Lecture 8. Random Walk on Graphs and Spectral Theory

Yuan Yao

HKUST

Recall: Laplacian Eigenmap and Diffusion Map

- Given a graph $G(V, E)$ with weight matrix W and $D = \mathbf{diag}(D_{ii})$ with $D_{ii} = \sum_j w_{ij}$.
- Define unnormalized Laplacian $L = D - W$
- Define the normalized Laplacian $\mathcal{L} = D^{-1/2} L D^{-1/2}$
- Define the row Markov matrix $P = D^{-1} W$
  - eigenvectors of L or $\mathcal{L}$;
  - generalized eigenvectors of L

$$Lv = \lambda D v$$

  - or equivalently, right eigenvectors of P

$$Pv = (1 - \lambda) v$$

- Which eigenvectors shall we choose as embeddings?

Random Walk (Markov Chain) on Graphs

- Perron-Frobenius Vector and Google's PageRank: this is about primary eigenvectors, as stationary distributions of Markov chains; application examples include Google's PageRank.

- Fiedler Vector, Cheeger's Inequality, and Spectral Bipartition: this is about the second eigenvector of graph Laplacians, characterizing the topological connected components and the basis for spectral clustering.

- Lumpability/Metastability: this is about multiple piecewise constant right eigenvectors of Markov matrices, widely used for diffusion map, Laplacian eigenmap, and Multiple spectral clustering ("MNcut" by Maila-Shi, 2001), etc.

- Mean first passage time, commute time distance: a connection with diffusion distances.

Random Walk (Markov Chain) on Graphs

- Transition Path Theory: this is about starting from a source set toward a target set, the stochastic transition paths on the graph
- Semi-supervised learning: this is about with partially labeled nodes on a graph, inferring the information on unlabeled points
- They are equivalent in the sense that they satisfy the same unnormalized Laplacian equations with Dirichelet boundary condition.

Outline

Nonnegative Matrix

- Given $A_{n\times n}$, we define

$$A > 0 \Leftrightarrow \ A \text{ is } \textbf{positive matrix} \ \Leftrightarrow A_{ij} > 0 \quad \forall i,j$$

$$A \geq 0 \Leftrightarrow A \text{ is } \textbf{nonnegative matrix} \ \Leftrightarrow A_{ij} \geq 0 \quad \forall i,j.$$

- Note that this definition is different from positive definiteness:

$$A \succ 0 \Leftrightarrow \ A \text{ is positive-definite} \ \Leftrightarrow x^T Ax > 0 \quad \forall x \neq 0$$

$$A \succeq 0 \Leftrightarrow \ A \text{ is semi-positive-definite} \ \Leftrightarrow x^T Ax \geq 0 \quad \forall x \neq 0$$

Perron Vector of Positive Matrix

Theorem (Perron Theorem for Positive Matrix)

Assume that $A > 0$, *i.e.*a positive matrix. Then

(1) $\exists \lambda^* > 0, \nu > 0, \|\nu\|_2 = 1, s.t.\ A\nu = \lambda^*\nu$, ν is a right eigenvector
$(\exists \lambda^* > 0, \omega > 0, \|\omega\|_2 = 1, s.t.\ (\omega^T)A = \lambda^*\omega^T$, left eigenvector)

(2) $\forall$ other eigenvalue λ of A, $|\lambda| < \lambda^*$

(3) ν is unique up to rescaling or λ^* is simple

(4) Collatz-Wielandt Formula

$$\lambda^* = \max_{x \geq 0, x \neq 0} \min_{x_i \neq 0} \frac{[Ax]_i}{x_i} = \min_{x>0} \max \frac{[Ax]_i}{x_i}.$$

Remark

- Such eigenvectors (ν and ω) will be called **Perron vectors**.
- An extension to nonnegative matrices is given by Perron.

Perron Vectors of Nonnegative Matrix

Theorem (Perron Theorem for Nonnegative Matrix)

Assume that $A \geq 0$, *i.e.*nonnegative. Then

(1') $\exists \lambda^* > 0, \nu \geq 0, \|\nu\|_2 = 1, s.t.\ A\nu = \lambda^* \nu$ (similar to left eigenvector)

(2') $\forall$ other eigenvalue λ of A, $|\lambda| \leq \lambda^*$

(3') ν is NOT unique

(4) Collatz-Wielandt Formula

$$\lambda^* = \max_{x \geq 0, x \neq 0} \min_{x_i \neq 0} \frac{[Ax]_i}{x_i} = \min_{x > 0} \max \frac{[Ax]_i}{x_i}$$

Remark

Notice the changes in (1'), (2'), and (3'):

- Perron vectors are nonnegative rather than positive.
- In the nonnegative situation what we lose is the uniqueness in λ^* (2') and ν (3').
- Can we add more conditions such that the loss can be remedied?
- The answer is yes, if we add the concepts of **irreducible** and **primitive** matrices.

Irreducible Matrix

Definition (Irreducible)

The following definitions are equivalent:

(1) For any $1 \leq i, j \leq n$, there is an integer $k \in \mathbb{Z}$, s.t. $A^k(i,j) > 0$. $\Leftrightarrow$

(2) Graph $G = (V, E)$ ($V = \{1, \ldots, n\}$ and $(i,j) \in E$ iff $A_{ij} > 0$) is (path-)connected, *i.e.* $\forall \{i, j\} \subseteq V$, there is a path

$$(x_0, x_1, \ldots, x_t) \in V^{n+1}, \text{where } x_0 = i,\ x_t = j \text{ and } (x_k, x_{k+1}) \in E,$$

that connects i and j.

Remark

- Irreducibility exactly describes the case that the induced graph from A is connected, *i.e.* every pair of nodes are connected by a path of arbitrary length.

- However primitivity strengthens this condition to k-connected, *i.e.* every pair of nodes are connected by a path of length k.

Primitive Matrix

Definition (Primitive)

The following characterizations hold:

1. There is an integer $k \in \mathbb{Z}$, such that $\forall i, j,\ A^k_{ij} > 0$; $\Leftrightarrow$
2. Any node pair $\{i, j\} \in E$ are connected with a path of length no more than k; $\Leftrightarrow$
3. A has unique $\lambda^* = \max |\lambda|$; $\Leftarrow$
4. A is **irreducible** and $A_{ii} > 0$**, for some** i.

Remark

- Note that condition (4) is sufficient for primitivity but not necessary.
- All the first three conditions are necessary and sufficient for primitivity.
- Primitive matrices ensure the uniqueness of eigenvalue in module λ^*.
- In comparison, irreducible matrices have a simple primary eigenvalue λ^* and 1-dimensional primary (left and right) eigenspace, with unique left and right eigenvectors up to a sign. However, there might be other eigenvalues whose absolute values (module) equal to the primary eigenvalue, i.e. $\lambda^* e^{i\omega}$.

Remark

- When A is a primitive matrix, A^k becomes a positive matrix for some k, then we can recover (1), (2) and (3) for positivity and uniqueness.

- This leads to the following Perron-Frobenius theorem.

Perron-Frobenius Theory of Primitive Matrix

Theorem (Nonnegative Matrix, Perron-Frobenius)

Assume that $A \geq 0$ and A is primitive. Then

1. $\exists \lambda^* > 0, \nu > 0, \|\nu\|_2 = 1, s.t.\ A\nu = \lambda^*\nu$ (right eigenvector) and $\exists \omega > 0, \|\omega\|_2 = 1, s.t.\ \omega^T A = \lambda^* \omega^T$ (left eigenvector)
2. $\forall$ other eigenvalue λ of A, $|\lambda| < \lambda^*$
3. ν is unique
4. Collatz-Wielandt Formula

$$\lambda^* = \max_{x>0} \min \frac{[Ax]_i}{x_i} = \min_{x>0} \max \frac{[Ax]_i}{x_i}$$

Such eigenvectors and eigenvalue will be called as Perron-Frobenius or primary eigenvectors/eigenvalue.

Example: Markov Chain on Graph

- Given a graph $G = (V, E)$, consider a random walk on G with transition probability $P_{ij} = \mathbf{Prob}(x_{t+1} = j | x_t = i) \geq 0$, a nonnegative matrix. Thus P is a row-stochastic or row-Markov matrix i.e. $P \cdot \mathbf{1} = \mathbf{1}$ where $\mathbf{1} \in \mathbb{R}^V$ is the vector with all elements being 1.

- From Perron theorem for nonnegative matrices, we know
  - $\nu^* = \overrightarrow{1} > 0$ is a right Perron eigenvector of P;
  - $\lambda^* = 1$ is a Perron eigenvalue and all other eigenvalues $|\lambda| \leq 1 = \lambda^*$;
  - $\exists$ left P-eigenvector π such that $\pi^T P = \pi^T$ where $\pi \geq 0, \quad 1^T \pi = 1$; such π is called an invariant/equilibrium distribution;
  - P is irreducible (G is connected) $\Rightarrow \pi$ unique;

Example: Markov Chain on Graph

From Perron-Frobenious theorem for primitive matrices, we know

- P is primitive (G connected by paths of length $\leq k$) $\Rightarrow |\lambda| = 1$ unique,
$$\Rightarrow \lim_{k\to\infty} \pi_0^T P^k \to \pi^T \quad \forall \pi_0 \geq 0, 1^T \pi_0 = 1$$

- This means when we take powers of P, *i.e.* P^k, all rows of P^k will converge to the stationary distribution π^T.

- Such a convergence only holds when P is primitive. If P is not primitive, *e.g.* $P = [0, 1; 1, 0]$ (whose eigenvalues are 1 and -1), P^k always oscillates and never converges.

Example: Markov Chain on Graph

What's the rate of the convergence?

- Let $\pi_t^T = \pi_0^T P^t$ and

$$\gamma = \max\{|\lambda_2|, \cdots, |\lambda_n|\}, \quad \lambda_1 = 1,$$

- Roughly speaking we have

$$\|\pi_t - \pi\|_1 \sim O(e^{-\gamma t}).$$

  This type of rates will be seen in various mixing time estimations.

Example: PageRank

- Consider a directed weighted graph $G = (V, E, W)$ whose weight matrix decodes the webpage link structure:

$$w_{ij} = \begin{cases} \#\{\text{link}: \quad i \mapsto j\}, & (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

- Define an out-degree vector $d_i^o = \sum_{j=1}^n w_{ij}$, which measures the number of out-links from i. A diagonal matrix $D = \mathbf{diag}(d_i^o)$ and a row Markov matrix $P_1 = D^{-1}W$, assumed for simplicity that all nodes have non-empty out-degree.

- This P_1 accounts for a random walk according to the link structure of webpages. One would expect that stationary distributions of such random walks will disclose the importance of webpages: the more visits, the more important.

Example: PageRank

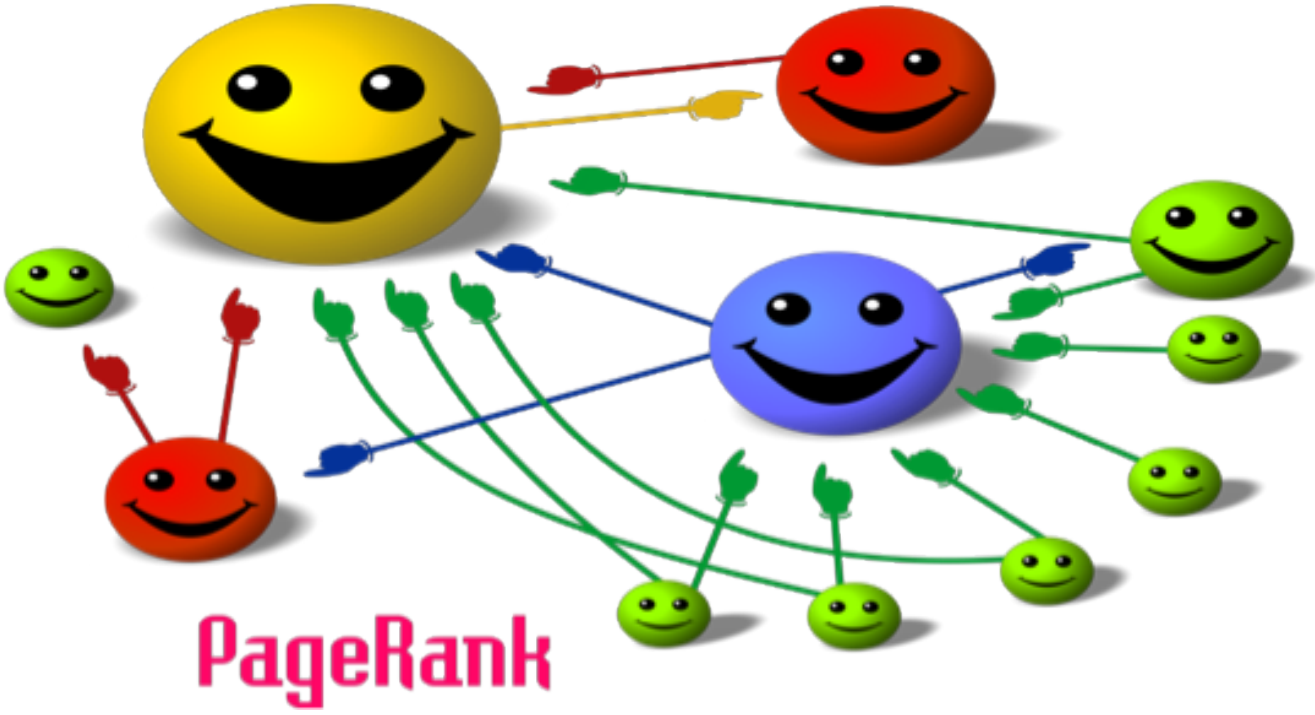


Figure: An illustration of weblink driven random walks and pagerank.

Example: PageRank

- However Perron-Frobenius above tells us that to obtain a unique stationary distribution, we need a **primitive** Markov matrix!

- Google's PageRank does the following trick. Let

$$P_\alpha = \alpha P_1 + (1-\alpha)E,$$

where $E = \frac{1}{n} 1 \cdot 1^T$ is a *random surfer model*, *i.e.* one can jump to any other webpage uniformly.

- So in the model P_α, a browser will play a dice: he will jump according to link structure with probability α or randomly surf with probability $1-\alpha$. For $1 > \alpha > 0$, P_α is a positive matrix, hence primitive (there exists a unique π: $\pi^T P_\alpha = \pi^T$).

- Google choose $\alpha = 0.85$ and in this case π gives PageRank scores..

Cheating the PageRank

- If there are many cross links between a small set of nodes (for example, Wikipedia), those nodes must appear to be high in PageRank.

- Now you probably can figure out how to cheat PageRank. This phenomenon actually has been exploited by spam webpages, and even scholar citations. After learning the nature of PageRank, we should be aware of such mis-behaviors.

Cheating the PageRank

- If there are many cross links between a small set of nodes (for example, Wikipedia), those nodes must appear to be high in PageRank.
- Now you probably can figure out how to cheat PageRank. This phenomenon actually has been exploited by spam webpages, and even scholar citations. After learning the nature of PageRank, we should be aware of such mis-behaviors.
- Above we just consider out-degree $d^{(o)}$. How about in-degree $d_k^{(i)} = \sum_j w_{jk}$?

Kleinberg's HITS algorithm

- High out-degree webpages can be regarded as *hubs*, as they provide more links to others. On the other hand, high in-degree webpages are regarded as *authorities*, as they were cited by others intensively. Basically in/out-degrees can be used to rank webpages, which gives relative ranking as authorities/hubs.
  - $d^{(o)}(i) = \sum_k w_{ik}$
  - $d^{(i)}(j) = \sum_k w_{kj}$

- Finally we discussed a bit on Jon Kleinberg's HITS algorithm, which is based on singular value decomposition (SVD) of link matrix W. It turns out Kleinberg's HITS algorithm gives pretty similar results to in/out-degree ranking.

HITS-Authority Algorithm

Definition (HITS-authority)

We use primary **right singular vector** of W as scores to give the ranking. To understand this, define $L_a = W^T W$.

- Primary right singular vector of W is just a primary eigenvector of nonnegative symmetric matrix L_a.

- Since $L_a(i,j) = \sum_k W_{ki} W_{kj}$, thus it counts the number of references which cites both i and j, *i.e.* $\sum_k \#\{i \leftarrow k \rightarrow j\}$. The higher value of $L_a(i,j)$ the more references received on the pair of nodes. Therefore Perron vector tend to rank the webpages according to authority.

HITS-Hub Algorithm

Definition (HITS-hub)

We use primary **left** singular vector of W as scores to give the ranking.

- Define $L_h = WW^T$, where primary left singular vector of W is just a primary eigenvector of nonnegative symmetric matrix L_h.

- Similarly $L_h(i,j) = \sum_k W_{ik}W_{jk}$, which counts the number of links from both i and j, hitting the same target, *i.e.*$\sum_k \#\{i \to k \leftarrow j\}$. Therefore the Perron vector L_h gives hub-ranking.

Outline

Simple Graph

- Let $G = (V, E)$ be an undirected, unweighted simple graph (*simple graph* means for every pair of nodes there are at most one edge associated with it; and there is no self loop on each node).

- We use $i \sim j$ to denote that node $i \in V$ is a neighbor of node $j \in V$, *i.e.* $(i, j) \in E$.

Adjacency Matrix

Definition (Adjacency Matrix)

$$A_{ij} = \begin{cases} 1 & i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

- We can use the weight of edge $i \sim j$ to define $A_{ij} = W_{ij}$ if the graph is weighted. That indicates $A_{ij} \in \mathbb{R}^+$.
- We can also extend A_{ij} to $\mathbb{R}$ which involves both positive and negative weights, like correlation graphs. But the theory below can not be applied to such weights being positive and negative.
- Define a diagonal matrix $D = \mathbf{diag}(d_i)$, where d_i is the degree of node i: $d_i = \sum_{j=1}^{n} A_{ij}$.

Unnormalized Graph Laplacians

Definition (Graph Laplacian)

$$L_{ij} := D - A = \begin{cases} d_i & i = j, \\ -1 & i \sim j \\ 0 & otherwise \end{cases}$$

- We often called it *unnormalized graph Laplacian*, as a distinction from the normalized graph Laplacian below.
- For weighted graphs, $L = D - W$.

Example: Linear Chain Graph

Example

$V = \{1, 2, 3, 4\}$, $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$. This is a linear chain with four nodes.

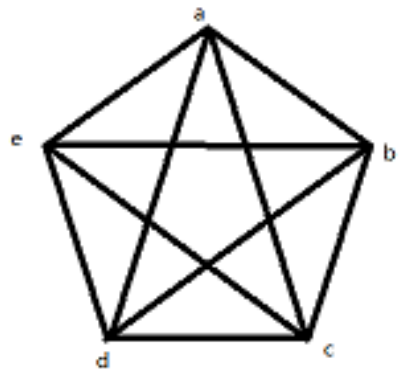
$$L = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

Example: Complete Graph

Example

A complete graph of n nodes, K_n. $V = \{1, 2, 3...n\}$, every two points are connected, as the figure above with $n = 5$.

$$L = \begin{pmatrix} n-1 & -1 & -1 & ... & -1 \\ -1 & n-1 & -1 & ... & -1 \\ -1 & ... & -1 & n-1 & -1 \\ -1 & ... & -1 & -1 & n-1 \end{pmatrix}.$$



Spectrum of L

- L is symmetric, so has an orthonormal eigen-system.
- L is positive semi-definite ($L \succeq 0$), since

$$\begin{aligned} v^T L v &= \sum_i \sum_{j:j\sim i} v_i(v_i - v_j) = \sum_i \left(d_i v_i^2 - \sum_{j:j\sim i} v_i v_j \right) \\ &= \sum_{i\sim j} (v_i - v_j)^2 \geq 0, \qquad \forall v \in \mathbb{R}^n. \end{aligned}$$

  so L has nonnegative eigenvalues.

A Square Root of L: Boundary Map

- $L \succeq 0 \Rightarrow L = BB^T$ for some B.
- In fact, one can choose $B \in \mathbb{R}^{|V| \times |E|}$:

$$B(i,(j,k)) = \begin{cases} 1, & i = j \text{ (start)}, \\ -1, & i = k \text{ (end)}, \\ 0, & \text{otherwise} \end{cases}$$

- B is called *incidence matrix* between a vertex $i \in V$ and an **oriented** edge $(j,k) = -(k,j) \in E$ (or *boundary map* in algebraic topology):
  - if the boundary vertex i meets the start of an edge, then returns 1;
  - if boundary vertex i meets the end of an edge, then -1;
  - otherwise the vertex is not on the boundary of the edge, 0.

Fiedler Theorem

Theorem (Fiedler)

Let L has n eigenvectors

$$Lv_i = \lambda_i v_i, \quad v_i \neq 0, \quad i = 0, \ldots, n-1$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. For the second smallest eigenvector v_1, define

$$N_- = \{i : v_1(i) < 0\},$$
$$N_+ = \{i : v_1(i) > 0\},$$
$$N_0 = V - N_- - N_+.$$

We have the following results:

1. $\#\{i, \lambda_i = 0\} = \#\{$**connected components** of $G\}$;
2. If G is connected, then both N_- and N_+ are connected. $N_- \cup N_0$ and $N_+ \cup N_0$ might be disconnected if $N_0 \neq \emptyset$.

Algebraic Connectivity

- Fiedler Theorem tells us that the second smallest eigenvalue can be used to tell us if the graph is topologically connected, *i.e.* G is connected if and only $\lambda_1 \neq 0$. In other words,
  A. $\lambda_1 = 0 \Leftrightarrow$ there are at least two connected components;
  B. $\lambda_1 > 0 \Leftrightarrow$ the graph is connected;

- When $N_0 = \emptyset$, the second smallest eigenvector can be used to bipartite the graph into two connected components by taking N_- and N_+.

- The second smallest eigenvalue λ_1 is often called as *Fiedler value*, or the *algebraic connectivity*; v_1 is called *Fiedler vector*.

A Sketchy Proof of the First Claim

Proof of Part I.

Let (λ, v) be a pair of eigenvalue-eigenvector, *i.e.*$Lv = \lambda v$. Since $L1 = 0$, so the constant vector $1 \in \mathbb{R}^n$ is always the eigenvector associated with $\lambda_0 = 0$. In general,

$$\lambda = \frac{v^T L v}{v^T v} = \frac{\sum_{i \sim j} (v_i - v_j)^2}{\sum_i {v_i}^2}.$$

Note that

$$0 = \lambda_1 \Leftrightarrow v_i = v_j \text{ (for } j \text{ is path connected with } i).$$

Therefore v is a piecewise constant function on connected components of G. If G has k components, then there are k independent piecewise constant vectors in the span of characteristic functions on those components, which can be used as eigenvectors of L. In this way, we proved the first part of the theory. □

Outline

Normalized Graph Laplacian

Definition (Normalized Graph Laplacian)

$$\mathcal{L}_{ij} = D^{-1/2} L D^{-1/2} = \begin{cases} 1 & i = j, \\ -\frac{1}{\sqrt{d_i d_j}} & i \sim j, \\ 0 & otherwise. \end{cases}$$

- $\mathcal{L} = D^{-1/2} L D^{-1/2} = D^{-1/2}(D - A)D^{-1/2} = I - D^{-1/2} A D^{-1/2}$.
- For eigenvectors $\mathcal{L}v = \lambda v$, we have

  $$(D^{-1/2} L D^{-1/2})v = \lambda v \Leftrightarrow Lu = \lambda D u, \quad u = D^{-1/2} v.$$

  Hence eigenvectors of $\mathcal{L}$, v, after rescaling by $D^{-1/2}v$, become generalized eigenvectors of L.

Algebraic Connectivity

Similar to Fiedler value,

$$\#\{\lambda_i(\mathcal{L}) = 0\} = \#\{\textbf{connected components of } G\}.$$

- Using the Rayleigh Quotient,

$$\begin{aligned} \lambda = \frac{v^T \mathcal{L} v}{v^T v} &= \frac{v^T D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}}v}{v^v} \\ &= \frac{u^T L u}{u^T D u} \\ &= \frac{\sum_{i \sim j} (u_i - u_j)^2}{\sum_j u_j{}^2 d_j}. \end{aligned}$$

Spectrum of Random Walks: Eigenvalues

(A) $I - \mathcal{L}$ is similar to the transition matrix of random walks:

$$P = D^{-1}A = D^{-1/2}(I - \mathcal{L})D^{1/2}.$$

(B) Therefore, their eigenvalues satisfy $\lambda_i(P) = 1 - \lambda_i(\mathcal{L})$.

Spectrum of Random Walks: Eigenvectors

(C) Consider the left eigenvector ϕ and right eigenvector ψ of P.

$$\phi^T P = \lambda \phi^T,$$

$$P\psi = \lambda \psi.$$

and

- $v = D^{-1/2}\phi$
- $v = D^{1/2}\psi$.

Then v is eigenvector of $I - \mathcal{L}$, $\mathcal{L}v = (1-\lambda)v$, since

$$\phi^T P = \lambda \phi^T \Leftrightarrow (\phi^T D^{-1/2})(D^{-1/2}(I-L)D^{-1/2}) = \lambda(\phi^T D^{-1/2})$$

$$P\psi = \lambda\psi \Leftrightarrow D^{-1/2}(I-L)D^{-1/2}(D^{1/2}\psi) = \lambda D^{1/2}\psi$$

Connections

If P is primitive,

- $\exists!\quad \lambda^*(P) = 1$
- $\phi^* \sim \pi(i) = d_i / \sum_i d_i$
- $\pi_i P_{ij} = A_{ij}/c = A_{ji}/c = \pi_j P_{ji}$, so P is reversible
- $\psi^* \sim \mathbf{1}$
- $\lambda_0(\mathcal{L}) = 0$
- $v_0 = v^*(i) \sim \sqrt{d_i}$
- Eigenvectors of $\mathcal{L}$ are orthonormal: $v_i^T v_j = \delta_{ij}$
- Left/right eigenvectors of P are bi-orthonormal: $\phi_i^T \psi_j = \delta_{ij}$

Normalized Cut

Let G be a graph, $G = (V, E)$ and S is a subset of V whose complement is $\bar{S} = V - S$. We define $Vol(S)$, $CUT(S)$ and $NCUT(S)$ as below.

$$Vol(S) = \sum_{i \in S} d_i.$$

$$CUT(S) = \sum_{i \in S, j \in \bar{S}} A_{ij}.$$

$$NCUT(S) = \frac{CUT(S)}{\min(Vol(S), Vol(\bar{S}))}.$$

$NCUT(S)$ is called **normalized-cut**.

Cheeger Constant

- We define the **Cheeger constant**

$$h_G = \min_S NCUT(S).$$

Finding minimal normalized graph cut is **NP-hard**.

- It is often defined that

$$\text{Cheeger ratio (expander):} \quad h_S := \frac{CUT(S)}{Vol(S)}$$

and

$$\text{Cheeger constant:} \quad h_G := \min_S \max\left\{h_S, h_{\bar{S}}\right\}.$$

Cheeger Inequality

Theorem (Cheeger Inequality)

For every undirected graph G,

$$\frac{h_G^2}{2} \leq \lambda_1(\mathcal{L}) \leq 2h_G.$$

- Cheeger Inequality says the second smallest eigenvalue provides both upper and lower bounds on the minimal normalized graph cut. Its proof gives us a constructive polynomial algorithm to achieve such bounds.

Proof of Upper Bound

- Assume the function f realizes the optimal normalized cut,

$$f(i) = \begin{cases} \frac{1}{Vol(S)} & i \in S, \\ \frac{-1}{Vol(\bar{S})} & i \in \bar{S}, \end{cases}$$

Using the Rayleigh Quotient, we get

$$\begin{aligned} \lambda_1 = \inf_{g \perp D^{1/2} e} \frac{g^T \mathcal{L} g}{g^T g} &\leq \frac{\sum_{i \sim j} (f_i - f_j)^2}{\sum f_i^2 d_i} \\ &= \frac{(\frac{1}{Vol(S)} + \frac{1}{Vol(\bar{S})})^2 CUT(S)}{Vol(S)\frac{1}{Vol(S)^2} + Vol(\bar{S})\frac{1}{Vol(\bar{S})^2}} \\ &= (\frac{1}{Vol(S)} + \frac{1}{Vol(\bar{S})}) CUT(S) \\ &\leq \frac{2CUT(S)}{\min(Vol(S), Vol(\bar{S}))} =: 2h_G. \end{aligned}$$

span

Proof of Lower Bound (Fan Chung 2014)

[Short Proof of Lower Bound]

- The proof is based on the fact that

$$h_G = \inf_{f \neq 0} \sup_{c \in \mathbb{R}} \frac{\sum_{x \sim y} |f(x) - f(y)|}{\sum_x |f(x) - c| d_x}$$

where the supreme over c is reached at $c^* = \text{median}(f(x) : x \in V)$.

Proof of Lower Bound (Fan Chung 2014) ?

$$
\begin{aligned}
\lambda_1 &= R(f)|_{f=\nu_1} = \sup_c \frac{\sum_{x\sim y}(f(x)-f(y))^2}{\sum_x (f(x)-c)^2 d_x}, \\
&\geq \frac{\sum_{x\sim y}(g(x)-g(y))^2}{\sum_x g(x)^2 d_x}, \quad g(x) = f(x) - c \\
&= \frac{(\sum_{x\sim y}(g(x)-g(y))^2)(\sum_{x\sim y}(g(x)+g(y))^2)}{(\sum_{x\in V} g^2(x) d_x)(\sum_{x\sim y}(g(x)+g(y))^2)} \\
&\geq \frac{(\sum_{x\sim y}|g^2(x)-g^2(y)|)^2}{(\sum_{x\in V} g^2(x) d_x)(\sum_{x\sim y}(g(x)+g(y))^2)}, \quad \text{Cauchy-Schwartz} \\
&\geq \frac{(\sum_{x\sim y}|g^2(x)-g^2(y)|)^2}{2(\sum_{x\in V} g^2(x) d_x)^2}, \quad (g(x)+g(y))^2 \leq 2(g^2(x)+g^2(y)) \\
&\geq \frac{h_G^2}{2}.
\end{aligned}
$$

This ends the proof of lower bound. □

Approximate NCut

In fact,

$$\frac{h_G^2}{2} \leq \frac{h_{\nu_1}^2}{2} \leq \lambda_1(\mathcal{L}) \leq 2h_G.$$

- h_f: the minimum Cheeger ratio determined by a sweep of f
  - order the nodes: $f(v_1) \geq f(v_2) \dots f(v_n)$
  - $S_i := \{v_1, \dots, v_i\}$
  - $h_f := \min_i h_{S_i}$

- This gives a constructive approximate NCut algorithm, as spectral bi-clustering.

Extensions

- Cheeger Inequality for directed graph: Chung-Lu (2005)
- High Order Cheeger Inequality for Multiple Eigenvectors of Graph Laplacians: James R. Lee, Shayan Oveis Gharan, Luca Trevisan (2011)
- High Order Cheeger Inequality for Connection Laplacians: Afonso S. Bandeira and Amit Singer (2012)
- High Order Cheeger Inequality on Simplicial Complexes: John Steenbergen, Caroline Klivans, Sayan Mukherjee (2012).

Outline

Coarse Grained Markov Chains

- Let P be the transition matrix of a Markov chain on graph $G = (V, E)$ with $V = \{1, 2, \cdots, n\}$, i.e. $P_{ij} = \Pr\{x_t = j : x_{t-1} = i\}$.

- Assume that V admits a partition Ω:

$$V = \cup_{i=1}^{k} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \ i \neq j.$$

$$\Omega = \{\Omega_s : s = 1, \cdots, k\}.$$

- Observe a sequence$\{x_0, x_1, \cdots, x_t\}$ sampled from the Markov chain with initial distribution π_0. Relabel $x_t \mapsto y_t \in \{1, \cdots, k\}$ by

$$y_t = \sum_{s=1}^{k} s\chi_{\Omega_s}(x_t),$$

where χ is the characteristic function. Thus we obtain a sequence (y_t) which is a coarse-grained representation of original sequence.

Lumpability

- Question: is the coarse-grained sequence y_t still Markovian?

Definition (Lumpability, Kemeny-Snell 1976)

P is **lumpable with respect to partition** Ω if the sequence $\{y_t\}$ is Markovian. In other words, the transition probabilities do not depend on the choice of initial distribution π_0 and history, *i.e.*

$$\begin{aligned} & \text{Prob}_{\pi_0}\{x_t \in \Omega_{k_t} : x_{t-1} \in \Omega_{k_{t-1}}, \cdots, x_0 \in \Omega_{k_0}\} \\ = \quad & \text{Prob}\{x_t \in \Omega_{k_t} : x_{t-1} \in \Omega_{k_{t-1}}\}. \end{aligned}$$

The lumpability condition above can be rewritten as

$$\text{Prob}_{\pi_0}\{y_t = k_t : y_{t-1} = k_{t-1}, \cdots, y_0 = k_0\} = \text{Prob}\{y_t = k_t : y_{t-1} = k_{t-1}\}. \tag{1}$$

Criteria for Lumpability

I. (Kemeny-Snell 1976) P is lumpable with respect to partition Ω $\Leftrightarrow \forall \Omega_s, \Omega_t \in \Omega$, $\forall i, j \in \Omega_s$, $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$, where $\hat{P}_{i\Omega_t} = \sum_{j \in \Omega_t} P_{ij}$.

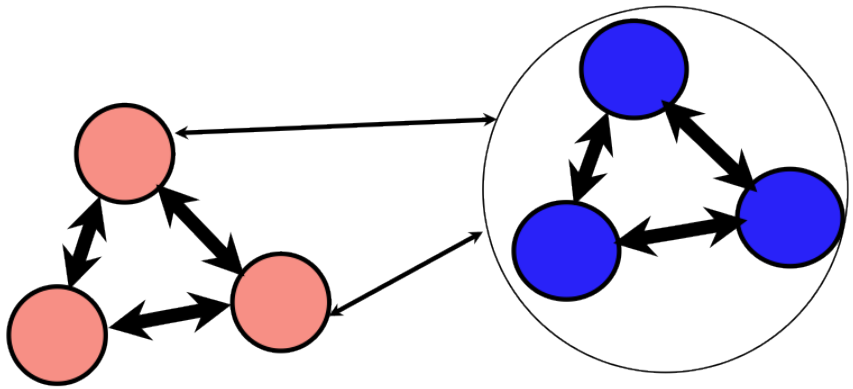


Figure: Lumpability condition $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$

Spectral Criteria for Lumpability

II. (Meila-Shi 2001) P is lumpable with respect to partition Ω and $\hat{P}$ ($\hat{p}_{st} = \sum_{i\in\Omega_s, j\in\Omega_t} p_{ij}$) is nonsingular $\Leftrightarrow$ P has k independent piecewise constant right eigenvectors in $\mathrm{span}\{\chi_{\Omega_s} : s = 1, \cdots, k\}$.

- So k-dimensional diffusion map (right eigenvectors of P) maps lumpable states into a simplex.

Example

- Consider a linear chain with $2n$ nodes whose adjacency matrix and degree matrix are given by

$$A = \begin{bmatrix} 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 0 & 1 & \\ & & & 1 & 0 & \end{bmatrix}, \quad D = \mathrm{diag}\{1, 2, \cdots, 2, 1\}$$

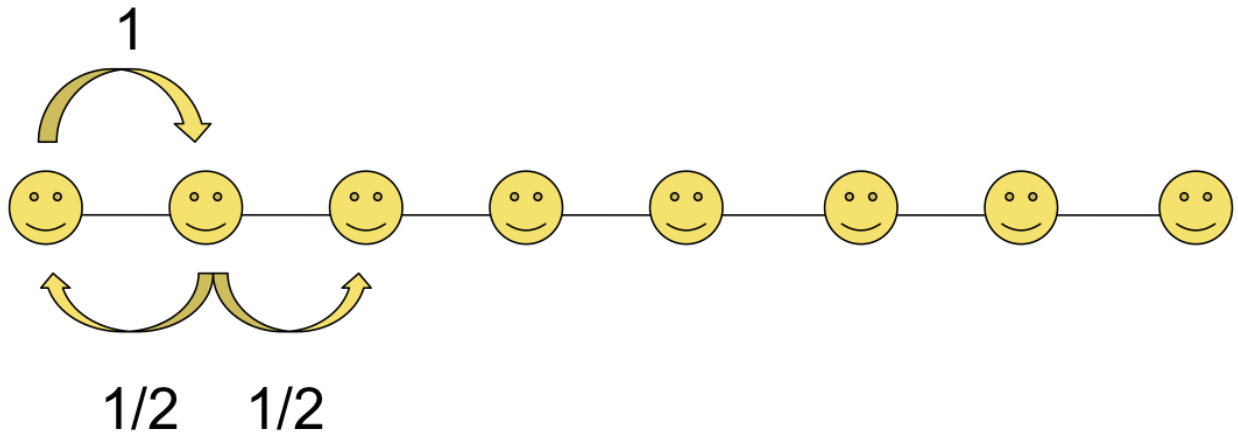


Figure: A linear chain of $2n$ nodes with a random walk.

Example

- So the transition matrix $P = D^{-1}A$ illustrated in Figure has a spectrum including two eigenvalues of magnitude 1, *i.e.*$\lambda_0 = 1$ and $\lambda_{n-1} = -1$. P is *lumpable* with respect to partition that $\Omega_1 = \{$odd nodes$\}$, $\Omega_2 = \{$even nodes$\}$. We can check that I and II are satisfied.

- To see I, note that for any two even nodes, say $i = 2$ and $j = 4$, $\hat{P}_{i\Omega_2} = \hat{P}_{j\Omega_2} = 1$ as their neighbors are all odd nodes, hence I is satisfied.

- To see II, note that ϕ_0 (associated with $\lambda_0 = 1$) is a constant vector while ϕ_1 (associated with $\lambda_{n-1} = -1$) is constant on even nodes and odd nodes respectively. Figure 4 shows the lumpable states when $n = 4$ in the left.

Lumpable $\neq$ Optimal NCut

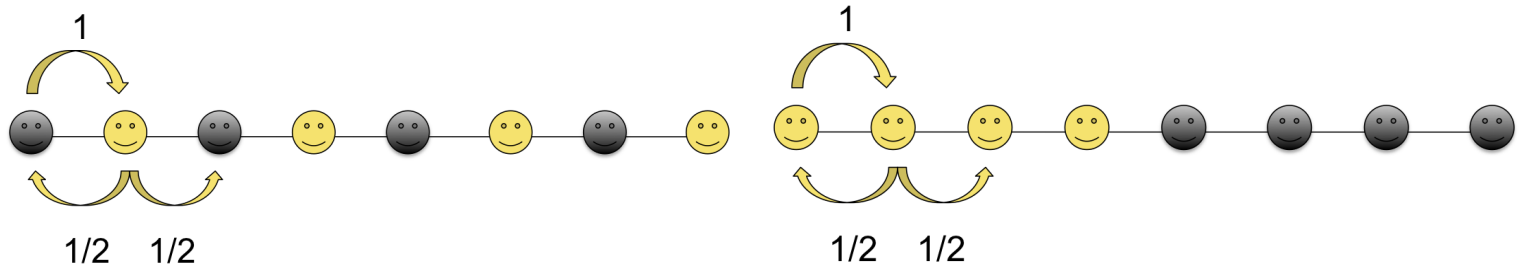


Figure: Left: two lumpable states; Right: optimal-bipartition of Ncut.

- Note that lumpable states might not be optimal bi-partitions in $NCUT = Cut(S)/\min(vol(S), vol(\bar{S}))$.

- In this example, the optimal bi-partition by Ncut is given by $S = \{1, \ldots, n\}$, shown in the right of Figure. In fact the second largest eigenvalue $\lambda_1 = 0.9010$ whose eigenvector

  $$v_1 = [0.4714, 0.4247, 0.2939, 0.1049, -0.1049, -0.2939, -0.4247, -0.4714],$$

  gives the optimal bi-partition.

Example: Uncoupled Markov Chains

- Uncoupled Markov chains are lumpable, e.g.

$$P_0 = \begin{bmatrix} \Omega_1 & & \\ & \Omega_2 & \\ & & \Omega_3 \end{bmatrix}, \ \hat{P}_{it} = \hat{P}_{jt} = 0.$$

Example: Nearly Uncoupled Markov Chains

- A markov chain $\tilde{P} = P_0 + O(\epsilon)$ is called nearly uncoupled Markov chain. Such Markov chains can be approximately represented as uncoupled Markov chains with *metastable states*, $\{\Omega_s\}$, where within metastable state transitions are fast while cross metastable states transitions are slow. Such a separation of scale in dynamics often appears in many phenomena in real lives, such as protein folding,

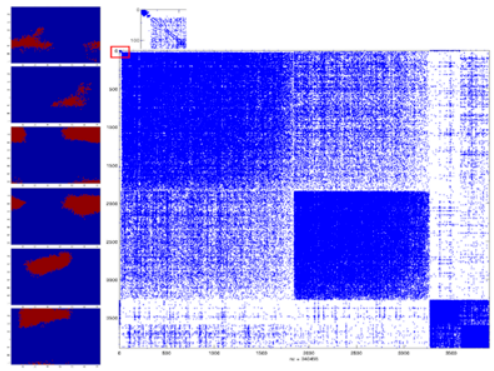


Figure: Nearly uncoupled Markov Chain for six metastable states in Alanine-dipeptide.

Illustration

- One's life transitions among metastable states: $primary\ schools \mapsto middle\ schools \mapsto high\ schools \mapsto college/university \mapsto work\ unit$, etc.

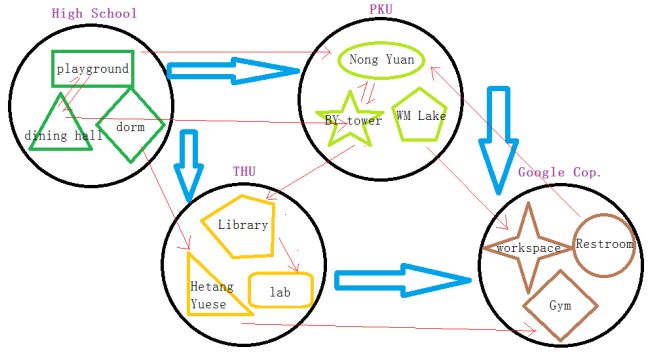


Figure: Metastable states of life transitions.

Application: MNcut

Meila-Shi (2001) calls the following algorithm as MNcut, standing for *modified Ncut*. Due to the theory above, perhaps we'd better to call it *multiple spectral clustering*.

1) Find top k right eigenvectors,

$$P\psi_i = \lambda_i \psi_i,\ i = 1, \cdots, k,\ \lambda_i = 1 - o(\epsilon).$$

2) Embedding $Y^{n\times k} = [\psi_1, \cdots, \psi_k]$.

3) k-means (or other suitable clustering methods) on Y to k-clusters.

- Note: k lumpable states are mapped to a k-simplex.

Example: Spectral Clustering

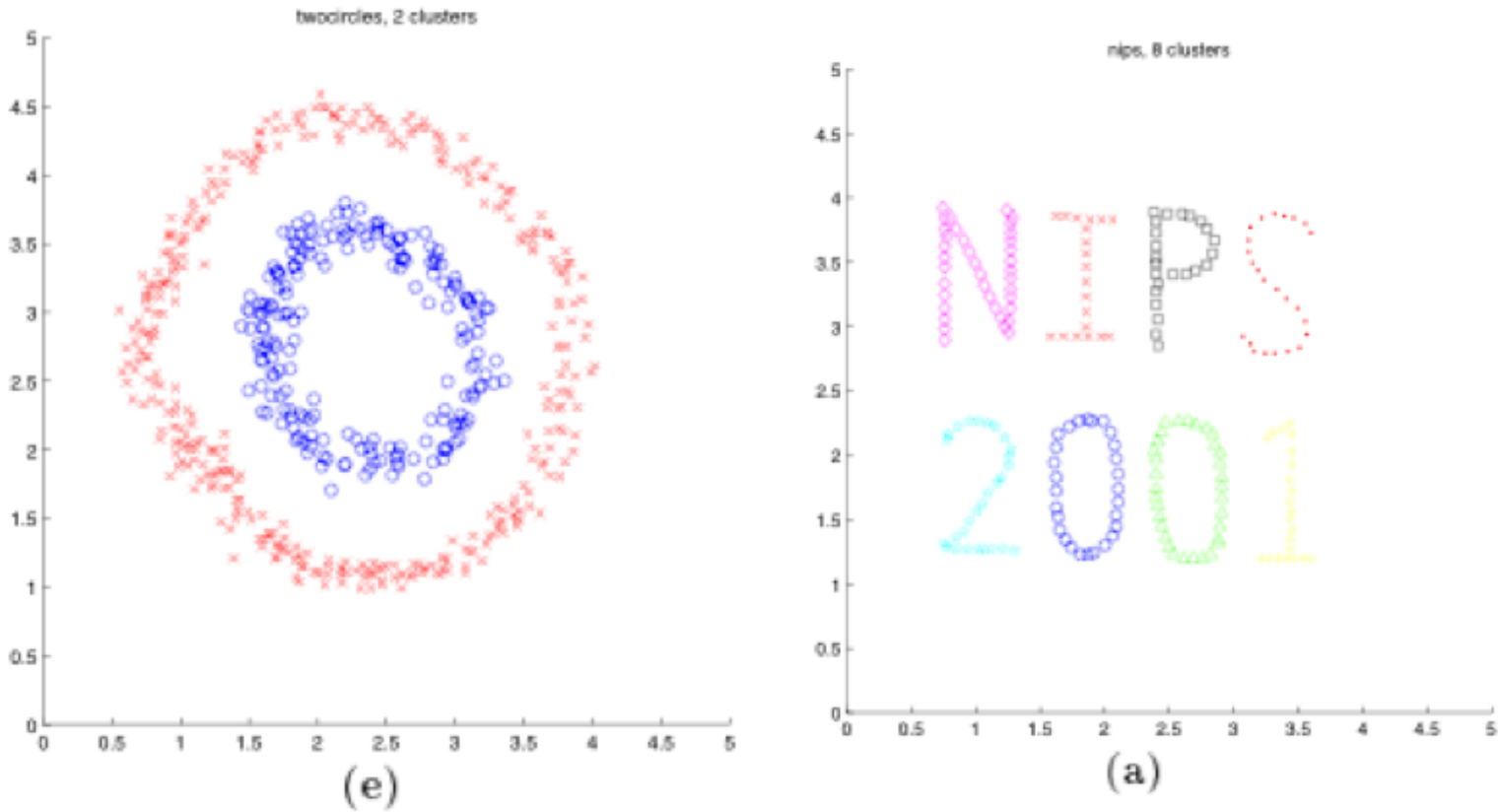


Figure: Spectral clustering of point cloud data in 2-D plane

Proof of Theorem

- Before the proof of the theorem, we note that condition I is in fact equivalent to

$$VUPV = PV, \tag{2}$$

where U is a k-by-n matrix where each row is a uniform probability that

$$U_{is}^{k\times n} = \frac{1}{|\Omega_s|}\chi_{\Omega_s}(i), \qquad i \in V,\ s \in \Omega,$$

and V is a n-by-k matrix where each column is a characteristic function on Ω_s,

$$V_{sj}^{n\times k} = \chi_{\Omega_s}(j).$$

- With this we have $\hat{P} = UPV$ and $UV = I$. Such a matrix representation will be useful in the derivation of condition II. Now we give the proof of the main theorem.

Proof of Claim I

- **I.** "⇒" To see the necessity, P is lumpable w.r.t. partition Ω, then it is necessary that

$$\underset{\pi_0}{\mathbf{Prob}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \underset{\pi_0}{\mathbf{Prob}}\{y_1 = t : y_0 = s\} = \hat{p}_{st}$$

which does not depend on π_0. Now assume there are two different initial distribution such that $\pi_0^{(1)}(i) = 1$ and $\pi_0^{(2)}(j) = 1$ for $\forall i, j \in \Omega_s$. Thus

$$\begin{aligned} \hat{p}_{i\Omega_t} &= \underset{\pi_0^{(1)}}{\mathbf{Prob}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} \\ = \quad \hat{p}_{st} &= \underset{\pi_0^{(2)}}{\mathbf{Prob}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \hat{p}_{j\Omega_t}. \end{aligned}$$

Proof of Claim I

- "$\Leftarrow$" To show the sufficiency, we are going to show that if the condition is satisfied, then the probability

$$\underset{\pi_0}{\mathbf{Prob}}\{y_t = t : y_{t-1} = s, \cdots, y_0 = k_0\}$$

depends only on $\Omega_s, \Omega_t \in \Omega$. Probability above can be written as $\mathbf{Prob}_{\pi_{t-1}}(y_t = t)$ where π_{t-1} is a distribution with support only on Ω_s which depends on π_0 and history up to $t-1$. But since $\mathbf{Prob}_i(y_t = t) = \hat{p}_{i\Omega_t} \equiv \hat{p}_{st}$ for all $i \in \Omega_s$, then $\mathbf{Prob}_{\pi_{t-1}}(y_t = t) = \sum_{i\in\Omega_s} \pi_{t-1}\hat{p}_{i\Omega_t} = \hat{p}_{st}$ which only depends on Ω_s and Ω_t.

Proof of Claim II

- **II.** "$\Rightarrow$": Since $\hat{P}$ is nonsingular, let $\{\psi_i,\, i = 1, \cdots, k\}$ are independent right eigenvectors of $\hat{P}$, *i.e.* $\hat{P}\psi_i = \lambda_i \psi_i$. Define $\phi_i = V\psi_i$, then ϕ_i are independent piecewise constant vectors in $\mathrm{span}\{\chi_{\Omega_i}, i = 1, \cdots, k\}$. We have

$$P\phi_i = PV\psi_i = VUPV\psi_i = V\hat{P}\psi_i = \lambda_i V\psi_i = \lambda_i \phi_i,$$

  i.e. ϕ_i are right eigenvectors of P.

Proof of Claim II

- **II.** "$\Leftarrow$": Let $\{\phi_i, i = 1, \cdots, k\}$ be k independent piecewise constant right eigenvectors of P in $\mathrm{span}\{\chi_{\Omega_i}, i = 1, \cdots, k\}$. There must be k independent vectors $\psi_i \in \mathbb{R}^k$ that satisfied $\phi_i = V\psi_i$. Then

$$P\phi_i = \lambda_i\phi_i \Rightarrow PV\psi_i = \lambda_i V\psi_i,$$

Multiplying VU to the left on both sides of the equation, we have

$$VUPV\psi_i = \lambda_i VUV\psi_i = \lambda_i V\psi_i = PV\psi_i, \ (UV = I),$$

which implies

$$(VUPV - PV)\Psi = 0, \quad \Psi = [\psi_1, \ldots, \psi_k].$$

Since Ψ is nonsingular due to independence of ψ_i, whence we must have $VUPV = PV$. □

Outline

Mean First Passage Time

- Consider a Markov chain P on graph $G = (V, E)$. In this section we study the *mean first passage time* between vertices, which exploits the unnormalized graph Laplacian and will be useful for commute time map against diffusion map.

Definitions

Definition.

1. *First passage time (or hitting time)*: $\tau_{ij} := \inf(t \geq 0 | x_t = j, x_0 = i)$;
2. *Mean First Passage Time*: $T_{ij} = \mathbf{E}_i \tau_{ij}$;
3. $\tau_{ij}^+ := \inf(t > 0 | x_t = j, x_0 = i)$, where τ_{ii}^+ is also called *first return time*;
4. $T_{ij}^+ = \mathbf{E}_i \tau_{ij}^+$, where T_{ii}^+ is also called *mean first return time*.

Here $\mathbf{E}_i$ denotes the conditional expectation with fixed initial condition $x_0 = i$.

Unnormalized Graph Laplacian

Theorem

Assume that P is irreducible. Let $L = D - W$ be the unnormalized graph Laplacian with Moore-Penrose inverse $L^\dagger$, where $D = \mathbf{diag}(d_i)$ with $d_i = \sum_{j:j\sim i} W_{ij}$ being the degree of node i. Then

1. Mean First Passage Time is given by

$$T_{ii} = 0,$$

$$T_{ij} = \sum_k L^\dagger_{ik} d_k - L^\dagger_{ij} vol(G) + L^\dagger_{jj} vol(G) - \sum_k L^\dagger_{jk} d_k, \quad i \neq j.$$

2. Mean First Return Time is given by

$$T^+_{ii} = \frac{1}{\pi_i}, \quad T^+_{ij} = T_{ij}.$$

Commute Time Distance

- As $L^\dagger$ is a positive semi-definite matrix, this leads to the following corollary.

Corollary

$$T_{ij} + T_{ji} = vol(G)(L^\dagger_{ii} + L^\dagger_{jj} - 2L^\dagger_{ij}). \tag{3}$$

Therefore the average commute time between i and j leads to an Euclidean distance metric

$$d_c(x_i, x_j) := \sqrt{T_{ij} + T_{ji}}$$

often called *commute time distance*.

Commute Time Embedding

- Assume the eigen-decomposition of L is $L\nu_i = \lambda_i \nu_i$ where $0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1}$.

- Define the **commute time map** by

$$\Psi(x) = \left(\frac{1}{\sqrt{\lambda_1}} \nu_1(x), \cdots, \frac{1}{\sqrt{\lambda_{n-1}}} \nu_{n-1}(x) \right)^T \in \mathbb{R}^{n-1}.$$

- Then $L^+_{ii} + L^+_{jj} - 2L^+_{ij} = ||\Psi(x_i) - \Psi(x_j))||^2_{l^2}$, and we call $d_r(x_i, x_j) = \sqrt{L^+_{ii} + L^+_{jj} - 2L^+_{ij}}$ the *resistance distance*. So we have $d_c(x_i, x_j) = \sqrt{T_{ij} + T_{ji}} = \sqrt{vol(G)} d_r(x_i, x_j)$.

Diffusion Map vs. Commute Time Map

Table: Comparisons between diffusion map and commute time map. Here $x \sim y$ means that x and y are in the same cluster and $x \nsim y$ for different clusters.

Diffusion Map	Commute Time Map
P's right eigenvectors	L's eigenvectors
scale parameters: α, ε, and t	scale: Gaussian ε
$\exists t$ s.t. $x \sim y$, $d_t(x,y) \to 0$	*
and $x \nsim y$, $d_t(x,y) \to \infty$	*

(*) Recently, Radl, von Luxburg and Hein showed that commute time distance between two points may not reflect the clustering information of these points, but just local densities at these points.

Outline

Setting

- The transition path theory was originally introduced in the context of continuous-time Markov process on continuous state space by Weinan E and Eric Vanden-Eijnden (2006) and later for discrete state space by Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden (2009). An application of discrete transition path theory for molecular dynamics is by Frank Noè et al. (2009). See E and Vanden-Eijnden (2010) for a review.
- The following material is adapted to the setting of discrete time Markov chain with transition probability matrix P in E, Lu, and Yao (2012). We assume reversibility in the following presentation, which can be extended to non-reversible Markov chains.

Setting

- Assume that an irreducible Markov Chain on graph $G = (V, E)$ admits the following decomposition $P = D^{-1}W = \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$.
  Here $V_l = V_0 \cup V_1$ denotes the labeled vertices with source set V_0 (e.g. reaction state in chemistry) and sink set V_1 (e.g. product state in chemistry), and V_u is the unlabeled vertex set (intermediate states). That is,
  - $V_0 = \{i \in V_l : f_i = f(x_i) = 0\}$
  - $V_1 = \{i \in V_l : f_i = f(x_i) = 1\}$
  - $V = V_0 \cup V_1 \cup V_u$ where $V_l = V_0 \cup V_1$

Remarks

- Given two sets V_0 and V_1 in the state space V, the transition path theory tells how these transitions between the two sets happen (mechanism, rates, etc.).

- If we view V_0 as a reactant state and V_1 as a product state, then one transition from V_0 to V_1 is a reaction event. The reactve trajectories are those part of the equilibrium trajectory that the system is going from V_0 to V_1.

- Let the hitting time of V_l be

$$\tau_i^k = \inf\{t \geq 0 : x(0) = i, x(t) \in V_k\}, \quad k = 0, 1.$$

Committor Function

- The central object in transition path theory is the committor function. Its value at $i \in V_u$ gives the probability that a trajectory starting from i will hit the set V_1 first than V_0, *i.e.*, the success rate of the transition at i.

Proposition

For $\forall i \in V_u$, define the *committor function*

$$q_i := \mathbf{Prob}(\tau_i^1 < \tau_i^0) = \mathbf{Prob}(\text{trajectory starting from } i \in V \text{ hit } V_1 \text{ before } V_0)$$

which satisfies **Laplacian equation with Dirichlet** boundary conditions

$$(Lq)(i) = [(I - P)q](i) = 0, \quad i \in V_u$$

$$q(V_0)|_{i \in V_0} = 0,\ q(V_1)|_{i \in V_1} = 1.$$

The solution is

$$q_u = (D_u - W_{uu})^{-1} W_{ul} q_l. \tag{4}$$

Remark

- The committor function provides natural decomposition of the graph. If $q(x)$ is less than 0.5, i is more likely to reach V_0 first than V_1; so that $\{i \mid q(x) < 0.5\}$ gives the set of points that are more attached to set V_0.

- Once the committor function is given, the statistical properties of the reaction trajectories between V_0 and V_1 can be quantified. We state several results characterizing transition mechanism from V_0 to V_1.

Remark

- By a reaction (transition) trajectory, we mean a sequence of transitions from V_0 to V_1, *i.e.* $(x_{t_1}, x_{t_1+1}, \ldots, x_{t_2})$ such that $x_{t_1} \in V_0$, $x_{t_2} \in V_1$, and $x_{t_k} \in V - (V_0 \cup V_1)$ for $t_1 < t_k < t_2$.

- Denote by R the set of such reaction trajectories.

Proposition

Proposition (Probability distribution of reactive trajectories)

The probability distribution of reactive trajectories

$$\pi_R(x) = \mathbb{P}(X_n = x, n \in R) \tag{5}$$

is given by

$$\pi_R(x) = \pi(x)q(x)(1 - q(x)). \tag{6}$$

- The distribution π_R gives the equilibrium probability that a reactive trajectory visits x. It provides information about the proportion of time the reactive trajectories spend in state x along the way from V_0 to V_1.

Proposition (Reactive current from V_0 to V_1)

The reactive current from $A = V_0$ to $B = V_1$, defined by

$$J(xy) = \mathbb{P}(X_n = x, X_{n+1} = y, \{n, n+1\} \subset R), \tag{7}$$

is given by

$$J(xy) = \begin{cases} \pi(x)(1 - q(x))P_{xy}q(y), & x \neq y; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

- The reactive current $J(xy)$ gives the average rate the reactive trajectories jump from state x to y. From the reactive current, we may define the effective reactive current on an edge and transition current through a node which characterizes the importance of an edge and a node in the transition from A to B, respectively.

Effective Reactive Current

Definition

The *effective current* of an edge xy is defined as

$$J^+(xy) = \max(J(xy) - J(yx), 0). \tag{9}$$

The *transition current* through a node $x \in V$ is defined as

$$T(x) = \begin{cases} \sum_{y \in V} J^+(xy), & x \in A = V_0 \\ \sum_{y \in V} J^+(yx), & x \in B = V_1 \\ \sum_{y \in V} J^+(xy) = \sum_{y \in V} J^+(yx), & x \notin A \cup B \end{cases} \tag{10}$$

- The effective reactive current on an edge and transition current through a node characterize the importance of an edge and a node in the transition from A to B, respectively.

Effective Reactive Current

- In applications one often examines partial transition current through a node connecting two communities $V^- = \{x : q(x) < 0.5\}$ and $V^+ = \{x : q(x) \geq 0.5\}$, *e.g.* $\sum_{y \in V^+} J^+(xy)$ for $x \in V^-$, which shows relative importance of the node in bridging communities.

Reaction Rate

Proposition (Reaction rate)

The reaction rate from $A = V_0$ to $B = V_1$ is given by

$$\nu = \sum_{x \in A, y \in V} J(xy) = \sum_{x \in V, y \in B} J(xy). \tag{11}$$

- The reaction rate ν, defined as the number of transitions from V_0 to V_1 happened in a unit time interval, can be obtained from adding up the probability current flowing out of the reactant state. This is stated by the next proposition.

Time Portion from A and B

- Finally, the committor functions also give information about the time proportion that an equilibrium trajectory comes from $A = V_0$ (the trajectory hits A last rather than $B = V_1$).

Proposition

The proportion of time that the trajectory comes from $A = V_0$ (resp. from $B = V_1$) is given by

$$\rho^A = \sum_{x \in V} \pi(x) q(x), \quad \rho^B = \sum_{x \in V} \pi(x)(1 - q(x)). \tag{12}$$

Example: Karate Club network

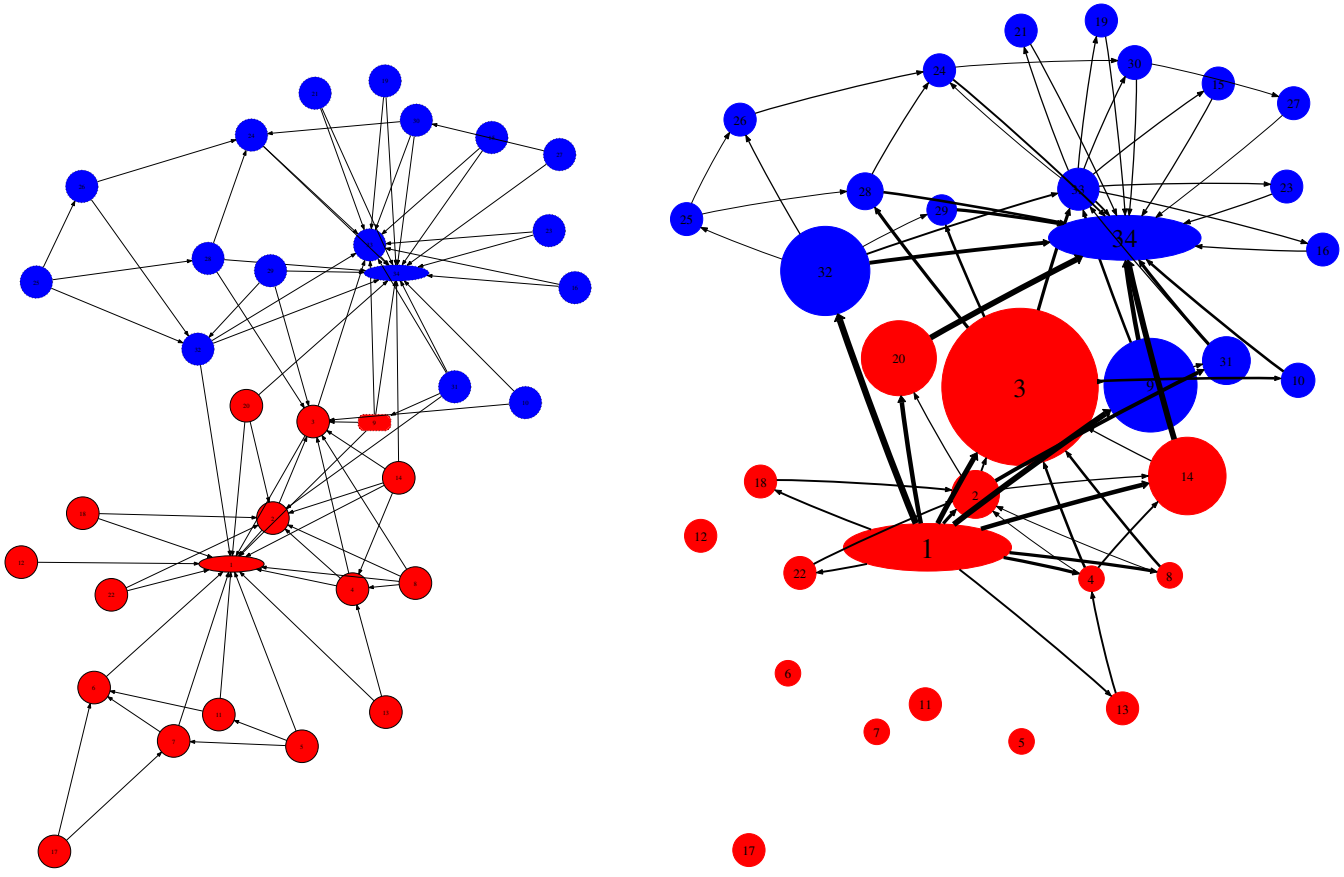


Figure: Effective Transition Current

Semi-supervised Learning

- Problem: $x_1, x_2, ..., x_l \in V_l$ are labeled data, that is data with the value $f(x_i), f \in V \to \mathbb{R}$ observed. $x_{l+1}, x_{l+2}, ..., x_{l+u} \in V_u$ are unlabeled. Our question is how to fully exploit the information provided in the labeled and unlabeled data to find the unobserved labels.

Semi-supervised Learning as Harmonic Extension

- Suppose the whole graph is $G = (V, E, W)$, where $V = V_l \cup V_u$ and weight matrix is partitioned into blocks $W = \begin{pmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{pmatrix}$. As before, we define $D = \mathbf{diag}(d_1, d_2, ..., d_n) = \mathbf{diag}(D_l, D_u),\ d_i = \sum_{j=1}^{n} W_{ij},\ L = D - W$.

- The goal is to find $f_u = (f_{l+1}, ..., f_{l+u})^T$ such that

$$\begin{aligned} \min \quad & f^T L f \\ s.t. \quad & f(V_l) = f_l \end{aligned}$$

where $f = \begin{pmatrix} f_l \\ f_u \end{pmatrix}$. This is a **Laplacian equation with Dirichlet boundary condition**.

Semi-supervised Learning and Committor Function

- Note that

$$f^T L f = (f_l^T, f_u^T) L \begin{pmatrix} f_l \\ f_u \end{pmatrix} = f_u^T L_{uu} f_u + f_l^T L_{ll} f_l + 2 f_u^T L_{ul} f_l$$

  So we have:

$$\frac{\partial f^T L f}{\partial f_u} = 0 \Rightarrow 2 L_{uu} f_u + 2 L_{lu} f_u = 0$$

$$\Rightarrow f_u = -L_{uu}^{-1} L_{ul} f_l = (D_u - W_{uu})^{-1} W_{ul} f_l$$

- This is the same equation as *committor function* (4) without probability constraints on f.

Summary

- We have introduced random walk on graphs with spectral characterization:
  - Perron-Frobenius Theory for primary eigenvector: e.g. PageRank
  - Fiedler Theory for Unnormalized Laplacian: e.g. algebraic connectivity and spectral partition
  - Cheeger Inequality for Normalized Laplacian: e.g. Approximate Normalized Cut and spectral clustering
  - Lumpability of Markov Chains: e.g. multiple spectral clustering

Summary

- More:
  - Mean First Passage Time and Commute Time Distance: e.g. pseudo-inverse of unnormalized Laplacian
  - Transition Path Theory: Dirichlet boundary problem for unnormalized Laplacian equations, e.g. semi-supervised learning