Introduction and Abstract

Spectral clustering is a common clustering primitive, but in spite of widespread usage there are few theoretical guarantees about the quality of the clustering output. Here we focus on spectral partitioning of graphs. In particular, we outline [DRS14] and show that given a sufficiently large gap in the spectrum between the $k$-th and $k+1$-th eigenvalues of the normalized graph Laplacian, a simple greedy algorithm gives provably good results.

Measuring Partition Quality

Intuitively, a graph has a "good" partition into $k$ clusters if we can separate the graph into $k$ pieces of approximately the same size by cutting relatively few edges. Further, those pieces should be "hard" to further bisect into two large pieces. This intuition is formalized as follows:

Let $G$ be a graph on $n$ vertices with vertex set $V(G)$, and let $S \subseteq V(G)$. For a subset $S \subseteq V$, the external conductance and internal conductance are defined to be

$$\phi_{\text{ext}}(S, G) := \frac{|E(S, V(G) \setminus S)|}{\text{vol}(S)}$$

$$\phi_{\text{int}}(S) := \min_{S \subseteq S \subseteq \overline{S}} \phi_{\text{ext}}(S, G[S])$$

where $E(S, V(G) \setminus S)$ denotes the set of edges which are incident to vertices in $S$ and $V(G) \setminus S$ and vol($S$) is the sum total of vertex degrees for vertices in $S$.

We define a $k$-partition to be a partition $\mathcal{A} = \{A_1, \ldots, A_k\}$ of $V(G)$ into $k$ disjoint subsets. We say that $\mathcal{A}$ is $(\alpha_n, \alpha_{\text{ext}})$-good, for some $\alpha_n, \alpha_{\text{ext}} \geq 0$, if for all $i \in \{1, \ldots, k\}$, we have

$$\phi_{\text{int}}(A_i) \geq \alpha_n$$

$$\phi_{\text{ext}}(A_i, A_j) \leq \alpha_{\text{ext}}$$

otherwise.

The Normalized Graph Laplacian

Label the vertices in $V(G)$ arbitrarily so that $V(G) = \{v_1, v_2, \ldots, v_n\}$, then for any $i, j \in \{1, \ldots, n\}$, the $i, j$-th entry of the normalized graph Laplacian of $G$, $L_G$, is

$$(L_G)_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ \frac{\deg(v_i)}{\deg(v_i) + \deg(v_j)} & \text{if } i \neq j \text{ and } v_j \text{ is adjacent to } v_i \\ 0 & \text{otherwise.} \end{cases}$$

where $\deg(v)$ denotes the degree of vertex $v$ in $G$.

Existential Guarantees

Oveis Gharan & Trevisan recently gave a condition which is sufficient for the existence of a good partition:

**Theorem [OT14]** Let $G$ be a graph and let $\lambda_k(L_G)$ denote the $k$-th eigenvalue of $L_G$ when ordered from smallest to largest (possibly with repetition). There exists a universal constant $c > 0$ such that for any graph $G$ with $\lambda_k(L_G) > c^2 \lambda_{k+1}(L_G)$, there exists a $k$-partition of $G$ that is $(0, \alpha_{\text{ext}})$-good, for $\alpha_{\text{ext}} = \epsilon^2$.

It is a natural question to consider how well spectral clustering algorithms work in practice when the input is guaranteed to have a good clustering. We take the following simple greedy $k$-Clustering algorithm as our prototype.

Algorithm: Greedy $k$-Clustering

**Input:** Graph $G$, $k \in \mathbb{Z}^+$, $R \in \mathbb{R}$

**Output:** Partition $C = \{C_1, \ldots, C_k\}$ of $V(G)$

Let $\xi_1, \ldots, \xi_k$ be the $k$ first eigenvectors of $L_G$ with each $\xi_i$ corresponding to $\lambda_i(L_G)$.

Let $f: V(G) \rightarrow \mathbb{R}^k$ where for any $u \in V(G)$,

$$f(u) = (\langle \xi_1(u), \deg(u)^{-1/2}\rangle, \ldots, \langle \xi_k(u), \deg(u)^{-1/2}\rangle)$$

$$V_i = f^{-1}(\{u \in V(G) : f(u) \cdot f(u) \leq R\})$$

for $i = 1, \ldots, k - 1$

$$u_i = \arg\min_{V_{i+1}} \|f(u) - f(V_{i+1})\|_2$$

$$C_i = f^{-1}(\{u \in V_{i+1} : f(u) \cdot f(V_{i+1}) \leq 2R\})$$

$$V_k = V_{k-1} \cup C_k$$

**Greedy $k$-Clustering in Action**

To help understand the function of the above algorithm observe the following example.

Greedy $k$-Clustering is run on the 1-skeleton of a triangulated mesh for $k = 3$ and a hand-tuned value of $R$. The left shows the result of the clustering procedure, as well as a graph of $\lambda_k(L_G)$ for $k \in \{1, \ldots, 20\}$. The right shows the same clustering on the spectral embedding of $G$. Here, the ball with the greatest mass comes from the center of the embedding (red), while the subsequent iteration corresponds to one lower leg.

Analysis of Greedy $k$-Clustering

Let $\mathcal{A}$ denote the clustering given by [OT14]. Each of the $k$ clusters in $\mathcal{A}$ concentrate around $k$ distinct points in this spectral embedding.

In particular, let $d_{\text{max}}$ denote the maximum degree of $G$, let $k \geq 1$, and $\tau > 0$ such that $\lambda_k(L_G) > \tau \lambda_{k+1}(L_G)$. Then any unit vector $x \in \text{span}(D^{-1/2} \xi_1, \ldots, D^{-1/2} \xi_k)$ there is a vector $x$ which is constant on $A_i \in \mathcal{A}$ and for which $\|x - \hat{x}\| \leq \Delta$ where

$$\Delta = 1/c + \epsilon' \sum_{i=1}^k |a_i|^2.$$ 

Further, the clusters are sufficiently separated. Since concentration is in terms of $k$-mass we proceed in a way that permits outliers. We argue that there exist $k$ distinct points $p_1, \ldots, p_k \in \mathbb{R}^k$ such that we define a new clustering $\mathcal{A}' = \{A_1', \ldots, A_k\}$ where $A_i' = \text{ball}(p_i, R)$ then for $R = R_0 := (d_{\text{max}}^2 - 2k\sqrt{R})/13kR$ we have for $i \neq j$, $\|p_i - p_j\| > 6R$ and $\mathcal{A}'$ differs from $\mathcal{A}$ in symmetric difference by at most

$$O(d_{\text{max}}k + n d_{\text{max}}k^2/\log n).$$

By establishing an upper bound on symmetric difference of $\mathcal{A}'$ and the output of Greedy $k$-Clustering we observe the following example.

A Provable Guarantee

Let $\mathcal{A}$ denote the clustering given by [OT14], then greedy $k$-Clustering of $G$ with $R = R_0$ outputs a partition $\hat{C}$ that is provably close to $\mathcal{A}$. In particular,

$$|A \Delta \hat{C}| = O(d_{\text{max}}k + n d_{\text{max}}k^2/\log n).$$

Sample $k$-Clusterings

Here we run the algorithm on a graph with multiple significant spectral gaps. Above are graphs with 5 groups of nodes. Edges within groups appear with highest probability. Additional edges are much more likely to appear among the three clusters on the right and between the two on the left than otherwise. This induces two prominent spectral gaps: $k = 2$ and $k = 6$.

References
