COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 18

SUMMARY

This Class: Spectral Clustering

- Finding good cuts via Laplacian eigenvectors.
- Start analysis via the stochastic block model.

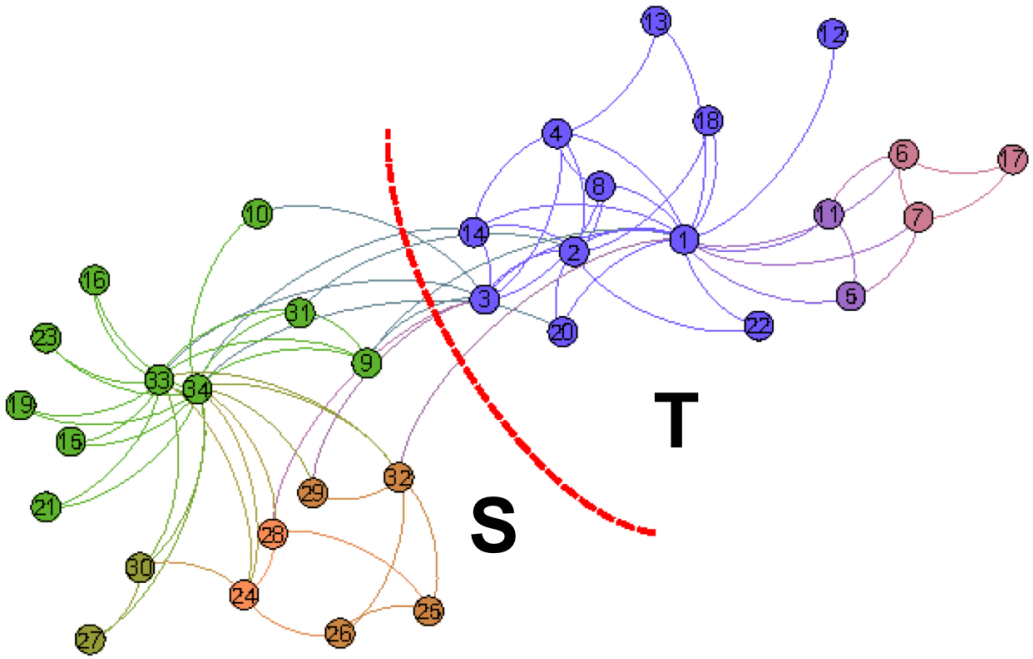
GRAPH CLUSTERING

SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

Community detection in naturally occurring networks.

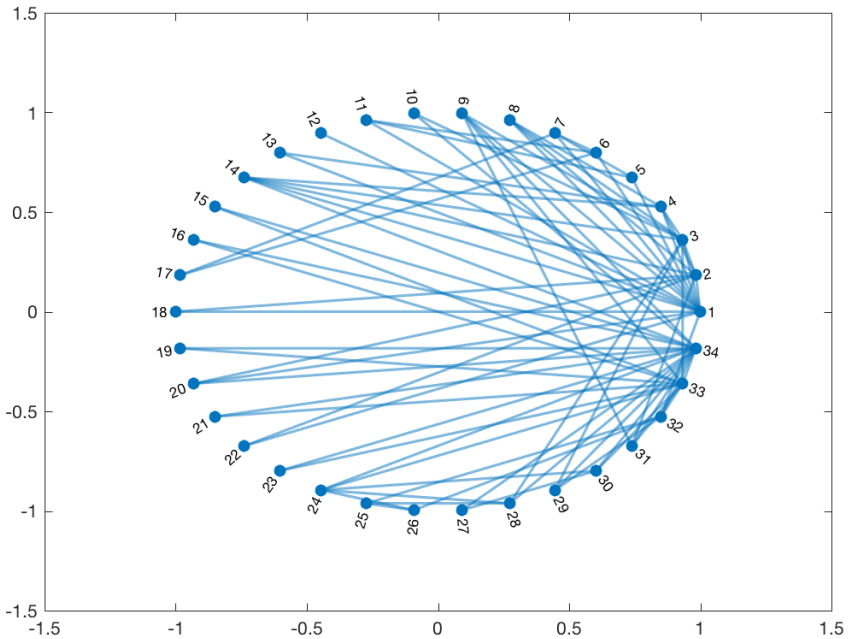


(a) Zachary Karate Club Graph

SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

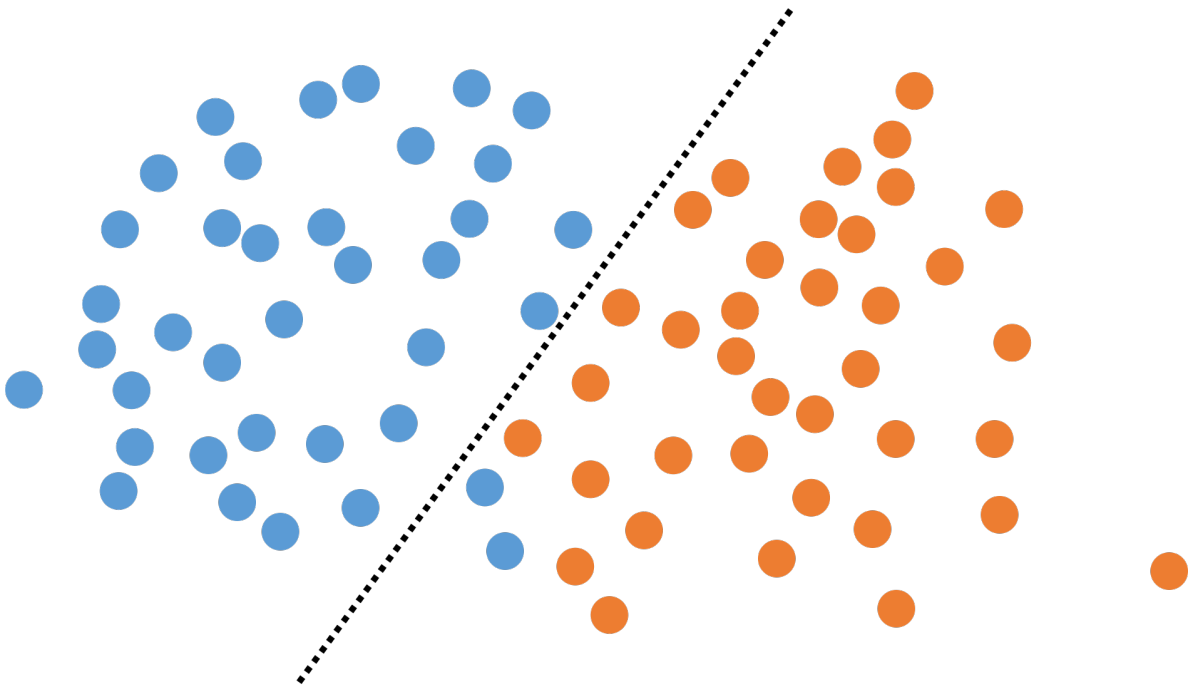
Community detection in naturally occurring networks.



SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

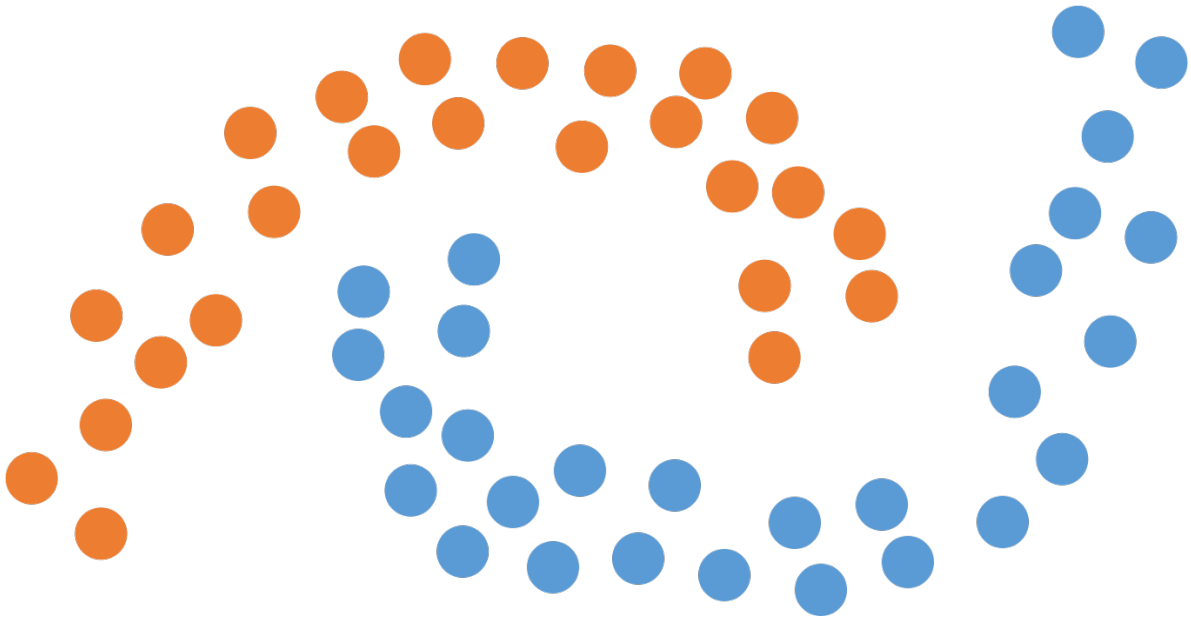
Linearly separable data.



SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

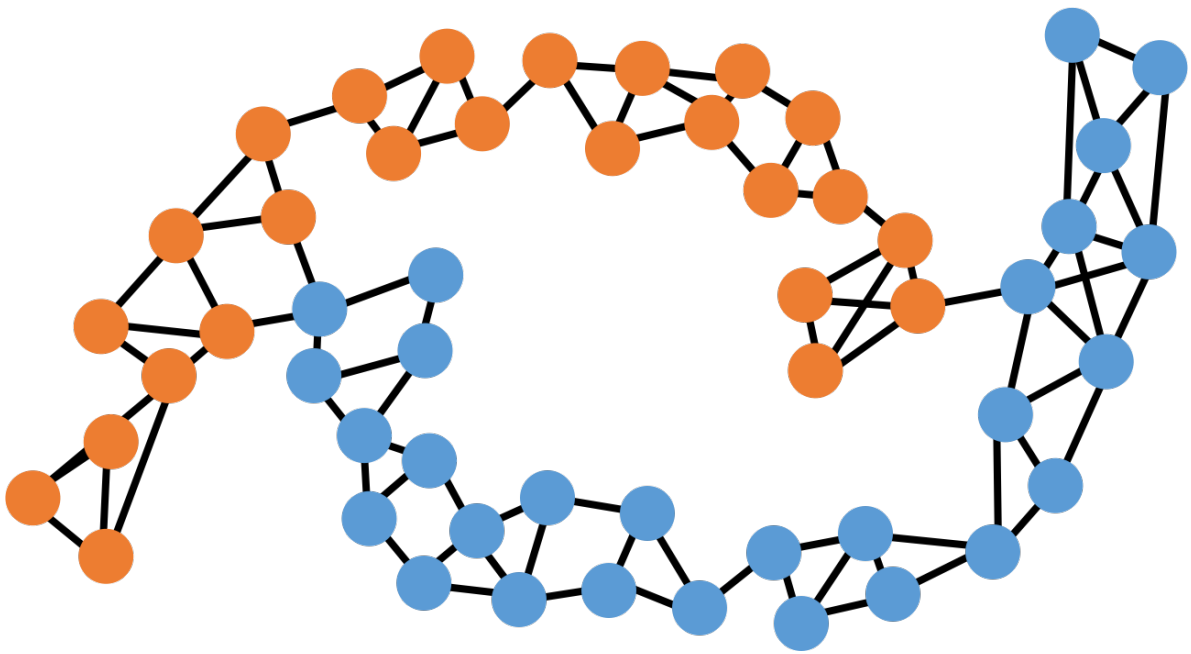
Non-linearly separable data k-nearest neighbor graph.



SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

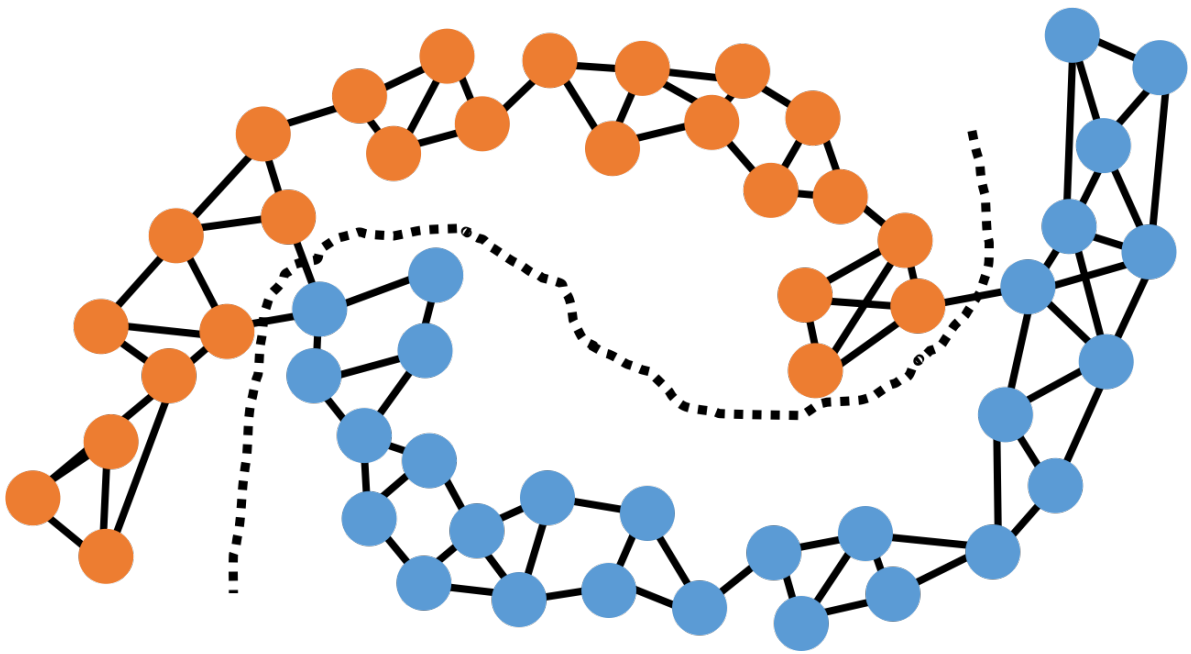
Non-linearly separable data k-nearest neighbor graph.



SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

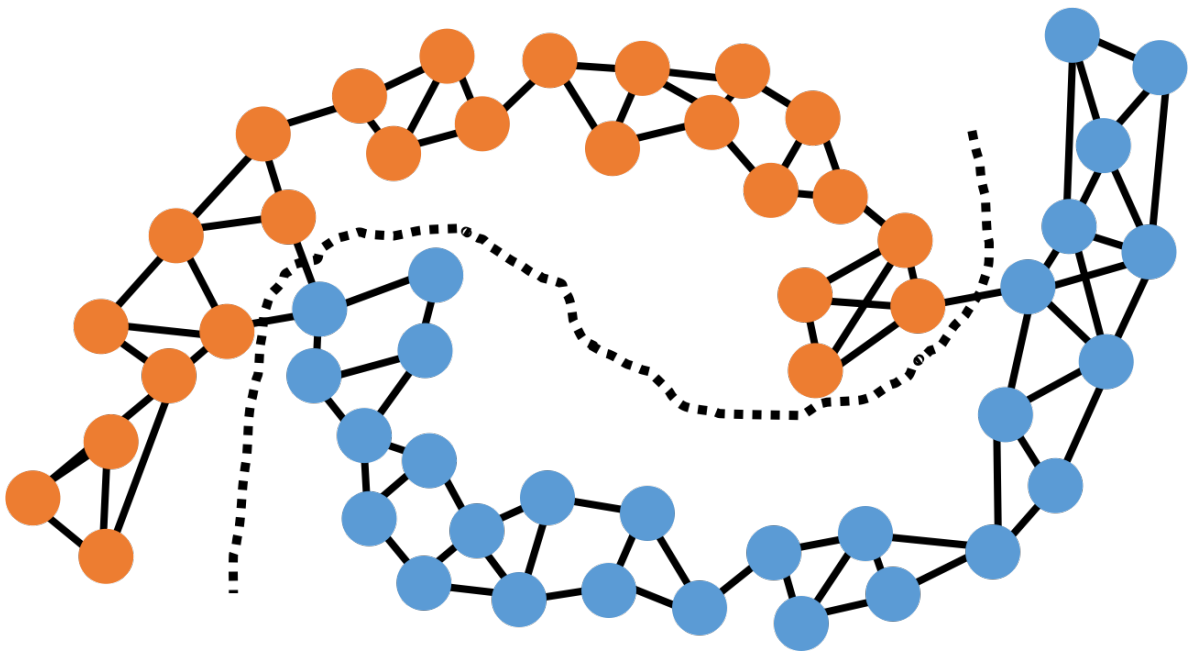
Non-linearly separable data k-nearest neighbor graph.



SPECTRAL CLUSTERING

A very common task is to partition or cluster vertices in a graph based on similarity/connectivity.

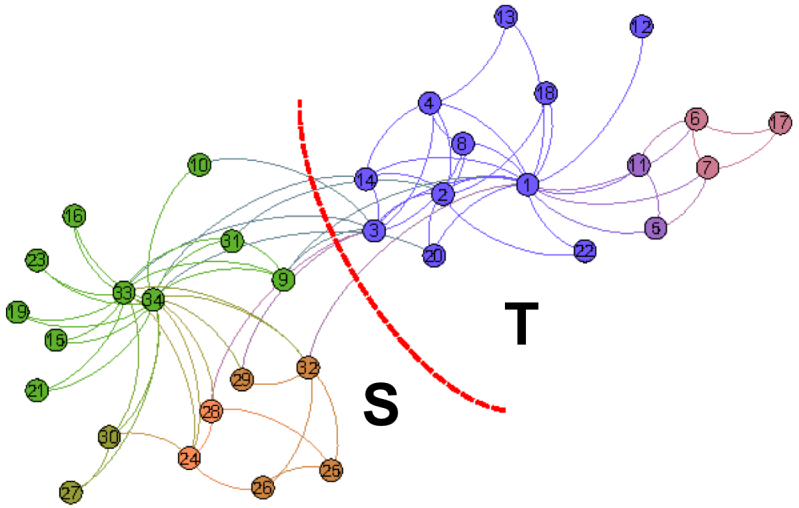
Non-linearly separable data k-nearest neighbor graph.



Can find this cut using eigendecomposition!

CUT MINIMIZATION

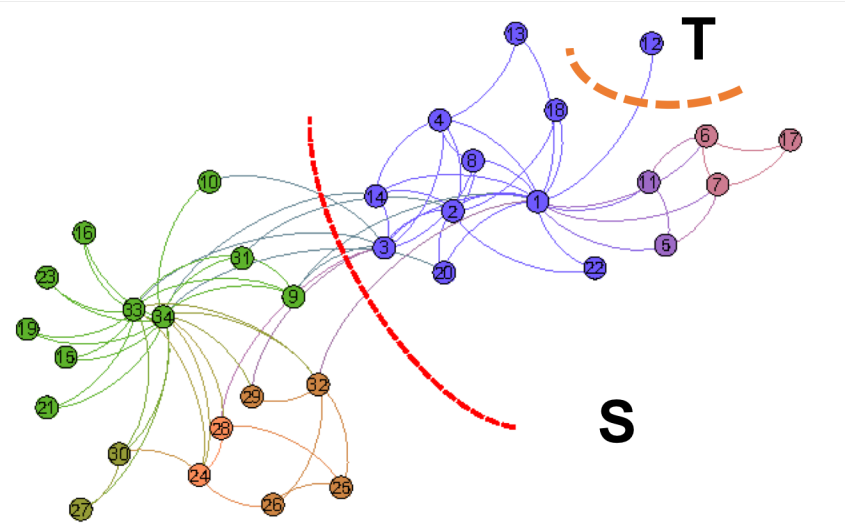
Simple Idea: Partition clusters along minimum cut in graph.



(a) Zachary Karate Club Graph

CUT MINIMIZATION

Simple Idea: Partition clusters along minimum cut in graph.

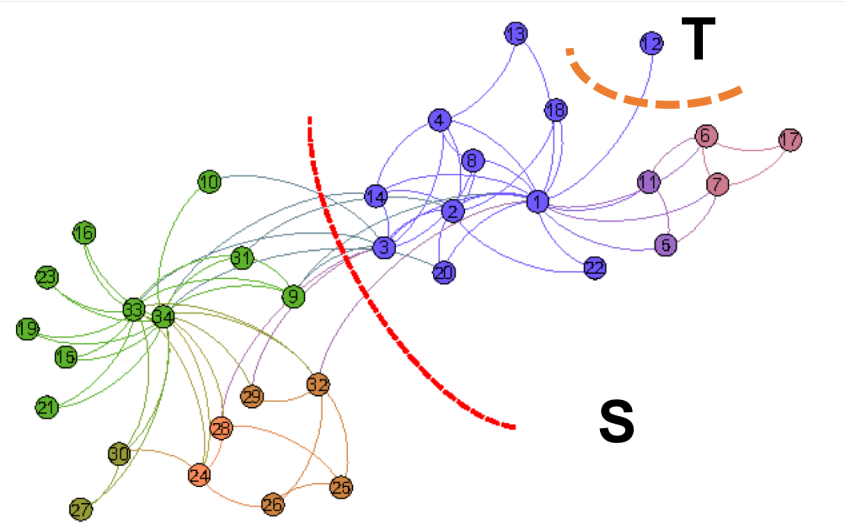


(a) Zachary Karate Club Graph

Small cuts are often not informative.

CUT MINIMIZATION

Simple Idea: Partition clusters along minimum cut in graph.



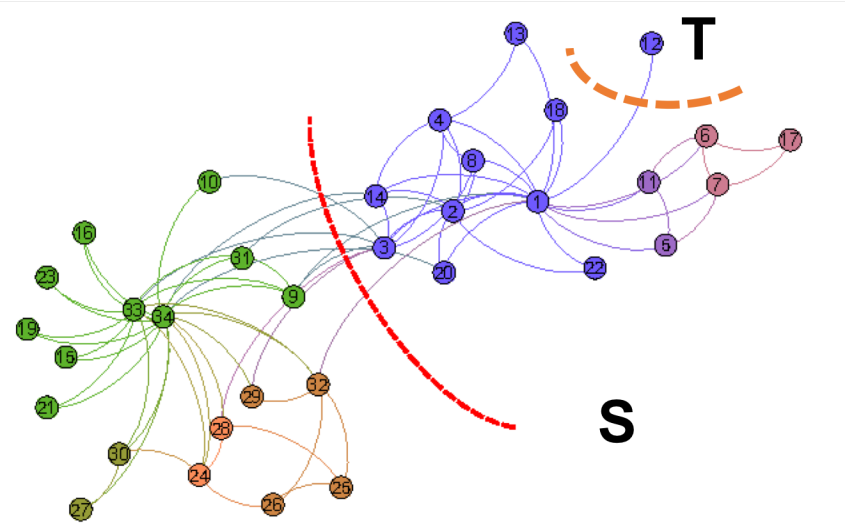
(a) Zachary Karate Club Graph

Small cuts are often not informative.

Solution: Encourage cuts that separate large sections of the graph.

CUT MINIMIZATION

Simple Idea: Partition clusters along minimum cut in graph.



(a) Zachary Karate Club Graph

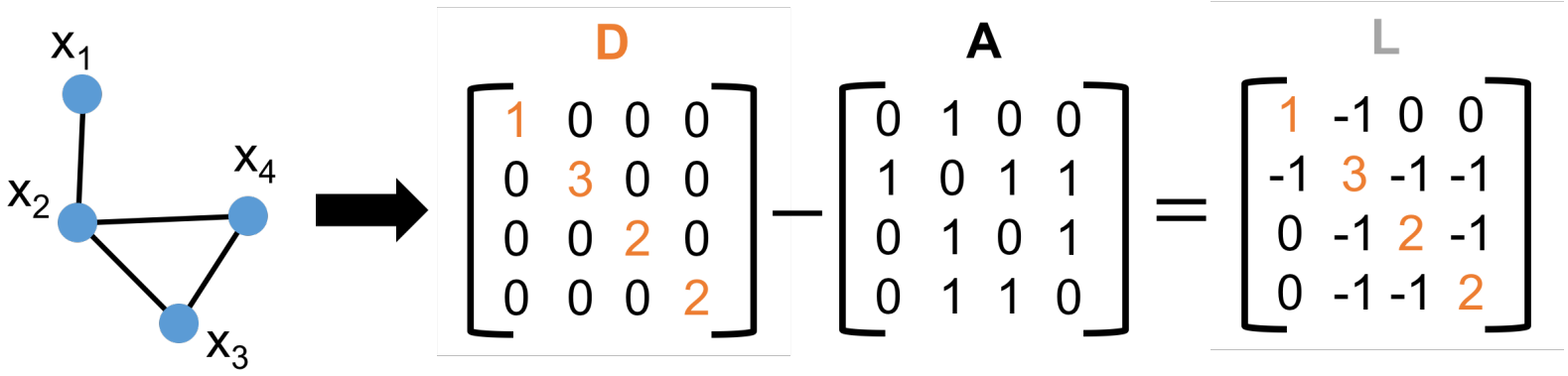
Small cuts are often not informative.

Solution: Encourage cuts that separate large sections of the graph.

- Let $\vec{v} \in \mathbb{R}^n$ be a cut indicator: $\vec{v}(i) = 1$ if $i \in S$. $\vec{v}(i) = -1$ if $i \in T$. Want $\vec{v}$ to have roughly equal numbers of 1s and -1s. I.e., $\vec{v}^T \vec{1} \approx 0$.

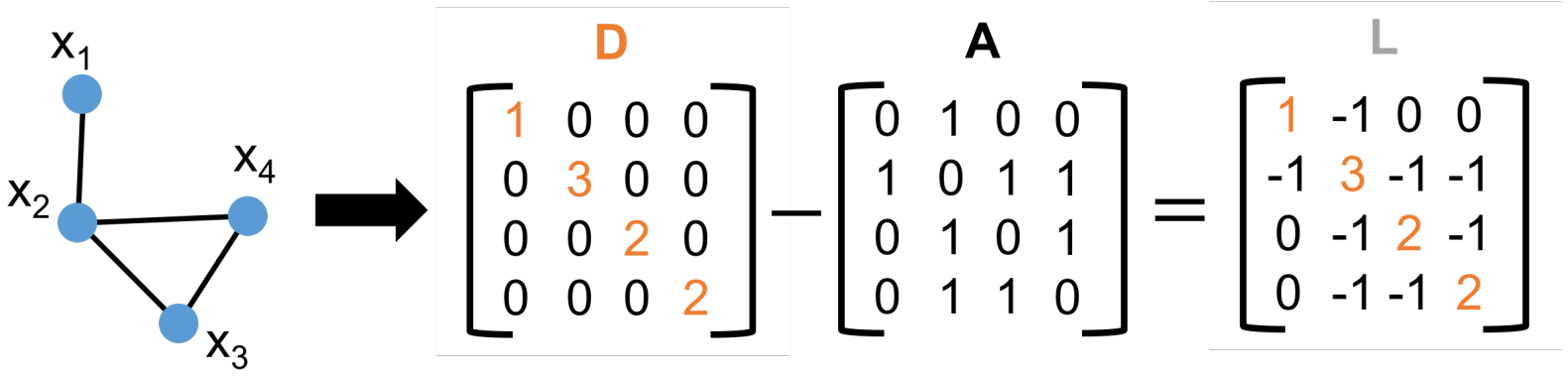
THE LAPLACIAN VIEW

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



THE LAPLACIAN VIEW

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



For any vector $\vec{v}$, its 'smoothness' over the graph is given by:

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

THE LAPLACIAN VIEW

For a cut indicator vector $\vec{v} \in \{-1, 1\}^n$ with $\vec{v}(i) = -1$ for $i \in S$ and $\vec{v}(i) = 1$ for $i \in T$:

1. $\vec{v}^T \mathbf{L} \vec{v} = \sum_{(i,j) \in E} (\vec{v}(i) - \vec{v}(j))^2 = 4 \cdot cut(S, T)$.

THE LAPLACIAN VIEW

For a cut indicator vector $\vec{v} \in \{-1, 1\}^n$ with $\vec{v}(i) = -1$ for $i \in S$ and $\vec{v}(i) = 1$ for $i \in T$:

1. $\vec{v}^T \mathbf{L} \vec{v} = \sum_{(i,j) \in E} (\vec{v}(i) - \vec{v}(j))^2 = 4 \cdot cut(S, T)$.
2. $\vec{v}^T \vec{1} = |T| - |S|$.

THE LAPLACIAN VIEW

For a cut indicator vector $\vec{v} \in \{-1, 1\}^n$ with $\vec{v}(i) = -1$ for $i \in S$ and $\vec{v}(i) = 1$ for $i \in T$:

1. $\vec{v}^T \mathbf{L} \vec{v} = \sum_{(i,j) \in E} (\vec{v}(i) - \vec{v}(j))^2 = 4 \cdot cut(S, T)$.
2. $\vec{v}^T \vec{1} = |T| - |S|$.

Want to minimize both $\vec{v}^T \mathbf{L} \vec{v}$ (cut size) and $\vec{v}^T \vec{1}$ (imbalance).

THE LAPLACIAN VIEW

For a cut indicator vector $\vec{v} \in \{-1, 1\}^n$ with $\vec{v}(i) = -1$ for $i \in S$ and $\vec{v}(i) = 1$ for $i \in T$:

1. $\vec{v}^T \mathbf{L} \vec{v} = \sum_{(i,j) \in E} (\vec{v}(i) - \vec{v}(j))^2 = 4 \cdot cut(S, T)$.
2. $\vec{v}^T \vec{1} = |T| - |S|$.

Want to minimize both $\vec{v}^T \mathbf{L} \vec{v}$ (cut size) and $\vec{v}^T \vec{1}$ (imbalance).

Next Step: See how this dual minimization problem is naturally solved by eigendecomposition.

SMALLEST LAPLACIAN EIGENVECTOR

Assuming the graph is connected, the smallest eigenvector of the Laplacian is:

$$\vec{v}_n = \frac{1}{\sqrt{n}} \cdot \vec{1} = \underset{v \in \mathbb{R}^n \text{ with } \|\vec{v}\|=1}{\arg\min} \vec{v}^T L \vec{v}$$

with eigenvalue $\vec{v}_n^T L \vec{v}_n = 0$.

n: number of nodes in graph, $\mathbf{A} \in \mathbb{R}^{n \times n}$: adjacency matrix, $\mathbf{D} \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $\mathbf{L} \in \mathbb{R}^{n \times n}$: Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$.

SMALLEST LAPLACIAN EIGENVECTOR

Assuming the graph is connected, the smallest eigenvector of the Laplacian is:

$$\vec{v}_n = \frac{1}{\sqrt{n}} \cdot \vec{1} = \underset{v \in \mathbb{R}^n \text{ with } \|\vec{v}\|=1}{\arg\min} \vec{v}^T L \vec{v}$$

with eigenvalue $\vec{v}_n^T L \vec{v}_n = 0$. Why?

n: number of nodes in graph, $\mathbf{A} \in \mathbb{R}^{n \times n}$: adjacency matrix, $\mathbf{D} \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $\mathbf{L} \in \mathbb{R}^{n \times n}$: Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$.

SECOND SMALLEST LAPLACIAN EIGENVECTOR

By Courant-Fischer, the second smallest eigenvector is given by:

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

SECOND SMALLEST LAPLACIAN EIGENVECTOR

By Courant-Fischer, the second smallest eigenvector is given by:

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

If $\vec{v}_{n-1}$ were in $\left\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right\}^n$ it would have:

- $\vec{v}_{n-1}^T L \vec{v}_{n-1} = \frac{4}{n} \cdot cut(S, T)$ as small as possible given that

$$\vec{v}_{n-1}^T \vec{v}_n = \frac{1}{\sqrt{n}} \vec{v}_{n-1}^T \vec{1} = \frac{|T| - |S|}{\sqrt{n}} = 0 \ .$$

SECOND SMALLEST LAPLACIAN EIGENVECTOR

By Courant-Fischer, the second smallest eigenvector is given by:

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

If $\vec{v}_{n-1}$ were in $\left\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right\}^n$ it would have:

- $\vec{v}_{n-1}^T L \vec{v}_{n-1} = \frac{4}{n} \cdot cut(S, T)$ as small as possible given that

$$\vec{v}_{n-1}^T \vec{v}_n = \frac{1}{\sqrt{n}} \vec{v}_{n-1}^T \vec{1} = \frac{|T| - |S|}{\sqrt{n}} = 0 \ .$$

- I.e., $\vec{v}_{n-1}$ would indicate the smallest perfectly balanced cut.

SECOND SMALLEST LAPLACIAN EIGENVECTOR

By Courant-Fischer, the second smallest eigenvector is given by:

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

If $\vec{v}_{n-1}$ were in $\left\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right\}^n$ it would have:

- $\vec{v}_{n-1}^T L \vec{v}_{n-1} = \frac{4}{n} \cdot cut(S, T)$ as small as possible given that

$$\vec{v}_{n-1}^T \vec{v}_n = \frac{1}{\sqrt{n}} \vec{v}_{n-1}^T \vec{1} = \frac{|T| - |S|}{\sqrt{n}} = 0 \ .$$

- I.e., $\vec{v}_{n-1}$ would indicate the smallest perfectly balanced cut.
- The eigenvector $\vec{v}_{n-1} \in \mathbb{R}^n$ is not generally binary, but still satisfies a 'relaxed' version of this property.

CUTTING WITH THE SECOND LAPLACIAN EIGENVECTOR

Find a good partition of the graph by computing

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

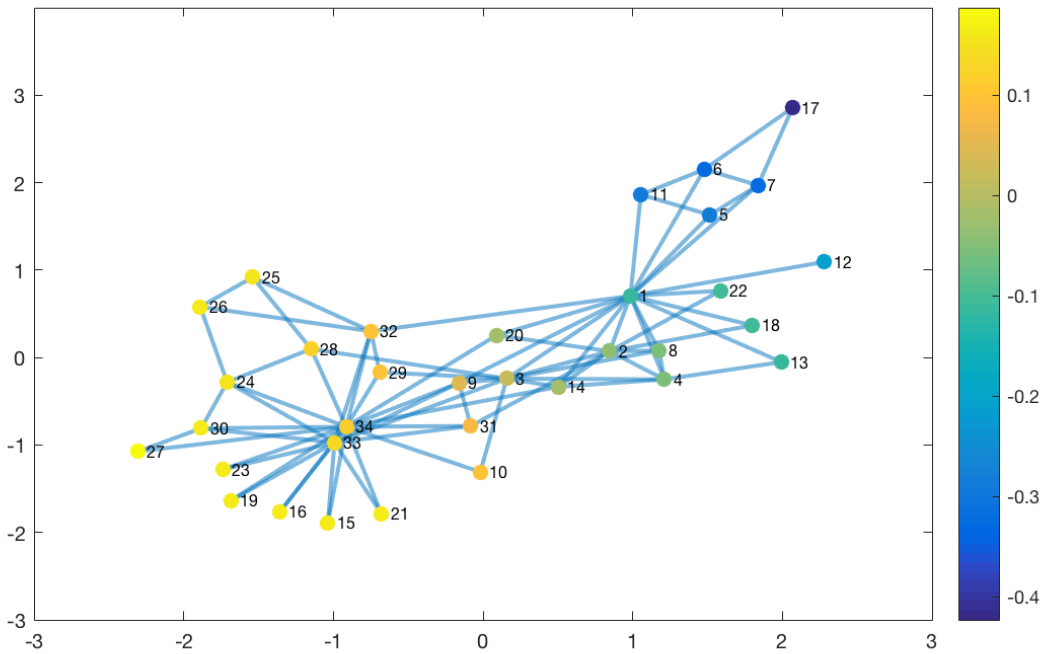
Set S to be all nodes with $\vec{v}_{n-1}(i) < 0$, T to be all with $\vec{v}_{n-1}(i) \geq 0$.

CUTTING WITH THE SECOND LAPLACIAN EIGENVECTOR

Find a good partition of the graph by computing

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

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

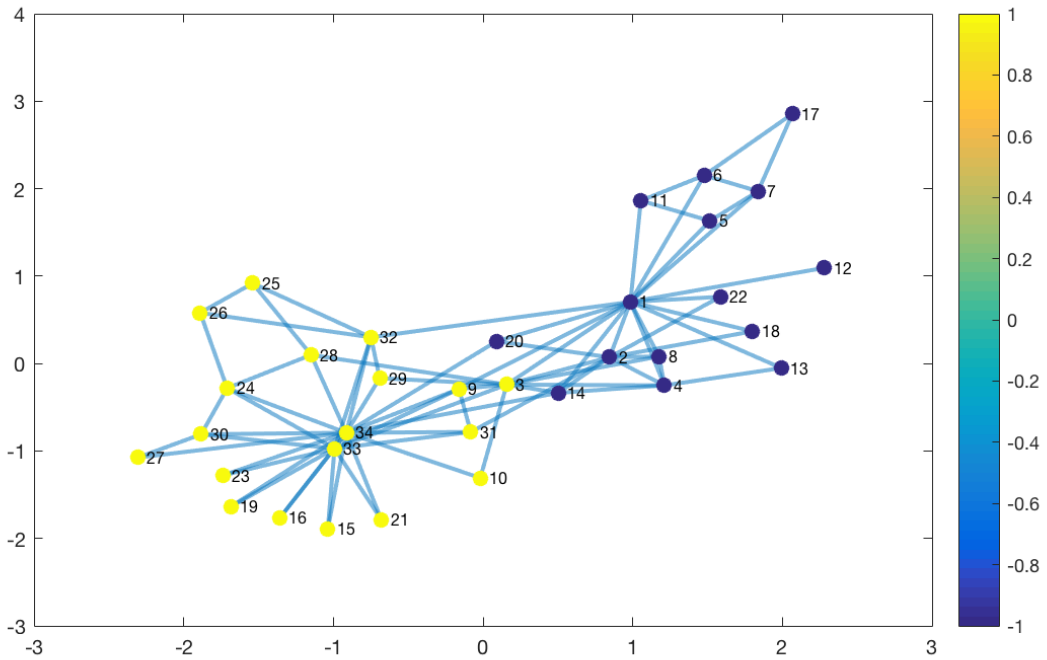


CUTTING WITH THE SECOND LAPLACIAN EIGENVECTOR

Find a good partition of the graph by computing

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

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



- **Summary:** To partition a graph, find the eigenvector of the Laplacian with the second smallest eigenvalue. Partition nodes based on whether corresponding value in eigenvector is positive/negative.

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

SPECTRAL CLUSTERING WITH GUARANTEES

- **Summary:** To partition a graph, find the eigenvector of the Laplacian with the second smallest eigenvalue. Partition nodes based on whether corresponding value in eigenvector is positive/negative.

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

- We argued this "should" partition graph along a small cut that separates the graph into large pieces.

SPECTRAL CLUSTERING WITH GUARANTEES

- **Summary:** To partition a graph, find the eigenvector of the Laplacian with the second smallest eigenvalue. Partition nodes based on whether corresponding value in eigenvector is positive/negative.

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

- We argued this "should" partition graph along a small cut that separates the graph into large pieces.
- Haven't given formal guarantees; it's difficult for general input graphs. But can consider randoms "natural" graphs...

SPECTRAL CLUSTERING WITH GUARANTEES

- **Summary:** To partition a graph, find the eigenvector of the Laplacian with the second smallest eigenvalue. Partition nodes based on whether corresponding value in eigenvector is positive/negative.

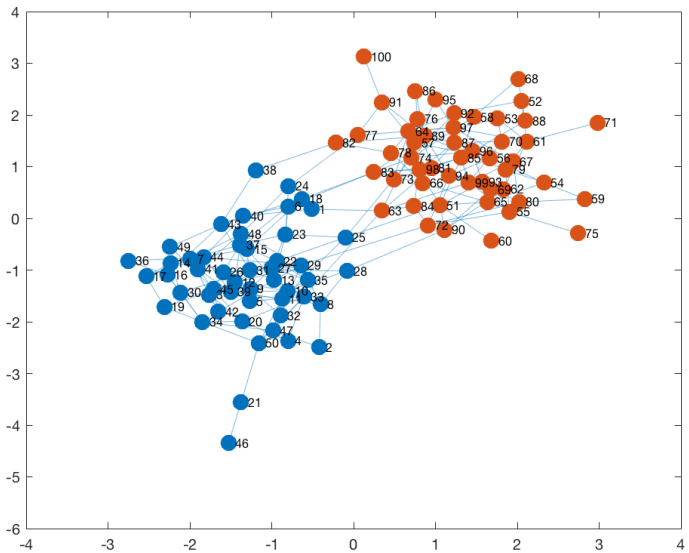
$$\vec{v}_{n-1} = \underset{\vec{v}\in\mathbb{R}^n, \|\vec{v}\|=1, \vec{v}^T\vec{1}=0}{\arg\min} \vec{v}^T L \vec{v}$$

- We argued this "should" partition graph along a small cut that separates the graph into large pieces.
- Haven't given formal guarantees; it's difficult for general input graphs. But can consider randoms "natural" graphs...
- **Common Approach:** Give a natural generative model for random inputs and analyze how the algorithm performs on inputs drawn from this model. Can be used to justify ℓ_2 linear regression, k-means clustering, 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.

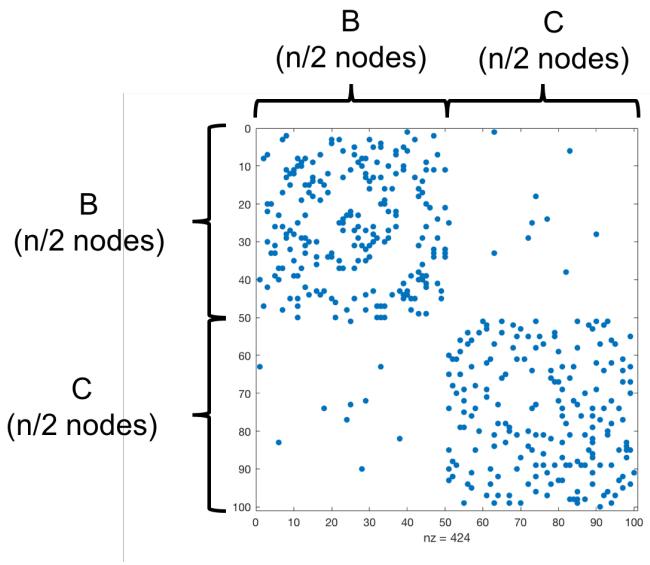
- Any two nodes in the same group are connected with probability p (including self-loops).
- Any two nodes in different groups are connected with prob. $q < p$.
- Connections are independent.

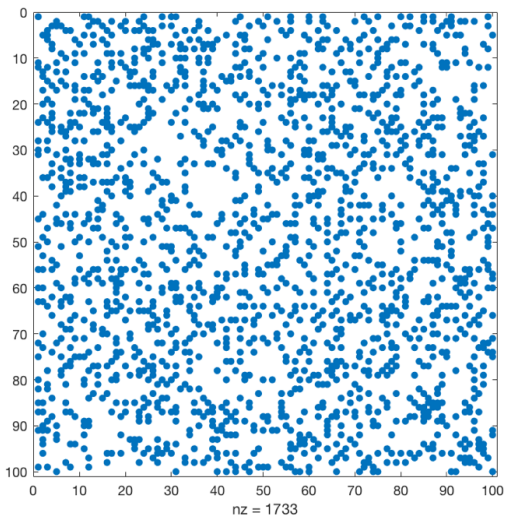


LINEAR ALGEBRAIC VIEW

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.

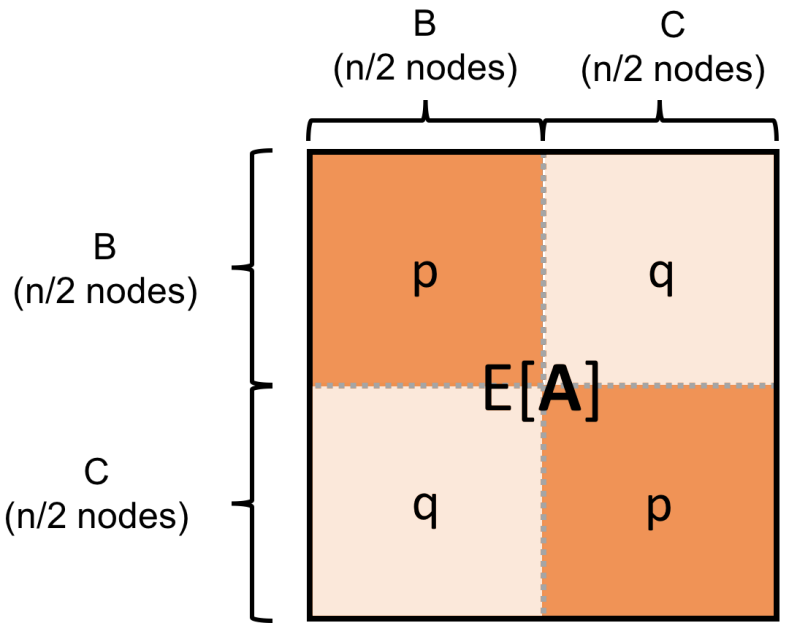




$G_n(p, q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the same group, and probability q between nodes not in the same group.

EXPECTED ADJACENCY SPECTRUM

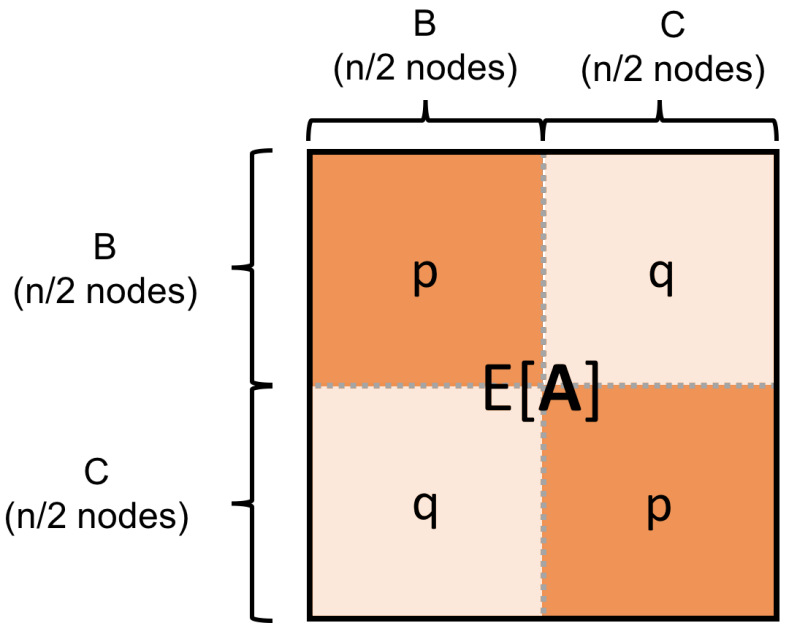
Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i, j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.



$G_n(p, q)$: 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

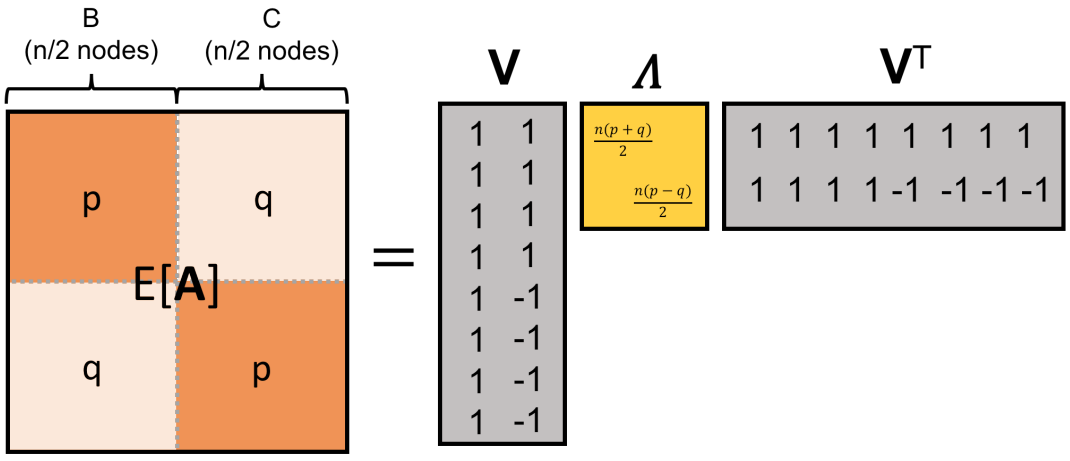
Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i, j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.



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

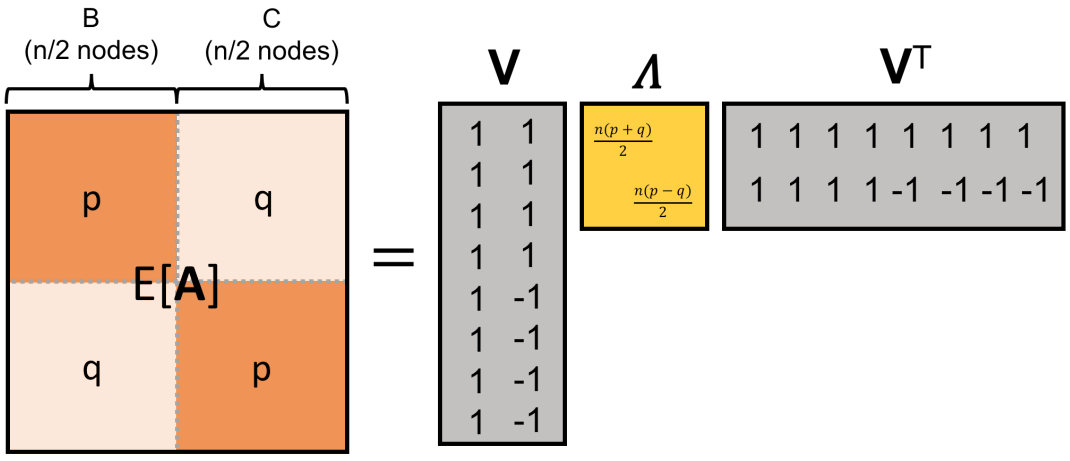
$G_n(p, q)$: stochastic block model distribution. B, C: groups with $n/2$ nodes each. Connections are independent with probability p between nodes in the same group, and probability q between nodes not in the same group.

EXPECTED ADJACENCY SPECTRUM



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

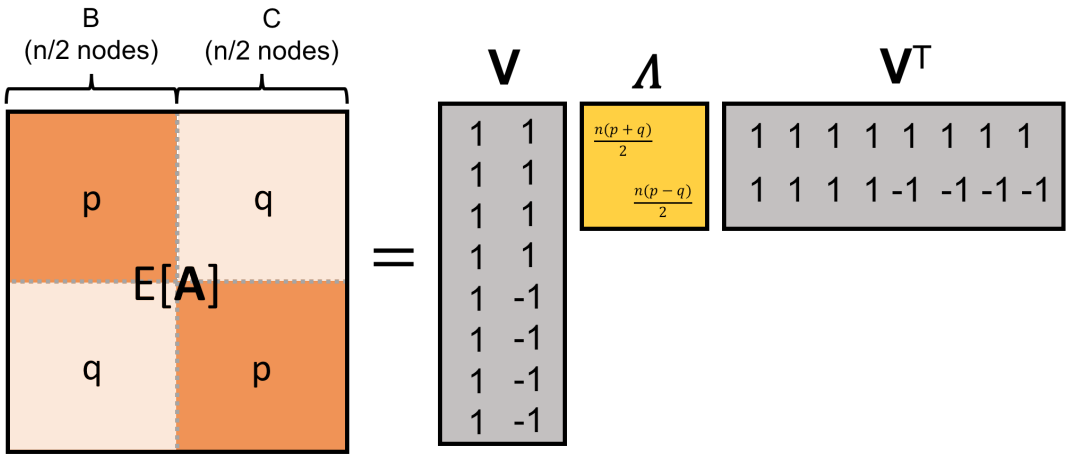
EXPECTED ADJACENCY SPECTRUM



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 $\mathbb{E}[\mathbf{A}]$ in some appropriate sense (matrix concentration inequality).

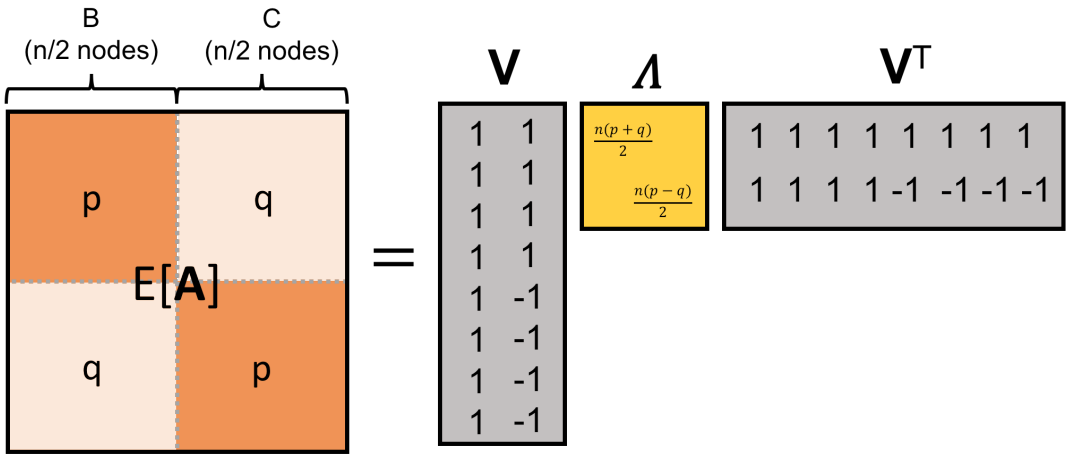
EXPECTED ADJACENCY SPECTRUM



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 $\mathbb{E}[\mathbf{A}]$ in some appropriate sense (matrix concentration inequality).
- Second eigenvector of A is close to $[1, 1, 1, \ldots, -1, -1, -1]$ and gives a good estimate of the communities.

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 $\mathbb{E}[\mathbf{A}]$ in some appropriate sense (matrix concentration inequality).
- Second eigenvector of A is close to $[1, 1, 1, \ldots, -1, -1, -1]$ and gives a good estimate of the communities.

When rows/columns aren't sorted by ID, second eigenvector is e.g., $[1, -1, 1, -1, \ldots, 1, 1, -1]$ and entries give 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 $\mathbb{E}[\mathbf{L}]$?

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

and so if $\mathbb{E}[\mathbf{A}]\vec{x} = \lambda\vec{x}$ then

$$\mathbb{E}[\mathbf{L}]\vec{x} = (n(p+q)/2 - \lambda)\vec{x}$$

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

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

and so if $\mathbb{E}[\mathbf{A}]\vec{x} = \lambda\vec{x}$ then

$$\mathbb{E}[\mathbf{L}]\vec{x} = (n(p+q)/2 - \lambda)\vec{x}$$

Therefore the first and second eigenvalues of $\mathbb{E}[\mathbf{A}]$ are the second and first eigenvectors 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.

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 matrices **A** and **L** were exactly equal to their expectation, partitioning using this eigenvector (i.e., spectral clustering) would exactly recover the two communities B and C.

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 matrices **A** and **L** were exactly equal to their 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 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

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

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

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 $\mathbf{A}, \overline{\mathbf{A}} \in \mathbb{R}^{d \times d}$ 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's eigenvalues with similar magnitudes.

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq 2} |\lambda_2 - \lambda_j|}$$

$\mathbf{A}$ 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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq 2} |\lambda_2 - \lambda_j|}$$

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 $i \geq 3$.

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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j\neq 2}|\lambda_2 - \lambda_j|}$$

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 $i \geq 3$.

$$\min_{j\neq 2}|\lambda_2 - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$

$\mathbf{A}$ 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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq 2} |\lambda_2 - \lambda_j|}$$

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 $i \geq 3$.

$$\min_{j \neq 2} |\lambda_2 - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$

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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j\neq 2}|\lambda_2 - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p-q)n/2} = O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

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 $i \geq 3$.

$$\min_{j\neq 2}|\lambda_2 - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$

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

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

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?

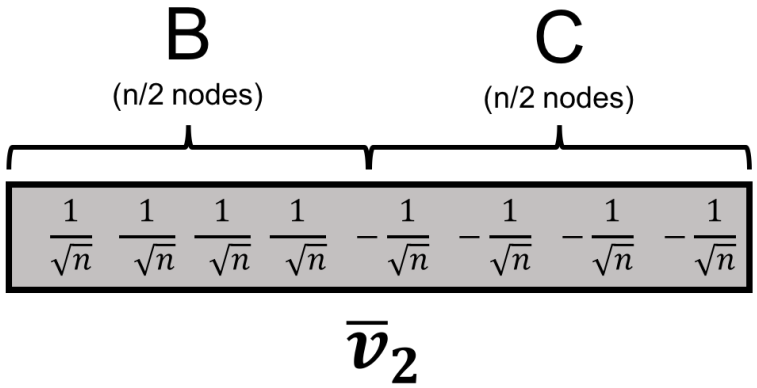
- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).

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.

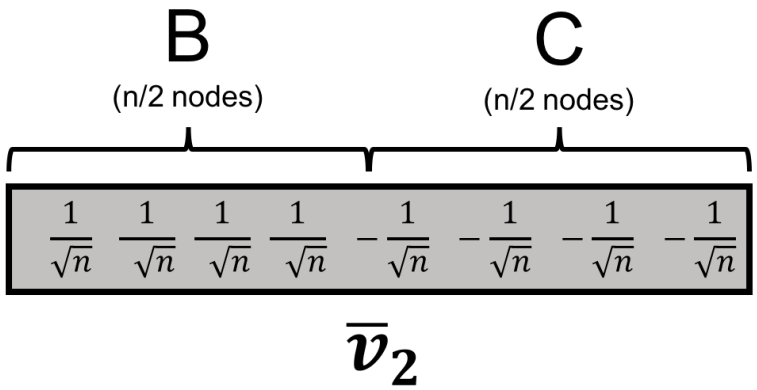


A adjacency matrix of random stochastic block model graph. p: connection probability within clusters. $q < p$: connection probability between clusters. n: number of nodes. $v_2, \bar{v}_2$: second eigenvectors of **A** and $\mathbb{E}[\mathbf{A}]$ respectively.

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.



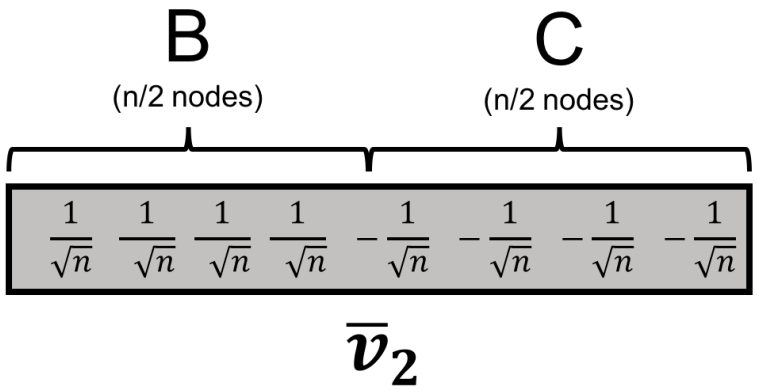
- Every i where $v_2(i)$, $\bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.

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.

APPLICATION TO STOCHASTIC BLOCK MODEL

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.