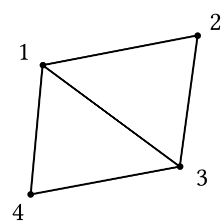
[CS1961: Lecture 15] Random Walk, Cheeger's Inequality

1 Markov Chain

In this lecture, we will introduce another convenient normalization of weighted adjacency matrix. This can be best described as a random walk on the graph. We will introduce some basic terminologies on random walk and Markov chains.

1.1 Random Walk on Undirected Graph

Consider a random walk on the following undirected graph. We start at $X_0 = 1$ and move to a neighbor of the current vertex u.a.r. at each step. The distribution of the next position X_{t+1} is determined only by the current state. This random walk is a simple *Markov chain.*



Definition 1 (Markov Chain)**.** *A sequence of random variables* $X_0, X_1, \dots, X_t, X_{t+1}, \dots$ *is a Markov chain if for any* $t \in \mathbb{N}$ *and any states* $j_0, j_1, \dots, j_t, j$*,*

$$\mathbf{Pr}\left[X_{t+1} = j \mid X_t = j_t, X_{t-1} = j_{t-1}, \dots, X_0 = j_0\right] = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = j_t\right].$$

A Markov chain can be characterized by a matrix $P = \left(p_{ij}\right)_{i,j\in\Omega} \in [0,1]^{\Omega\times\Omega}$ where $p_{ij} = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = i\right]$. The transition matrix P is a stochastic matrix since $\sum_{j\in\Omega} p_{ij} = 1$ for all $i \in \Omega$. For example, in the above random walk, we have $\Omega = [4]$ and

$$p_{ij} = \begin{cases} 0, & \text{if } i \nsim j \\ \frac{1}{\deg(i)}, & \text{if } i \sim j \end{cases}.$$

Sometimes we will simply denote the transition matrix P as the Markov chain for convenience.

Let $A_G = \left(w_{ij}\right)_{i,j\in[n]}$ be a weighted graph where every $w_{ij} \geq 0$. We can normalize it into a random walk P_G:

- For every $i \in [n]$, let $w_i = \sum_j w_{ij}$;
- For every $i, j \in [n]$, let $P_G(i,j) = \frac{w_{ij}}{w_i}$.

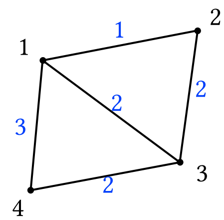
We also define a distribution π over $[n]$ as $\pi(i) = \frac{w_i}{\sum_{j\in[n]} w_j}$.

As a result, the P_G constructed above satisfies that for every $i, j \in [n]$,

$$\pi(i) \cdot P_G(i,j) = \pi(j) \cdot P_G(j,i) (= w_{ij}).$$

This is called *reversibility* of P_G. The distribution π is called a *stationary distribution* of P_G.

For example, in the following graph, we have $P_G(1,2) = \frac{1}{1+2+3} = \frac{1}{6}$, $P_G(1,3) = \frac{1}{3}$ and $P_G(1,4) = \frac{1}{2}$.



2 The Spectrum of Markov Chains

2.1 Spectral Decomposition

Another advantage to use reversible chains is that their transition matrices are symmetric in some sense. Suppose P is reversible with respect to π. Let $\Pi = \mathrm{diag}(\pi)$ be the diagonal matrix with $\Pi(i,i) = \pi(i)$. Define $Q = \Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$, then we can verify that Q is symmetric:

$$Q(i,j) = \pi(i)^{\frac{1}{2}} P(i,j) \pi(j)^{-\frac{1}{2}} = \pi(j)^{\frac{1}{2}} P(j,i) \pi(i)^{-\frac{1}{2}} = Q(j,i).$$

So we can apply the spectral decomposition theorem for Q, which yields

$$Q = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\top},$$

where $\lambda_1 \geq \cdots \geq \lambda_n$ are eigenvalues of Q with corresponding orthonormal eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. If we let $\mathbf{v}_i := \Pi^{-\frac{1}{2}} \mathbf{u}_i$, then the above is equivalent to

$$P = \sum_{i=1}^{n} \lambda_i \Pi^{-\frac{1}{2}} \mathbf{u}_i \mathbf{u}_i^{\top} \Pi^{\frac{1}{2}} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\top} \Pi.$$

We claim that $\lambda_1, \ldots, \lambda_n$ are eigenvalues of P with corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. To see this, we have for any $j \in [n]$:

$$\begin{aligned} P\mathbf{v}_j &= \sum_{i=1}^{n} \lambda_i \Pi^{-\frac{1}{2}} \mathbf{u}_i \mathbf{u}_i^{\top} \Pi^{\frac{1}{2}} \mathbf{v}_j \\ &= \sum_{i=1}^{n} \lambda_i \Pi^{-\frac{1}{2}} \mathbf{u}_i \mathbf{u}_i^{\top} \Pi^{\frac{1}{2}} \Pi^{-\frac{1}{2}} \mathbf{u}_j \\ &= \lambda_j \mathbf{v}_j. \end{aligned}$$

Everything looks nice if we equip $\mathbb{R}^n$ with the inner product $\langle \cdot, \cdot \rangle_\Pi$ defined as $\langle \mathbf{x}, \mathbf{y} \rangle_\Pi = \mathbf{x}^\top \Pi \mathbf{y} = \sum_{i=1}^n \pi(i)\mathbf{x}(i)\mathbf{y}(i)$. It is clear that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are orthonormal with respect to the inner product:

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_\Pi = \begin{cases} 0, & \text{if } i \neq j; \\ 1, & \text{if } i = j. \end{cases}$$

2.2 *Graph Expansion*

We want to measure the connectivity of a graph in terms of its spectrum. Let $G = (V, E)$ be a weighted graph with nonnegative weights $\left(w_{ij}\right)_{i,j \in [n]}$. For any $S \subseteq V$, we define the expansion of S as

$$Q(S, \overline{S}) := \sum_{i \in S, j \in V \setminus S} \pi(i) P(i, j).$$

Furthermore, we define the *expansion* of S as

$$\Phi(S) = \frac{Q(S, \overline{S})}{\pi(S)},$$

where $\pi(S) = \sum_{i \in S} \pi(i)$. Suppose $X_t \sim \pi$, then $\Phi(S) = \mathbf{Pr}\left[X_{t+1} \notin S \mid X_t \in S\right]$, which is the probability of escaping S. Equivalently, it is the ratio between the weight of edges connecting S and $\overline{S}$ and the weight of edges incident to S.

The expansion of G is the smallest $\Phi(S)$ over all S with measure at most 1/2, i.e., $\Phi(G) = \min_{S \subseteq V: \pi(S) \leq \frac{1}{2}} \Phi(S)$. We have the following celebrated connection between $\Phi(G)$ and λ_2.

Theorem 2 (Cheeger's Inequality)**.** $\frac{1-\lambda_2}{2} \leq \Phi(P) \leq \sqrt{2(1-\lambda_2)}$.

Moreover, λ_2 also carries the information on how to partition G into the hardest $(S, \overline{S})$.

3 *Cheeger's Inequality*

Now we prove Cheeger's ineqaulity. We present the proof in terms of a reversible Markov chain P with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Consider the Laplacian matrix $L = I - P$ with eigenvalues $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_n$. We have shown that $\gamma_1 = 0$ and $\gamma_i = 1 - \lambda_i$ for all $i \in [n]$. Then the Cheeger's inequality can be written in terms of γ_2.

Theorem 3 (Cheeger's Inequality)**.** $\frac{\gamma_2}{2} \leq \Phi(P) \leq \sqrt{2\gamma_2}$.

We prove $\frac{\gamma_2}{2} \leq \Phi(P)$ (①) and $\Phi(P) \leq \sqrt{2\gamma_2}$ (②) respectively.

Proof of ①. We relate $\Phi(P)$ with γ_2 using the variational characterization. Note that

$$\gamma_2 = \min_{\substack{U \subseteq \mathbb{R}^n \\ \dim(U)=2}} \max_{\mathbf{x} \in U \setminus \{0\}} \frac{\langle \mathbf{x}, L\mathbf{x} \rangle_\Pi}{\langle \mathbf{x}, \mathbf{x} \rangle_\Pi}.$$

Let S be the subset of V such that $\Phi(P) = \max\left\{\Phi(S), \Phi(\overline{S})\right\}$. Let $U = \text{span}\left(\mathbf{1}_S, \mathbf{1}_{\overline{S}}\right)$. For any $\mathbf{x} \in U$, we can write $\mathbf{x}$ as $a\mathbf{1}_S + b\mathbf{1}_{\overline{S}}$ for some constants a and b. Then

$$\frac{\langle \mathbf{x}, L\mathbf{x}\rangle_\Pi}{\langle \mathbf{x}, \mathbf{x}\rangle_\Pi} = \frac{\sum_{\{i,j\}\in E} \pi(i)P(i,j)(x_i - x_j)^2}{\sum_i \pi(i)x_i^2} = \frac{\sum_{i\in S, j\in \overline{S}} \pi(i)P(i,j)(a-b)^2}{\pi(S)a^2 + \pi(\overline{S})b^2}$$
$$\leq \frac{2\sum_{i\in S, j\in \overline{S}} \pi(i)P(i,j)(a^2+b^2)}{\pi(S)a^2 + \pi(\overline{S})b^2}$$
$$\leq 2\max\left\{\frac{\sum_{i\in S, j\in \overline{S}} \pi(i)P(i,j)}{\pi(S)}, \frac{\sum_{i\in S, j\in \overline{S}} \pi(i)P(i,j)}{\pi(\overline{S})}\right\} = 2\Phi(P)$$

where the second inequality follows from the fact that for positive real numbers z_1, z_2, y_1, y_2, $\frac{z_1+z_2}{y_1+y_2} \leq \max\left\{\frac{z_1}{y_1}, \frac{z_2}{y_2}\right\}$. □

By definition, $\Phi(P) = \min_{\substack{S\subseteq V \\ \pi(S)\leq \frac{1}{2}}} \Phi(S)$. To prove ②, we only need to find a $S \subseteq V$ such that $\Phi(S) \leq \sqrt{2\gamma_2}$. Such S can be generated using the Fiedler's algorithm. With input $\mathbf{x} \in \mathbb{R}^V$:

- sort V according to $\mathbf{x}$, get $V = \{v_1, \ldots, v_n\}$ where $\mathbf{x}(v_1) \leq \mathbf{x}(v_2) \leq \mathbf{x}(v_n)$;
- for each $i \in [n]$, let $S_i = \{v_1, \ldots, v_i\}$;
- return the S_i with the minimum $\Phi(S_i) \vee \Phi(\overline{S}_i)$.[1]

[1] $a \vee b$ means $\max\{a, b\}$.

Theorem 4. *For any* $\mathbf{x} \perp \mathbf{1}$, *assume the Fiedler's algorithm returns* S *with input* $\mathbf{x}$. *Then* $\Phi(S) \leq \sqrt{2R_L(\mathbf{x})}$.

With Theorem 4, the proof of ② is straightforward. Note that $\mathbf{v}_2$, the eigenvector of L corresponding to eigenvalue γ_2, is the minimizer of $R_L(\mathbf{x})$ on the constraint that $\mathbf{x} \perp \mathbf{1}$. We can divide the graph into different blocks where each block is well connected inside. Intuitively, to get smaller $\sum_{\{i,j\}\in E} \pi(i)P(i,j)(x_i - x_j)^2$, we tend to assign the same value to the x_i's in the same block. The Fiedler's algorithm will return a partition that divides the blocks into two groups. This indicates that $\mathbf{v}_2$ contains the information to find the bottleneck of the graph.

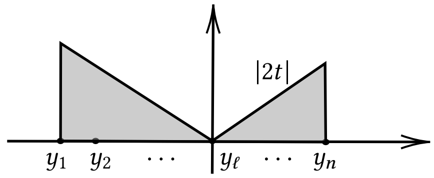
Proof of ②. Run the Fiedler's algorithm with input $\mathbf{x} = \mathbf{v}_2$ and get output S. By Theorem 4, $\Phi(S) \leq \sqrt{2R_L(\mathbf{v}_2)} = \sqrt{2\gamma_2}$.

□

It remains to prove Theorem 4.

Proof of Theorem 4. Input $\mathbf{x}$ and run the Fiedler's algorithm. W.l.o.g., assume $x_1 \leq \cdots \leq x_n$.

Define ℓ be the minimum k such that $\sum_{i=1}^k \pi_i \geq \frac{1}{2}$. Let $\mathbf{y} = (y_1, \ldots, y_n) = \mathbf{x} - x_\ell \cdot \mathbf{1}$. That is, $y_i = x_i - x_\ell$ for all $i \in [n]$. Rescale $\mathbf{y}$ such that $y_1^2 + y_n^2 = 1$. We



randomly pick $t \in [y_1, y_n]$ with density $2|t|$ and set $S_t = \{i \in [n] \mid y_i \le t\}$. Then

$$\max\left\{\Phi(S_t), \Phi(\overline{S}_t)\right\} = \frac{\sum_{i\in S_t, j\in \overline{S}_t} \pi(i)P(i,j)}{\min\left\{\pi(S_t), \pi(\overline{S}_t)\right\}}.$$

Let $A := \sum_{i\in S_t, j\in \overline{S}_t} \pi(i)P(i,j)$ and $B := \min\left\{\pi(S_t), \pi(\overline{S}_t)\right\}$. We claim that $\frac{\mathbf{E}[A]}{\mathbf{E}[B]} \le \sqrt{2R_L(\mathbf{x})}$.

By definition,

$$\mathbf{E}[A] = \sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j)\mathbf{Pr}\left[i \in S_t, j \in \overline{S}_t\right]. \tag{1}$$

Note that $\mathbf{Pr}\left[i \in S_t, j \in \overline{S}_t\right]$ is the probability that $t \in [y_i, y_j]$, which can be calculated directly by integration:

$$\begin{aligned}
\text{Equation (1)} &= \sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j) \int_{y_i}^{y_j} 2|t|\,\mathrm{d}t \\
&= \sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j)\left(\operatorname{sgn}(y_j)y_j^2 - \operatorname{sgn}(y_i)y_i^2\right) \\
&\le \sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j)\left(|y_i| + |y_j|\right)\left(y_j - y_i\right) \\
&= \sum_{\substack{\{i,j\}\in E\\ i<j}} \left(\pi(i)P(i,j)\right)^{\frac{1}{2}}\left(|y_i| + |y_j|\right) \cdot \left(\pi(i)P(i,j)\right)^{\frac{1}{2}}\left(y_j - y_i\right). \qquad (2)
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\text{Equation (2)} &\le \sqrt{\sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j)\left(|y_i| + |y_j|\right)^2} \cdot \sqrt{\sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j)\left(y_j - y_i\right)^2} \\
&\le \sqrt{2\sum_{\substack{\{i,j\}\in E\\ i<j}} \pi(i)P(i,j)\left(y_i^2 + y_j^2\right)} \cdot \sqrt{\langle \mathbf{y}, L\mathbf{y}\rangle_\Pi} \\
&= \sqrt{2\langle \mathbf{y}, \mathbf{y}\rangle_\Pi} \cdot \sqrt{\langle \mathbf{y}, L\mathbf{y}\rangle_\Pi}
\end{aligned}$$

Recall that ℓ is a middle line of π. Therefore, if $t < 0$, $\pi(S_t) \le \pi(\overline{S}_t)$ and otherwise $\pi(S_t) > \pi(\overline{S}_t)$. Then we have

$$\mathbf{E}[B] = \underbrace{\mathbf{Pr}[t < 0]\,\mathbf{E}[\pi(S_t) \mid t < 0]}_{(3)} + \underbrace{\mathbf{Pr}[t \ge 0]\,\mathbf{E}\left[\pi(\overline{S}_t) \,\middle|\, t \ge 0\right]}_{(4)}.$$

Note that

$$(3) = \mathbf{Pr}[t < 0] \sum_{i=1}^{\ell-1} \pi(i) \mathbf{Pr}[i \in S_t \mid t < 0]$$
$$= \sum_{i=1}^{\ell-1} \pi(i) \int_{y_i}^{0} 2|t| \, \mathrm{d}t = \sum_{i=1}^{\ell-1} \pi(i) y_i^2.$$

Similarly, $(4) = \sum_{i=\ell}^{n} \pi(i) y_i^2$. Summing up the two terms, we have $\mathbf{E}[B] = \sum_{i=1}^{n} \pi(i) y_i^2 = \langle \mathbf{y}, \mathbf{y} \rangle_\Pi$. Therefore,

$$\frac{\mathbf{E}[A]}{\mathbf{E}[B]} \le \sqrt{\frac{2\langle \mathbf{y}, L\mathbf{y} \rangle_\Pi}{\langle \mathbf{y}, \mathbf{y} \rangle_\Pi}} = \sqrt{2R_L(\mathbf{y})}.$$

Since $\mathbf{y}$ is obtained by adding a constant offset to $\mathbf{x}$ and $\mathbf{x} \perp \mathbf{1}$, we have $\langle \mathbf{y}, \mathbf{y} \rangle_\Pi \ge \langle \mathbf{x}, \mathbf{x} \rangle_\Pi$ and $\langle \mathbf{y}, L\mathbf{y} \rangle_\Pi = \langle \mathbf{x}, L\mathbf{x} \rangle_\Pi$. Thus

$$\frac{\mathbf{E}[A]}{\mathbf{E}[B]} \le \sqrt{2R_L(\mathbf{y})} \le \sqrt{2R_L(\mathbf{x})},$$

or equivalently

$$\mathbf{E}\left[A - B\sqrt{2R_L(\mathbf{x})}\right] \le 0.$$

Therefore, the probability of choosing $t \in [y_1, y_n]$ such that $A - B\sqrt{2R_L(\mathbf{x})} \le 0$ is nonzero. This proves the existence of S_t that $\max\left\{\Phi(S_t), \Phi(\overline{S}_t)\right\} \le \sqrt{2R_L(\mathbf{x})}$and thus indicates the correctness of Theorem 4. □