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.

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

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

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

$G_n(p, q)$: stochastic block model distribution. $B, C$: groups with $n/2$ nodes each. Connections are independent with probability $p$ between nodes in the same group, and probability $q$ between nodes not in the same group.
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 $E[A]$ 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 $E[A]$ 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 $\mathbb{E}[L]$?
Upshot: The second smallest eigenvector of $\mathbb{E}[\mathbf{L}]$ 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 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 $\mathbb{E}[L]$ 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 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ and $\overline{\mathbf{v}}_1, \overline{\mathbf{v}}_2, \ldots, \overline{\mathbf{v}}_d$. Letting $\theta(\mathbf{v}_i, \overline{\mathbf{v}}_i)$ denote the angle between $\mathbf{v}_i$ and $\overline{\mathbf{v}}_i$, for all $i$:

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

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

$$\sin \theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|}.$$
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 $i \geq 3$.

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

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

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

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

\( A \): 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).
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.

$\bar{v}_2$ is 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 $E[A]$ 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 \( \|v_2 - \bar{v}_2\|_2^2 \).

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

- \( A \) 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 $\|v_2 - \bar{v}_2\|_2^2$.
- 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 $E[A]$ 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.