

COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Spring 2020.

Lecture 19

- Problem Set 3 due this upcoming Monday at 8pm.
- Final to be held on Zoom: May 6th from 1:00pm-3:00pm.

Last Class: Spectral Clustering

- Splitting a graph into communities is important in network analysis and non-linear data analysis.
- Want to find a **small cut** that is also **balanced**.
- Argued that the second smallest eigenvector of the graph Laplacian matrix can be used to find such a cut.
- Intuitive argument but not a formal proof that the identified cut is 'good'.

This Class: The Stochastic Block Model

- A simple clustered graph model where we can prove the effectiveness of spectral clustering.
- One of the most important random graph models.

So Far: Have argued that spectral clustering partitions a graph effectively, along a small cut that separates the graph into large pieces. But it is difficult to give any formal guarantee on the 'quality' of the partitioning in general graphs.

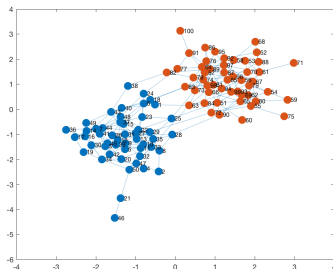
Common Approach: Give a natural **generative model** for random inputs and analyze how the algorithm performs on inputs drawn from this model.

- Very common in algorithm design for data analysis/machine learning (can be used to justify least squares regression, k -means clustering, PCA, etc.)

STOCHASTIC BLOCK MODEL

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.

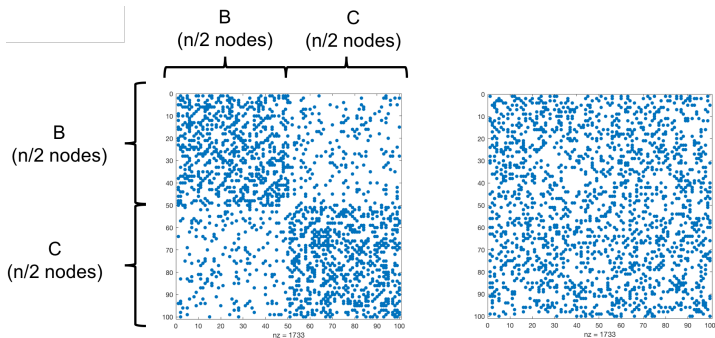
- Any two nodes in the **same group** are connected with probability p (including self-loops).
- Any two nodes in **different groups** are connected with prob. $q < p$.
- Connections are independent.



LINEAR ALGEBRAIC VIEW

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

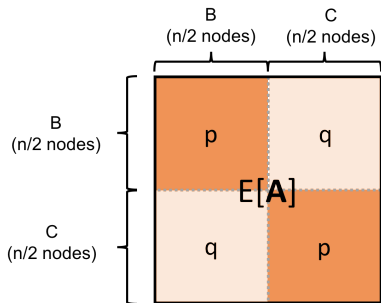
- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G , ordered in terms of group ID. What is $\mathbb{E}[\mathbf{A}]$?



$G_n(p, q)$: 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.

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 $\mathbb{E}[\mathbf{A}]$?

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

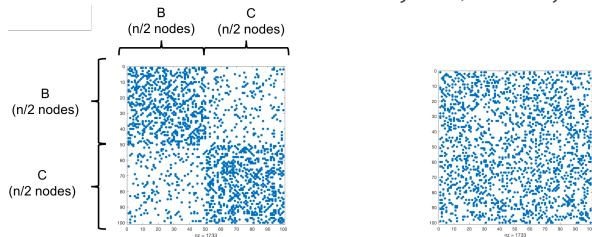
$$\begin{array}{c}
 \begin{array}{cc}
 \text{B} & \text{C} \\
 (n/2 \text{ nodes}) & (n/2 \text{ nodes})
 \end{array} \\
 \begin{array}{|c|c|}
 \hline
 \begin{array}{c} p \\ q \end{array} & \begin{array}{c} q \\ p \end{array} \\
 \hline
 \begin{array}{c} q \\ p \end{array} & \begin{array}{c} p \\ q \end{array} \\
 \hline
 \end{array}
 \mathbb{E}[\mathbf{A}]
 =
 \begin{array}{|c|}
 \hline
 \mathbf{V} \\
 \hline
 \begin{array}{c} 1 \ 1 \\ 1 \ 1 \\ 1 \ 1 \\ 1 \ 1 \\ 1 \ -1 \\ 1 \ -1 \\ 1 \ -1 \\ 1 \ -1 \end{array} \\
 \hline
 \end{array}
 \begin{array}{|c|}
 \hline
 \mathbf{\Lambda} \\
 \hline
 \begin{array}{c} \frac{n(p+q)}{2} \\ \frac{n(p-q)}{2} \end{array} \\
 \hline
 \end{array}
 \begin{array}{|c|}
 \hline
 \mathbf{V}^T \\
 \hline
 \begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{array} \\
 \hline
 \end{array}
 \end{array}$$

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 \mathbf{A} is close to $[1, 1, 1, \dots, -1, -1, -1]$ 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:** The first two eigenvectors of \mathbf{PAP}^T are $\mathbf{P}\vec{v}_1$ and $\mathbf{P}\vec{v}_2$.
- $\mathbf{P}\vec{v}_2 = [1, -1, 1, -1, \dots, 1, 1, -1]$ gives community ids.

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 $\mathbb{E}[\mathbf{L}]$?

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 $\mathbb{E}[\mathbf{L}]$?

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

- If the random graph G (equivalently \mathbf{A} and \mathbf{L}) were exactly equal to its expectation, partitioning using this eigenvector 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.

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

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?

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, \dots, v_d and $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_d$. 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 $\lambda_1, \dots, \lambda_d$ are the eigenvalues of $\bar{\mathbf{A}}$.

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

$$\begin{array}{c} \mathbf{A} \\ \begin{array}{|c|c|} \hline 1+\varepsilon & 0 \\ \hline 0 & 1 \\ \hline \end{array} \end{array} - \begin{array}{c} \bar{\mathbf{A}} \\ \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 1+\varepsilon \\ \hline \end{array} \end{array} = \begin{array}{c} \mathbf{A}-\bar{\mathbf{A}} \\ \begin{array}{|c|c|} \hline \varepsilon & 0 \\ \hline 0 & \varepsilon \\ \hline \end{array} \end{array}$$

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 $i \geq 3$.

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

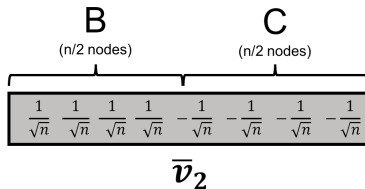
Typically, $\frac{(p-q)n}{2}$ will be the minimum of these two gaps.

An 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?

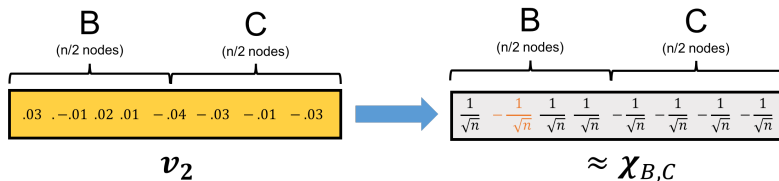
- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- \bar{v}_2 is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.



- 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 $\mathbb{E}[A]$ respectively.

Upshot: If G is a stochastic block model graph with adjacency matrix 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.