# Report: Analysis of Karate Club Network

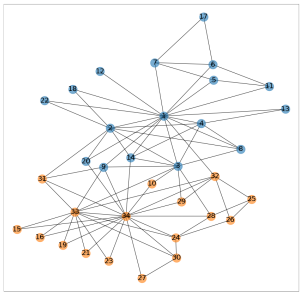
**Qing Du, Jiahao He**

## Abstract

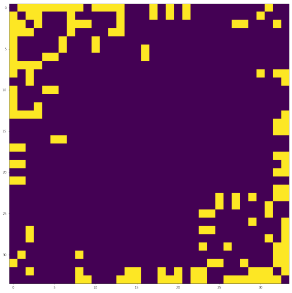
We analyse Karate Club Network via four different approaches: (1) classical network flow algorithm to produce a minimum cut, (2) spectral clustering methods based on the eigenvectors of the (normalised) graph Laplacian, (3) path transition theory of random walk, (4) a hierarchical decomposition scheme proposed in [1]. All four methods perform well in bipartitioning the graph based purely incidence relation so as to match the ground truth separation and are rather consistent in the nodes that are hard to classify, despite their different underlying rationale. For each method, we summarize their procedure, report their results, and form comparison with the other methods.

## 1 Problem Definition

In this project, we analyse Karate Clue Network via 4 different approaches. As shown in Figure 1a, each node represents a club member and each edge represents the affinity relationship between the two members. The conflicts on instruction fee between the coach (node 1) and the president (node



(a) Karate Club Network



(b) Adjacent Matrix (yellow=1, purple=0)

34) result in a separation of the club into two: the blue one led by node 1 and the orange one led by 34. We adopt the following methods to bipartite the graph based purely the adjacent relation and compare the bipartition with the ground-truth separation:

1. **Minimum-cut Bipartition.** If we know in advance that node 1 and node 34 are in different partition, we can directly apply minimum cut algorithm with 1 as the source and 34 as the sink to provide a bipartition. Without prior knowledge on the source-sink pair, we perform pair-wise minimum cut algorithm (or maximum flow algorithm by duality) and search for the pair that provides the largest minimum cut (the largest maximum flow).
2. **Spectral Methods.** The second smallest eigenvectors of both the unnormalized or normalized graph Laplacian can be used to bipartite the graph where the class of a node is

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

determined by the sign of its corresponding entry in the eigenvector. Moreover, both induced graphs on the two clusters are guaranteed to be connected.

3. **Transition Path Theory.** A random walk $\{X_n : n \geq 0\}$ is defined on the graph where at each step, the next node is equally likely to be any neighbour of the current node. Assume we know that node 1 and 34 are in different clusters, we use $P(\tau_{34} < \tau_1 | X_0 = x)$, where $\tau_v$ is the first hitting time to node $v$, to indicate whether node $x$ is closer to 34 or 1 and to further bipartite the graph.
4. **Hierarchical Critical Nodes Decomposition.** [1] proposes a hierarchical method for network analysis in which the vertex set can be disjointly decomposed into *critical nodes* of different index level and *attract basin* of those critical nodes based on any injective function from the vertex set to $\mathbb{R}$. We analyse the decomposition based on the negative logarithm of node degree.

Throughout the discussion, we use $G = (V, E)$ to refer to the graph in Figure 1a with $V$ as the vertex set and $E$ as the edge set.

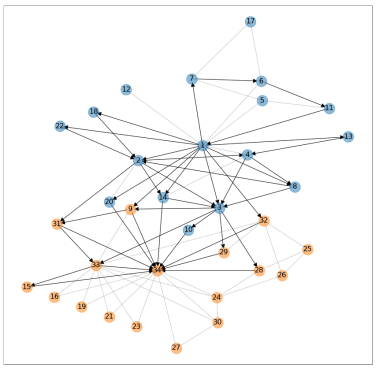
## 2 Minimum-cut bipartition

### 2.1 With given source and sink

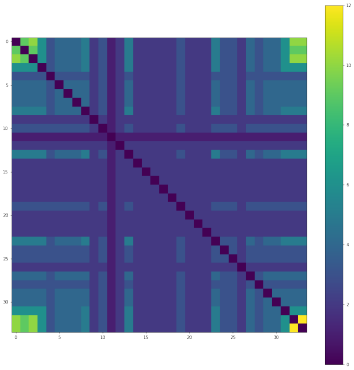
We treat nodes 1 and 34 as the source and sink in a maximum flow problem and we set the flow capacity on each edge to be 1. Then, in any optimal solution, the set of nodes reachable from 1 in the directed residual graph forms a minimum cut that separates the two nodes can be found, that is, we solve the following pair of dual problems:

$$\max_{f_{ij} \geq 0 : \{i,j\} \in E} \sum_j f_{1j} \qquad\qquad \min_{S \in V} \sum_{u \in S, v \in S^c} 1_{\{\{u,v\} \in E\}}$$
$$s.t. \quad \sum_j f_{ij} = \sum_j f_{ji}, \forall i \neq 1, 34. \qquad\qquad s.t. \quad 1 \in S, 34 \in S^c.$$

We present one solution in Figure 2a. The solid directed edges represent flows (all of 1 unit) on those edges and the two colors represent the partition given by the minimum cut. Comparing with the ground truth, we find that only nodes 9 and 10 are misclassified.



(a) Max-Flow Min-Cut



(b) Pairwise Min-Cut

### 2.2 Without given source and sink

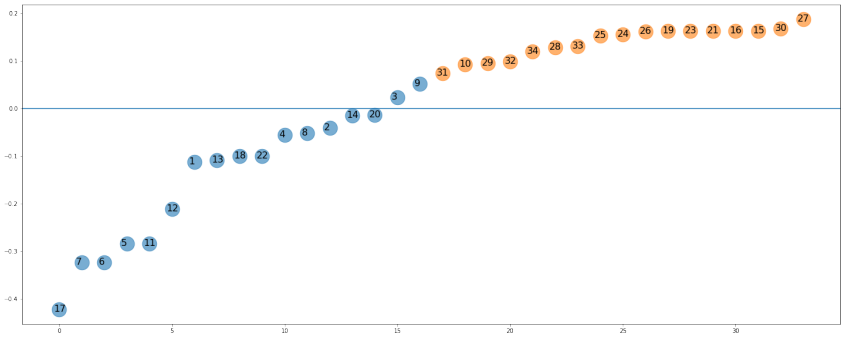
Without given source or sink, we set each pair of nodes as source and sink and for each pair, we find a minimum cut and show the result in Figure 2b. We observe that the pair $(1, 34)$ gives the third largest

minimum cut value among all possible pairs. The other larger pairs are all in the form $(u, v)$ such that $u$ is a node with a high degree and $v$ is an adjacent node to $u$ and the partitions formed by these cuts are highly imbalanced (with one node in one bipartite and the rest of the nodes in the other). Thus, seeking for minimum cut with a large cut value may be an efficient and acceptable option for bipartition.

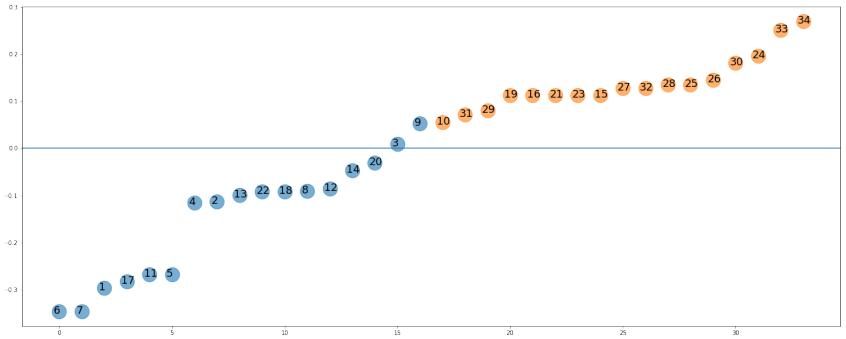
# 3 Spectral bipartition

## 3.1 Unnormalized spectral bipartition

Let $A$ be the adjacent matrix of $G$, i.e. $A_{ij} = 1$ if $\{i, j\} \in E$ and $A_{ij} = 0$ otherwise, and $D = diag(d_i)$. The Laplacian is $L = D - A$. $L \succeq 0$ and the smallest eigenvalue is 0 with the all one vector as an eigenvector. Since $G$ is connected, Fiedler's Theorem asserts that the second smallest eigenvector $\lambda_2^F$ (or the Fiedler value) is positive and if we let $F_0$ and $F_1$ be the sets of vertices such that their corresponding entries in the eigenvector $\boldsymbol{v}^F$ of $\lambda_2^F$ is negative and non-negative, respectively, then both induced subgraphs are connected. For $G$, we find that $\lambda_2^F \approx 0.46852$. Figure 3a shows the eigenvector entry of each vertex in an ascending order, where the label on each node represents its label in $G$. The color of each node represents their ground-truth class. Nodes above (below) 0 form $F_0$ ($F_1$). One observes that except for nodes 3 and 9, the classification is correct, and nodes in the blue cluster always have smaller entry in the eigenvector than those in the orange cluster. Moreover, we observe that nodes $15, 16, 19, 21, 23$ are equal in value, this reflect the fact that they are in equivalent position in $G$ or we can form endomorphism of $G$ by simply permuting these nodes.




(a) Eigenvector of $\lambda_2^F$




(b) Eigenvector of $\lambda_2^C$

Figure 3: Spectral Method

### 3.2 Normalized spectral bipartition

Let $L_{sym} = D^{-1/2}LD^{-1/2}$ be the symmetrically normalized graph Laplacian. We repeat the same procedure as in the unnormalized spectral bipartition except that $L$ is replaced by $L_{sym}$. We find that the second smallest eigenvalue or Cheeger value $\lambda_2^C \approx 0.13227$ and its eigenvector $\boldsymbol{v}^C$ is shown in Figure 3b. We observe that $\boldsymbol{v}^C$ gives exactly the same clustering as $\boldsymbol{v}^F$. However, the ordering of the nodes are different under the two methods, and the extreme nodes 1 and 34 are closer to median in $\boldsymbol{v}^F$ than that in $\boldsymbol{v}^C$. Note that in both methods, nodes 3 and 10 which are misclassified by the minimum-cut bipartition, are also very close to the threshold 0. This reflects the hardness nature of classifying those nodes.

## 4 Transition Path Analysis

$P = D^{-1}A$ defines a probability transition matrix and hence a Markov chain $\{X_n : n \geq 0\}$ with state space $V$. Let $\tau_S := \{n > 0 : X_n \in S\}$ be the first hitting time to set $S \subseteq V$. Since we know nodes 1 and 34 belong to different club, we use the *committor function* $q_i := P(\tau_{34} < \tau_1 | X_0 = i)$ to measure node $i$'s affinity to node 34 over node 1. Conditioned on the first step transition, it follows that $q_1 = 0$, $q_{34} = 1$ and $q_i = \sum_{j \in V} P_{ij} q_j$ for $i \neq 1, 34$. Figure 4 plots the value of $q_i$ in an

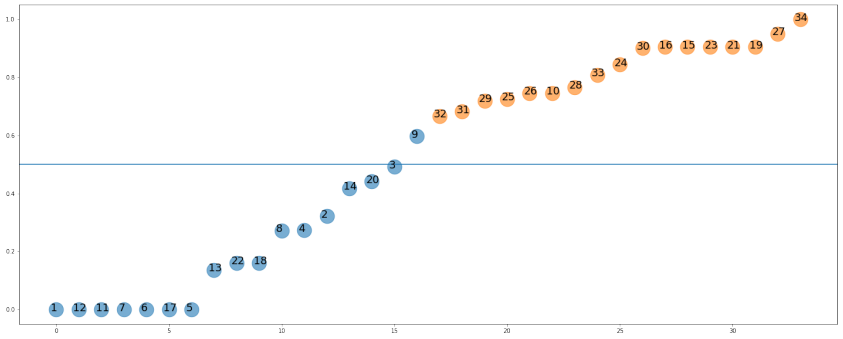
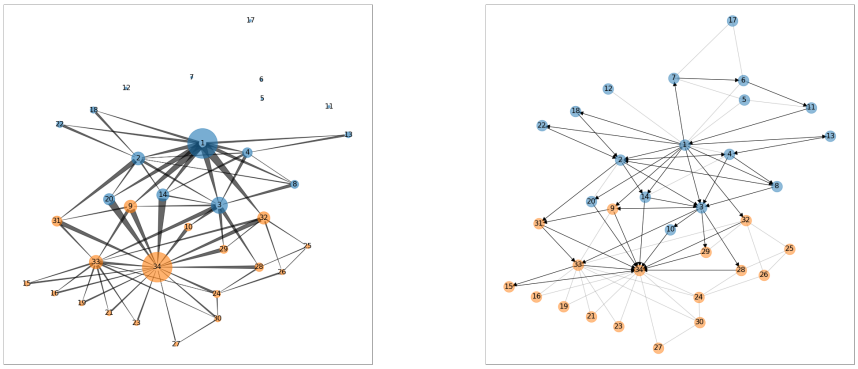



Figure 4: $q_i = P(\tau_{34} < \tau_1 | X_0 = i)$

ascending order and the color of each node represents its the ground-truth class. We observe that each node in the same cluster as 1 has a lower $q$ value than those in the opposite cluster and setting the threshold to be the median of $q$ values perfectly recovers the two clusters since in ground truth, each cluster are equal in size. Without knowing the ground-truth cluster size in advance, we may set 0.5 as a hard threshold since $q$ values are probabilities. In that case, only node 9 is misclassified and $q_3 \approx 0.49215$ is indicating a swinging state between the two clubs. Note that $q_i = 0$ for $i = 1, 12, 11, 7, 6, 17, 5$. This means that these nodes cannot reach node 34 without passing through node 1.

Since the degree (row) vector $d = (d_i)_{i \in V}$ satisfies $dP = dD^{-1}A = \mathbf{1}A = d$, the stationary distribution $\pi$ is proportional to $d$, i.e., $\pi_i = d_i \left( \sum_{j \in V} d_j \right)^{-1}$. In steady state, the *reactive current* $J_{ij} := \pi_i (1 - q_i) P_{ij} q_j$, $j \neq i$, is the rate at which a jump from $i$ to $j$ occurs on trajectory segments starting from 1 and end with 34. Cancelling out reactive current of opposite directions, we obtain the *effective current* $J_{ij}^+ := (J_{ij} - J_{ji})^+$. $J^+$ satisfies flow conservation constraint except at nodes 1 and 34 and defines a flow from 1 to 34. Let $T_i := \sum_j J_{ij}^+$, $i \neq 1, 34$, be the *transition current* (total inflow/outflow) through node $i$ and for node 1 (node 34 resp.), the transition current through it is defined as the total outflow from (inflow to) that node. The effective/transition current of an edge/node reflects the effective (removing back and forth jump at the same edge) frequency that the edge/node is visited in a trajectory segment from 1 to 34.

In Figure 5, we compare the effective flow from the random walk with the flow from the unit-capacity maximum-flow solution. In Figure 5a, we use wedge-style arrow to indicate the direction and volume



(a) Effective and Transition Current (b) Max-Flow

Figure 5: Effective current vs. Max-Flow

of the current on an edge: the current is from the edge's fat tail to its slim head and a wider end indicates a larger $J^+$ value. Comparing with the maximum flow solution, we find that for all edges that are used by both flows, the direction of flow on that edge is the same under both flows. The direction information can be regarded as an ordering structure of the graph, e.g. a directed flow from node $8$ to node $2$ indicates that $8$ is more likely to be in the same cluster as $1$ than $2$ does. The comparison suggests that the ordering induced by effective current is consistent with that by maximum flow solution.

## 5 Hierarchical Decomposition of Critical nodes

In this section, we follow [1] and decompose $G$ in a hierarchical fashion. For each node $v \in V$, let the *energy* of a node $v$, $h(v) = -d_v + \varepsilon_v$ be the negative of $v$'s degree, perturbed by a noise $\varepsilon_v$ to ensure that $h$ is injective.

- The set of *non-degenerate index-0 critical nodes* is defined as the set of local minima $\mathcal{C}_0 := \{v : h(v) < h(u),\ \forall u \in \mathcal{N}(v)\}$, where $\mathcal{N}(v)$ is the set of neighbours of $v$.
- For each locally minimal $v$, the *attraction basin* $\mathcal{A}_0(v)$ is defined to be the set of vertices (including $v$) such that $v$ is the unique local minimal that they can reach through an energy-descending path.
- Nodes that are neither a local minimal nor in the attraction basin of a local minimal form the *level-0 boundary* $\mathcal{B}_0$.

Perform the same decomposition as above on the induced subgraph of $G$ on $\mathcal{B}_0$ to obtain $\bar{\mathcal{C}}_1$ (the set of *non-degenerate index-1 critical nodes*), $\mathcal{A}_1(v)$ the *attraction basin* for each node $v \in \bar{\mathcal{C}}_1$, and the *level-1 boundary* $\mathcal{B}_1 = \mathcal{B}_0 \setminus \cup_{v \in \bar{\mathcal{C}}_1} \mathcal{A}_1(v)$. Repeat the procedure until the boundary at certain level becomes empty. The procedure produces a disjoint decomposition of $V$. We present the decomposition in Figure 6.

Direction of an edge indicates the gradient flow of $h$, in this example, the direction is from the node with a lower degree to its neighbour which has a higher degree (with random tie-breaking). Nodes that are more transparent are in the attraction basin of the node that is of the same color and shape but is more opaque. In this example, there are only two levels: the circles are level-0 critical nodes with their attraction basin, while the diamonds are level-0. Nodes $1$ and $34$ are the local minima, i.e., with the largest degrees among their neighbours. The usually misclassified nodes via previous approaches, nodes $9$ and $3$, are both on the level-0 boundary, i.e. they can reach both local minima through energy-descending paths: $9 \to 1$, $9 \to 34$, $3 \to 1$, and $3 \to 33 \to 34$. All the nodes on the level-0 boundary, nodes $9, 3, 31, 20, 14, 32$, are the nodes that are closest to the threshold $0.5$ as

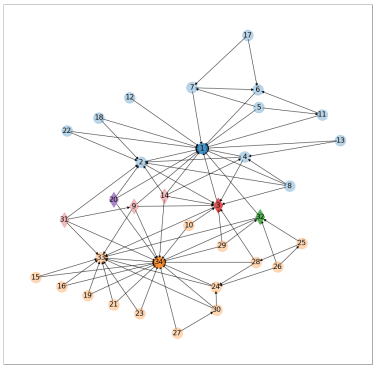


Figure 6: Critical Nodes, Attraction Basins and Boundaries

shown in Figure 5a, and are also very close to the threshold $0$ in the spectral methods as shown in Figure 3. Thus, the decomposition method provides an alternative language to explain how likely a person is in each class: those in the attraction basin are attracted while those on the boundary are swinging. We also try to give a heuristic interpretation for the choice of energy function $h$. The decomposition method tends to cluster the nodes around low energy (local minima) while the degree of a node reflects how resourceful it is inside this circle. If one member is linked with only one of the most resourceful person (instructor/president) through a sequence of more resourceful members, he is very likely to join the club led by that person.

## 6 Conclusion and Potential Direction

In this project, we examine four alternative methods to analyse Karate Club network. All four methods perform well in general and are consistent in illustrating the hardness of classifying certain nodes, e.g. nodes $3$ and $9$, as they are really on the "boundary" of the two clusters. The first method can only output a bipartition, while the latter three methods are able to provide additional description on how likely each node belongs to a cluster: either by how far away it is from a numerical threshold, or by whether it resides in the attraction basin or on the boundary. Thus, it may be a potential direction to consider reasonable capacity assignment in minimum-cut problem to produce confidence level for the classification of each node. Both the minimum-cut bipartition and the transition path analysis are of semi-supervised nature: we may take advantage of the given knowledge that some two nodes are in distinct cluster and incorporate this information into the algorithm, while it is unclear how to encode this information into spectral method and hierarchical method as prior knowledge.

## References


[1] Weinan, E., Jianfeng Lu, and Yuan Yao. "The landscape of complex networks." arXiv preprint arXiv:1204.6376 (2012).

[2] Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.