CS168, Spring 2024

Mini-Project #6

Due by 11am on Thursday, May 16

Instructions

- You can work in groups of up to four students. If you work in a group, please submit one assignment via Gradescope (with all group members' names).
- Detailed submission instructions can be found on the course website (`https://web.stanford.edu/class/cs168`) under "Coursework - Assignments" section.
- Use 12pt or higher font for your writeup.
- Make sure the plots you submit are easy to read at a normal zoom level.
- If you've written code to solve a certain part of a problem, or if the part explicitly asks you to implement an algorithm, you must also include the code in your pdf submission. You do not need to mark the pages having the code when you mark pages containing answers to questions on Gradescope.
- Code marked as Deliverable should be pasted into the relevant section. Keep variable names consistent with those used in the problem statement, and with general conventions. No need to include import statements and other scaffolding, if it is clear from context. Use the `verbatim` or minted environments to paste code in LaTeX.

```
def example():
     print "Your code should be formatted like this."
```

- **Reminder:** No late assignments will be accepted, but we will drop your lowest assignment grade.

Part 1: Spectral Methods Intuition

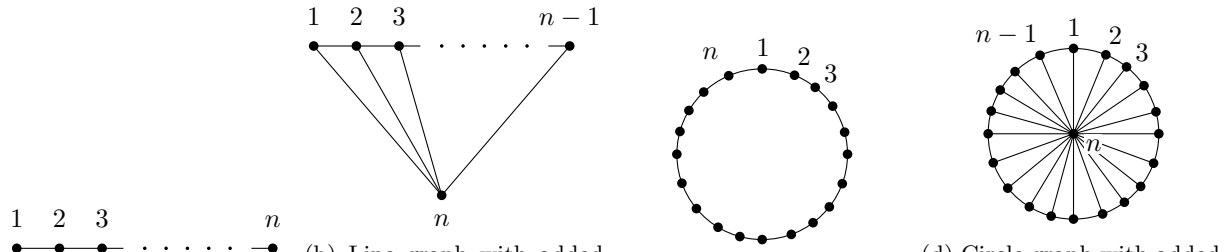
Goal:
In this exercise you will build some intuition for the eigenvectors of various simple graphs.

Description:

(a) (6 points) Consider the graphs as given in Figure 1. For this part, take $n = 5$. For each graph write down the Laplacian matrix $L = D - A$ where D is the diagonal matrix with entry $D_{i,i}$ being the degree of the ith node, and A is the adjacency matrix, with entry $A_{i,j} = 1$ if there is an edge between nodes i and j, and $A_{i,j} = 0$ otherwise. Your answer should be in the form of actual matrices (i.e., not just English descriptions of matrices).

(b) (9 points) For each of the graphs of question (a), compute the eigenvectors and eigenvalues of the Laplacian matrix L and the adjacency matrix A, when there are $n = 100$ vertices. For both L and A, plot the eigenvectors corresponding to the two smallest and two largest eigenvalues. Please include eight plots; two for each graph with one corresponding to the eigenvectors of L and one corresponding to A. Clearly label the four eigenvectors on each plot[1]. In light of the interpretation of $\mathbf{v}^t L\mathbf{v} = \frac{1}{2}\sum_{(i,j)\in E}(\mathbf{v}(i) - \mathbf{v}(j))^2$, explain why these eigenvectors make sense. One brief (2-4 sentence) explanation is sufficient.

[1]When plotting an eigenvector $\mathbf{v}$, the x-axis ranges from 1 through n, and the ith point is plotted at location $(i, \mathbf{v}(i))$.



(a) Line graph.

(b) Line graph with added point.

(c) Circle graph.

(d) Circle graph with added point.

Figure 1: The graphs for question 1a.

(c) (6 points) For this question, you will consider each of the 4 graphs of part (a) in the case that $n = 100$. For each such $n = 100$ node graph, plot the embedding of the graph onto the eigenvectors corresponding to the 2nd and 3rd smallest eigenvalues of the Laplacian. That is: if $\mathbf{v}_2$ is the second eigenvector and $\mathbf{v}_3$ is the third eigenvector of the Laplacian, create a scatter plot with the points $(\mathbf{v}_2(i), \mathbf{v}_3(i))$ for $i \in \{1, \ldots 100\}$. Overlay the edges of the graph, i.e. for every pair of points i, j, that are connected in the graph G, draw an edge between $(\mathbf{v}_2(i), \mathbf{v}_3(i))$ and $(\mathbf{v}_2(j), \mathbf{v}_3(j))$.

(d) (4 points) Pick 500 random points in the unit square by independently choosing their x and y coordinates uniformly at random from the interval $[0, 1]$. Form a graph by adding an edge between every pair of points whose Euclidean distance is at most $1/4$. Compute the eigenvectors of the Laplacian of this graph. Plot the embedding of this graph onto the second and third eigenvectors (i.e. those corresponding to the 2nd and 3rd smallest eigenvalues). Do *not* overlay the edges of the graph, just plot the vertices. For all points in the original graph with x and y coordinates both less than $1/2$, plot their embeddings in a different color. Are these points clustered together in the embedding? Why does this make sense?

(e) (2 points) Repeat the previous part, but instead plot the embedding of this graph onto the *largest* two eigenvectors. As in the previous part, for all points in the original graph with x and y coordinates both less than $1/2$, plot their images in a different color. Are these points clustered together in the embedding? Why does this make sense?

Deliverables: For part (a) the 4 Laplacian matrices; for part (b) the code, the description and seven plots for each of the 4 graphs (*not* eight, since one of them would be trivial), and your discussion; for part (c) for each of the 4 graphs a plot of the spectral embedding; for (d) and (e) plots in two colors and discussion.

Part 2: Finding Friends

Goal: Experience the magic of graph spectra: use the eigenvectors of a (tiny) subset of the facebook graph to find large, insular groups of friends.

Description: In this part you will play with a part of the Facebook friend graph to get some appreciation for using spectral methods on a real dataset. The data come from the Stanford Network Analysis Project (SNAP)[2]. The file `cs168mp6.csv` is part of the `ego-Facebook` dataset on the SNAP website. Facebook friendships are represented naturally by a node for each person, and an edge if and only if two people are friends on Facebook. In the dataset each row represents a friendship (edge) between two people (nodes) as identified by unique identifiers.

(a) *[do not hand in]* Load in the datafile and make sure that you have 61796 rows, and 1495 unique persons.

[2] `http://snap.stanford.edu/`

(b) (2 points) Compute the smallest 12 eigenvalues and corresponding eigenvectors of the Laplacian of the friendship graph.[3] Note: be sure to use the Laplacian of the graph, NOT the adjacency matrix. Print a list of the smallest 12 eigenvalues, rounded to the nearest 0.001.

(c) (7 points) How many connected components does this graph have, *and what are the sizes of the connected components*? Justify your answer using the eigenvalues and eigenvectors of the Laplacian. (Keep in mind that your linear algebra package might have some numerical issues, and an eigenvalue of 10^{-12} should probably be regarded as 0.) Explain how you came up with your answer.

(d) (10 points) The *conductance* of a set of nodes in a graph is a natural measure of how tightly knit/insular that set is, with a lower conductance indicating a more tightly knit set. Given a graph $G = (V, E)$ with adjacency matrix A, and a subset of the nodes $S \subset V$, the conductance is defined as:

$$cond(S) = \frac{\sum_{i \in S, j \in V \setminus S} A_{i,j}}{\min(A(S), A(V \setminus S))},$$

where $A(S)$ is the sum of degrees of vertices in set S. For example, if G is the circle graph (Figure 1c) and S is $n/2$ consecutive points (for simplicity assume n is a multiple of 2), then:

$$cond(S) = \frac{\sum_{i \in S, j \in V \setminus S} A_{i,j}}{\min(A(S), A(V \setminus S))} = \frac{2}{2 \cdot (n/2)} = \frac{2}{n}$$

In the context of a friend graph, the conductance of a set of individuals corresponds to the ratio of the number of friendships between the outside world and that set, to the total number of friendships involving that set.[4] If $cond(S) = 0$, then that set is disconnected from the rest of the graph, and if $cond(S) = 1$, then there are no internal friendships among members of that set.

Find at least 3 sets, S_1, S_2, and S_3 of people in the friendship graph, such that each set has at least 150 people, and no more than $n/2 \approx 750$ people, and each set has conductance at most 0.1. The three sets should also be disjoint (nearly disjoint is also okay, though its definitely not okay if one of the sets is a subset of another). For each set, report its size, 10 of its members, and the conductance. Explain how you found each set; in particular, for each set, if you used an eigenvectors to identify that set, please include a plot of those eigenvector, say which eigenvector it is (e.g. the vector corresponding to the 17th smallest eigenvalue) and explain how you identified the set of nodes in the component from that eigenvector (e.g. "This set corresponds to the vertices whose values in the 16th eigenvector are in the range $0.02 \pm .001$.")

[Hint: If the smallest s eigenvalues are zero, then you should probably look at the eigenvectors corresponding to the $(s+1)^{st}$ smallest and higher eigenvalues. Also, you might not be able to use a single eigenvector to find all three sets—its worth looking at a number of eigenvectors, even the 20th or 30th might still have some nice information about the clusters of friends.]

(e) (4 points) Now select a random set of 150 nodes, and compute the conductance of that set. Do the sets you found in part (d) seem tight-knit compared to this benchmark?

Deliverables: Your code for part (b); for part (c) the number of connected components and sizes of the components, and discussion; for part (d) three plots of eigenvectors showing the 3 sets S_1, S_2, S_3, and the sizes, conductances, and 10 members of the sets; for (e) the conductance and discussion.

[3] Keep in mind that if a graph's eigenvalue λ has multiplicity $\rho_\lambda > 1$, then the corresponding eigenvectors are not unique and will depend on the software you use—for this reason your eigenvectors might be different than those of your classmates, though the eigenvalues should be identical up to numerical precision.

[4] So researchers that look for tightly knit groups look for sets with a low conductance.