Chapter 2

Graphs and Clustering

2.1 Graphs

A graph $G(V,E)$ is a collection of nodes $V = \{1,\dots,N\}$ and edges $E \subset V \times V$. We will use the convention that the number of nodes in a graph is $N = |V|$ and the number of edges in the graph is $M = |E|$. We will generally consider undirected, unweighted graphs.

2.2 Path Components and Union-Find

2.3 Spectral Clustering

2.3.1 The Graph Laplacian

First, we recall the definition of the incidence matrix $B \in \mathbb{R}^{N\times N}$, which encodes the relationships between nodes and edges:

$$\begin{cases} B[i,k] & = -1 \qquad e_k = (i,j) \\ B[j,k] & = +1 \qquad e_k = (i,j) \\ B[\cdot,k] & = 0 \qquad \text{otherwise} \end{cases} \tag{2.1}$$

For an undirected graph we can choose the sign arbitrarily for the edge $e_k = (i,j)$, as long as $B[i,k] = -B[j,k]$.

The graph Laplacian $L = BB^T$.

Exercise 2.3.1. *The graph Laplacian L can also be written $L = D - A$ where D is the diagonal degree matrix of the graph G, and A is the indicdence matrix of G.*

Proposition 2.3.1. *L satisfies the following properties:*

1. *$x^T Lx = \sum_{(i,j)\in E}(x_j - x_i)^2$*
2. *L is symmetric, positive semi-definite*
3. *The null eigenspace of L is spanned by indicators on connected components.*

Proof. Item 1: $x^T Lx = (B^T x)^T(B^T x)$. $B^T x$ is a M-dimensional vector whose k-th entry is $(B^T x)[k] = x_j - x_i$. The quadratic form $x^T Lx$ is just the inner product of this vector with itself, so

$$x^T Lx = \sum_{(i,j)\in E}(x_j - x_i)^2 \tag{2.2}$$

Item 2: symmetry is easy, since $L = BB^T$. Positive semi-definite means that $x^T Lx \geq 0$ for any $x \in \mathbb{R}^N$. From item 1, we know that this is the sum of squares $\sum_{(i,j)\in E}(x_j - x_i)^2$. Every entry in the sum is non-negative, so the sum is non-negative.

Item 3: because L is symmetric its eigenvalues are real and there exists an orthogonal eigenbasis $\{(v_i, \lambda_i)\}_{i=1}^N$ for L, so $Lv_i = \lambda_i v_i$, and $v_i^T v_j = 0$ if $i \neq j$. Let $\mathbb{I}_C \in \mathbb{R}^N$ denote the indicator vector on a path-connected component $C \subseteq V$:

$$\mathbb{I}_C[i] = \begin{cases} 1 & i \in C \\ 0 & i \notin C \end{cases} \tag{2.3}$$

Note that $v_i^T L v_i = \lambda_i \|v_i\|_2^2$, and $v_j^T L v_i = 0$. As a result, if $x^T L x = 0$, then x is in the null eigenspace of L. We can verify that

$$\begin{aligned} \mathbb{I}_C^T L \mathbb{I}_C^T &= \sum_{(i,j)\in E} (\mathbb{I}_C[j] - \mathbb{I}_C[i])^2 \\ &= \sum_{(i,j)\in E} (0)^2 \\ &= 0 \end{aligned}$$

because any edge (i, j) connects two vertices in the same path component. Either $i, j \in C$, in which case $\mathbb{I}_C[j] - \mathbb{I}_C[i] = 1 - 1 = 0$, or thare are in a different path component and $\mathbb{I}_C[j] - \mathbb{I}_C[i] = 0 - 0 = 0$. We conclude that $\mathbb{I}_C$ is in the null eigenspace of L.

Now, suppose that x is a vector in the null eigenspace of L. Then $x^T L x = 0$. This means the sum $\sum_{(i,j)\in E} (x_j - x_i)^2 = 0$. Because the sum is zero, and all terms in the sum are non-negative, every term in the sum must be zero. This means $x[j] - x[i] = 0$ for all $(i, j) \in E$, which implies that $x[j] = x[i]$ for any two vertices in the same path component. As a result, any eigenvector in the nullspace is in the span of indicators of connected components.

Furthermore, since vertices belong to a unique path component, if $C, D \subseteq V$ are distinct path components then the vectors $\mathbb{I}_C$ and $\mathbb{I}_D$ are orthogonal. □

As s a result of proposition 2.3.1, we can identify path-connected components of a graph by computing the null eigenspace of the graph Laplacian.

2.3.2 Clustering within Path Components

We'll now restrict our attention to a graph G which has a single path component. In this case, the null eigenspace is spanned by the constant vector $\mathbb{I}$. We would generally like to be able to cluster within path components. What we would like to do is to partition the vertex set V into $S, \bar{S} \subset V$, where $S \cup \bar{S} = V$ and $S \cap \bar{S} = \emptyset$ as to minimize the number of edges that connect the two sets. Let

$$E(S, \bar{S}) = \{(i, j) \in E \mid i \in S, j \in \bar{S} \text{ or } i \in \bar{S}, j \in S\} = (S \times \bar{S} \cup \bar{S} \times S) \cap E$$

denote the set of edges that connect S and $\bar{S}$. Let $v_S = (\mathbb{I}_S - \mathbb{I}_{\bar{S}})/2$, and note that

$$\begin{aligned} v_S^T L v_S &= \sum_{(i,j)\in E} ((\mathbb{I}_S[j] - \mathbb{I}_S[i] + \mathbb{I}_{\bar{S}}[i] - \mathbb{I}_{\bar{S}}[j])/2)^2 \\ &= \sum_{(i,j)\in S\times\bar{S}\cap E} (1)^2 + \sum_{(i,j)\in \bar{S}\times S\cap E} (1)^2 \\ &= |E(S, \bar{S})| \end{aligned}$$

Note that if we're seeking to minimize $|E(S, \bar{S})|$ that we're trying to minimize this quadratic form subject to some constraints. One way to approach this is to look at the eigenvector v_1 associated with the smallest non-zero eigenvalue of the graph Laplacian, λ_1 and to partition the graph based on the sign of the entries in the vector v_i.

$$S = \{i \mid v_1[i] > 0\} \tag{2.4}$$

Note that there is a sign ambiguity in eigenvectors, but it doesn't matter. We recover the same partition $S, \bar{S}$ either way.

The Cheeger inequality gives a notion of how well a cut based on the smallest non-zero eigenvalue approximates an optimal cut.