Spectral Clustering

在这部分，我们研究了谱方法的一个重要类别,从而在全局层次内理解网络。“谱”是指从图得出的矩阵的谱或特征值，这可以使我们深入了解图本身的结构。特别是在理解探索频谱聚类算法时，该算法利用这些工具对图中的节点进行聚类。

频谱聚类算法通常包括三个基本阶段。

1. 预处理：构造图的矩阵表示形式，例如邻接矩阵（但我们将探索其他选项）
2. 分解：计算矩阵的特征向量和特征值，并使用它们创建低维表示空间
3. 分组：根据集群在该空间中的表示将点分配给集群

Graph Partitioning

让我们公式化我们要解决的任务。我们从无向图 $G(V, E)$ 开始。我们的目标是用某种方法将 V 分为两个不相交的组 A, B (即 $A \cap B = \emptyset$ 且 $A \cup B = V$)，使组内部的连接数最大化，并使两个组之间的连接数最小。

为了进一步公式化目标，下面介绍一些术语:

- **Cut(割)**: 表示两个不相交的节点集之间有多少连接。 $cut(A, B) = \sum_{i \in A, j \in B} w_{ij}$ 其中 w_{ij} 是节点 i 和 j 之间边的权重。
- **Minimum cut(最小割):** $\arg\min_{A,B} cut(A, B)$

由于我们要尽量减少 A 和 B 之间的连接数, 我们可能会决定以**最小割**为我们的优化目标。但是我们发现这种方式最终会产生非常不直观的集群——我们通常可以简单地设置 A 为一个几乎没有传出连接的单节点，B 为网络中的其它部分，从而获得一个很小的**割**。而我们需要的是一种衡量内部集群连接性的方法。

引入**传导性**(**conductance**)可以平衡组内和组间连接性的问题。我们定义传导性为 $\phi(A, B) = \frac{cut(A,B)}{min(vol(A), vol(B))}$ 其中$vol(A) = \sum_{i \in A} k_i$ 是节点 A 的总（加权）度。可以粗略地认为传导性类似于表面积与体积之比：分子为 A 和 B 共享曲面的面积，同时分母努力确保 A 和 B 之间具有相似的体积。由于采取这种方法，选择 A 和 B 并且最小化它们的传导性，相比最小化割具有更均衡的分区。 由于要最大程度地减小电导是一个NP-hard问题，因此如何有效地找到一个良好的分区是现在需要面临的挑战。

Spectral Graph Partitioning

频谱图分割是一种允许我们使用特征向量确定传导性的方法。我们将从介绍频谱图理论的一些基本技术开始。

频谱图理论的目的是分析代表图形的矩阵的“频谱”。所谓频谱是指表示图的矩阵，按照其幅值大小排序及其对应的特征值 λ_i 的集合 $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ 。比如d-正则图的邻接矩阵的最大特征向量/特征值对是全一向量 $x = (1, 1, \ldots, 1)$, 并且特征值 $\lambda = d$。练习：具有两个分量（每个分量为d-regular）的不连续图的特征向量是什么？注意，根据谱定理，邻接矩阵（是实数和对称的）具有正交特征向量的完整谱。

我们可以使用频谱图理论分析哪些矩阵?

1. 邻接矩阵: 由于该矩阵与图结构直接相关，因此它是一个很好的切入点。它还具有对称的重要特性，这意味着它具有完整的实值正交特征向量谱
2. 拉普拉斯矩阵 L: 定义 $L = D - A$ ，其中 D 是对角矩阵， D_{ii} 表示节点 i 的度。 A 是图的邻接矩阵。拉普拉斯矩阵使我们离图的直接结构更远，但是又具有一些有趣的特性，这些特性使我们更加关注于图的更深层次结构方面的内容。我们注意到，全1向量是特征值为0的拉普拉斯矩阵的特

征向量。最后，由于 L 是半正定的, 这意味着它有三个等效条件：它的特征值都是非负的，对于某些矩阵 N 有$L=N^TN$ 并且对于每个向量 x 有 $x^TLx\geq 0$ 。这个属性使我们可以使用线性代数工具来理解 L ，从而理解原始图。

特别的， λ_2 作为 L 第二小的特征值，对它的研究使我们在理解图聚类方面取得了长足的进步。根据瑞利商理论，我们有$\lambda_2=\min_{x:x^Tw_1=0}\frac{x^TLx}{x^Tx}$ 其中 w_1 是特征值 λ_1 对应的特征向量；换句话说，我们将向量子空间中与第一个特征向量正交的目标最小化，以便找到第二个特征向量，(L 是对称的，因此具有特征值的正交基)。在高层次上，瑞利商将特征向量搜索构架为一个优化问题，使我们可以运用优化技术。注意，目标值并不依赖于 x 的大小，因此可以将其大小限制为1。另外请注意我们知道的 L 的第一个向量 是特征值为0的全为一的向量。所以说 x 正交于这个向量等于说 $\sum_i x_i=0$。

使用 L 的这些属性和定义可以写出对于 λ_2 更具体的公式：

$$\lambda_2=\min_x\frac{\sum_{(i,j)\in E}(x_i-x_j)^2}{\sum_i x_i^2}$$
$$\text{subject to}\quad \sum_i x_i=0$$

如果我们另外限制 x 为单位长度，目标函数将会转换为 $\min_x\sum_{(i,j)\in E}(x_i-x_j)^2$.

λ_2 与我们找到图的最佳分割的最初目标有何关系？让我们将分区 (A,B) 表示为向量 y ，并且 $y_i=1$ if $i\in A$ and $y_i=-1$ if $i\in B$。 我们先尝试在执行分区大小平衡问题 ($|A|=|B|$) 的同时尽量减少割，而不是使用传导性，这就相当于 $\sum_i y_i=0$。基于这个大小限制，可以最小化分区的割。比如寻找 y 最小化 $\sum_{(i,j)\in E}(y_i-y_j)^2$ ， y 的值必须是 $+1$ 或者 -1 ，这样会使得 y 的长度是固定的。这个优化问题看起来很像 λ_2 的定义，事实上根据上述发现，我们可以通过最小化拉普拉斯矩阵的 λ_2 达成这一目标，并且最佳聚类 y 由其对应的特征向量（称为Fiedler向量）给出。

现在，我们已经在 L 的特征值和图划分之间建立了联系，让我们进一步推动连接，看看是否可以摆脱硬约束 $|A|=|B|$ ，也许更灵活的传导性度量与 λ_2 之间存在某种关系。在这里我们重新定义传导性：如果图 G 被分为 A 和 B 且 $|A|\leq|B|$ ，那么割的传导性定义为$\beta=cut(A,B)/|A|$。这将 β 和λ_2 建立了关系：特别的 $\frac{\beta^2}{2k_{max}}\leq\lambda_2\leq 2\beta$，其中 k_{max}是图中的最大节点度，这个不等式称之为Cheeger不等式。由于我们需要最小化传导性 β，因此 λ_2 的上界在图分割中非常有用，该不等式可以使我们能够很好地估计传导性 β。相应的特征向量被定义为:

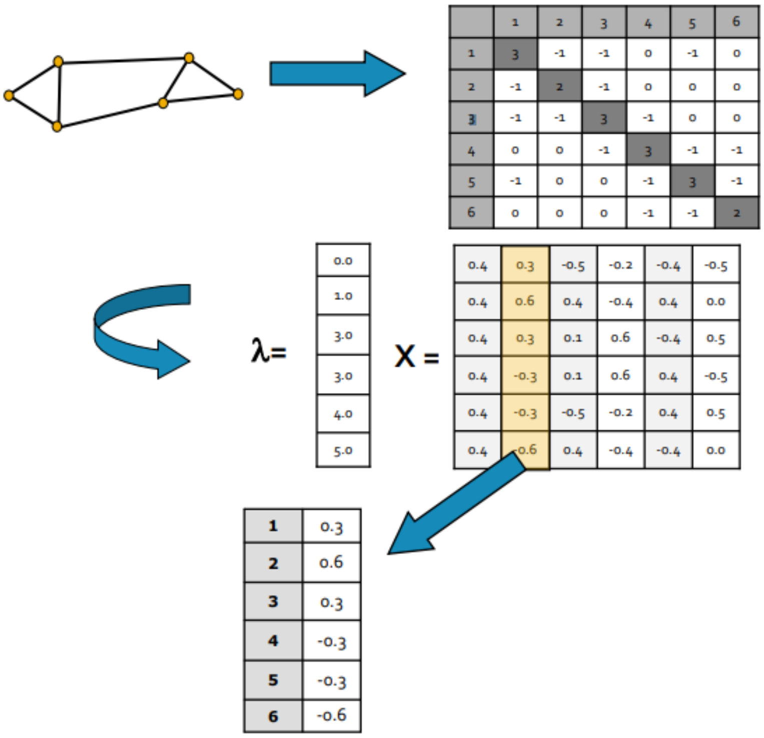
$$x_i=\begin{cases}-\frac{1}{a} & \text{if } i\in A\\ +\frac{1}{b} & \text{if } i\in B\end{cases}$$

x_i 的符号对应于每个节点的分配。

Spectral Partitioning Algorithm

将所有已知汇总起来说明频谱分割算法。

1. 预处理：构建图的拉普拉斯矩阵 L
2. 分解：将顶点映射到第二个特征向量中的相应值
3. 分组：对这些值进行排序，并将列表一分为二，以得出图分区



一些需要考虑的实际情况

- 如何在第3步中选择分割点？这里比较灵活——既可以使用简单的方法（例如零分割或中值分割），也可以使用更复杂的方法（例如最小化一维的标准化切割）。
- 如何将图划分为两个以上的集群？可以将图先分为两个簇，然后再细分这些簇，依此类推（Hagen等人，92）但这可能是效率低下且不稳定的。取而代之的是，可以使用多个特征向量进行聚类，让每个节点由其在这些特征向量中的组成表示，然后对这些表示进行聚类，例如通过k-means（Shi-Malik '00）聚类，这种方法在最近的论文中经常使用。从某种意义上说，该方法在原理上也更为可靠，它近似于最佳的K-way归一化切割，强调了内部聚类并将点映射到一个充分分离的嵌入式空间。此外，使用特征向量可尽量减少丢失的信息，因为我们可以选择使（更多信息）分量与更大的特征值相对应。
- 如何选择簇数？我们可以尝试选择聚类数 k 以最大化**eigengap**，**eigengap**即两个连续特征值之间的绝对差（按降序排列）。

Motif-Based Spectral Clustering

如果我们想通过比原始边更高级别的模式进行聚类怎么办？我们可以将图形 **Motif** 聚类为“模块”。并以类似的方式做所有事情。先从提出关于割，体积和传导性的类似定义开始：

- $cut_M(S)$ 是 Motif 的数量，其中Motif中的某些节点位于割的一侧，而其余节点在割的另一侧
- $vol_M(S)$ 是Motif M 中终点在 S 的端点数
- 定义 $\phi(S) = cut_M(S)/vol_M(S)$

我们如何找到Motif簇? 给定一个Motif M 和图 G，我们需要找到一组节点 S 从而最小化 $\phi_M(S)$。 这是一个NP-hard 问题，因此我们将再次使用谱方法，即**Motif谱聚类**:

1. 预处理: 创建一个矩阵 $W^{(M)}$， $W_{ij}^{(M)}$ 表示边 (i,j) 在 M 中出现的次数
2. 分解: 对矩阵 $W^{(M)}$ 使用标准谱聚类
3. 分组: 与标准谱聚类相同

同样，我们可以有一个 Cheeger 不等式的 motif 形式：$\phi_M(S) \leq 4\sqrt{\phi_M^*}$，其中 ϕ_M^* 是最佳传导性。

基于Motif的谱聚类方法可以应用于食物网（motif由生物学决定）和基因调控网络。