum_{j \in V} J^{+}ij = \sum_{j \in V} J^{+}ji, & i \in V - A - B \end{cases} \tag{12}$$

The equivalence in the last formula is resulted from the definition of the *committor* function. Obviously, a node with high transition current through it plays a key role in the transition from A to B. Similarly, we can adopt this approach to identify the key nodes who bridge two communities of nodes. For example, in the thresholding scheme V_0 and V_1 we mentioned previously, the transition function

$$T(i) = \sum_{j \in V} J^{+}ij, \; i \in V_0 \tag{13}$$

measures the contribution of every node in V_0 in the connection of V_0 and V_1 .

3 Evaluation on Karate Club network

As shown in Fig. 1, in the Zachary's karate club network, there are 34 nodes representing 32 members, the coach $A = \{1\}$, and the president $B = \{34\}$. See Fig. 1. The undirected and unweighted edges represent the affinity relation between club members. The story behind the network is: The coach would like to raise the instruction fee, while the president does not allow it. The conflicts finally result in a fission of the club – the coach leaves the club with his fans, who are marked in red, and sets up his own club; the other members, who are marked in blue, remain in the old club with president.

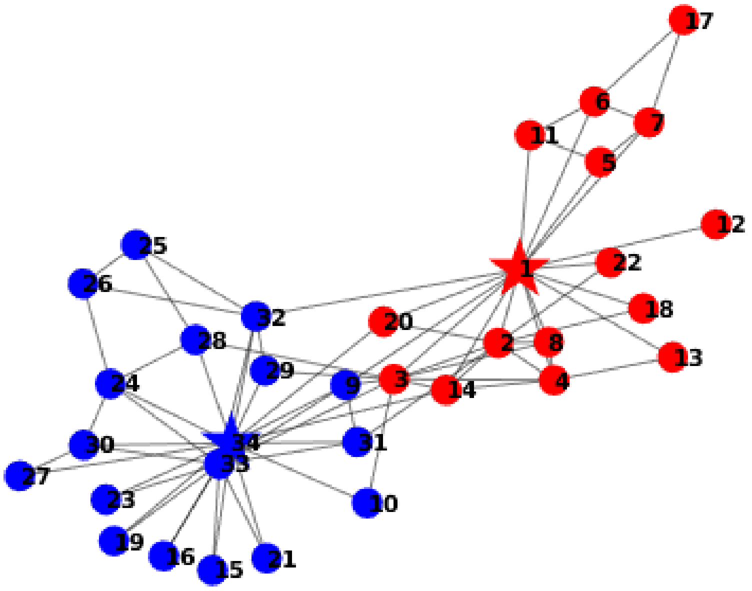


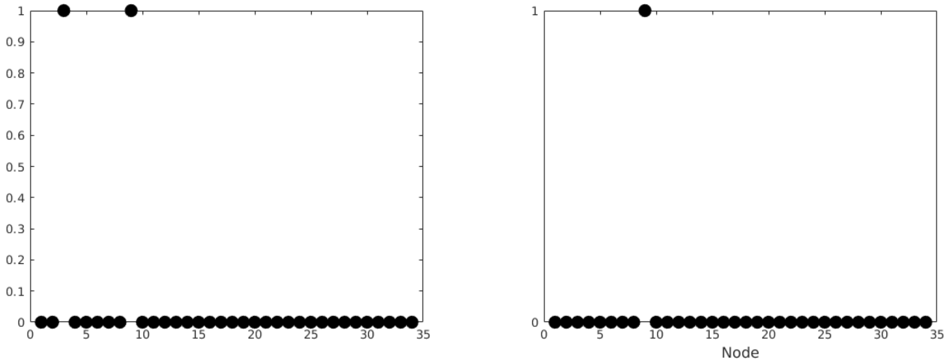
Figure 1: The Zachary's Karate Club network. Node highlighted with stars are $A = \{1\}, B = \{34\}$ are local minimal, nodes are distinguished by red and blue color into two clusters.

3.1 Spectral Clustering via Cheeger Vector

We apply the method described in Sec. 2.1 to the graph, the $Cheeger$ vector, second eigenvector of normalized graph Laplacian $\mathcal{L}$ is calculated, and we cluster the nodes into two clusters via thresholding to obtain the clustering vector c_s

$$c_s(i) = \begin{cases} 0, & Cheeger(i) > 0, \\ 1, & Cheeger(i) \leq 0. \end{cases} \tag{14}$$

Then the spectral clustering vector c, is compared to the true fission c_0, the difference is calculated as $\delta_s = c_0 - c_s$, which is plotted in Fig. 2a.



(a) Difference between spectral clustering with true fission.

(b) Difference between *committor* function clustering with true fission.

Figure 2: Clustering results compared to true fission in spectral clustering and path transition analysis.

3.2 Markov Chain and Stationary Distribution

We assume that from each node a random walker will jump to its neighbors with equal probability, the row Markov matrix is defined as $P = D^{-1}A$, with who the the stationary distribution is obtained by solving left eigenvector in Eqn. 2.2.1, the result is illustrated in Fig. 3.

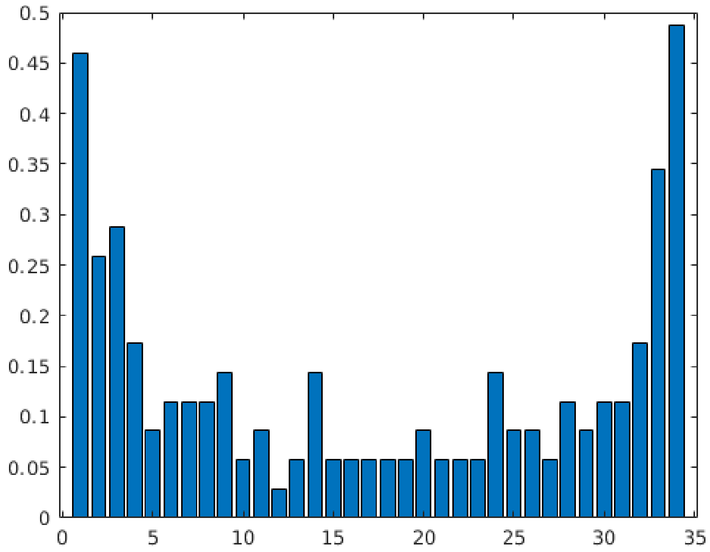


Figure 3: The stationary distribution of the Markov chain.

3.3 Committor Function

We compute the *committor* function $q(i)$ by Eqn. 2.2.2 for every node $i \in V$. We also run the thresholding scheme based on the *committor* function $V_1 = \{i \in V | q(i) \geq 0.5\}$ and $V_0 = \{i \in |q(i) < 0.5\}$ to bipartite the graph. By the definition of $q(i)$, we regard V_1 as members stay in the old club and regard V_0 as the members join the coach's new club. We measure the difference between the bi-partition V_1 and V_0 and the true fission $\delta_s = c_0 - c_s$. The result is presented in Fig. 2b. It turns out the thresholding scheme classifies nodes accurately, except node 9 whose decision about the clubs is independent of the connections to the other nodes.

Therefore, we can say the thresholding scheme based on the *committor* function is probably a better approach for the Zachary's karate club network, in comparison to the spectral clustering via Cheeger vector in Fig. 2a.

3.4 Effective Current and Transition Current

We then calculate the reactive current $J(ij)$, effective current $J^+(ij)$ and transition current $T(i)$ for every node $i \in V$ and edge $ij \in E$. The results are drawn and the node size is coded according to the *committor* function $q(i)$ in Fig. 4, where we noticed that the source of the current $A = \{1\}$ and the sink $B = \{34\}$ which is marked as stars in Fig. 4 also naturally has large node size. The edges on the short trajectories from A to B, for example $(1, 20, 34)$ and $(1, 18, 2, 20, 34)$ may have strong effective currents relatively. The nodes {5,6,7,11,12,17} which are isolated in the transition current graph usually do not contribute on bridging the two clusters.

The directed edges in Fig.4 represent effective currents $J^+(ij)$, for all $1 \leq i, j \leq 34$ such that $J^+(ij) > 0$. Recall that in the case $J^+(ij) > 0, J^+(ji) = 0$ and hence every current travels in at most one direction on an edge. For every edge, the arrow indicates the direction of the current, while the width is proportional to the strength of the current i.e. the value of $J^+(ij)$.

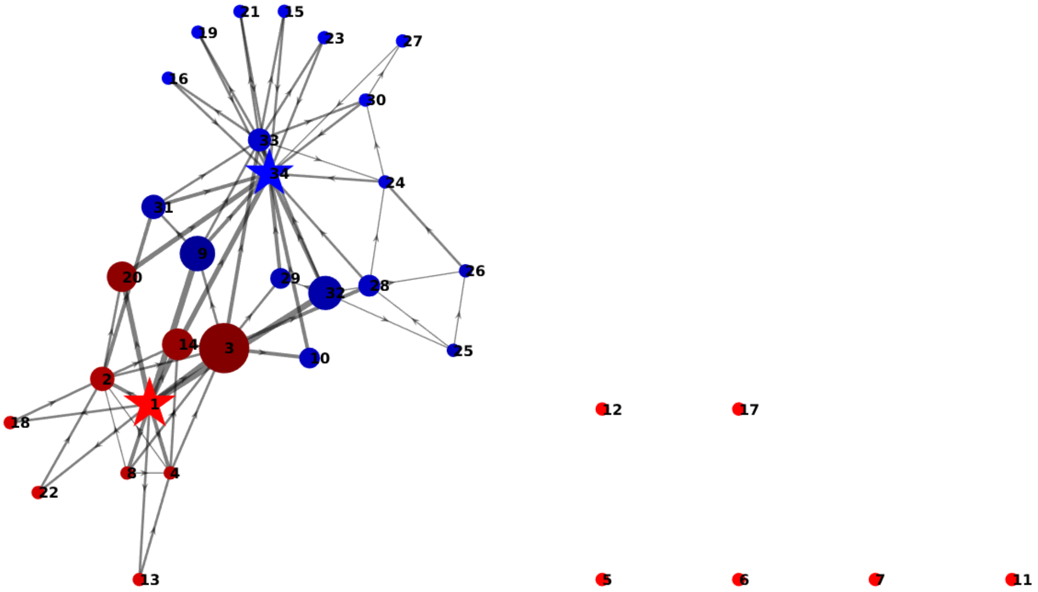


Figure 4: The *committor* function value at every node and the thresholding scheme (presented by the color of nodes), the effective current of every edge (presented by the arrow and width of edges), and the transition current through every node (presented 7 by the size of nodes). The nodes {5,6,7,11,12,17} are isolated in the transition current graph which means these nodes plays no importance on bridging the two clusters.

4 Evaluation on More Networks

In this section, we perform spectral clustering and transition path analysis on two more networks, the social network of Les Misérables, and the LAO protein binding transition network. Following the similar methodology and evaluation process, we directly give the result analysis after the network description and introduction to preprocessing.

4.1 The Social Network of Les Misérables

The social network of Les Misérables, collected by Knuth, consists of 77 main characters in the novel by Victor Hugo. The edge weight w_{ij} record the number of co-occurrence of two characters i and j in the same scene.

The original network exhibits a single local (global) minimum, Valjean, who is the central character as the whole novel was written around his experience. However, dropping those edges whose weights are no more than a threshold value (7 here), there appears a subnetwork which is closely associated with the Paris uprising on the 5th and 6th of June 1832. To obtain this subnetwork, the unconnected nodes or isolated nodes should be removed after applied edge weight thersholding, then after dropping the tiny subnetwork {Myriel, Napoleon, MlleBaptistine}, the target network we want to investigate is shown in Fig. 5. This subnetwork consists of two local minima, Enjoras and Valjean, which is highlighted as with stars in 5, the former being the leader of the revolutionary students called Friends of the ABC, the Abaissé. Led by Enjolras, its other principal members are Courfeyrac, Combeferre, and Laigle (nicknamed Bossuet) et al., who fought and died in the insurrection. So from the graph construction and local minimum analysis, we can find the local minimas have there true meaning in the practical social network.

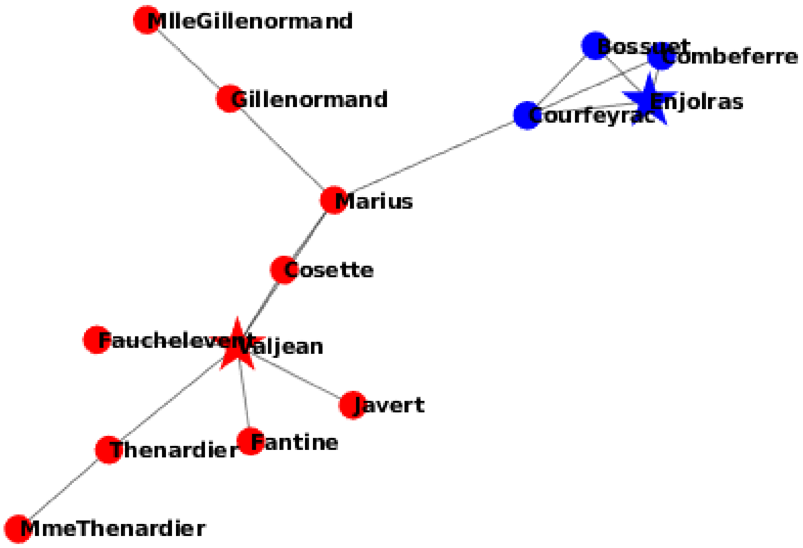


Figure 5: The subnetwork of the Les Misérables social network. Two local minima, Enjoras and Valjean are highlighted with star markers, and the node color is coded by the same clustering result on two different clustering methods.

Afterwards, the spectral clustering and *committor* function clustering are performed on the obtain subnetwork, the clustering results c_s and c_t of these two approaches are illustrated using circle marks and check markers as shown in Fig. 6. Without true fission, from this clustering result we can know both methods give the uniform clustering, given the two local minimas. These results are applied in the visualization the subnetwork in Fig. 5.

Finally, the effective and transition currents are computed following the same clue with Karate Club network, the transition path graph is shown in Fig. 7, from which we know that the significant

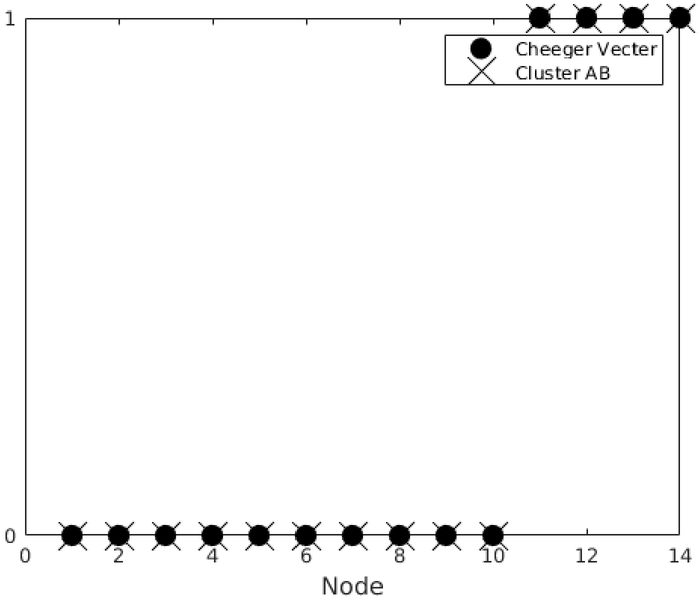


Figure 6: Clustering results using spectral clustering and *committor* function clustering.

effective current lie in node Courfeyrac and Marius. Practically in the network, Courfeyrac, a law student and often seen as the heart of the group, who introduced Marius to the Friends of ABC. Marius, a descend of the Gillenormands, though badly injured in the battle, was saved by the main character Valjean when the barricade fell and married to Cosette, the adopted daughter of Valjean.

We conclude that the knowledge derived from the transition path analysis is accords with our piror knowledge in practical network.

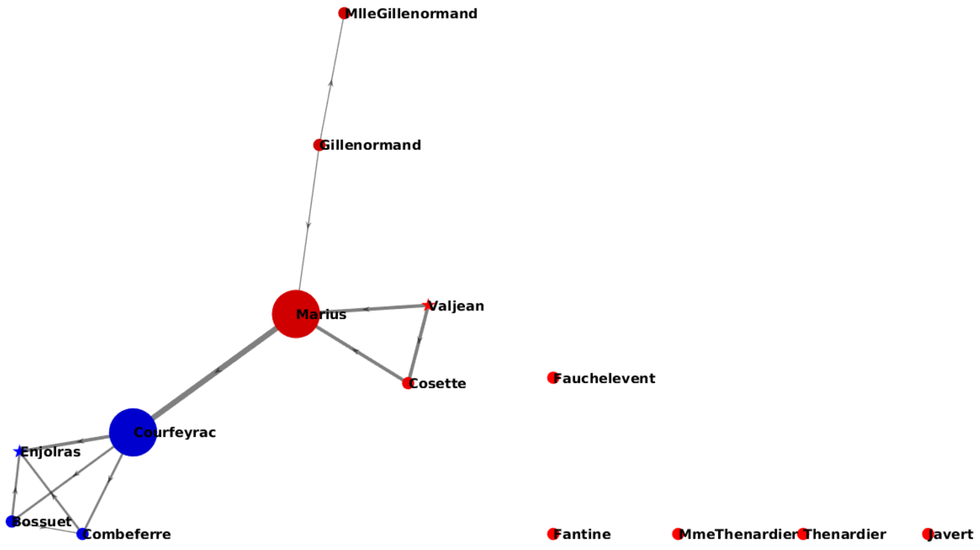


Figure 7: The *committor* function value at every node and the thresholding scheme (presented by the color of nodes), the effective current of every edge (presented by the arrow and width of edges), and the transition current through every node. The isolated nodes in the transition current graph play no importance on bridging the two clusters.

4.2 LAO Protein Binding Transition Network

This application examines the binding of Lysine-, Arginine-, Ornithine-binding (LAO) protein to its ligand. The critical node analysis provides us a concise summary of global structure of networks while preserving important pathways, which enables us to reach a more thorough description than previous approximate analysis.

The Markov state model was constructed with 54 metastable states, using data obtained from molecular dynamics simulation, more information about this network is describe in the reference paper. We threshold this graph to an undirected graph by keeping those edges $\{i, j\}$ such that $(w_{ij} + w_{ji})/2 > 40$, to avoid small numbers of transitions which may be heavily influenced by the noise caused by the way of counting the transition. The graph representation by removing the isolated states is shown in Fig. 8, where actual two local mininas are found, node 12 and nod 49, which are not the same with the local minima calculation result in reference paper. The main reason of this difference maybe the way of constructing the undirected graph by thresholding.

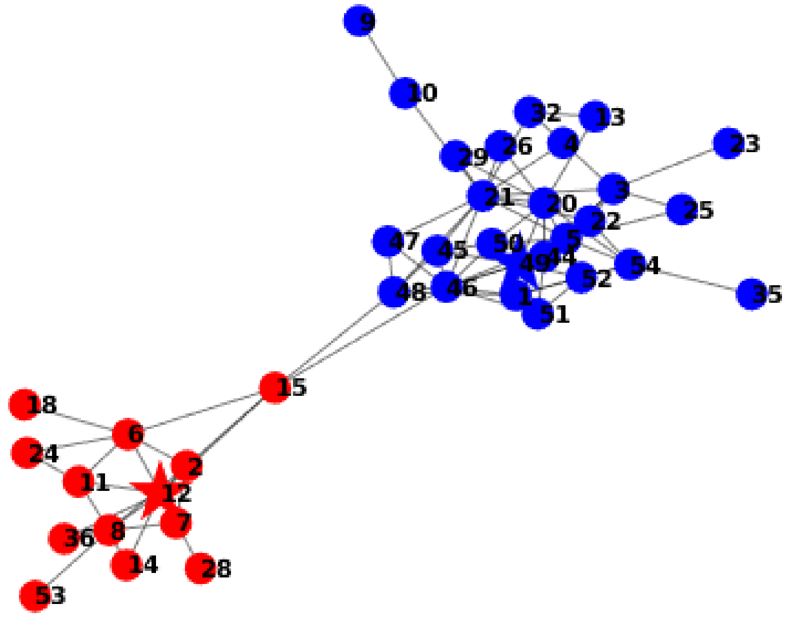


Figure 8: Undirected graph visulization of LAO network, two local minimas are found are thresholding and dropping isolated states. The transition states are clustered into two clustering using spectral clustering and *committor* function.

By applying spectral clustering and *committor* function on the preprocessed graph, the clustering results c_s and c_t are shown in Fig. 9, which also showing the same clustering in this case.

Finally, the effective and transition currents are computed, the node color is coded with *committor* function and the size of the node is coded by transition currents. From Fig. 4, we know the most import bridging edges within the two clusters are node 15 and 48.

References

[1] [1] Weinan, E., Jianfeng Lu, and Yuan Yao. "The Landscape of Complex Networks: Critical Nodes and A Hierarchical Decomposition." Methods and Applications of Analysis 20.4 (2013): 383.

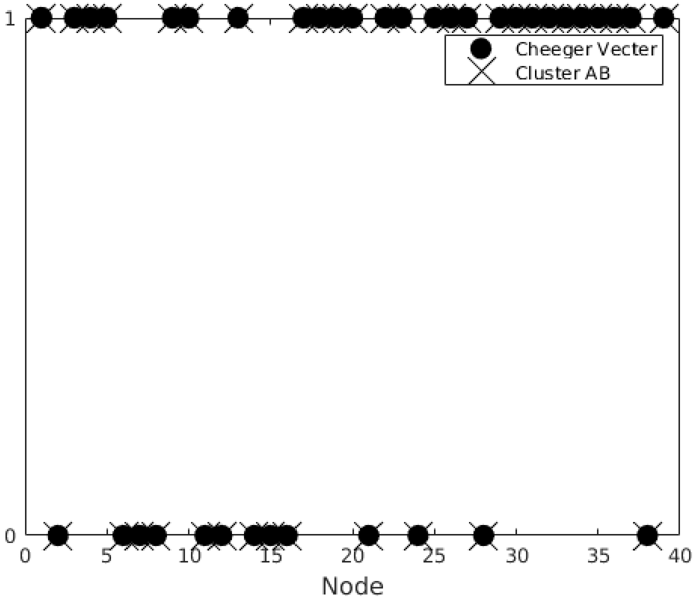


Figure 9: Clustering results by applying spectral clustering and *committor* function on the preprocessed graph.

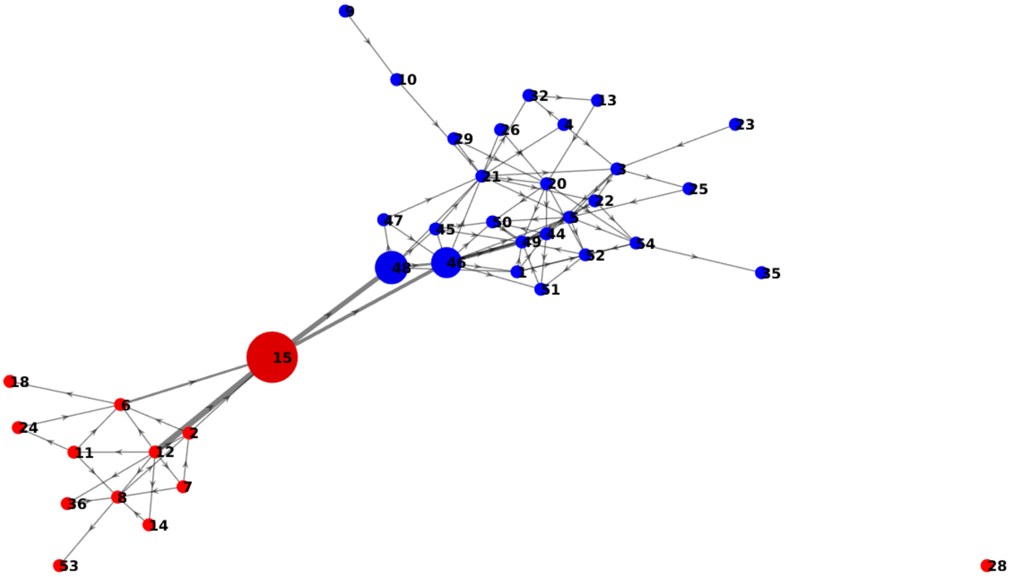


Figure 10: The *committor* function value at every node and the thresholding scheme (presented by the color of nodes) of LAO transition network, the effective current of every edge (presented by the arrow and width of edges), and the transition current through every node. The isolated nodes in the transition current graph play no importance on bridging the two clusters.