# COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2020.

Lecture 19

- Week 10 Quiz is due Monday at 8pm.

## Last Class: Spectral Graph Theory

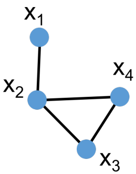
- View of a graph in terms of adjacency matrix and Laplacian.
- Spectral embedding for non-linear dimensionality reduction.
- Start on graph clustering for community detection and non-linear clustering.
- Idea of finding small cuts that separate large sets of nodes.

## This Class: Spectral Clustering and the Stochastic Block Model

- Spectral clustering: finding good cuts via Laplacian eigenvectors.
- Stochastic block model: A simple clustered graph model where we can prove the effectiveness of spectral clustering.

# 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.




$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}}_{\mathbf{D}} - \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}}_{\mathbf{A}} = \underbrace{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}}_{\mathbf{L}}$$

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

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

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. $\vec{v}^T \vec{1} = |V| - |S|$.

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 (sort of) by eigendecomposition.

The smallest eigenvector of the Laplacian is:

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

with eigenvalue $\vec{v}_n^T \mathbf{L} \vec{v}_n = 0$. 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 $\mathbf{L} = \mathbf{A} - \mathbf{D}$.

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

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

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

- $\vec{v}_{n-1}^T L \vec{v}_{n-1} = \frac{4}{\sqrt{n}} \cdot cut(S, T)$ as small as possible given that $\vec{v}_{n-1}^T \vec{v}_n = \frac{1}{\sqrt{n}} \vec{v}_{n-1}^T \vec{1} = \frac{|T|-|S|}{n} = 0$.
- 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.

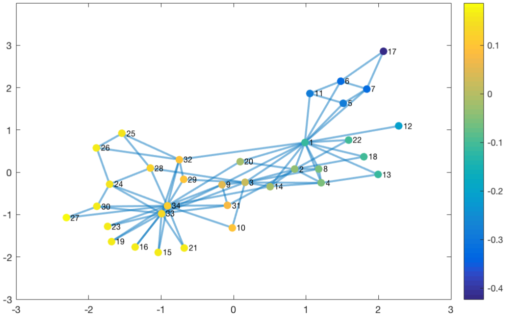
$n$: number of nodes in graph, $\mathbf{A} \in \mathbb{R}^{n \times n}$: adjacency matrix, $\mathbf{D} \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $\mathbf{L} \in \mathbb{R}^{n \times n}$: Laplacian matrix $\mathbf{L} = \mathbf{A} - \mathbf{D}$. $S, T$: vertex sets on different sides of cut.

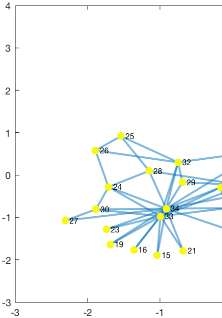
# CUTTING WITH THE SECOND LAPLACIAN EIGENVECTOR

Find a good partition of the graph by computing

$$\vec{v}_2 = \underset{v \in \mathbb{R}^d \text{ with } \|\vec{v}\|=1,\ \vec{v}_2^T \vec{1}=0}{\arg\min} \vec{v}^T \mathbf{L} \vec{v}$$

Set $S$ to be all nodes with $\vec{v}_2(i) < 0$, $T$ to be all with $\vec{v}_2(i) \geq 0$.

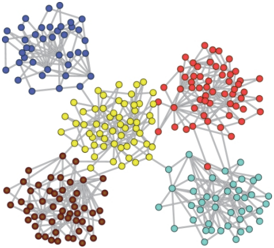


# SPECTRAL PARTITIONING IN PRACTICE

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

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



**Spectral Clustering:**

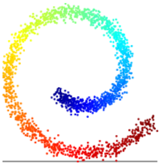
- Compute smallest $k$ nonzero eigenvectors $\vec{v}_{n-1}, \ldots, \vec{v}_{n-k}$ of $\bar{\mathbf{L}}$.

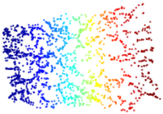
# LAPLACIAN EMBEDDING

The smallest eigenvectors of $\mathbf{L} = \mathbf{D} - \mathbf{A}$ give the orthogonal 'functions' that are smoothest over the graph. I.e., minimize

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

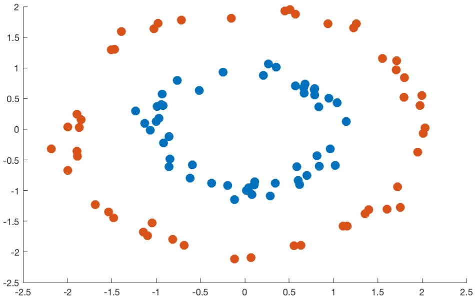
Embedding points with coordinates given by $[\vec{v}_{n-1}(j), \vec{v}_{n-2}(j), \ldots, \vec{v}_{n-k}(j)]$ ensures that coordinates connected by edges have minimum total squared Euclidean distance.



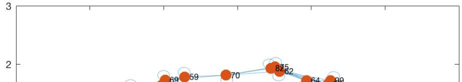


- Spectral Clustering
- Laplacian Eigenmaps
- Locally linear embedding
- Isomap
- Node2Vec, DeepWalk, etc. (variants on Laplacian)

**Original Data:** (not linearly separable)




**$k$-Nearest Neighbors Graph:**



**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.

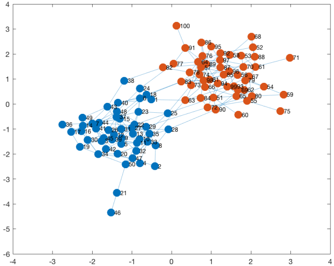
**Common Approach:** Give a natural generative model for random inputs and analyze how the algorithm performs on inputs drawn from this model.

- 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

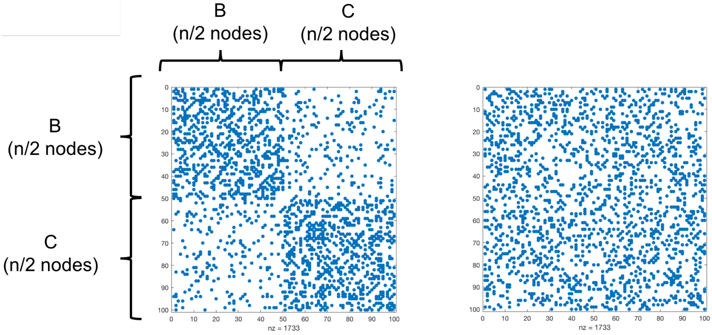
**Stochastic Block Model (Planted Partition Model):** Let $G_n(p, q)$ be a distribution over graphs on $n$ nodes, split randomly into two groups $B$ and $C$, each with $n/2$ nodes.

- 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.



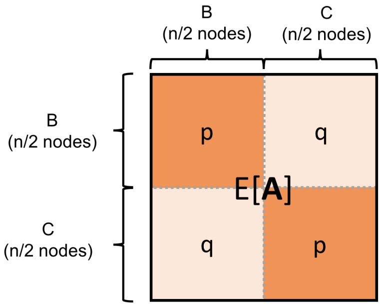
Let $G$ be a stochastic block model graph drawn from $G_n(p, q)$.

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




$G_n(p, q)$: 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.