COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Spring 2020.

Lecture 19

LOGISTICS

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

SUMMARY

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’.

$$\underbrace{v^T L v}_{\text{size cut}} \quad \text{s.t.} \quad \underbrace{v^T \mathbb{1} = 0}_{\text{cut is balanced}}$$

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.

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 (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.

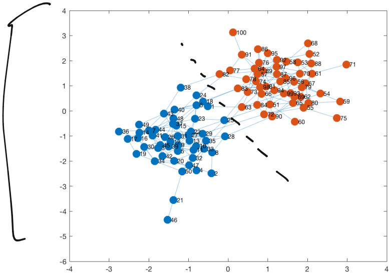
Stochastic Block Model (Planted Partition Model): Let $G_n(p,q)$ be a distribution over graphs on n nodes, split randomly into two groups B and C, each with $n/2$ nodes.

- 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.

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. $\mathbb{E}\deg(v_i) = p \cdot \frac{n}{2} + q \cdot \frac{n}{2}$

- 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.

V_{n-1}



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

$G_n(p, q)$: 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.

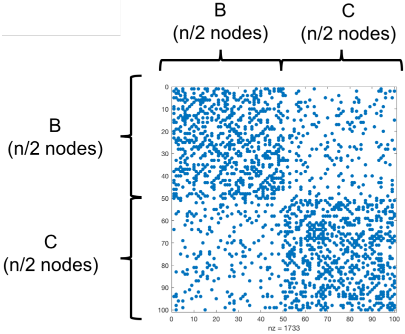
Let G be a stochastic block model graph drawn from $G_n(p, q)$.

- 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 same group, and probability q between nodes not in the same group.

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

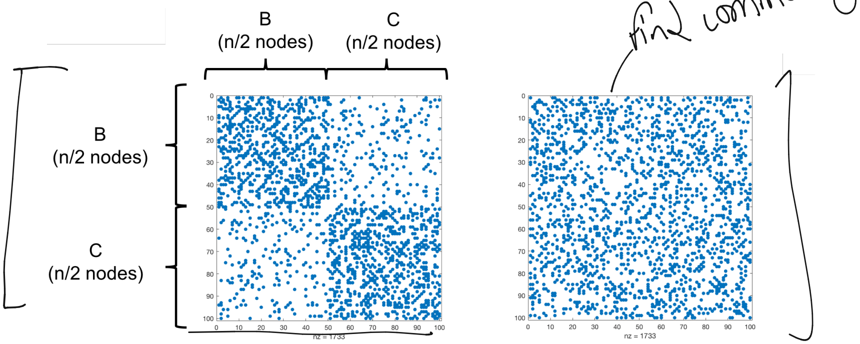
- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G, ordered in terms of group ID.



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

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.



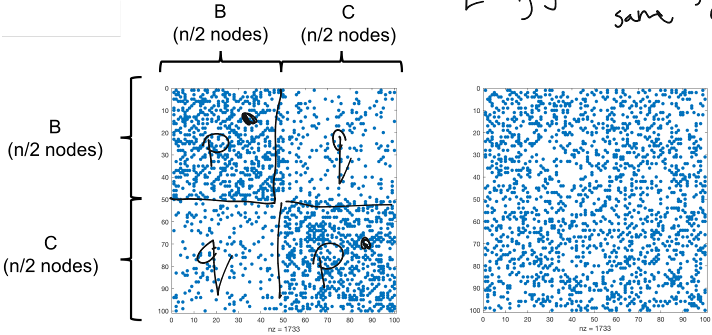
$G_n(p,q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the same group, and probability q between nodes not in the same group.

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

$\mathbb{E}[A_{ij}]$ v_i, v_j are in same community.

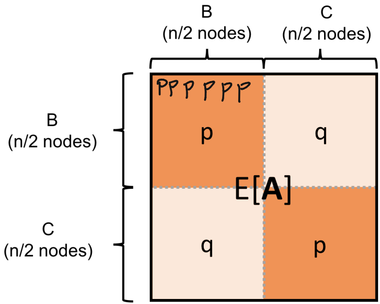


> $G_n(p, q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the same group, and probability q between nodes not in the same group.

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.

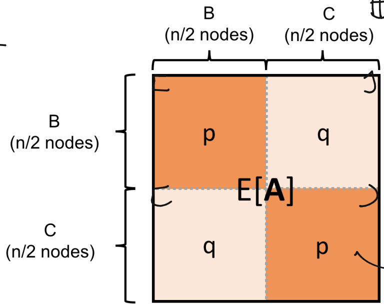
$\mathbb{E} A_{ii} = p$



$G_n(p, q)$: 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. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i, j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.

$$\mathbb{E}[A] = \begin{bmatrix} | & & \\ v_1 \, v_2 \, v_3 \ldots \\ | & & \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \begin{bmatrix} - v_1^T - \\ \vdots \\ - v_n^T - \end{bmatrix}$$



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

$\mathbb{E}[A] = V \Lambda V^T$

rank$(\mathbb{E}(A)) = 2$

nonzero eigenvalues = 2.

$G_n(p, q)$: 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 $\mathbb{E}[\mathbf{A}]$?

$n=2$: $\begin{bmatrix} p & q \\ q & p \end{bmatrix}$

$$\mathbb{E}[A] = \begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix} \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}$$

$$v_1 = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \Big\} n$$

$$\begin{bmatrix} p & q \\ q & p \end{bmatrix} v_1 = \frac{1}{\sqrt{n}} \begin{bmatrix} p & q \\ q & p \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{(p+q)n}{2} \\ \frac{(p+q)n}{2} \\ \vdots \end{bmatrix} = \frac{(p+q)n}{2} \cdot v_1$$

$$\lambda_1 = \frac{(p+q)n}{2}$$

$$v_2 = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ \vdots \\ -1 \\ \vdots \\ -1 \end{bmatrix} \begin{matrix} \}\frac{n}{2} \\ \}\frac{n}{2} \end{matrix}$$

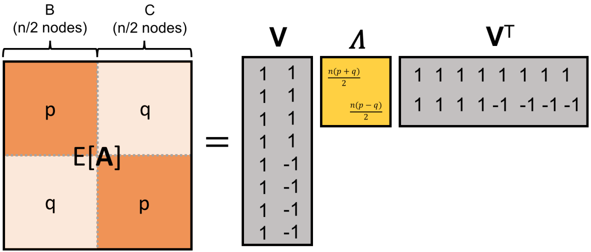
$$v_2^T v_1 = 0$$

$$\begin{bmatrix} p & q \\ q & p \end{bmatrix} v_2 = \frac{1}{\sqrt{n}} \begin{bmatrix} p & q \\ q & p \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ -1 \\ \vdots \\ -1 \end{bmatrix} = \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{(p-q)n}{2} \\ \frac{(p-q)n}{2} \\ -\frac{(p-q)n}{2} \\ \vdots \end{bmatrix} \begin{matrix} \}\frac{n}{2} \\ \}\frac{n}{2} \end{matrix} = \frac{(p-q)n}{2} \cdot v_2$$

$$\lambda_2 = \frac{(p-q)n}{2}$$

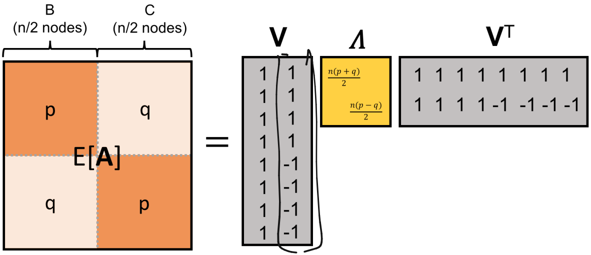
Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

EXPECTED ADJACENCY SPECTRUM



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

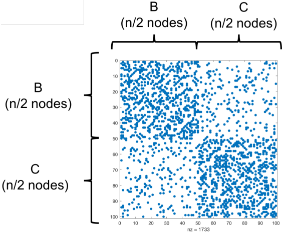
EXPECTED ADJACENCY SPECTRUM

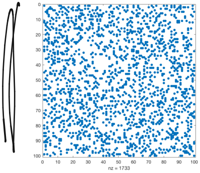


If we compute $\vec{v}_2$ then we recover the communities B and C!

- 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.

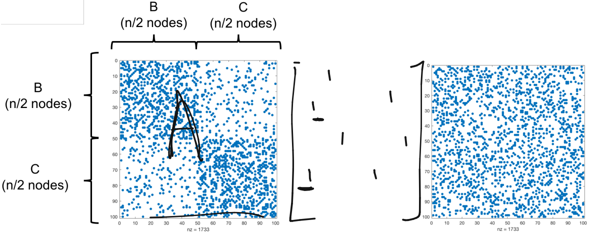
Goal is to recover communities – so adjacency matrix won't be ordered in terms of community ID (or our job is already done!)





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 $\mathbf{P}\vec{v}_2$.
- $\mathbf{P}\vec{v}_2 = [1, -1, 1, -1, \ldots, 1, 1, -1]$ gives community ids.

$P^T P = I$

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

$$L = \begin{bmatrix} D & & \\ & \ddots & \\ & & \ddots \end{bmatrix} - \begin{bmatrix} & A & \end{bmatrix}$$

$$\mathbb{E}[L] = \mathbb{E}[D] - \mathbb{E}[A] = \frac{(p+q)n}{2} \cdot I - \mathbb{E}[A] = \mathbb{E}[L]$$

$$\begin{bmatrix} \frac{(p+q)n}{2} & & \\ & \frac{(p+q)n}{2} & \\ & & \ddots \end{bmatrix} - \begin{bmatrix} p & q \\ q & p \end{bmatrix}$$

$$\mathbb{E}[L]v = \frac{(p+q)n}{2} v - \mathbb{E}[A]v$$

$v_i \rightarrow$ eigenvector of A and of L

$$\mathbb{E}[L] v_i = \left(\frac{(p+q)n}{2} - \lambda_i(A)\right)$$

$\longrightarrow$

$\frac{p+q}{2}n \;\_\_\; \cdots \;\_\_\; \frac{p+q}{2}n \quad qn \quad 0$

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

$\mathbb{E}[A]$: $v_1, v_2, v_3 \cdots v_n$ $\quad \lambda_i = 0 \;\; \forall i > 2.$

$\lambda_1 = \left(\frac{p+q}{2}\right)n \quad \lambda_2 = \left(\frac{p-q}{2}\right)n$

$\mathbb{E}[L]$: eigenvale of $\mathbb{E}[L]$ correspond to v_1

$\mathbb{E}[L]v_1 = \left(\frac{p+q}{2}\right)n\, v_1 - \mathbb{E}[A]v_1 = \left(\frac{p+q}{2}\right)n\, v_1 - \left(\frac{p+q}{2}\right)n\, v_1 = 0$

$\mathbb{E}[L]v_2 = \left(\frac{p+q}{2}\right)n\, v_2 - \left(\frac{p-q}{2}\right)n\, v_2 = (qn)\, v_2$

second smallest eigenvector

$\mathbb{E}[L]v_i = \left(\frac{p+q}{2}\right)n\, v_i - 0 = \left(\frac{p+q}{2}\right)n\, v_i$

$\begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}$

[1 1 1 -1 -1 -1]

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

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 (equivilantly $\mathbf{A}$ and $\mathbf{L}$) were exactly equal to its expectation, partitioning using this eigenvector would exactly recover the two communities B and C.

EXPECTED LAPLACIAN SPECTRUM

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

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.

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?

Davis-Kahan Eigenvector Perturbation Theorem: Suppose $\mathbf{A}, \overline{\mathbf{A}} \in \mathbb{R}^{d\times d}$ are symmetric with $\|\mathbf{A} - \overline{\mathbf{A}}\|_2 \leq \epsilon$ 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 $\overline{\mathbf{A}}$.

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

EIGENVECTOR PERTURBATION

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

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

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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

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

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 $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\sin \theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

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 $i \geq 3$.

$\mathbf{A}$ 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.

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

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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

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

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

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.

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 $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j\neq i}|\lambda_i - \lambda_j|}$$

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.

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.

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

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.

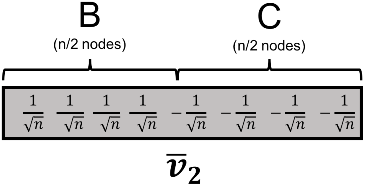
So Far: $\sin\theta(v_2, \bar{v}_2) \leq O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$. What does this give us?

- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).

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.

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?

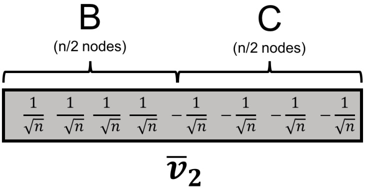
- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{v}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.



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.

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.



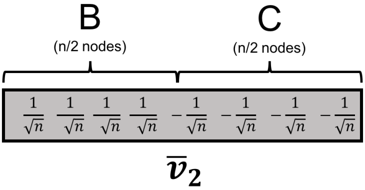
- Every i where $v_2(i), \bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.

> **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

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?

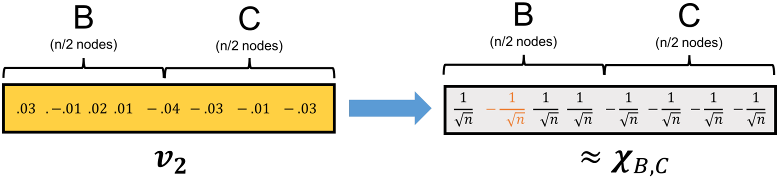
- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{v}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.



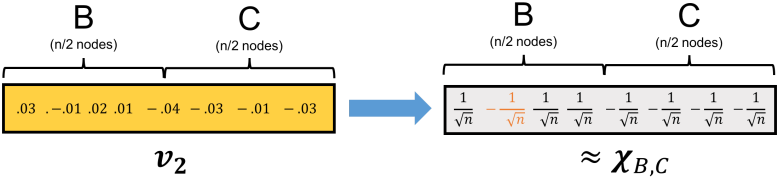
- Every i where $v_2(i), \bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.
- So they differ in sign in at most $O\left(\frac{p}{(p-q)^2}\right)$ positions.

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.

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.

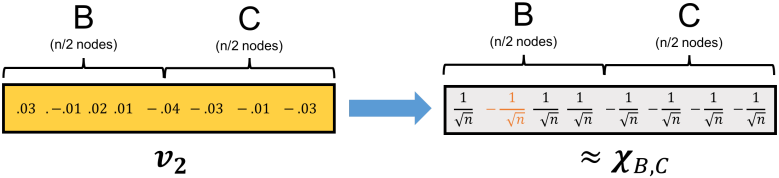


Upshot: If G is a stochastic block model graph with adjacency matrix $\mathbf{A}$, if we compute its second large eigenvector v_2 and assign nodes to communities according to the sign pattern of this vector, we will correctly assign all but $O\left(\frac{p}{(p-q)^2}\right)$ nodes.



- Why does the error increase as q gets close to p?

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.