Homework 3. Random Walks on Graphs

Instructor: Yuan Yao

1. Let $v \in \mathbb{R}^n$. Find the eigenvectors and eigenvalues of the $n \times n$ cyclic matrix C whose entries are given by
$$C_{ij} = \begin{cases} v(i-j+1) & \text{for } 1 \le j \le i \le n, \\ v(i-j+n+1) & \text{for } 1 \le i < j \le n. \end{cases}$$
Suppose v is symmetric in the sense that it satisfies
$$v(i) = v(n+1-i) \quad \text{for } i = 1, \ldots, n.$$
What can be said about the eigenvalues and eigenvector of C in the symmetric case?

2. Suppose $G = (V, E)$ is a bipartite graph, that is, $V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$ and $E \subset (V_1 \times V_2) \cup (V_2 \times V_1)$ (i.e., $E \cap (V_k \times V_k) = \emptyset$, for $k = 1, 2$). Show that $\lambda = -1$ is an eigenvalue of the random walk (row) Markov matrix P associated with G.

3. Define the normalized K-cut of a graph with n vertices into $K \ge 2$ disjoint sets $A_1, A_2, \ldots, A_K$ as
$$\mathrm{NCut}(A_1, A_2, \ldots, A_K) = \sum_{k=1}^{K} \frac{\mathrm{cut}(A_k)}{\mathrm{vol}(A_k)}.$$
Let the indicator vectors $h_1, h_2, \ldots, h_K$ be given by
$$h_k(i) = \begin{cases} \frac{1}{\sqrt{\mathrm{vol}(A_k)}} & \text{for } i \in A_k \\ 0 & \text{for } i \notin A_k \end{cases},$$
and let H be the $n \times K$ matrix whose columns are the h_k's, i.e., $H = [h_1\ h_2\ \ldots\ h_K]$.
   (a) Show that the normalized K-cut can be written as
   $$\mathrm{NCut}(A_1, A_2, \ldots, A_K) = \mathrm{Tr}(H^T L H),$$
   where L is the unnormalized graph Laplacian.
   (b) Show that $H^T D H = I$, where I is the $K \times K$ identity matrix.
   (c) Conclude a spectral relaxation algorithm for the K-cut problem, like the generalized eigenvector for bipartition problem.

4. Let $G = (V, E)$ be a connected simple graph. For two nodes i and j, the *distance* between i and j is the number of edges in a shortest path joining i and j. The *diameter* of graph G is the maximum distance between any two nodes of G, denoted by D. Show that
$$\lambda_1(\mathcal{L}) \ge \frac{1}{D vol(G)},$$
where $\lambda_1(\mathcal{L})$ is the second smallest eigenvalue of normalized graph Laplacian $\mathcal{L}$ (Hint: Lemma 1.9, Chung, Spectral Graph Theory). Consider the linear chain graph on n nodes $\{1, 2, \ldots, n\}$ with $n-1$ edges given by $(i, i+1)$ for $i = 1, \ldots, n-1$. Which lower bound is tighter for chain graph, $\lambda_1 \ge \frac{h_G^2}{2}$ and the one above? Can you give an example of a graph for which the other lower bound is tighter?

5. Consider a random walk that starts at node i. Show that the probability $\pi_t(j)$ to find the random walker at time t at node j satisfies

$$|\pi_t(j) - \pi_\infty(j)| \le \sqrt{\frac{d(j)}{d(i)}}\,(1 - \lambda_2(\mathcal{L}))^t,$$

where $d(i)$ and $d(j)$ are the degrees of nodes i and j, π_∞ is the stationary distribution, and $\lambda_2(\mathcal{L})$ is the second eigenvalue of the normalized Laplacian (Assume that $|1 - \lambda_2(\mathcal{L})| = \max_{i \neq 1} |1 - \lambda_i(\mathcal{L})|$. Reference: Section 1.5, Chung, Spectral Graph Theory).

6. Consider a lazy random walk on the ring graph on n nodes with $A_{i,i} = \frac{1}{2}$ and $A_{i,i+1} = A_{i,i-1} = \frac{1}{4}$. Use MATLAB to compute the squared diffusion distances $D_t^2(i,j)$ and the squared commute time distances $T_{ij} + T_{ji}$ for several values of n and t. Compare these distances with the geodesic distance $|i-j|$. Can you find an analytic expression for $D_t^2(i,j)$? Proceed as far as possible.

7. Suppose $P = D^{-1}W$ is irreducible, and consider the first passage time τ_{ij} (with the convention that $\tau_{ii} = 0$).
   (a) Find the second moment $\mathbb{E}[\tau_{ij}^2]$ and conclude the variance $\operatorname{Var}\tau_{ij}$ (Reference: Section 4.5, Kemeny Snell 1976, Finite Markov Chains.).
   (b) Is $\mathbb{E}[\tau_{ij}^2] + \mathbb{E}[\tau_{ji}^2]$ a (squared) distance? If so, find the embedding.
   (c) Is $\operatorname{Var}\tau_{ij} + \operatorname{Var}\tau_{ji}$ a (squared) distance? If so, find the embedding.

8. Suppose $d : V \times V \to \mathbb{R}$ is a distance function.
   (a) Is d^2 a distance function? Prove or give a counterexample.
   (b) Is $\sqrt{d}$ a distance function? Prove or give a counterexample.

9. Suppose that $S \in \mathcal{R}^{n \times n}$ is a positive semidefinite (PSD) matrix, i.e. for any $v \in \mathcal{R}^n$, $v^T S v \ge 0$.
   (a) Let
   $$d_S^2(i,j) = S_{ii} + S_{jj} - 2S_{ij}.$$
   Show that d_S is a distance function.
   (b) Show that S is PSD iff $S = A^T A$ for some A.
   (c) Let H be the Hadamard square of S, i.e. $H_{ij} = [S_{ij}]^2$ (in MATLAB notation: `H = S .* S`). Prove that H is also PSD.

10. Download the mainland Chinese university weblink data at

    http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/univ_cn.mat

    Compare stability of PageRank when α changes as well as HITS authority and hub ranking.

11. Design a spectral clustering method to decompose the Karate Club network into 2 classes, and compare your result with the ground truth given by vector `c0` in data. Data can be found at

    http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/karate.mat

12. Choose arbitrary THREE digits from

    http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.digits/

    Design a spectral clustering algorithm to cluster all of your data into 3 classes. Compute the confusion matrix of your algorithm: $C \in \mathcal{R}^{3 \times 3}$ such that $C(i,j)$ denotes the number of samples that your algorithmic outputs i while the genuine class label is j.