COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor
Lecture 20
Spectral Graph Partitioning

• Focus on separating graphs with small but relatively balanced cuts.
• Connection to second smallest eigenvector of graph Laplacian.
• Today: Provable guarantees for stochastic block model.
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.

\[
\tilde{v}_{n-1} = \arg \min_{\tilde{v} \in \mathbb{R}^n, \|\tilde{v}\|=1, \tilde{v}^T \mathbf{1} = 0} \tilde{v}^T L \tilde{v}
\]
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.

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

We argued this “should” partition graph along a small cut that separates the graph into large pieces.
SPECTRAL CLUSTERING WITH GUARANTEES

• 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.

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

• 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…
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.
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 $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[A])_{i,j} = p$ for $i,j$ in same group, $(\mathbb{E}[A])_{i,j} = q$ 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 $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[A])_{i,j} = p$ for $i, j$ in same group, $(\mathbb{E}[A])_{i,j} = q$ otherwise.

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

$G_n(p, q)$: 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.
If we compute $\vec{v}_2$ then we recover the communities $B$ and $C$!
If we compute $\tilde{v}_2$ then we recover the communities $B$ and $C$!

- Can show that for $G \sim G_n(p, q)$, $A$ is “close” to $E[A]$ in some appropriate sense (matrix concentration inequality).
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 $\mathbb{E}[A]$ 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.
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 $\mathbf{E}[A]$ 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.
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 $E[L]$?
Expected Laplacian Spectrum

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 $E[L]$?

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

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

$$E[L]\vec{x} = \left( \frac{n(p + q)}{2} - \lambda \right) \vec{x}$$

Therefore the first and second eigenvalues of $E[L]$ are the second and first eigenvectors of $E[L]$. 

7
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 $E[L]$?

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

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

$$E[L] \vec{x} = \left(\frac{n(p + q)}{2} - \lambda\right) \vec{x}$$

Therefore the first and second eigenvalues of $E[A]$ are the second and first eigenvectors of $E[L]$. 
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 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 Markov’s, 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 $\mathbb{E}[L]$ 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 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

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

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

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

For the stochastic block model application, we want to show that the second eigenvectors of $A$ and $\mathbb{E}[A]$ 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 $\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_d$. 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 $\overline{\mathbf{A}}$.

The errors get large if there’s 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 \( p \geq O \left( \frac{\log^4 n}{n} \right) \),

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

\( A \): 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 \( p \geq O \left( \frac{\log^4 n}{n} \right) \),
\[
\sin \theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq 2} |\lambda_2 - \lambda_j|}
\]

Recall: \( \mathbb{E}[A] \) has eigenvalues \( \lambda_1 = \frac{(p+q)n}{2} \), \( \lambda_2 = \frac{(p-q)n}{2} \), \( \lambda_i = 0 \) for \( i \geq 3 \).
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 $p \geq O\left(\frac{\log^4 n}{n}\right)$,

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

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

$$\min_{j \neq 2} |\lambda_2 - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$
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 $p \geq O\left(\frac{\log^4 n}{n}\right)$,

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

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

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

Typically, $\frac{(p-q)n}{2}$ will be the minimum of these two gaps.
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 $p \geq O\left(\frac{\log^4 n}{n}\right)$,
\[
\sin \theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq 2} |\lambda_2 - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p - q)n/2} = O\left(\frac{\sqrt{p}}{(p - q)\sqrt{n}}\right)
\]

Recall: $\mathbb{E}[A]$ 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 $\mathbb{E}[A]$ 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 \( \|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2n}\right) \) (exercise).

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

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

A adjacency matrix of random stochastic block model graph. \( p \): connection probability within clusters. \( q \leq p \): 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 \( \|v_2 - \bar{v}_2\|_2^2 \).

\[ B \quad (n/2 \text{ nodes}) \quad C \quad (n/2 \text{ nodes}) \]

\[ \bar{v}_2 \]

\( A \) 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.
**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.
- 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.
Suppose $A$ is a $2 \times 2$ symmetric matrix with orthonormal eigenvectors $\vec{v}_1, \vec{v}_2$ and $\lambda_1 = 1, \lambda_2 = 1/2$.

Let $\vec{x} = \frac{1}{2} \vec{v}_1 + \frac{1}{2} \vec{v}_2$.

Then $A p \vec{x} = \lambda_1 \frac{1}{2} \vec{v}_1 + \lambda_2 \frac{1}{2} \vec{v}_2 = \frac{1}{2} \vec{v}_1 + \frac{1}{2} \vec{v}_2$

$p \to \frac{1}{2}

And $\|A p \vec{x}\|_2^2 = \sqrt{\frac{1}{2}^2 + \frac{1}{2}^2} = \frac{1}{2}$

Furthermore $(A p \vec{x}) / \|A p \vec{x}\|_2 \to \vec{v}_1$.
Suppose $A$ is a $2 \times 2$ symmetric matrix with orthonormal eigenvectors $\vec{v}_1, \vec{v}_2$ and $\lambda_1 = 1, \lambda_2 = 1/2$
Suppose \( A \) is a 2 \( \times \) 2 symmetric matrix with orthonormal eigenvectors \( \mathbf{v}_1, \mathbf{v}_2 \) and \( \lambda_1 = 1, \lambda_2 = 1/2 \)

Let \( \mathbf{x} = \mathbf{v}_1/2 + \mathbf{v}_2/2 \)
• Suppose $A$ is a $2 \times 2$ symmetric matrix with orthonormal eigenvectors $\vec{v}_1, \vec{v}_2$ and $\lambda_1 = 1, \lambda_2 = 1/2$

• Let $\vec{x} = \vec{v}_1/2 + \vec{v}_2/2$

• Then $A^p\vec{x} = \lambda_1^p \vec{v}_1/2 + \lambda_2^p \vec{v}_2/2 = \vec{v}_1/2 + \vec{v}_2/2^{p+1} \rightarrow \vec{v}_1/2$
Suppose $A$ is a $2 \times 2$ symmetric matrix with orthonormal eigenvectors $\vec{v}_1, \vec{v}_2$ and $\lambda_1 = 1, \lambda_2 = 1/2$.

Let $\vec{x} = \vec{v}_1/2 + \vec{v}_2/2$.

Then $A^p \vec{x} = \lambda_1^p \vec{v}_1/2 + \lambda_2^p \vec{v}_2/2 = \vec{v}_1/2 + \vec{v}_2/2^{p+1} \rightarrow \vec{v}_1/2$.

And $\|A^p \vec{x}\|_2 = \sqrt{1/2^2 + 1/2^{2p+2}} \rightarrow 1/2$. 
• Suppose $\mathbf{A}$ is a $2 \times 2$ symmetric matrix with orthonormal eigenvectors $\vec{v}_1, \vec{v}_2$ and $\lambda_1 = 1, \lambda_2 = 1/2$

• Let $\vec{x} = \vec{v}_1/2 + \vec{v}_2/2$

• Then $\mathbf{A}^p \vec{x} = \lambda_1^p \vec{v}_1/2 + \lambda_2^p \vec{v}_2/2 = \vec{v}_1/2 + \vec{v}_2/2^{p+1} \rightarrow \vec{v}_1/2$

• And $\|\mathbf{A}^p \vec{x}\|_2 = \sqrt{1/2^2 + 1/2^{2p+2}} \rightarrow 1/2$

• Furthermore

$$\left(\mathbf{A}^p \vec{x}\right)/\|\mathbf{A}^p \vec{x}\|_2 \rightarrow \vec{v}_1$$