Final-Project - CSIC5011
Spectral Clustering and Transition Path Analysis on the Karate Club Network

Cao Guiyu (SID: 20392411), LO Yi-Su (SID: 20399988)
Department of Mathematics, School of Science
The Hong Kong University of Science and Technology
gcaoaa@connect.ust.hk, yloab@connect.ust.hk

Abstract

In this work, we apply the spectral clustering via the Cheeger vector and the transition path analysis to Zachary's karate club network. The two approaches are introduced and the bipartitions of the network obtained from the two approaches separately are compared with the ground truth. Some further information behind the network is also investigated and discussed to make us understand the nature of the network better.

1 Problem Statement

In the **Zachary's karate club network**, there are 34 nodes representing 32 members, the coach (node 1), and the president (node 34). See Figure 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.

Our main question and task in this project is: Is it possible we can bipartite the nodes according to the edge structure between them, and obtain a bipartition close to the ground truth fission in Figure 1? Moreover, if we can also discover more information about this network?

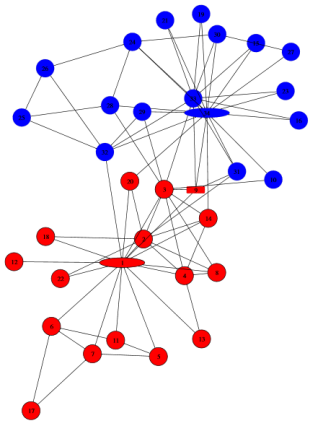

Figure 1: Karate Club Network

2 Methods

To address the problem above, we will firstly apply the spectral clustering via the Cheeger vecor to bipartite the network. Then, a Markov chain according to the network structure is defined and the stationary distribution is computed. Lastly, transition path analysis is performed as the second approach to bipartite the network and as a new way to examine the network in order to get more insight into it.

2.1 Spectral clustering via the Cheeger vector

The normalized graph Laplacian is constructed as

$$L = D^{-1/2}(D - A)D^{1/2},$$

where A is adjacency matrix and D is the a diagonal Matrix $D = \text{diag}(d_i)$,

$$A_{ij} = \begin{cases} 1, \text{ node } i \backsim \text{node } j, \\ 0, \text{ otherwise}, \end{cases}$$

$$d_i = \sum_{j=1}^{n} A_{ij}.$$

Let L has n eigenvalues as

$$Lv_i = \lambda_i v_i, \quad v_i \neq 0, \quad i = 0, \cdots, n-1,$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. For the second smallest eigenvector v_1, which is also known as the *Cheeger vector*, define

$$N_- = \{i : v_1(i) < 0\},$$
$$N_+ = \{i : v_1(i) > 0\}.$$

So, the graph can be bipartited into two components as N_- and N_+.

2.2 Transition path theory

In this section, we mostly follow the instruction given in [1].

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

$$\pi^T P = \pi^T$$

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

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

If P is irreducible, then π is unique.

Committor function

In a network, let $A \subset V$ and $B \subset V$ be two subsets of nodes, which can be referred to as two states. For example, in the Zachary's karate club network, $A = \{1\}$ represents the coach and $B = \{34\}$ represents the president. For a Markov chain defined above, we have the stationary distribution $\{\pi(i) : i \in V\}$ and stationary trajectories along the edges between nodes. The *committor function* $q : V \to [0, 1]$ measuring the probability that every trajectory starting from i hits B ahead of A is required to satisfy

$$\begin{cases} q(i) = 0, & i \in A, \\ q(i) = 1, & i \in B, \\ \sum\limits_{y \in V} P_{ij} q(j) - q(i) = 0, & i \in V - (A \cup B), \end{cases}$$

where P_{ij} is an entry in the transition matrix P defined above. Those equations are reasonable requirements that the first two equations simply follow the definition of q and the last one asks the probability at every node equals to the sum of the probability at the other nodes times the probability of moving to those nodes. One can also reorganize those equations into a linear system

$$Lq = 0, \quad \text{where} \quad q(A) = 0, \quad q(B) = 1.$$

For example, in the Zachary's karate club network, we have

$$L = \begin{pmatrix} P_{2,1} & -1 & P_{2,3} & \cdots & P_{2,34} \\ P_{3,1} & P_{3,2} & -1 & \cdots & P_{3,34} \\ \vdots & \vdots & \vdots & & \vdots \\ P_{33,1} & P_{33,2} & P_{33,3} & \cdots & P_{33,34} \end{pmatrix} \quad \text{and} \quad q = \begin{pmatrix} q_1 = q(1) = 0 \\ q_2 = q(2) \\ q_3 = q(3) \\ \cdots \\ q_{34} = q(34) = 1 \end{pmatrix},$$

in which the committor function value at every node could be easily solved for.

As an immediate application of the committor function, the thresholding scheme $V^+ = \{i \in V \mid q_i \geq 0.5\}$ and $V^- = \{i \in V \mid q(i) < 0.5\}$ provide a decomposition of the graph by predicting which state the trajectory starting from node i is more likely to reach first.

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

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

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.

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

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

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

$$T(i) = \begin{cases} \sum\limits_{j \in V} J^+(ij), & \text{if } i \in A, \\ \sum\limits_{j \in V} J^+(ji), & \text{if } i \in B, \\ \sum\limits_{j \in V} J^+(ij) = \sum\limits_{j \in V} J^+(ji), & \text{if } i \in V - (A \cup B). \end{cases}$$

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^- and V^+ we mentioned previously, the transition function

$$T(i) = \sum_{j \in V^+} J^+(ij), \quad i \in V^-$$

measures the contribution of every node in V^- in the connection of V^- and V^+.

3 Results and Discussion

3.1 Data and code

The `Matlab` data for Zachary's Karate club network used in this section is accessible at

`https://math.stanford.edu/~yuany/course/data/karate.mat`

An introduction to this data set is already given in Section 1. The `Matlab` source code for the following experiments could be downloaded at

`https://goo.gl/RnsHgs`

3.2 Experiment A: Spectral clustering via the Cheeger vector

We apply the method described in Section 2.1 to the network. To distinguish the difference between the Cheeger vector and the true fission, we define the difference $\triangle$ as

$$\triangle(i) = C_0(i) - Cheeger(i), \quad i = 1, 2, \cdots 34,$$

where C_0 is the true fission and $Cheeger$ is calculated from the Cheeger vector, as

$$Cheeger(i) = \begin{cases} 0, \ i \in N_-, \\ 1, \ i \in N_+, \end{cases}$$

where N_- represents the members stay in old club and N_+ represents the members leave for the coach's club. The scatter picture for $\triangle$ is shown in Figure 2.

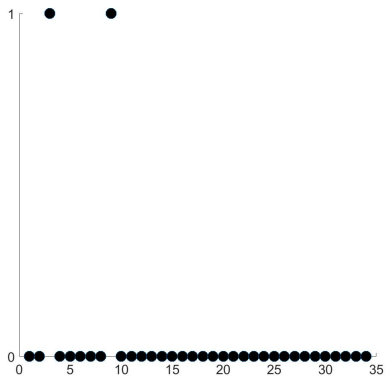

Figure 2: Difference between true fission and bi-partition by Cheeger vector.

From Figure 2, it is illustrated that Cheeger vector method misleads **node 3** and **node 9**. Actually, node 3 and node 9 stay in old club, but Cheeger vector method regards them in the coach's club. As indicated in [1], node 9 is a special case in which the guy joined the coach's club due to the necessity to finish a course. Node 9 could be thought of as a critical point which is pretty difficult to bipartite. On the whole, the Cheeger vector method agrees well with the true fission.

3.3 Experiment B: Transition path analysis

In this section, we work with transition path analysis introduced in Section 2.2 and perform several experiments to discover more information in the network. Firstly, let $G = (V, E), V = \{1, 2, \cdots, 34\}$ be the graph of the karate club network and let $A = \{1\}$ (the coach) and $B = \{34\}$ (the president) be two states in it. We then examine the transition from A to B, which corresponds to the switch from the coach's new club to the old club led by the president.

Markov chain and the stationary distribution

We assume that from each node a random walker will jump to its neighbors with equal probability, i.e.

$$P = D^{-1}A,$$

where A and D are defined in Section 2.1. With the transition matrix P, the stationary distribution is obtained by solving

$$\pi^T P = \pi^T$$

and results in

$$\begin{aligned}\pi^T = [\ & 0.1026\ 0.0577\ 0.0641\ 0.0385\ 0.0192\ 0.0256\ 0.0256\ 0.0256\ 0.0321\ 0.0128\ 0.0192 \\ & 0.0064\ 0.0128\ 0.0321\ 0.0128\ 0.0128\ 0.0128\ 0.0128\ 0.0128\ 0.0192\ 0.0128\ 0.0128 \\ & 0.0128\ 0.0321\ 0.0192\ 0.0192\ 0.0128\ 0.0256\ 0.0192\ 0.0256\ 0.0256\ 0.0385\ 0.0769\ 0.1090].\end{aligned}$$

Note that π is normalized here. Figure 3 shows a histogram of π.

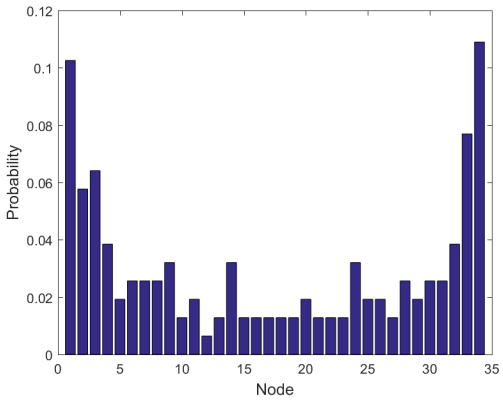

Figure 3: The stationary distribution of the given Markov chain.

Committor function

We then compute the committor function $q(i)$, 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 in Figure 6 and explained in the following. As a comparison, the true fission of the network is shown again in Figure 5.

In Figure 6, we also run the thresholding scheme based on the committor function $V^+ = \{i \in V \mid q(i) \geq 0.5\}$ and $V^- = \{i \in \mid q(i) < 0.5\}$ to bipartite the graph. By the definition of $q(i)$, we regard V^+ as members stay in the old club and regard V^- as the members join the coach's new club. Moreover, in V^+, lighter blue nodes have committor function values near 1, such as node 34; while darker blue nodes have values near 0.5, such as node 32. Similarly, in V^-, lighter red nodes have committor function values near 0, such as node 1; while the darker red nodes have values near 0.5, such as node 3.

We further measure the difference between the bi-partition V^+ and V^- and the true fission with the same approach as Section 3.2. The result is presented in Figure 4. 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. The reason is already given in [1] and the previous experiment. Therefore, we can say the thresholding scheme based on the committor function is probably a perfect approach for the Zachary's karate club network, in comparison to the spectral clustering via Cheeger vector (see Figure 2).

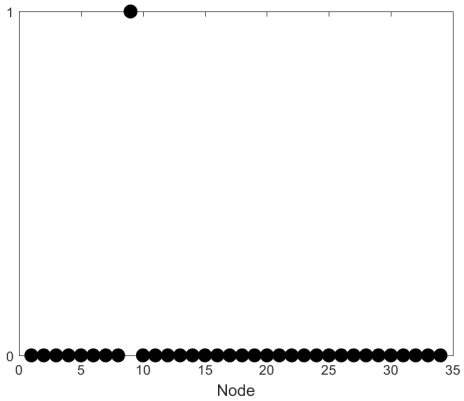

Figure 4: Difference between true fission and bi-partition by the committor function.

Effective current

The directed edges in Figure 6 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 $J^+(ij)$. We discuss several interesting observations below.

- $A = \{1\}$ is a source of the currents, and $B = \{34\}$ is a sink. It is a natural result due to the definitions of the effective current and reactive current.
- A shorter trajectory from A to B usually enjoys larger effective currents than a longer trajectory. For instance, the trajectories $(1, 20, 34)$ and $(1, 18, 2, 20, 34)$.
- There are no effective currents between nodes $5, 6, 7, 11, 12, 17$, and their neighbors. Therefore, they appear as isolated nodes in Figure 6. In fact, it is because $q(i) = 0$ for $i = 5, 6, 7, 11, 12, 17$. The cause of such an interesting phenomenon is clear when we look at the original network: A is the only one neighbor of both node 12 and the small community $5, 6, 7, 11, 17$. That is, every trajectory starting from those nodes must reach A before they go to B, which leads to zero committor function values.

Transition current

The size of every node is proportional to the strength of the transition current going though it i.e. the value of $T(i)$. It is obvious that $A = \{1\}$ and $B = \{34\}$ are the largest nodes. Besides them, nodes $3, 9, 32$ are the next largest nodes which receive the most effective currents and therefore are important nodes for the transition from A to B. We can also conclude that those nodes are key members for the coach's club to interact with the president of the old club.

4 Conclusion

We apply the spectral clustering via the Cheeger vector and the transition path analysis to bipartite the network respectively. The results agree with the true fission well. In Transition path analysis, we also get some further information about the nature of the network.

References

[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.

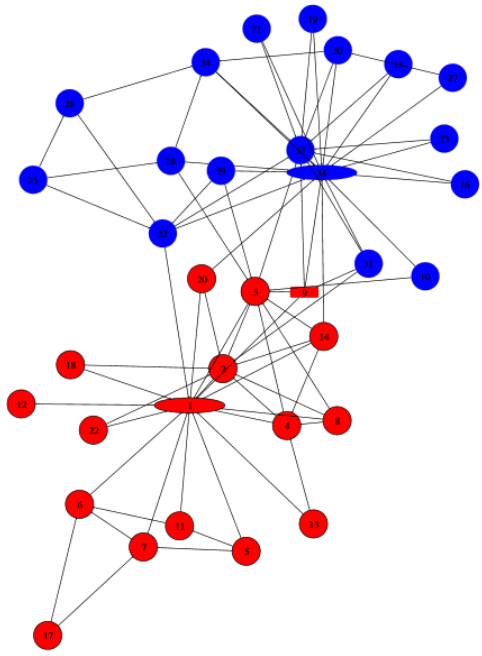

Figure 5: The Zachary's karate club network.

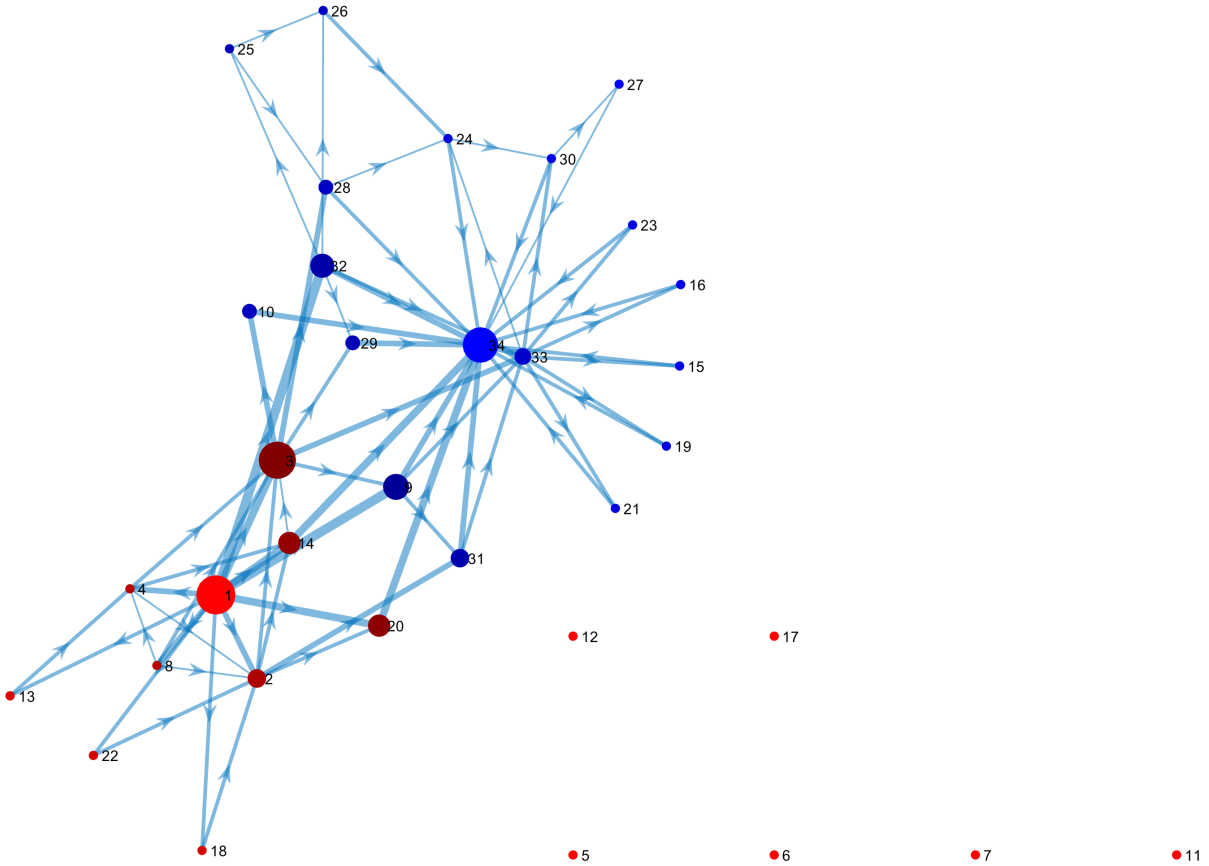

Figure 6: The committor function value at every node and the thresholding scheme (presented by the color of nodes), the effective current of every edge (presented by the arrow and width of edges), and the transition current through every node (presented by the size of nodes).