k-way conductance to find k-clusters

Anmol Anand
aanand@tamu.edu
Advisor: Dr. Nate Veldt
Department of Computer Science
Texas A&M University

Abstract

Dividing a graph into k-clusters can be defined as dividing the vertex set into k mutually-disjoint subsets such that each induced subgraph is dense and sparsely connected to the rest of the original graph. For $k = 2$, the conductance function helps us measure how good a 2-clustering is [7] [2] [6]. This paper defines and explores a novel k-way conductance function that will help us measure how good a k-clustering is.

1 Definitions and goals

Define $\mathscr{S}_k$ as a k-way partition of V into non-empty, mutually disjoint subsets $\{S_i\}_{i=1}^k$:

$$\begin{aligned} & \mathscr{S}_k = \{S_1, S_2, ..., S_k\} \\ \text{s.t.} \quad & S_i \subseteq V \text{ and } S_i \neq \emptyset \; \forall \, i \\ & \bigcup_{i=1}^k S_i = V \\ & S_i \bigcap S_j = \emptyset, \forall \, i \neq j \end{aligned}$$

We define a novel k-way conductance function ψ over $\mathscr{S}_k$ as follows:

$$\psi(\mathscr{S}_k) = \frac{1}{k} \cdot \frac{\sum_{i=1}^k cut(S_i)}{\min_{i=1}^k vol(S_i)} \tag{1}$$

Ideally, the goal would be to find a good k-clustering in the form of an optimal k-way partition $\mathscr{S}_k^*$:

$$\mathscr{S}_k^* = \text{argmin}_{\mathscr{S}_k} \; \psi(\mathscr{S}_k)$$

It can be seen that for $k = 2$, the above problem is the minimum conductance problem.

<u>**Goals**</u>:

- Compare ψ with other existing k-way conductance functions. Present examples to compare k-clustering achieved by different k-way conductance functions.
- Analyse the complexity of the minimum k-way conductance problem.
- State and prove the bounds on the k-way conductance ψ (Cheeger's inequality equivalent).

2 Previous work

There have been other functions to measure how good a k-clustering is. One such measure is the following k-way conductance function [3]:

$$\phi(\mathscr{S}_k) = \max_{1 \leq i \leq k} \frac{cut(S_i)}{min\{vol(S_i), vol(\overline{S_i})\}} \tag{2}$$

Once again, the goal would be to find a good k-clustering in the form of an optimal k-way partition $\mathscr{S}_k^*$:

$$\mathscr{S}_k^* = \text{argmin}_{\mathscr{S}_k}\ \phi(\mathscr{S}_k)$$

i.e. we choose $\mathscr{S}_k^*$ to minimize the conductance of the maximum conductance subset.

It can be seen that for $k = 2$, once again, the above problem is the minimum conductance problem.

Similarities between ψ and ϕ: We can see that these two k-way conductance functions are very similar in the ways they are used. The goal in both cases is to find the k-way partition that minimizes these conductance functions. For $k = 2$, both of them become equal to the usual conductance function.

3 Complexity analysis

Both k-way conductance functions ψ in (1), and ϕ in (2) become equal to the usual conductance function if we set $k = 2$. And because minimizing conductance is an $\mathcal{NP}$-hard problem [8], minimizing k-way conductance ψ or ϕ is also an $\mathcal{NP}$-hard problem. Although, solutions to the decision version of both these k-way conductance problems are verifiable in polynomial time (to be precise, in $O(|V| + |E|)$ time for either of them). Therefore, both these k-way conductance problems are $\mathcal{NP}$-complete.

4 *k*-way Cheeger's inequality

Let $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_{|V|}$ be the eigenvalues of the normalized laplacian matrix $\mathcal{L}$ of graph $G(V, E)$. Then:

Theorem 1. ***Cheeger's inequality*** *[1] [3]*

For a two-way partition $\{S, \overline{S}\}$ on the graph G:

$$\frac{\lambda_2}{2} \leq \frac{cut(S)}{min\{vol(S), vol(\overline{S})\}} \leq \sqrt{2 \cdot \lambda_2} \tag{3}$$

Theorem 2. ***Cheeger's inequality over k-way conductance*** ϕ [***3***] [***5***] [***4***]

For any k-way partition $\mathscr{S}_k$ on the graph G:

$$\frac{\lambda_k}{2} \leq \phi(\mathscr{S}_k) \leq O(k^2)\sqrt{\lambda_k} \tag{4}$$

Theorem 3. ***Cheeger's inequality over k-way conductance*** ψ

For any k-way partition $\mathscr{S}_k$ on the graph G:

$$\frac{\lambda_2}{k \cdot (k-1)} \leq \psi(\mathscr{S}_k) \tag{5}$$

Theorem 3 defines only the lower-bound on the k-way conductance ψ. I could not complete the proof for an upper-bound. A novel proof of the lower bound in Theorem 3 is presented in the appendix.

As we can see, equations (4) and (5) are consistent with equation (3) if we set $k = 2$.

5 Comparison with examples

We computed the optimal k-way partitions of some graphs based on the k-way conductance functions ψ and ϕ respectively. Since both these problems are hard, we use a brute force approach where we compute k-way conductance for all possible k-way partitions. The results are decent and similar for both approaches i.e. while minimizing ψ and ϕ respectively.

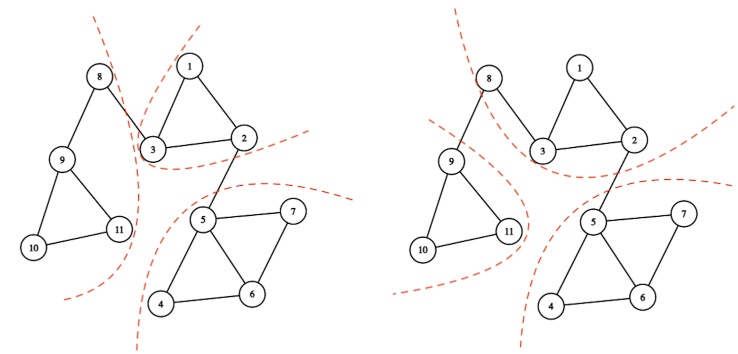

Figure 1: Optimal 3-way partitions of the same graph based on minimum ψ (left) and minimum ϕ (right) respectively.

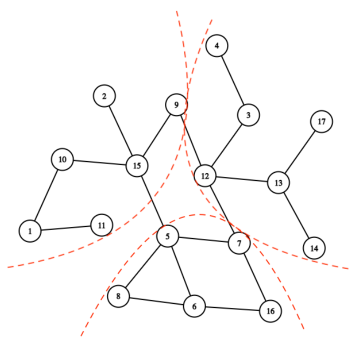

Figure 2: The above 3-way partition leads to the optimal value of both conductance functions ψ and ϕ respectively.

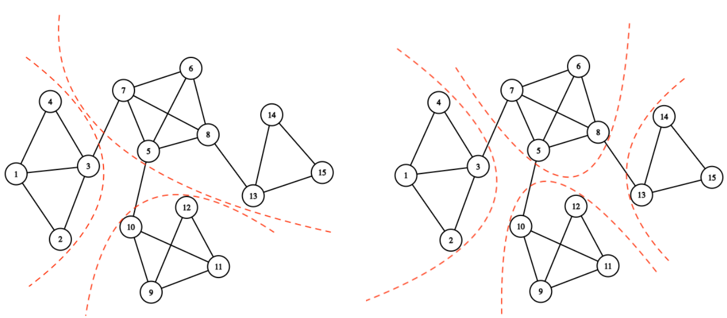

Figure 3: The above 3-way (left) and 4-way (right) partitions of the same graph lead to the optimal value of both conductance functions ψ and ϕ respectively.

References

[1] Fan RK Chung. Laplacians of graphs and cheeger's inequalities. *Combinatorics, Paul Erdos is Eighty*, 2(157-172):13–2, 1996.

[2] R Kannan, S Vempala, and A Vetta. On clusterings: Good, bad and spectral. In *Proceedings of the FOCS'00 Forty-First Annual Symposium on the Foundation of Computer Science*, pages 367–377. New York: IEEE Computer Society Press, 2000.

[3] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. Improved cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. *CoRR*, abs/1301.5584, 2013.

[4] James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities, 2011.

[5] Anand Louis, Prasad Raghavendra, Prasad Tetali, and Santosh S. Vempala. Many sparse cuts via higher eigenvalues. *CoRR*, abs/1111.0965, 2011.

[6] U von Luxberg. A tutorial on spectral clustering. In *Statistics and Computing 17(4)*, pages 395–416, 2007.

[7] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2001.

[8] Jiří Šíma and Satu Elisa Schaeffer. On the np-completeness of some graph cluster measures. In Jiří Wiedermann, Gerard Tel, Jaroslav Pokorný, Mária Bieliková, and Július Štuller, editors, *SOFSEM 2006: Theory and Practice of Computer Science*, pages 530–537, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Appendix

A novel proof of the lower bound in Theorem 3.

Define the Rayleigh quotient $R : \mathbb{C}^{|V|} \to \mathbb{R}_+$

$$R(\mathbf{y}) = \frac{\mathbf{y}^\dagger \mathbf{L} \mathbf{y}}{\mathbf{y}^\dagger \mathbf{D} \mathbf{y}} \tag{6}$$

where $\mathbf{y}^\dagger$ is the conjugate transpose of the complex vector $\mathbf{y}$, $\mathbf{L}$ is the laplacian matrix, and $\mathbf{D}$ is the degree matrix of the graph G.

Define $\mathbf{y}_{\mathscr{S}_k} \in \mathbb{C}^{|V|}$ using $\mathscr{S}_k = \{S_1, S_2, ..., S_k\}$ as follows:

$$\mathbf{y}_{\mathscr{S}_k}(u) = \frac{\omega^i}{vol(S_i)}, \text{ if } u \in S_i \subseteq V \tag{7}$$

where ω is the k^{th} root of unity.

The second largest eigenvalue (λ_2) of the normalized laplacian matrix of G ($\mathcal{L}$) can be computed as:

$$\lambda_2 = \min_{\mathbf{y}:\, \mathbf{y} \neq 0,\, \mathbf{y}^\dagger \mathbf{d} = 0} R(\mathbf{y}) \tag{8}$$

It can be seen that ${\mathbf{y}_{\mathscr{S}_k}}^\dagger \mathbf{d} = 0$:

$$\begin{aligned}
{\mathbf{y}_{\mathscr{S}_k}}^\dagger \mathbf{d} &= \sum_{u \in V} \overline{\mathbf{y}_{\mathscr{S}_k}(u)} \cdot d_u \\
\implies {\mathbf{y}_{\mathscr{S}_k}}^\dagger \mathbf{d} &= \sum_{i=1}^{k} \sum_{u \in S_i} \frac{\overline{\omega^i}}{vol(S_i)} \cdot d_u \\
\implies {\mathbf{y}_{\mathscr{S}_k}}^\dagger \mathbf{d} &= \sum_{i=1}^{k} \frac{\overline{\omega^i}}{vol(S_i)} \cdot \left(\sum_{u \in S_i} d_u \right) \\
\implies {\mathbf{y}_{\mathscr{S}_k}}^\dagger \mathbf{d} &= \sum_{i=1}^{k} \frac{\overline{\omega^i}}{vol(S_i)} \cdot vol(S_i) \\
\implies {\mathbf{y}_{\mathscr{S}_k}}^\dagger \mathbf{d} &= \sum_{i=1}^{k} \overline{\omega^i}
\end{aligned}$$

$$\implies \mathbf{y}_{\mathscr{S}_k}{}^{\dagger}\mathbf{d} = 0 \tag{9}$$

Since, $\mathbf{y}_{\mathscr{S}_k} \neq 0$ and $\mathbf{y}_{\mathscr{S}_k}{}^{\dagger}\mathbf{d} = 0$:

$$R(\mathbf{y}_{\mathscr{S}_k}) \geq \min_{\mathbf{y}:\, \mathbf{y}\neq 0,\, \mathbf{y}^{\dagger}\mathbf{d}=0} R(\mathbf{y})$$

The RHS is the second smallest eigenvalue of $\mathcal{L}$ (from (8)):

$$\implies R(\mathbf{y}_{\mathscr{S}_k}) \geq \lambda_2 \tag{10}$$

Computing $R(\mathbf{y}_{\mathscr{S}_k})$:

$$R(\mathbf{y}_{\mathscr{S}_k}) = \frac{(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k})}{(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{D}\, (\mathbf{y}_{\mathscr{S}_k})} \tag{11}$$

Computing $(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k})$:

$$(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{uv\in E} ||\mathbf{y}_{\mathscr{S}_k}(u) - \mathbf{y}_{\mathscr{S}_k}(v)||^2$$

$$\implies (\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{i<j} \sum_{\substack{uv\in E \\ uv\in S_i\times S_j}} \left|\left| \frac{\omega^i}{vol(S_i)} - \frac{\omega^j}{vol(S_j)} \right|\right|^2$$

$$\implies (\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{i<j} \left(\left|\left| \frac{\omega^i}{vol(S_i)} - \frac{\omega^j}{vol(S_j)} \right|\right|^2 \cdot \sum_{\substack{uv\in E \\ uv\in S_i\times S_j}} 1 \right)$$

$$\implies (\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) \leq \left(\sum_{i<j} \left|\left| \frac{\omega^i}{vol(S_i)} - \frac{\omega^j}{vol(S_j)} \right|\right|^2 \right) \cdot \left(\sum_{i<j} \sum_{\substack{uv\in E \\ uv\in S_i\times S_j}} 1 \right) \tag{12}$$

$$\sum_{i<j} \sum_{\substack{uv\in E \\ uv\in S_i\times S_j}} 1 = \frac{1}{2} \cdot \sum_{i\neq j} \sum_{\substack{uv\in E \\ uv\in S_i\times S_j}} 1$$

$$\implies \sum_{i<j} \sum_{\substack{uv\in E \\ uv\in S_i\times S_j}} 1 = \frac{1}{2} \cdot \sum_{i} cut(S_i) \tag{13}$$

From (12) and (13):

$$(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) \leq \left(\sum_{i<j} \left|\left| \frac{\omega^i}{vol(S_i)} - \frac{\omega^j}{vol(S_j)} \right|\right|^2 \right) \cdot \frac{1}{2} \cdot \sum_{i} cut(S_i)$$

For $a, b \in \mathbb{C}$, $||a-b||^2 \leq (||a|| + ||b||)^2$:

$$\implies (\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) \leq \left(\sum_{i<j} \left(\left|\left| \frac{\omega^i}{vol(S_i)} \right|\right| + \left|\left| \frac{\omega^j}{vol(S_j)} \right|\right| \right)^2 \right) \cdot \frac{1}{2} \cdot \sum_{i} cut(S_i)$$

$$\implies (\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{L}\, (\mathbf{y}_{\mathscr{S}_k}) \leq \left(\sum_{i<j} \left(\frac{1}{vol(S_i)} + \frac{1}{vol(S_j)} \right)^2 \right) \cdot \frac{1}{2} \cdot \sum_{i} cut(S_i) \tag{14}$$

Computing $(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{D}\, (\mathbf{y}_{\mathscr{S}_k})$:

$$(\mathbf{y}_{\mathscr{S}_k})^{\dagger}\, \mathbf{D}\, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{u\in V} d_u ||\mathbf{y}_{\mathscr{S}_k}(u)||^2$$

$$
\begin{aligned}
&\implies (\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{D} \, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{i=1}^{k} \sum_{u \in S_i} d_u \left|\left| \frac{\omega^i}{vol(S_i)} \right|\right|^2 \\
&\implies (\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{D} \, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{i=1}^{k} \sum_{u \in S_i} d_u \left(\frac{1}{vol(S_i)} \right)^2 \\
&\implies (\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{D} \, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{i=1}^{k} vol(S_i) \left(\frac{1}{vol(S_i)} \right)^2 \\
&\implies (\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{D} \, (\mathbf{y}_{\mathscr{S}_k}) = \sum_{i=1}^{k} \frac{1}{vol(S_i)}
\end{aligned} \tag{15}
$$

Let $Sum_1 = \sum_i \frac{1}{vol(S_i)}$ and let $Sum_2 = \sum_i \frac{1}{vol(S_i)^2}$. Then:

$$
\sum_{i<j} \left(\frac{1}{vol(S_i)} + \frac{1}{vol(S_j)} \right)^2 = (k-2) \cdot Sum_2 + (Sum_1)^2, \text{ where } k = |\mathscr{S}| \tag{16}
$$

Finding an upper bound on $(k-2) \cdot Sum_2$ as follows:

$$
\begin{aligned}
&Sum_2 \leq Sum_1 \cdot \max_i \left(\frac{1}{vol(S_i)} \right) \\
&\implies (k-2) \cdot Sum_2 \leq (k-2) \cdot Sum_1 \cdot \max_i \left(\frac{1}{vol(S_i)} \right)
\end{aligned} \tag{17}
$$

Finding an upper bound on $(Sum_1)^2$ as follows:

$$
\begin{aligned}
&Sum_1 \leq k \cdot \max_i \left(\frac{1}{vol(S_i)} \right) \\
&\implies (Sum_1)^2 \leq k \cdot Sum_1 \cdot \max_i \left(\frac{1}{vol(S_i)} \right)
\end{aligned} \tag{18}
$$

Adding (17) and (18):

$$
(k-2) \cdot Sum_2 + (Sum_1)^2 \leq (2k-2) \cdot Sum_1 \cdot \max_i \left(\frac{1}{vol(S_i)} \right) \tag{19}
$$

From (14), (16), and (19):

$$
\begin{aligned}
&(\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{L} \, (\mathbf{y}_{\mathscr{S}_k}) \leq (2k-2) \cdot Sum_1 \cdot \max_i \left(\frac{1}{vol(S_i)} \right) \cdot \frac{1}{2} \cdot \sum_i cut(S_i) \\
&\implies (\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{L} \, (\mathbf{y}_{\mathscr{S}_k}) \leq (k-1) \cdot \left(\sum_i \frac{1}{vol(S_i)} \right) \cdot \max_i \left(\frac{1}{vol(S_i)} \right) \cdot \sum_i cut(S_i)
\end{aligned} \tag{20}
$$

From (15) and (20):

$$
\begin{aligned}
&\frac{(\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{L} \, (\mathbf{y}_{\mathscr{S}_k})}{(\mathbf{y}_{\mathscr{S}_k})^\dagger \, \mathbf{D} \, (\mathbf{y}_{\mathscr{S}_k})} \leq (k-1) \cdot \max_i \left(\frac{1}{vol(S_i)} \right) \cdot \sum_i cut(S_i) \\
&\implies R(\mathbf{y}_{\mathscr{S}_k}) \leq (k-1) \cdot \frac{\sum_i cut(S_i)}{\min_i vol(S_i)}
\end{aligned} \tag{21}
$$

From (10) and (21):

$$
\begin{aligned}
&\frac{\lambda_2}{k \cdot (k-1)} \leq \frac{1}{k} \cdot \frac{\sum_i cut(S_i)}{\min_i vol(S_i)} \\
&\implies \frac{\lambda_2}{k \cdot (k-1)} \leq \psi(\mathscr{S}_k)
\end{aligned} \tag{22}
$$

Therefore, we have found a lower bound on the generalized conductance function.