COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2023.

Lecture 20

Logistics

- Problem Set 3 is due Friday at 11:59pm.
- Problem Set 4 will be released immediately and due 12/1.
- Next Tuesday will be the last class of the spectral algorithms unit. We will take a closer look at how eigenvectors/singular vectors are actually computed in practice.

Summary

Last Class: Spectral Clustering

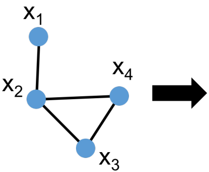
- Spectral clustering: finding good cuts via Laplacian eigenvectors.
- The second smallest eigenvector can be used to find a small but balanced cut.
- Heuristic argument. Mathematical motivation via Courant-Fischer, but no formal proofs.

This Class: The Stochastic Block Model

- A simple clustered graph model where we can prove the effectiveness of spectral clustering (i.e., clustering with the Laplacian eigenvectors)

Review

For a graph with adjacency matrix $\mathbf{A}$ and degree matrix $\mathbf{D}$, $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the graph Laplacian.



$$\overset{\mathbf{D}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}} - \overset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}} = \overset{\mathbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}}$$

How smooth any vector $\vec{v}$ is over the graph can be measured by:

$$\sum_{(i,j)\in E} (\vec{v}(i) - \vec{v}(j))^2 = \vec{v}^T \mathbf{L} \vec{v}.$$

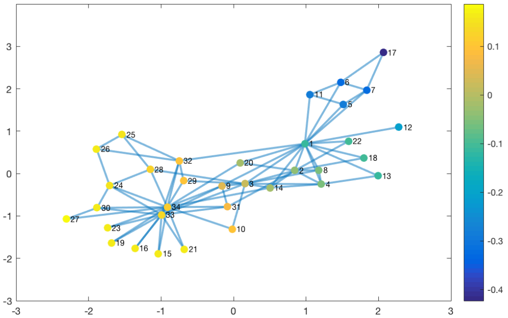
- The second smallest eigenvector $\vec{v}_{n-1}$ of $\mathbf{L}$, minimizes $\vec{v}_{n-1}^T \mathbf{L} \vec{v}_{n-1}$ subject to $\vec{v}_{n-1}^T \vec{1} = 0$.
- By thresholding this vector, we tend to find small cuts ($\vec{v}_{n-1}^T \mathbf{L} \vec{v}_{n-1}$ is small), that are well-balanced ($\vec{v}_{n-1}^T \vec{1} = 0$).

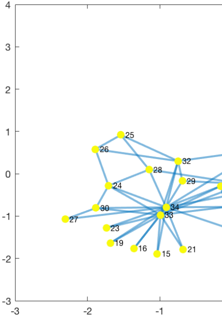
Cutting With the Second Laplacian Eigenvector

Find a good partition of the graph by computing

$$\vec{v}_{n-1} = \underset{v \in \mathbb{R}^d \text{with } \|\vec{v}\|=1,\ \vec{v}^T\vec{1}=0}{\arg\min} \vec{v}^T \mathbf{L} \vec{v}$$

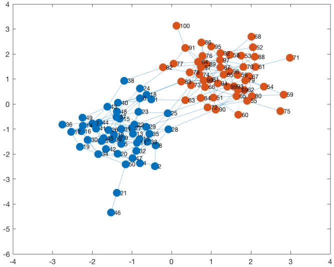
Set S to be all nodes with $\vec{v}_{n-1}(i) < 0$, T to be all with $\vec{v}_{n-1}(i) \geq 0$.





Stochastic Block Model (Planted Partition Model): Let $G_n(p, q)$ be a distribution over graphs on n nodes, split randomly into two groups B and C, each with $n/2$ nodes.

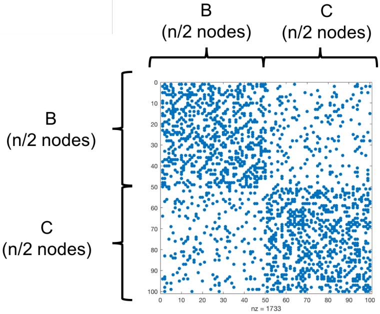
- Any two nodes in the same group are connected with probability p (including self-loops).
- Any two nodes in different groups are connected with prob. $q < p$.
- Connections are independent.

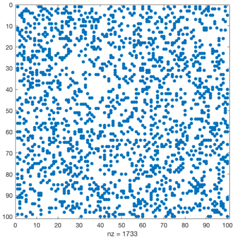


Linear Algebraic View

Let G be a stochastic block model graph drawn from $G_n(p,q)$.

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G, ordered in terms of group ID.





$G_n(p,q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the the same group, and probability q between nodes not in the same group.

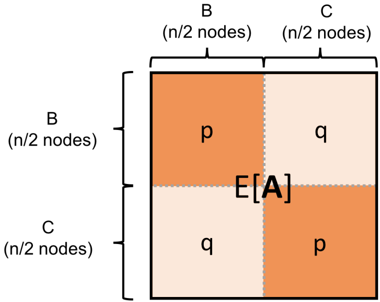
Expected Adjacency Matrix

Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix. What is $\mathbb{E}[\mathbf{A}]$?

$G_n(p, q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the same group, and probability q between nodes not in the same group.

Expected Adjacency Spectrum

Letting G be a stochastic block model graph drawn from $G_n(p,q)$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ be its adjacency matrix. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i,j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.



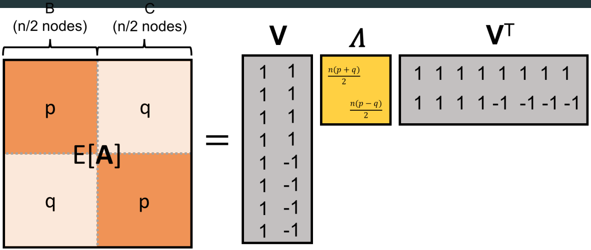
What is $\text{rank}(\mathbb{E}[\mathbf{A}])$? What are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

$G_n(p,q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the same group, and probability q between nodes not in the same group.

Expected Adjacency Spectrum

Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

Expected Adjacency Spectrum

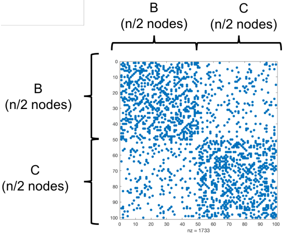


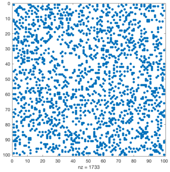
If we compute $\vec{v}_2$ then we recover the communities B and C!

- Can show that for $G \sim G_n(p, q)$, $\mathbf{A}$ is close to $\mathbb{E}[\mathbf{A}]$ with high probability (matrix concentration inequality).
- Thus, the true second eigenvector of A is close to $[1, 1, 1, \ldots, -1, -1, -1]$ and gives a good estimate of the communities.

Spectrum of Permuted Matrix

Goal is to recover communities – so adjacency matrix won't be ordered in terms of community ID (or our job is already done!)





- Actual adjacency matrix is $\mathbf{PAP}^T$ where $\mathbf{P}$ is a random permutation matrix and $\mathbf{A}$ is the ordered adjacency matrix.
- **Exercise (see Problem Set 3):** The first two eigenvectors of $\mathbf{PAP}^T$ are $\mathbf{P}\vec{v}_1$ and $\mathbf{P}\vec{v}_2$.
- $\mathbf{P}\vec{v}_2 = [1, -1, 1, -1, \ldots, 1, 1, -1]$ gives community ids.

Expected Laplacian Spectrum

Letting G be a stochastic block model graph drawn from $G_n(p, q)$, $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix and $\mathbf{L}$ be its Laplacian, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{L}]$?

Expected Laplacian Spectrum

Letting G be a stochastic block model graph drawn from $G_n(p,q)$, $\mathbf{A} \in \mathbb{R}^{n\times n}$ be its adjacency matrix and $\mathbf{L}$ be its Laplacian, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{L}]$?

Expected Laplacian Spectrum

Upshot: The second smallest eigenvector of $\mathbb{E}[\mathbf{L}]$ is $\chi_{B,C}$ – the indicator vector for the cut between the communities.

- If the random graph G (equivilantly $\mathbf{A}$ and $\mathbf{L}$) were exactly equal to its expectation, partitioning using this eigenvector (i.e., spectral clustering) would exactly recover the two communities B and C.

How do we show that a matrix (e.g., $\mathbf{A}$) is close to its expectation? Matrix concentration inequalities.

- Analogous to scalar concentration inequalities like Markovs, Chebyshevs, Bernsteins.
- Random matrix theory is a very recent and cutting edge subfield of mathematics that is being actively applied in computer science, statistics, and ML.

Everything after this slide is bonus material, if you are interested in how we formally prove that spectral clustering succeeds in the stochastic block model, using matrix concentration bounds.

Matrix Concentration

Matrix Concentration Inequality: If $p \geq O\left(\frac{\log^4 n}{n}\right)$, then with high probability

$$\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\|_2 \leq O(\sqrt{pn}).$$

where $\|\cdot\|_2$ is the matrix spectral norm (operator norm).

For any $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\|\mathbf{X}\|_2 = \max_{z\in\mathbb{R}^d:\|z\|_2=1} \|\mathbf{X}z\|_2$.

Exercise: Show that $\|\mathbf{X}\|_2$ is equal to the largest singular value of $\mathbf{X}$. For symmetric $\mathbf{X}$ (like $\mathbf{A} - \mathbb{E}[\mathbf{A}]$) show that it is equal to the magnitude of the largest magnitude eigenvalue.

For the stochastic block model application, we want to show that the second eigenvectors of $\mathbf{A}$ and $\mathbb{E}[\mathbf{A}]$ are close. How does this relate to their difference in spectral norm?

Eigenvector Perturbation

Davis-Kahan Eigenvector Perturbation Theorem: Suppose $\mathbf{A}, \bar{\mathbf{A}} \in \mathbb{R}^{d \times d}$ are symmetric with $\|\mathbf{A} - \bar{\mathbf{A}}\|_2 \leq \epsilon$ and eigenvectors $v_1, v_2, \ldots, v_d$ and $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_d$. Letting $\theta(v_i, \bar{v}_i)$ denote the angle between v_i and $\bar{v}_i$, for all i:

$$\sin[\theta(v_i, \bar{v}_i)] \leq \frac{\epsilon}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\bar{\mathbf{A}}$.

The errors get large if there are eigenvalues with similar magnitudes.

Eigenvector Perturbation

$$\underset{\mathbf{A}}{\begin{bmatrix} 1+\varepsilon & 0 \\ 0 & 1 \end{bmatrix}} - \underset{\bar{\mathbf{A}}}{\begin{bmatrix} 1 & 0 \\ 0 & 1+\varepsilon \end{bmatrix}} = \underset{\mathbf{A}-\bar{\mathbf{A}}}{\begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}}$$

Application to Stochastic Block Model

Claim 1 (Matrix Concentration): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\|_2 \leq O(\sqrt{pn}).$$

Claim 2 (Davis-Kahan): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p-q)n/2} = O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

Recall: $\mathbb{E}[\mathbf{A}]$, has eigenvalues $\lambda_1 = \frac{(p+q)n}{2}$, $\lambda_2 = \frac{(p-q)n}{2}$, $\lambda_i = 0$ for $i \geq 3$.

$$\min_{j \neq i} |\lambda_i - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$

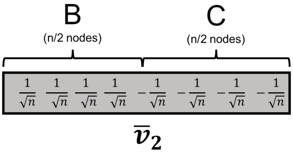
Typically, $\frac{(p-q)n}{2}$ will be the minimum of these two gaps.

> $\mathbf{A}$ adjacency matrix of random stochastic block model graph. p: connection probability within clusters. $q < p$: connection probability between clusters. n: number of nodes. $v_2, \bar{v}_2$: second eigenvectors of $\mathbf{A}$ and $\mathbb{E}[\mathbf{A}]$ respectively.

Application to Stochastic Block Model

So Far: $\sin\theta(v_2, \bar{v}_2) \leq O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$. What does this give us?

- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{v}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.

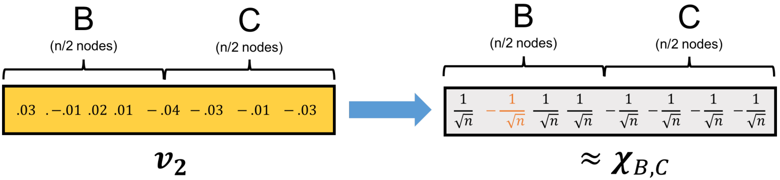


- Every i where $v_2(i)$, $\bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.
- So they differ in sign in at most $O\left(\frac{p}{(p-q)^2}\right)$ positions.

A adjacency matrix of random stochastic block model graph. p: connection probability within clusters. $q < p$: connection probability between clusters. n: number of nodes. $v_2, \bar{v}_2$: second eigenvectors of **A** and $\mathbb{E}[\mathbf{A}]$ respectively.

Application to Stochastic Block Model

Upshot: If G is a stochastic block model graph with adjacency matrix $\mathbf{A}$, if we compute its second large eigenvector v_2 and assign nodes to communities according to the sign pattern of this vector, we will correctly assign all but $O\left(\frac{p}{(p-q)^2}\right)$ nodes.



- Why does the error increase as q gets close to p?
- Even when $p - q = O(1/\sqrt{n})$, assign all but an $O(n)$ fraction of nodes correctly. E.g., assign 99% of nodes correctly.