CSC 2421H : Graphs, Matrices, and Optimization Lecture 2 : 17 Sep 2018

Clustering

Lecturer: Sushant Sachdeva *Scribe: Liwen Lu*

1 Clustering of a Graph and Eigenvalue

Informally, "clustering" is grouping vertices such that there are more connectivity inside each group compare to between groups.

1.1 Conductance

Definition 1.1. Volume*: let* $S \subseteq V$, $Vol(S) \stackrel{\text{def}}{=} \sum_{v \in S} d_v$ *where* $d_v = deg(v) = D_{v,v}$

Lemma 1.2. *Define the indicator of set* S *as*

$$\vec{\mathbf{1}}_S(v) = \begin{cases} 1, & v \in S \\ 0, & otherwise \end{cases}$$

Then

$$\vec{\mathbf{1}}_S^T D \vec{\mathbf{1}}_S = Vol(S)$$

Definition 1.3. *Define* Boundary of S *as*

$$E(S, \overline{S}) = \{(u,v) | (u,v) \in E, u \in S, v \in V \setminus S\}$$

, where $\overline{S}$ *is the complement set.*

Definition 1.4. *Define* conductance of S measured in graph G *to be*

$$\phi_G(S) \stackrel{\text{def}}{=} \frac{|E(S, \overline{S})|}{min\{Vol(S), Vol(V \setminus S)\}},$$

and define conductance of graphc G *to be*

$$\phi(G) \stackrel{\text{def}}{=} \min_{\substack{S \subseteq V \\ S \neq \emptyset, V}} \phi_G(S)$$

Computing $\phi(G)$ is called "minimum conductance problem", and it is famously NP hard. However, we can connect conductance and eigenvalues.

1.2 Normalized Laplacian Matrix

Let $\nu_2 = \min_{y^T D \vec{\mathbf{1}}} \left(\frac{y^T L y}{y^T D y} \right)$.

Define $x = D^{\frac{1}{2}} y$ (note $D^{\frac{1}{2}}$ is replacing each of the diagonal entries in diagonal matrix D with square root), we can write $y^T D y = x^T x$.

Note we are assuming the graph G is connected, then diagonal of D is strictly non-negative, so the mapping between x and y is a bijection, which gives us $y^T D y = x^T x \Leftrightarrow y = D^{-\frac{1}{2}} x$. Then we can rewrite ν_2 as

$$\nu_2 = \min_{x^T \left(D^{\frac{1}{2}} \vec{\mathbf{1}} \right)} \frac{x^T D^{-\frac{1}{2}} L D^{-\frac{1}{2}} x}{x^T x}$$

Definition 1.5. *Define* Normalized Laplacian Matrix *as*

$$\boldsymbol{N} \stackrel{\text{def}}{=} D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

We also claim that:

$$\begin{aligned} \lambda_1(\boldsymbol{N}) &= 0 \\ \psi_1 &= D^{\frac{1}{2}} \vec{\mathbf{1}} \\ \nu_2 &= \min_{x^T \psi_1 = 0} \frac{x^T \boldsymbol{N} x}{x^T x} \\ &= \lambda_2(\boldsymbol{N}) \end{aligned}$$

1.3 Cheegar's Inequality

Theorem 1.6.

$$\frac{\nu_2}{2} \leq \phi(G) \leq \sqrt{2\nu_2}, \nu_2 \leq 2$$

For example, if $\nu_2 = 0.01 \Rightarrow 0.05 \leq \phi(G) \leq 0.14$
We'll first prove the left hand side, $\frac{\nu_2}{2} \leq \phi(G)$.

Proof. By definition, we know $\exists S s.t. \phi(G) = \phi_G(S) = \frac{|E(S,\overline{S})|}{Vol(S)}$ and $Vol(S) \leq Vol(\overline{S})$.
Recall $\nu_2 = \min\limits_{y^T D \vec{\mathbf{1}}} \left(\frac{y^T L y}{y^T D y} \right)$ and $\vec{\mathbf{1}}_S^T D \vec{\mathbf{1}}_S = Vol(S)$, we have

$$(\vec{1})_S^T L (\vec{1})_S = \sum_{(u,v) \in E} \left(\vec{\mathbf{1}}_S(u) - \vec{\mathbf{1}}_S(v) \right)^2$$

note

$$\left(\vec{\mathbf{1}}_S(u) - \vec{\mathbf{1}}_S(v) \right)^2 = \begin{cases} 1, & \vec{\mathbf{1}}_S(v) \neq \vec{\mathbf{1}}_S(v) \Leftrightarrow (u,v) \in E(S,\overline{S}) \\ 0, & \text{otherwise} \end{cases}$$

therefore

$$(\vec{1})_S^T L (\vec{1})_S = \sum_{(u,v) \in E} \vec{\mathbf{1}}[(u,v) \in E(S,\overline{S})] = |E(S,\overline{S})|$$

If we set $y = \vec{\mathbf{1}}_S$, it would minimize $\frac{y^T L y}{y^T D y}$, however it may not satisfy $y^T D \vec{\mathbf{1}}$. In order to make y satisfy $y^T D \vec{\mathbf{1}}$, let $y = \vec{\mathbf{1}}_S + c\vec{\mathbf{1}}$, note we still have $y^T L y = |E(S,\overline{S})|$.

$$y^T D \vec{\mathbf{1}} = 0 \Leftrightarrow c = \frac{-\vec{\mathbf{1}}_S^T D \vec{\mathbf{1}}}{\vec{\mathbf{1}}^T D \vec{\mathbf{1}}} = -\frac{Vol(S)}{Vol(V)} \qquad \text{(by previous lemma)}$$

$$
\begin{aligned}
y^T D y &= \vec{\mathbf{1}}_S^T D \vec{\mathbf{1}}_S + 2c\vec{\mathbf{1}}_S^T \vec{\mathbf{1}} + c^2 \vec{\mathbf{1}}^T D \vec{\mathbf{1}} \\
&= \vec{\mathbf{1}}_S^T - \frac{\left(\vec{\mathbf{1}}_S^T D \vec{\mathbf{1}}\right)^2}{\vec{\mathbf{1}}^T D \vec{\mathbf{1}}} \\
&= \vec{\mathbf{1}}_S^T - \frac{Vol(S)^2}{Vol(V)} \\
&= \vec{\mathbf{1}}_S^T \left(1 - \frac{Vol(S)}{Vol(V)}\right) \\
&\geq \frac{Vol(S)}{2} && \text{(since } Vol(S) \leq Vol(\overline{S})\text{)}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\frac{y^T L y}{y^T D y} &= \frac{|E(S, \overline{S})|}{y^T D y} \\
&\leq \frac{2|E(S, \overline{S})|}{Vol(S)} \\
&= 2\phi_G(S) \\
&= 2\phi(G) && \text{(by assumption)} \\
\Rightarrow \lambda_2 &= \min_{y^T D \vec{\mathbf{1}}} \frac{y^T L y}{y^T D y} \leq 2\phi(G) \\
\Rightarrow \frac{\lambda_2}{2} &\leq \phi(G)
\end{aligned}
$$

□

Lemma 1.7. *Given y s.t. $y^T D \vec{\mathbf{1}} = 0$, we can find a distribution on t with $S_t \subseteq V$ s.t. $Vol(S_t) \leq \frac{Vol(V)}{2}$, and*

$$
\frac{\mathbb{E}_t |E(S_t, \overline{S_t})|}{\mathbb{E}_t Vol(S_t)} \leq \sqrt{\frac{2 y^T L y}{y^T D y}}
$$

Let t be independent choices, P_t be distribution of t, X_t, Y_t are variables depend on t with $Y_t \geq 0$.
We can prove that $\exists t$ s.t.

$$
\frac{X_t}{Y_t} \leq \frac{\mathbb{E}_t X_t}{\mathbb{E}_t Y_t} = \frac{\sum_t P_t X_t}{\sum_t P_t Y_t}
$$

The proof is left as exercise. This implies that $\exists t$, s.t.

$$
\frac{|E(S_t, \overline{S_t})|}{Vol(S_t)} \leq \sqrt{\frac{2 y^T L y}{y^T D y}}
$$

Furthermore, if y was the minimizing vector for ν_2, then

$$
\phi_S(S_t) = \frac{|E(S_t, \overline{S_t})|}{Vol(S_t)} \leq \sqrt{2\nu_2}
$$