COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor
Lecture 19
Last Class: Spectral Clustering

• Spectral clustering: finding good cuts via Laplacian eigenvectors.
Summary

Last Class: Spectral Clustering

- Spectral clustering: finding good cuts via Laplacian eigenvectors.

This Class: Stochastic Block Model

- Stochastic block model: A simple clustered graph model where we can prove the effectiveness of spectral clustering.
- Prove that clustering with the Laplacian eigenvectors (spectral clustering) finds communities in the stochastic block model.
For a graph with adjacency matrix $A$ and degree matrix $D$, $L = D - A$ is the graph Laplacian.

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

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

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

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

Let $S$ be nodes with $\vec{v}_{n-1}(i) < 0$, $T$ be nodes with $\vec{v}_{n-1}(i) \geq 0$. 
Find a good partition of the graph by computing

\[ \vec{v}_{n-1} = \arg \min_{\vec{v} \in \mathbb{R}^d \text{with } \|\vec{v}\|=1, \vec{v}^T \vec{1} = 0} \vec{v}^T L \vec{v} \]

Let \( S \) be nodes with \( \vec{v}_{n-1}(i) < 0 \), \( T \) be nodes with \( \vec{v}_{n-1}(i) \geq 0 \).
The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\overline{L} = D^{-1/2}LD^{-1/2}$.

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

The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\overline{L} = D^{-1/2} LD^{-1/2}$.

**Important Consideration:** What to do when we want to split the graph into more than two parts?

$n$: number of nodes in graph, $A \in \mathbb{R}^{n \times n}$: adjacency matrix, $D \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $L \in \mathbb{R}^{n \times n}$: Laplacian matrix $L = A - D$. 
The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\overline{L} = D^{-1/2}LD^{-1/2}$.

**Important Consideration:** What to do when we want to split the graph into more than two parts?

**Spectral Clustering:**

$n$: number of nodes in graph, $A \in \mathbb{R}^{n \times n}$: adjacency matrix, $D \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $L \in \mathbb{R}^{n \times n}$: Laplacian matrix $L = A - D$. 
The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\bar{L} = D^{-1/2}LD^{-1/2}$.

**Important Consideration:** What to do when we want to split the graph into more than two parts?

**Spectral Clustering:**

- Compute smallest $k$ nonzero eigenvectors $\tilde{v}_{n-1}, \ldots, \tilde{v}_{n-k}$ of $\bar{L}$.

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

4
SPECTRAL PARTITIONING IN PRACTICE

The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian \( \bar{L} = D^{-1/2}LD^{-1/2} \).

**Important Consideration:** What to do when we want to split the graph into more than two parts?

**Spectral Clustering:**

- Compute smallest \( k \) nonzero eigenvectors \( \vec{v}_{n-1}, \ldots, \vec{v}_{n-k} \) of \( \bar{L} \).
- Represent each node by its corresponding row in \( V \in \mathbb{R}^{n \times k} \) whose columns are \( \vec{v}_{n-1}, \ldots \vec{v}_{n-k} \).

**Notes:**

- \( n \): number of nodes in graph, \( A \in \mathbb{R}^{n \times n} \): adjacency matrix, \( D \in \mathbb{R}^{n \times n} \): diagonal degree matrix, \( L \in \mathbb{R}^{n \times n} \): Laplacian matrix \( L = A - D \).
The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\overline{L} = D^{-1/2}LD^{-1/2}$.

**Important Consideration:** What to do when we want to split the graph into more than two parts?

**Spectral Clustering:**

- Compute smallest $k$ nonzero eigenvectors $\vec{v}_{n-1}, \ldots, \vec{v}_{n-k}$ of $\overline{L}$.
- Represent each node by its corresponding row in $V \in \mathbb{R}^{n \times k}$ whose columns are $\vec{v}_{n-1}, \ldots, \vec{v}_{n-k}$.
- Cluster these rows using $k$-means clustering (or really any clustering method).

$n$: number of nodes in graph, $A \in \mathbb{R}^{n \times n}$: adjacency matrix, $D \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $L \in \mathbb{R}^{n \times n}$: Laplacian matrix $L = A - D$. 
**So Far:** Argued that spectral clustering partitions a graph, along a small cut that separates the graph into large pieces.
So Far: Argued that spectral clustering partitions a graph, along a small cut that separates the graph into large pieces.

• Haven’t given formal guarantees on ‘quality’ of the partitioning.
So Far: Argued that spectral clustering partitions a graph, along a small cut that separates the graph into large pieces.

- Haven’t given formal guarantees on ‘quality’ of the partitioning.
- This is difficult to do for general input graphs.
So Far: Argued that spectral clustering partitions a graph, along a small cut that separates the graph into large pieces.

- Haven’t given formal guarantees on ‘quality’ of the partitioning.
- This is difficult to do for general input graphs.

Common Approach: Give a natural generative model for random inputs and analyze how the algorithm performs on inputs drawn from this model.
**So Far:** Argued that spectral clustering partitions a graph, along a small cut that separates the graph into large pieces.

- Haven’t given formal guarantees on ‘quality’ of the partitioning.
- This is difficult to do for general input 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 $\ell_2$ linear 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.

- 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 $A \in \mathbb{R}^{n \times n}$ 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.
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.
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.
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$ and $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[A]$?
If we compute $\vec{v}_2$ then we recover the communities $B$ and $C$!
If we compute $\vec{v}_2$ then we recover the communities $B$ and $C$!

- Can show that for $G \sim G_n(p, q)$, $A$ is close to $E[A]$ 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!)
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 $PAP^T$ where $P$ is a random permutation matrix and $A$ is the ordered adjacency matrix.
- **Exercise:** The first two eigenvectors of $PAP^T$ are $P\vec{v}_1$ and $P\vec{v}_2$.
- $P\vec{v}_2 = [1, -1, 1, -1, \ldots, 1, 1, -1]$ gives community ids.
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$, $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix and $L$ be its Laplacian, what is $E[L]$?
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$, $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix and $L$ be its Laplacian, what are the eigenvectors and eigenvalues of $\mathbb{E}[L]$?
**Upshot:** The second smallest eigenvector of $\mathbb{E}[L]$ is $\chi_{B,C}$ – the indicator vector for the cut between the communities.
**Upshot:** The second smallest eigenvector of $\mathbb{E}[L]$ is $\chi_{B,C}$ – the indicator vector for the cut between the communities.

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

---

**EXPECTED LAPLACIAN SPECTRUM**

How do we show that a matrix (e.g., $A$) 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.
**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 $G$ (equivalently $\mathbf{A}$ 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.
Matrix Concentration Inequality: If $p \geq O\left(\frac{\log^4 n}{n}\right)$, then with high probability

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

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

For any $X \in \mathbb{R}^{n \times d}$, $\|X\|_2 = \max_{z \in \mathbb{R}^d : \|z\|_2 = 1} \|Xz\|_2$. 
Matrix Concentration Inequality: If \( p \geq O \left( \frac{\log^4 n}{n} \right) \), then with high probability

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

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

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

For the stochastic block model application, we want to show that the second eigenvectors of \( A \) and \( E[A] \) are close. How does this relate to their difference in spectral norm?
Davis-Kahan Eigenvector Perturbation Theorem: Suppose $A, \overline{A} \in \mathbb{R}^{d \times d}$ are symmetric with $\|A - \overline{A}\|_2 \leq \epsilon$ and eigenvectors $v_1, v_2, \ldots, v_d$ and $\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_d$. Letting $\theta(v_i, \overline{v}_i)$ denote the angle between $v_i$ and $\overline{v}_i$, for all $i$:

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

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

The errors get large if there are eigenvalues with similar magnitudes.
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 $p \geq O\left(\frac{\log^4 n}{n}\right)$,

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

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

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

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 $i \geq 3$. 

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

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

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

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

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 $i \geq 3$.

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

Typically, $\frac{(p-q)n}{2}$ will be the minimum of these two gaps.
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 \( p \geq O \left( \frac{\log^4 n}{n} \right) \),
\[
\sin \theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p - q)n/2} = O \left( \frac{\sqrt{p}}{(p - q)\sqrt{n}} \right)
\]

Recall: \( \mathbb{E}[A] \) has eigenvalues \( \lambda_1 = \frac{(p+q)n}{2} \), \( \lambda_2 = \frac{(p-q)n}{2} \), \( \lambda_i = 0 \) for \( i \geq 3 \).
\[
\min_{j \neq i} |\lambda_i - \lambda_j| = \min \left( qn, \frac{(p-q)n}{2} \right).
\]

Typically, \( \frac{(p-q)n}{2} \) will be the minimum of these two gaps.
So Far: $\sin \theta(v_2, \bar{v}_2) \leq O \left( \frac{\sqrt{p}}{(p-q)\sqrt{n}} \right)$.
**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 \(\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2n}\right)\) (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 \(\mathbb{E}[A]\) 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 \( \|v_2 - \bar{v}_2\|^2_2 \leq O\left(\frac{p}{(p-q)^2n}\right) \) (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.
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 \( \|v_2 - \bar{v}_2\|_2^2 \).

\[ \begin{array}{cccccccc}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} \\
\end{array} \]

B (n/2 nodes) \hspace{1cm} C (n/2 nodes)

\( \bar{v}_2 \)

**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 \( \mathbb{E}[A] \) 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 \( \|v_2 - \bar{v}_2\|^2 \leq O \left( \frac{p}{(p-q)^2 n} \right) \) (exercise).
- \( \bar{v}_2 \) is \( \frac{1}{\sqrt{n}} \chi_{B,C} \): the community indicator vector.

\[
\begin{array}{cccccccc}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\
\end{array}
\]

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