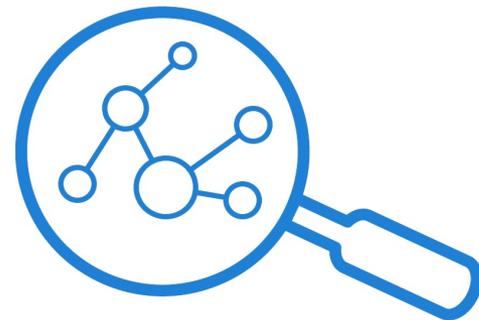


数据科学与大数据技术 的数学基础



第十一讲



计算机学院

余皓然

2024/5/27

课程内容

Part1 随机化方法

一致性哈希 布隆过滤器 CM Sketch方法 最小哈希
欧氏距离下的相似搜索 Jaccard相似度下的相似搜索

Part2 谱分析方法

主成分分析 奇异值分解 谱图论

Part3 最优化方法

压缩感知



谱图论

拉普拉斯矩阵



图数据

谱分析方法中的“谱”：矩阵特征值的集合

➤ 主成分分析、奇异值分解：

- 将 m 条 n 维数据用 $m \times n$ 矩阵 \mathbf{X} 表示，再分析 $\mathbf{X}^T \mathbf{X}$ 的特征值/特征向量或 \mathbf{X} 的奇异值/左右奇异向量，用于数据可视化、数据压缩、矩阵补全等



图数据

谱分析方法中的“谱”：矩阵特征值的集合

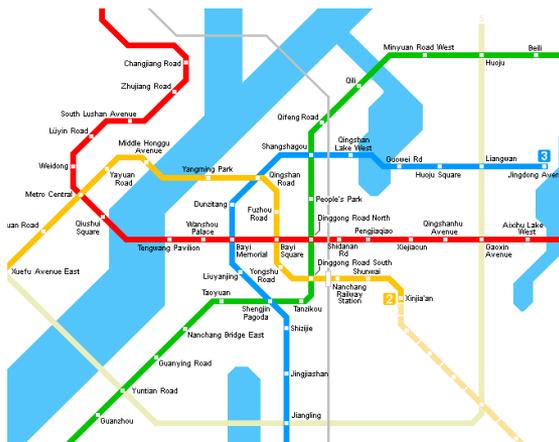
➤ 主成分分析、奇异值分解：

- 将 m 条 n 维数据用 $m \times n$ 矩阵 X 表示，再分析 $X^T X$ 的特征值/特征向量或 X 的奇异值/左右奇异向量，用于数据可视化、数据压缩、矩阵补全等

➤ 图数据是一类特殊的数据（可以用图表示的数据）



社交网络



交通网络

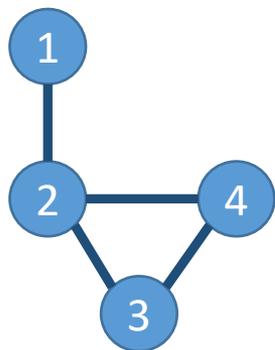


网页超链接网络
(PageRank算法)

图数据

➤ 图数据：用图 $G = (\mathcal{V}, \mathcal{E})$ 表示

- 集合 \mathcal{V} 包含图中所有顶点，记 $n = |\mathcal{V}|$ 为顶点个数
- 集合 \mathcal{E} 包含图中所有边



$$\mathcal{V} = \{1, 2, 3, 4\}, n = 4$$

$$\mathcal{E} = \{(1, 2), (2, 1), (2, 3), (3, 2), (2, 4), (4, 2), (3, 4), (4, 3)\}$$

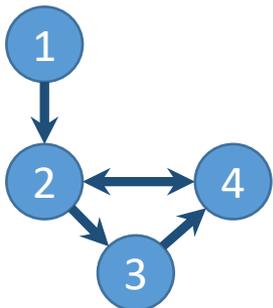
* (对无向图的) 另一种写法 $\mathcal{E} = \{(1, 2), (2, 3), (2, 4), (3, 4)\}$

默认 $(2, 1), (3, 2), (4, 2), (4, 3) \in \mathcal{E}$

图数据

本课程不涉及的拓展情况

包含有向边

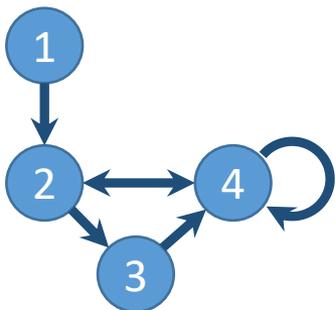


$$\mathcal{V} = \{1,2,3,4\}, n = 4$$

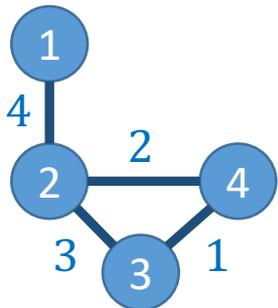
$$\mathcal{E} = \{(1,2), (2,3), (2,4), (4,2), (3,4)\}$$

例：关注/被关注；引用/被引用

包含自环



带权边

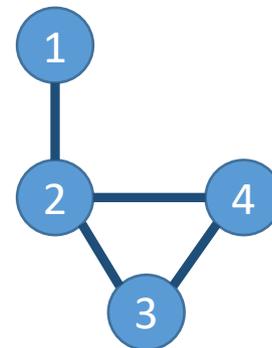


例：交通网络，各地间距离不同

图数据

➤ 图数据：用图 $G = (\mathcal{V}, \mathcal{E})$ 表示

- 集合 \mathcal{V} 包含图中所有顶点，记 $n = |\mathcal{V}|$ 为顶点个数
- 集合 \mathcal{E} 包含图中所有边



能否用矩阵表征图 G ，再通过分析矩阵的特征值/特征向量解决关于图数据的问题？

谱图论 (spectral graph theory)

spectral graph theory

About 1,610,000 results (0.18 sec)

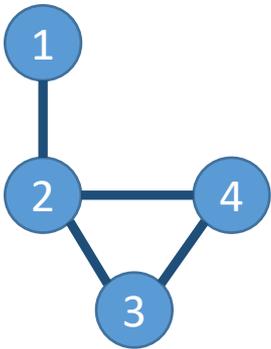
[BOOK] [Spectral graph theory](#)
FRK Chung, FC Graham - 1997 - books.google.com
Beautifully written and elegantly presented, this book is based on 10 lectures given at the CBMS workshop on **spectral graph theory** in June 1994 at Fresno State University. Chung's ...
☆ Save 📄 Cite Cited by 9938 Related articles All 5 versions 🔗

[PDF] [Spectral graph theory](#)
D Spielman - Combinatorial scientific computing, 2012 - Citeseer
Spectral graph theory is the study and exploration of graphs through the eigenvalues and eigenvectors of matrices naturally ... We then survey a few applications of **spectral graph theory**. ...
☆ Save 📄 Cite Cited by 247 Related articles All 6 versions 🔗

[Spectral graph theory and its applications](#)
DA Spielman - 48th Annual IEEE Symposium on Foundations ..., 2007 - ieeexplore.ieee.org
... to **graph** partitioning and expansion. In Section 6, I explain how **spectral graph theory** has ...
I discuss **spectral** approaches to testing **graph** isomorphism in Section 8. Finally, I conclude ...
☆ Save 📄 Cite Cited by 282 Related articles All 14 versions 🔗

矩阵表示

➤ $n \times n$ 的邻接矩阵A (adjacency matrix) : $A_{ij} = 1$ 当且仅当 $(i, j) \in \mathcal{E}$

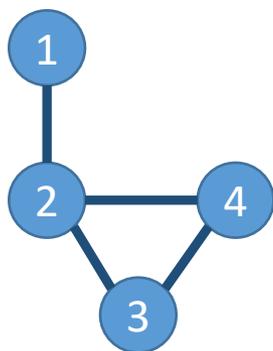


$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

矩阵表示

➤ $n \times n$ 的邻接矩阵A (adjacency matrix) : $A_{ij} = 1$ 当且仅当 $(i, j) \in \mathcal{E}$

➤ $n \times n$ 的拉普拉斯矩阵L: $L_{ij} = D_{ij} - A_{ij} = \begin{cases} \text{顶点 } i \text{ 的度数, 若 } i = j, \\ -1, \text{ 若 } (i, j) \in \mathcal{E}, \\ 0, \text{ 其它情况.} \end{cases}$



D (度矩阵)

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

A

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

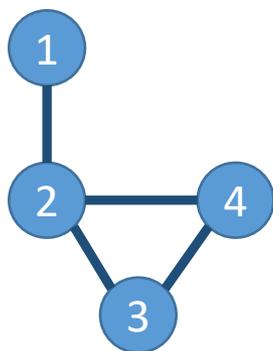
L

$$= \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

矩阵表示

➤ $n \times n$ 的邻接矩阵A (adjacency matrix) : $A_{ij} = 1$ 当且仅当 $(i, j) \in \mathcal{E}$

➤ $n \times n$ 的拉普拉斯矩阵L: $L_{ij} = D_{ij} - A_{ij} = \begin{cases} \text{顶点}i\text{的度数, 若}i=j, \\ -1, \text{若}(i, j) \in \mathcal{E}, \\ 0, \text{其它情况.} \end{cases}$


$$\begin{array}{ccc} \mathbf{D} \text{ (度矩阵)} & \mathbf{A} & \mathbf{L} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} & - \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} & = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} \end{array}$$

对本课程考虑的无向简单图, A和L都是**对称矩阵**, L的每行或每列元素之和都为0

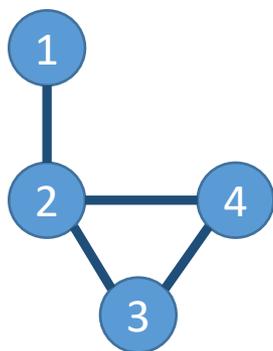
矩阵表示



拉普拉斯
(法国, 1749–1827)

➤ $n \times n$ 的邻接矩阵**A** (adjacency matrix) : $A_{ij} = 1$ 当且仅当 $(i, j) \in \mathcal{E}$

➤ $n \times n$ 的拉普拉斯矩阵**L**: $L_{ij} = D_{ij} - A_{ij} = \begin{cases} \text{顶点 } i \text{ 的度数, 若 } i = j, \\ -1, \text{ 若 } (i, j) \in \mathcal{E}, \\ 0, \text{ 其它情况.} \end{cases}$



D (度矩阵)

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

A

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

L

$$= \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

对本课程考虑的无向简单图, **A**和**L**都是**对称矩阵**, **L**的每行或每列元素之和都为0

矩阵表示

- 可通过分析 L 的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵 L 是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

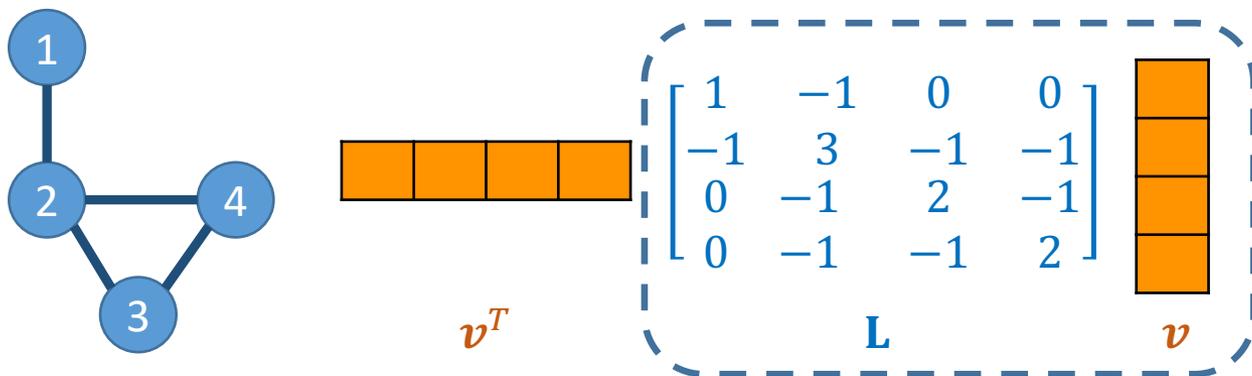
化简 $v^T L v$



矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

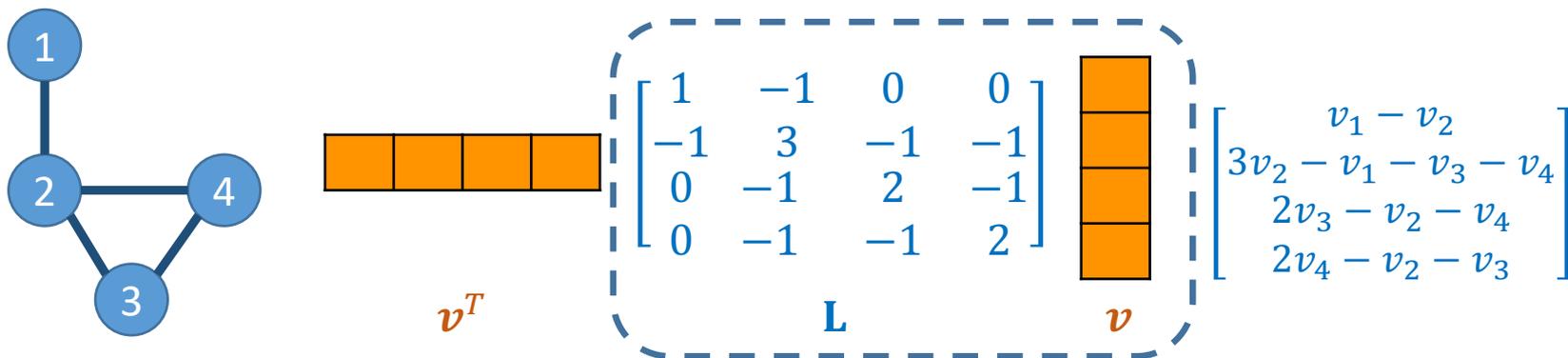
$$\text{化简 } v^T L v = \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right)$$



矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

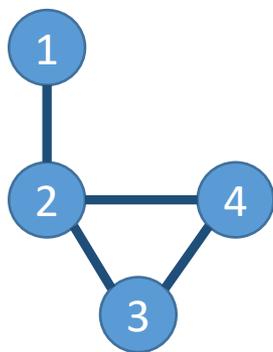
$$\text{化简 } v^T L v = \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right)$$



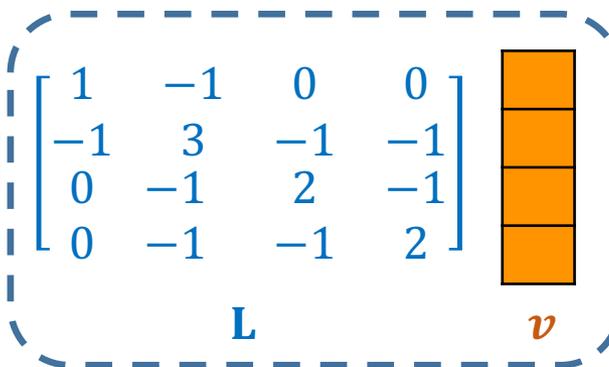
矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

$$\text{化简 } v^T L v = \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right)$$



v^T



L

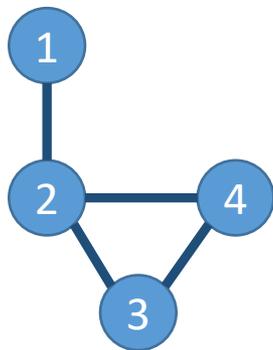
v

$$= \begin{bmatrix} v_1 - v_2 \\ 3v_2 - v_1 - v_3 - v_4 \\ 2v_3 - v_2 - v_4 \\ 2v_4 - v_2 - v_3 \\ v_1 - v_2 \\ v_2 - v_1 + v_2 - v_3 + v_2 - v_4 \\ v_3 - v_2 + v_3 - v_4 \\ v_4 - v_2 + v_4 - v_3 \end{bmatrix}$$

矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

$$\text{化简 } v^T L v = \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right)$$

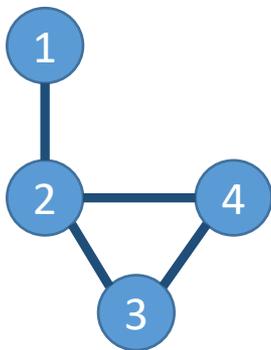


$$\begin{bmatrix} v_1 - v_2 \\ v_2 - v_1 + v_2 - v_3 + v_2 - v_4 \\ v_3 - v_2 + v_3 - v_4 \\ v_4 - v_2 + v_4 - v_3 \end{bmatrix}$$

矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

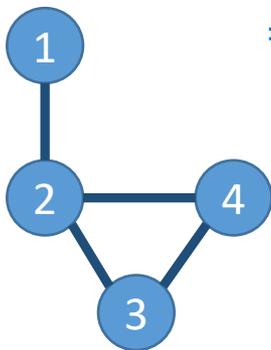
$$\begin{aligned} \text{化简 } v^T L v &= \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right) = \sum_{i=1}^n \sum_{j:(i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{(i,j) \in \mathcal{E}} v_i (v_i - v_j) \end{aligned}$$



矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

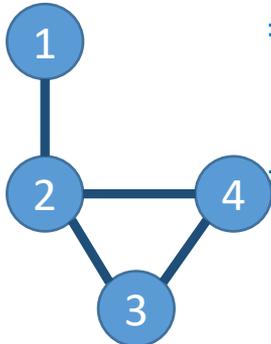
$$\begin{aligned} \text{化简 } v^T L v &= \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right) = \sum_{i=1}^n \sum_{j:(i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{(i,j) \in \mathcal{E}} v_i (v_i - v_j) = \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i > j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{j > i: (j,i) \in \mathcal{E}} v_j (v_j - v_i) \end{aligned}$$



矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

化简 $v^T L v$

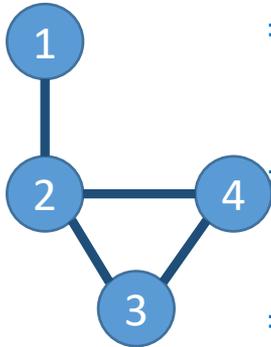
$$\begin{aligned} &= \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right) = \sum_{i=1}^n \sum_{j:(i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{(i,j) \in \mathcal{E}} v_i (v_i - v_j) = \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i > j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{j > i: (j,i) \in \mathcal{E}} v_j (v_j - v_i) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i < j: (i,j) \in \mathcal{E}} v_j (v_j - v_i) \end{aligned}$$


在无向图中，若 $(j, i) \in \mathcal{E}$ 则有 $(i, j) \in \mathcal{E}$

矩阵表示

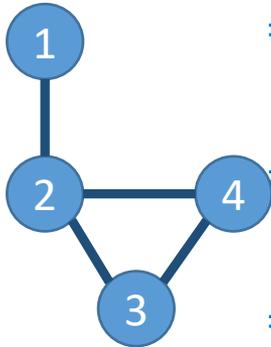
- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

化简 $v^T L v$

$$\begin{aligned} &= \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right) = \sum_{i=1}^n \sum_{j:(i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{(i,j) \in \mathcal{E}} v_i (v_i - v_j) = \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i > j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{j > i: (j,i) \in \mathcal{E}} v_j (v_j - v_i) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i < j: (i,j) \in \mathcal{E}} v_j (v_j - v_i) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)(v_i - v_j) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2 \end{aligned}$$


矩阵表示

- 可通过分析L的特征值、特征向量处理图数据
- 可证 $n \times n$ 的拉普拉斯矩阵L是半正定矩阵，即对任意向量 $v \in \mathbb{R}^n$ 有 $v^T L v \geq 0$

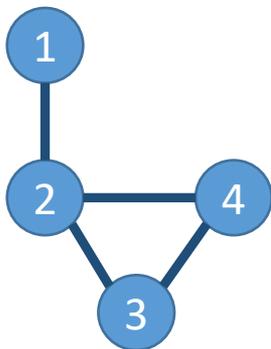

$$\begin{aligned} \text{化简 } v^T L v &= \sum_{i=1}^n v_i \left(\sum_{j:(i,j) \in \mathcal{E}} (v_i - v_j) \right) = \sum_{i=1}^n \sum_{j:(i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{(i,j) \in \mathcal{E}} v_i (v_i - v_j) = \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i > j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{j > i: (j,i) \in \mathcal{E}} v_j (v_j - v_i) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} v_i (v_i - v_j) + \sum_{i < j: (i,j) \in \mathcal{E}} v_j (v_j - v_i) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)(v_i - v_j) \\ &= \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2 \end{aligned}$$

恒为非负值

半正定矩阵L的特征值非负

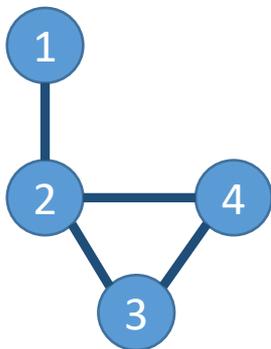
矩阵表示

$$v^T L v = \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2 \text{ 的含义?}$$



矩阵表示

$$v^T L v = \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2 \text{ 的含义}$$



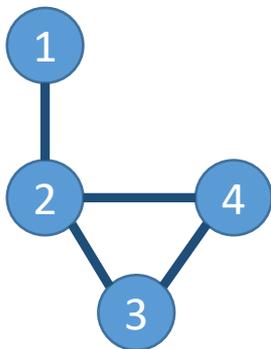
$$(v_1 - v_2)^2 + (v_2 - v_3)^2 + (v_2 - v_4)^2 + (v_3 - v_4)^2$$



矩阵表示

$$\mathbf{v}^T \mathbf{L} \mathbf{v} = \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2 \text{ 的含义}$$

- 若每个顶点 i 分配一个值 v_i ， $\mathbf{v}^T \mathbf{L} \mathbf{v}$ 刻画相邻顶点的值的差的平方和



$$(v_1 - v_2)^2 + (v_2 - v_3)^2 + (v_2 - v_4)^2 + (v_3 - v_4)^2$$

$\mathbf{v}^T \mathbf{L} \mathbf{v}$ 的值小说明相邻点有类似的 v_i 值

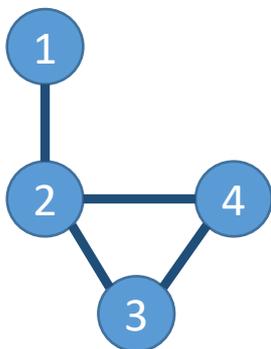
例：社交网络好友的年龄



矩阵表示

L的特征值与特征向量包含了图的许多特征

➤ 至少有一个特征值为0

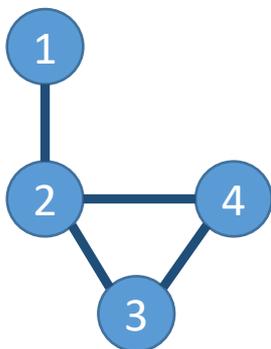


矩阵表示

L的特征值与特征向量包含了图的许多特征

➤ 至少有一个特征值为0

构造单位向量 $\boldsymbol{v} = \left[\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right]^T$ ，易证 $\mathbf{L}\boldsymbol{v} = 0\boldsymbol{v}$



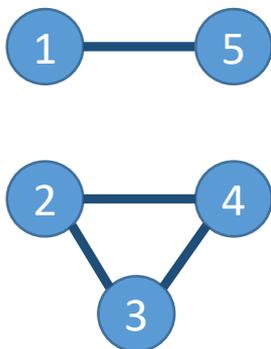
$$\mathbf{L}\boldsymbol{v} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} \frac{1}{\sqrt{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

回顾：无向简单图中L的每行之和都为0

矩阵表示

L的特征值与特征向量包含了图的许多特征

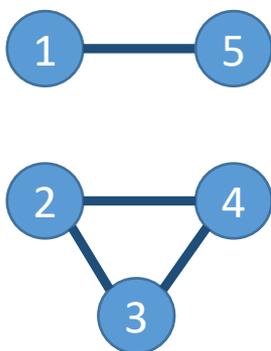
- 至少有一个特征值为0
- 零特征值的个数等于连通子图的个数



矩阵表示

L的特征值与特征向量包含了图的许多特征

- 至少有一个特征值为0
- 零特征值的个数等于连通子图的个数



$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 \\ 0 & 2 & -1 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

```
Lexample =
```

```
1 0 0 0 -1
0 2 -1 -1 0
0 -1 2 -1 0
0 -1 -1 2 0
-1 0 0 0 1
```

```
>> [Vexample,Dexample]=eig(Lexample)
```

```
Vexample =
```

```
-0.7071 0 -0.7071 0 0
0 0.5774 0 0.5164 -0.6325
0 0.5774 0 -0.8059 -0.1310
0 0.5774 0 0.2895 0.7634
-0.7071 0 0.7071 0 0
```

```
Dexample =
```

```
0 0 0 0 0
0 0.0000 0 0 0
0 0 2.0000 0 0
0 0 0 3.0000 0
0 0 0 0 3.0000
```

注意：特征向量不唯一

谱图论

拉普拉斯矩阵的特征值与特征向量



特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

由 L 是实对称矩阵可得：（证明参见Courant-Fischer最小最大定理）

$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

最优目标函数值 $v_1^T L v_1 = v_1^T \lambda_1 v_1 = \lambda_1$ （即最大特征值）

定理引自Christopher Musco <NYU CS-GY 6763: Algorithmic Machine Learning and Data Science>

另可参见Xianyi Zeng <UTEP MATH 5330: Computational Methods of Linear Algebra>

以及Wing-Kin Ma <CUHK ENGG 5781: Matrix Analysis and Computations>

特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

由 L 是实对称矩阵可得：（证明参见 Courant-Fischer 最小最大定理）

$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} v^T L v$$

$$v_3 = \arg \max_{\|v\|=1, v \perp v_1, v_2} v^T L v$$

⋮

$$v_n = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{n-1}} v^T L v$$

定理引自 Christopher Musco <NYU CS-GY 6763: Algorithmic Machine Learning and Data Science>

另可参见 Xianyi Zeng <UTEP MATH 5330: Computational Methods of Linear Algebra>

以及 Wing-Kin Ma <CUHK ENGG 5781: Matrix Analysis and Computations>

特征向量

用 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ 表示 \mathbf{L} 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

由 \mathbf{L} 是实对称矩阵可得：（证明参见Courant-Fischer最小最大定理）

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

$$\mathbf{v}_2 = \arg \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

$$\mathbf{v}_3 = \arg \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \mathbf{v}_2} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

⋮

$$\mathbf{v}_n = \arg \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \dots, \mathbf{v}_{n-1}} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

$$\mathbf{v}_1^T \mathbf{L} \mathbf{v}_1 = \mathbf{v}_1^T \lambda_1 \mathbf{v}_1 = \lambda_1 \quad (\text{即最大特征值})$$

$$\mathbf{v}_2^T \mathbf{L} \mathbf{v}_2 = \mathbf{v}_2^T \lambda_2 \mathbf{v}_2 = \lambda_2 \quad (\text{即第二大特征值})$$

$$\mathbf{v}_3^T \mathbf{L} \mathbf{v}_3 = \mathbf{v}_3^T \lambda_3 \mathbf{v}_3 = \lambda_3 \quad (\text{即第三大特征值})$$

$$\mathbf{v}_n^T \mathbf{L} \mathbf{v}_n = \mathbf{v}_n^T \lambda_n \mathbf{v}_n = \lambda_n \quad (\text{即最小特征值})$$

定理引自 Christopher Musco <NYU CS-GY 6763: Algorithmic Machine Learning and Data Science>

另可参见 Xianyi Zeng <UTEP MATH 5330: Computational Methods of Linear Algebra>

以及 Wing-Kin Ma <CUHK ENGG 5781: Matrix Analysis and Computations>

特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

由 L 是实对称矩阵可得：

$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

证明：

步骤一：用 L 的单位特征向量 v_1, v_2, \dots, v_n 表示任意单位向量 $v = \sum_i \langle v_i, v \rangle v_i$

步骤二：将 L 写成 $\sum_i \lambda_i v_i v_i^T$ ，利用正交性化简目标函数得 $\sum_i \lambda_i \langle v_i, v \rangle^2$

步骤三：因 v 是单位向量，有 $\sum_i \langle v_i, v \rangle^2 = 1$ ，得目标函数最大值为 λ_1 ，由 $v = v_1$ 取得



特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

由 L 是实对称矩阵可得：

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} v^T L v$$

证明：

步骤一：用 L 的单位特征向量 v_1, v_2, \dots, v_n 表示任意单位向量 $v = \sum_{i \neq 1} \langle v_i, v \rangle v_i$

步骤二：将 L 写成 $\sum_i \lambda_i v_i v_i^T$ ，利用正交性化简目标函数得 $\sum_{i \neq 1} \lambda_i \langle v_i, v \rangle^2$

步骤三：因 v 是单位向量，有 $\sum_i \langle v_i, v \rangle^2 = 1$ ，得目标函数最大值为 λ_2 ，由 $v = v_2$ 取得

特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量 (依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$)

由 L 是实对称矩阵可得:

$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} v^T L v$$

$$v_3 = \arg \max_{\|v\|=1, v \perp v_1, v_2} v^T L v$$

\vdots

$$v_n = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{n-1}} v^T L v$$

$$v_n = \arg \min_{\|v\|=1} v^T L v$$

$$v_{n-1} = \arg \min_{\|v\|=1, v \perp v_n} v^T L v$$

$$v_{n-2} = \arg \min_{\|v\|=1, v \perp v_n, v_{n-1}} v^T L v$$

\vdots

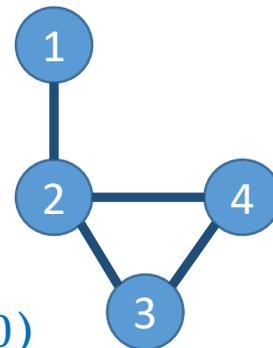
$$v_1 = \arg \min_{\|v\|=1, v \perp v_n, \dots, v_2} v^T L v$$

定理引自 Christopher Musco <NYU CS-GY 6763: Algorithmic Machine Learning and Data Science>

另可参见 Xianyi Zeng <UTEP MATH 5330: Computational Methods of Linear Algebra>

以及 Wing-Kin Ma <CUHK ENGG 5781: Matrix Analysis and Computations>

特征向量



用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量 (依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$)

回顾 $v^T L v = \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2$ (相邻顶点的值的差的平方和)

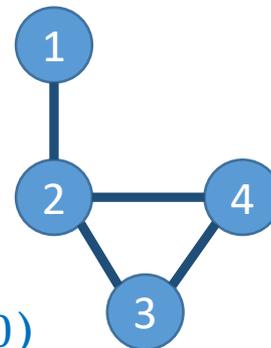
$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

$$v_n = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{n-1}} v^T L v$$

$$v_n = \arg \min_{\|v\|=1} v^T L v$$

$$v_1 = \arg \min_{\|v\|=1, v \perp v_n, \dots, v_2} v^T L v$$

特征向量



用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量 (依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$)

回顾 $v^T L v = \sum_{i < j: (i,j) \in E} (v_i - v_j)^2$ (相邻顶点的值的差的平方和)

➤ 特征向量 $v_1 = [v_{11}, v_{12}, \dots, v_{1n}]^T$: 对相邻点 i 与 j , v_{1i} 与 v_{1j} 相差较大

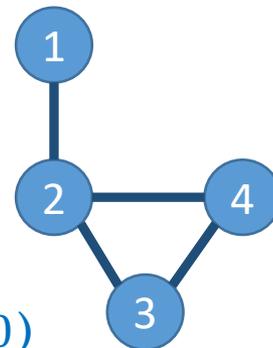
$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

$$v_n = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{n-1}} v^T L v$$

$$v_n = \arg \min_{\|v\|=1} v^T L v$$

$$v_1 = \arg \min_{\|v\|=1, v \perp v_n, \dots, v_2} v^T L v$$

特征向量



用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量 (依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$)

回顾 $v^T L v = \sum_{i < j: (i,j) \in \mathcal{E}} (v_i - v_j)^2$ (相邻顶点的值的差的平方和)

- 特征向量 $v_1 = [v_{11}, v_{12}, \dots, v_{1n}]^T$: 对相邻点 i 与 j , v_{1i} 与 v_{1j} 相差较大
- ...
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$: 对相邻点 i 与 j , $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$: 对相邻点 i 与 j , v_{ni} 与 v_{nj} 相等

可证 $v_n^T L v_n = v_n^T \lambda_n v_n = \lambda_n = 0$, 说明 $\sum_{i < j: (i,j) \in \mathcal{E}} (v_{ni} - v_{nj})^2 = 0$

$$v_1 = \arg \max_{\|v\|=1} v^T L v$$

$$v_n = \arg \max_{\|v\|=1, v \perp v_1, \dots, v_{n-1}} v^T L v$$

$$v_n = \arg \min_{\|v\|=1} v^T L v$$

$$v_1 = \arg \min_{\|v\|=1, v \perp v_n, \dots, v_2} v^T L v$$

特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

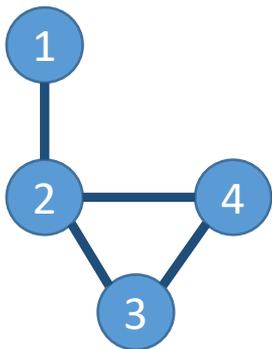
- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等



特征向量

用 v_1, v_2, \dots, v_n 表示L的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等



```
Lexample =  
    1   -1    0    0  
   -1    3   -1   -1  
    0   -1    2   -1  
    0   -1   -1    2  
  
>> [Vexample,Dexample]=eig(Lexample)  
  
Vexample =  
    0.5000    0.8165  -0.0000    0.2887  
    0.5000    0.0000  -0.0000   -0.8660  
    0.5000   -0.4082  -0.7071    0.2887  
    0.5000   -0.4082    0.7071    0.2887  
       $v_n$        $v_{n-1}$        $v_1$   
  
Dexample =  
    0.0000    0    0    0  
    0    1.0000    0    0  
    0    0    3.0000    0  
    0    0    0    4.0000
```

特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等

20个顶点连成环的图

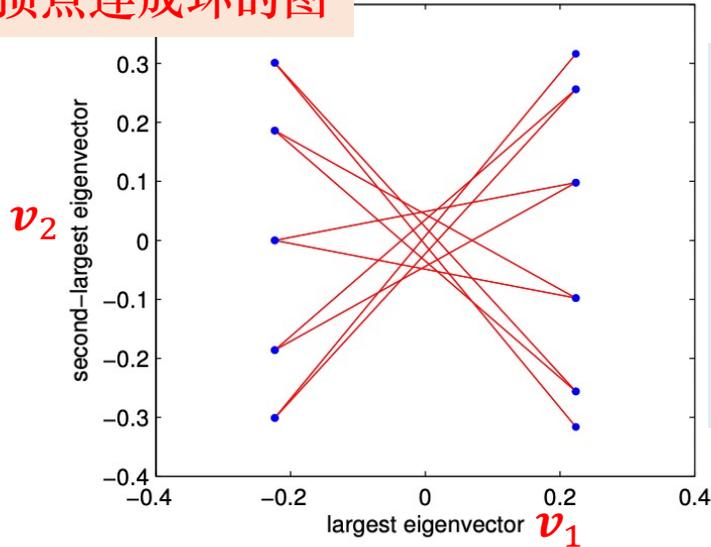


特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等

20个顶点连成环的图



计算 L 的特征向量 v_1 与 v_2

为20个点分别画出
 $(v_{11}, v_{21}), (v_{12}, v_{22}), \dots, (v_{1,20}, v_{2,20})$

用红线把实际图中相邻的点连接起来（每个点有两条红线）

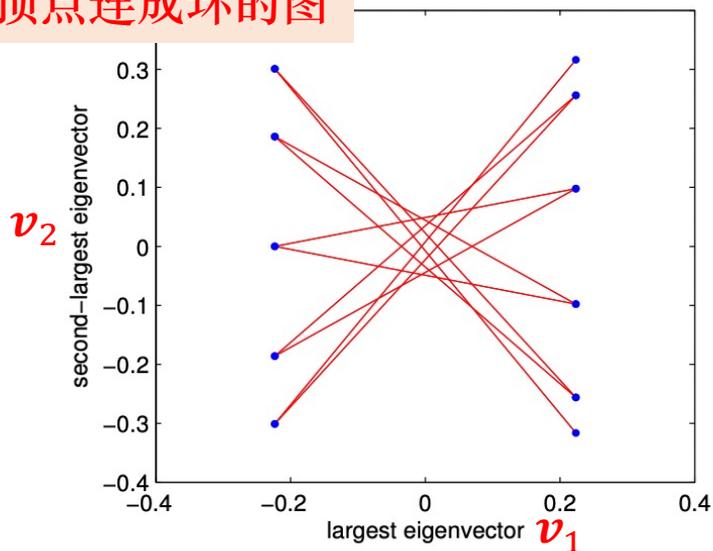
（存在值相同的情况）

特征向量

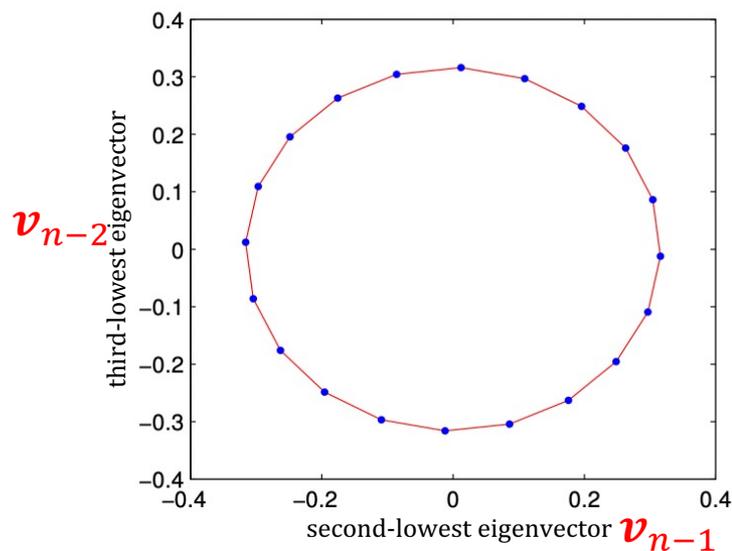
用 v_1, v_2, \dots, v_n 表示L的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等

20个顶点连成环的图



(存在值相同的情况)



(所有点的 v_{ni} 都相等，故不画 v_n)

特征向量

用 v_1, v_2, \dots, v_n 表示 L 的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等

400个顶点构成的20x20网格状的图

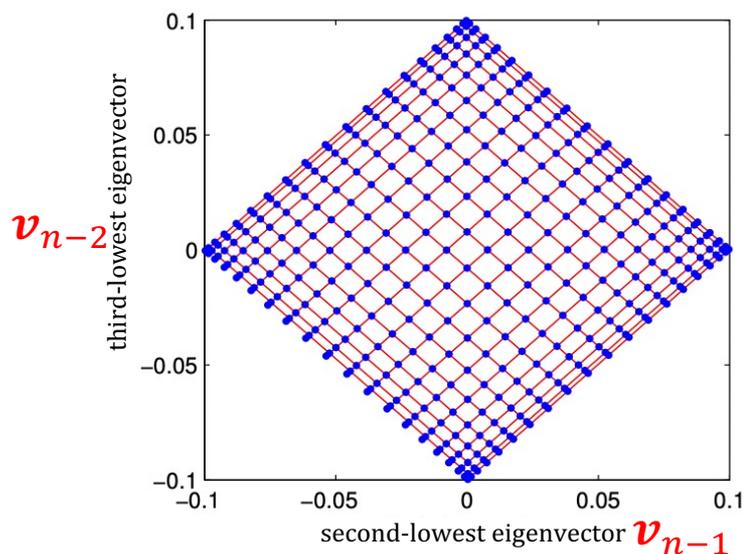
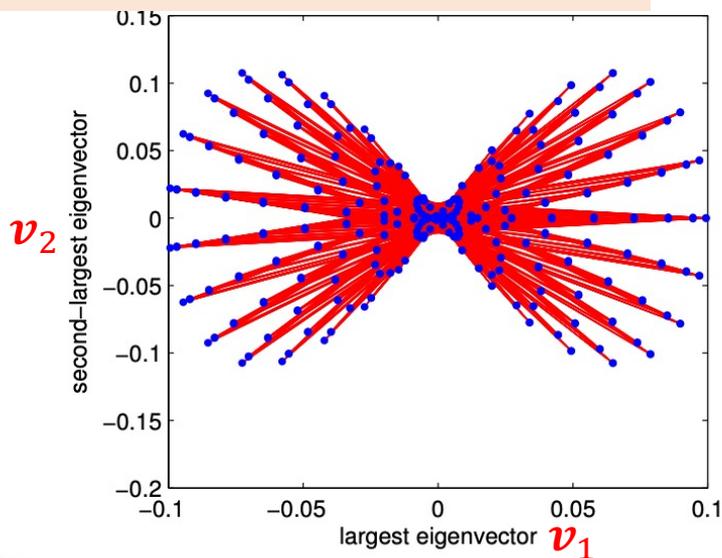


特征向量

用 v_1, v_2, \dots, v_n 表示L的单位特征向量（依次对应 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ）

- 特征向量 v_1 ：对相邻点 i 与 j ， v_{1i} 与 v_{1j} 相差较大
- 特征向量 $v_{n-1} = [v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,n}]^T$ ：对相邻点 i 与 j ， $v_{n-1,i}$ 与 $v_{n-1,j}$ 相差较小
- 特征向量 $v_n = [v_{n1}, v_{n2}, \dots, v_{nn}]^T$ ：对相邻点 i 与 j ， v_{ni} 与 v_{nj} 相等

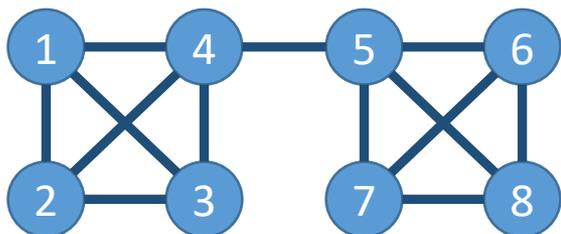
400个顶点构成的20x20网格状的图



谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

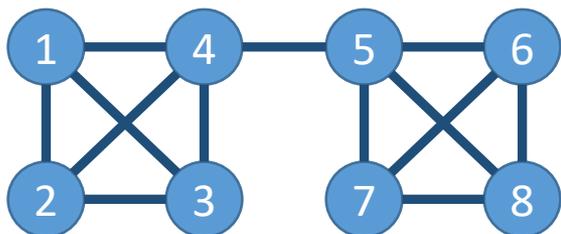
例: 每个顶点用一个实数表示 (每个顶点的嵌入属于 \mathbb{R})



谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用一个实数表示 (每个顶点的嵌入属于 \mathbb{R})

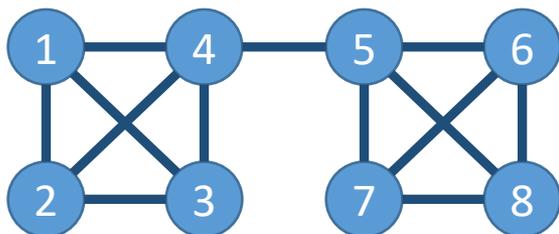


L =							
3	-1	-1	-1	0	0	0	0
-1	3	-1	-1	0	0	0	0
-1	-1	3	-1	0	0	0	0
-1	-1	-1	4	-1	0	0	0
0	0	0	-1	4	-1	-1	-1
0	0	0	0	-1	3	-1	-1
0	0	0	0	-1	-1	3	-1
0	0	0	0	-1	-1	-1	3

谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用一个实数表示 (每个顶点的嵌入属于 \mathbb{R})

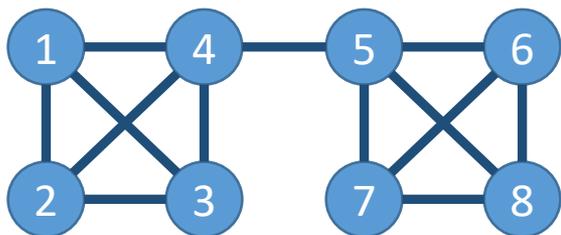


```
>> [V,D]=eig(L)
V =
  0.3536  -0.3825  0.0560  -0.1090  -0.0624  0.8288  0.0495  0.1426
  0.3536  -0.3825  0.3675  -0.4547  -0.3807  -0.4681  0.0495  0.1426
  0.3536  -0.3825  -0.6572  0.4485  -0.0808  -0.2576  0.0495  0.1426
  0.3536  -0.2470  0.2336  0.1152  0.5239  -0.1031  -0.1485  -0.6626
  0.3536  0.2470  0.2336  0.1152  0.5239  -0.1031  -0.1485  0.6626
  0.3536  0.3825  -0.4999  -0.5806  0.0196  0.0299  -0.3466  -0.1426
  0.3536  0.3825  0.2662  0.4654  -0.5435  0.0733  -0.3466  -0.1426
  0.3536  0.3825  0.0000  0.0000  -0.0000  0.0000  0.8416  -0.1426
      vn  vn-1
D =
 -0.0000  0  0  0  0  0  0  0
  0  0.3542  0  0  0  0  0  0
  0  0  4.0000  0  0  0  0  0
  0  0  0  4.0000  0  0  0  0
  0  0  0  0  4.0000  0  0  0
  0  0  0  0  0  4.0000  0  0
  0  0  0  0  0  0  4.0000  0
  0  0  0  0  0  0  0  5.6458
```

谱嵌入

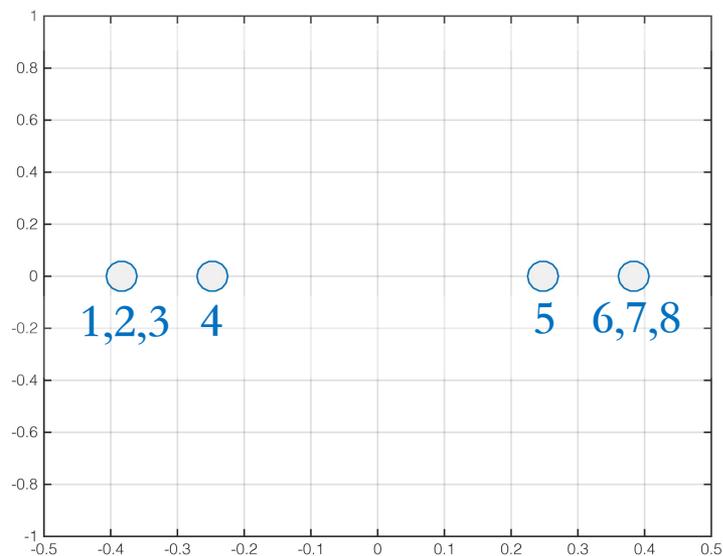
谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用一个实数表示 (每个顶点的嵌入属于 \mathbb{R})



-0.3825
-0.3825
-0.3825
-0.2470
0.2470
0.3825
0.3825
0.3825

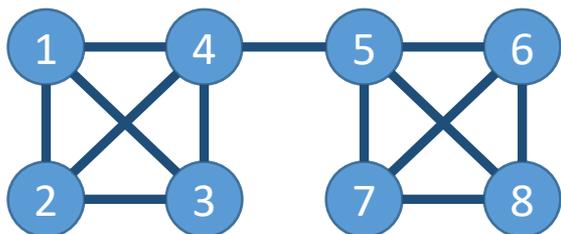
v_{n-1}



谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用两个实数表示 (每个顶点的嵌入属于 \mathbb{R}^2)

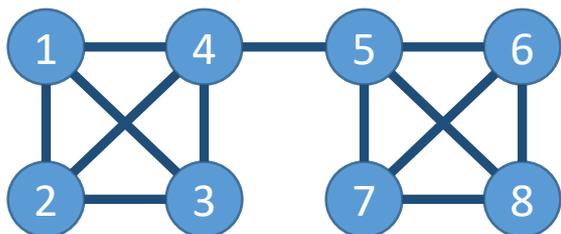


```
>> [V,D]=eig(L)
V =
  0.3536  -0.3825  0.0560  -0.1090  -0.0624  0.8288  0.0495  0.1426
  0.3536  -0.3825  0.3675  -0.4547  -0.3807  -0.4681  0.0495  0.1426
  0.3536  -0.3825  -0.6572  0.4485  -0.0808  -0.2576  0.0495  0.1426
  0.3536  -0.2470  0.2336  0.1152  0.5239  -0.1031  -0.1485  -0.6626
  0.3536  0.2470  0.2336  0.1152  0.5239  -0.1031  -0.1485  0.6626
  0.3536  0.3825  -0.4999  -0.5806  0.0196  0.0299  -0.3466  -0.1426
  0.3536  0.3825  0.2662  0.4654  -0.5435  0.0733  -0.3466  -0.1426
  0.3536  0.3825  0.0000  0.0000  -0.0000  0.0000  0.8416  -0.1426
      v_{n-1}  v_{n-2}
D =
 -0.0000  0  0  0  0  0  0  0
  0  0.3542  0  0  0  0  0  0
  0  0  4.0000  0  0  0  0  0
  0  0  0  4.0000  0  0  0  0
  0  0  0  0  4.0000  0  0  0
  0  0  0  0  0  4.0000  0  0
  0  0  0  0  0  0  4.0000  0
  0  0  0  0  0  0  0  5.6458
```

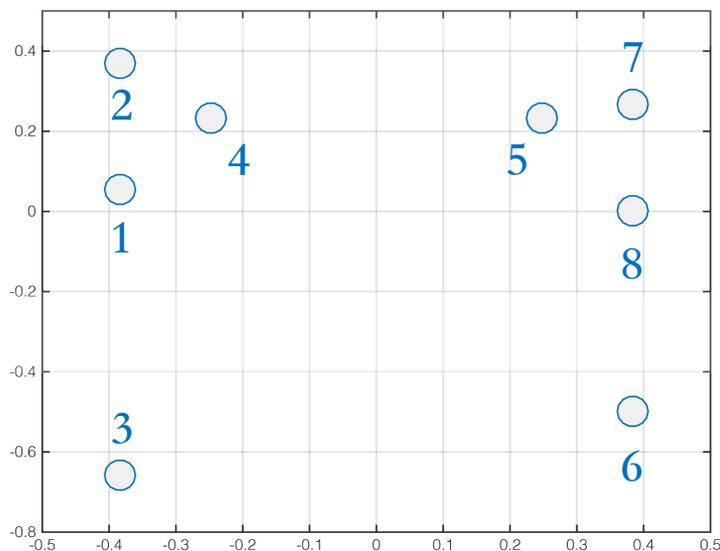
谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用两个实数表示 (每个顶点的嵌入属于 \mathbb{R}^2)



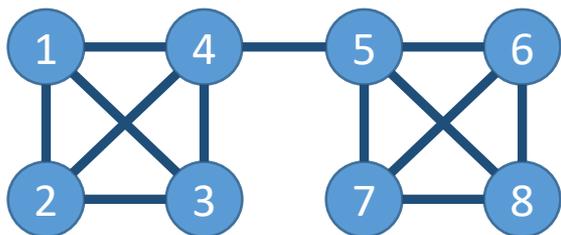
-0.3825	0.0560
-0.3825	0.3675
-0.3825	-0.6572
-0.2470	0.2336
0.2470	0.2336
0.3825	-0.4999
0.3825	0.2662
0.3825	0.0000
v_{n-1}	v_{n-2}



谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用三个实数表示 (每个顶点的嵌入属于 \mathbb{R}^3)



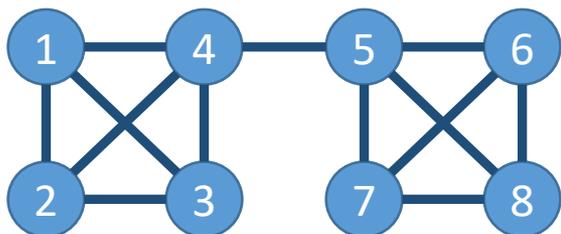
```
>> [V,D]=eig(L)
V =
0.3536 -0.3825 0.0560 -0.1090 -0.0624 0.8288 0.0495 0.1426
0.3536 -0.3825 0.3675 -0.4547 -0.3807 -0.4681 0.0495 0.1426
0.3536 -0.3825 -0.6572 0.4485 -0.0808 -0.2576 0.0495 0.1426
0.3536 -0.2470 0.2336 0.1152 0.5239 -0.1031 -0.1485 -0.6626
0.3536 0.2470 0.2336 0.1152 0.5239 -0.1031 -0.1485 0.6626
0.3536 0.3825 -0.4999 -0.5806 0.0196 0.0299 -0.3466 -0.1426
0.3536 0.3825 0.2662 0.4654 -0.5435 0.0733 -0.3466 -0.1426
0.3536 0.3825 0.0000 0.0000 -0.0000 0.0000 0.8416 -0.1426
D =
-0.0000 0 0 0 0 0 0 0
0 0.3542 0 0 0 0 0 0
0 0 4.0000 0 0 0 0 0
0 0 0 4.0000 0 0 0 0
0 0 0 0 4.0000 0 0 0
0 0 0 0 0 4.0000 0 0
0 0 0 0 0 0 4.0000 0
0 0 0 0 0 0 0 5.6458
```

v_{n-1} v_{n-2} v_{n-3}

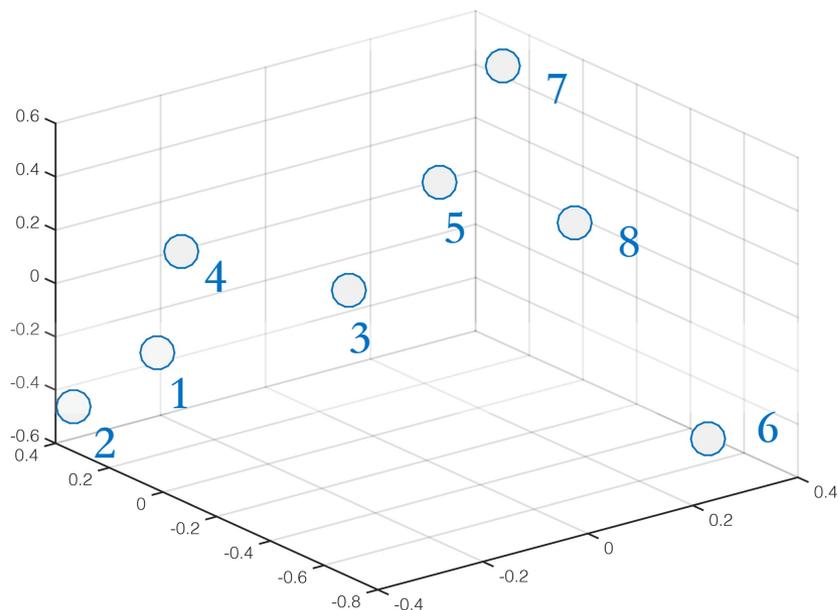
谱嵌入

谱嵌入 (spectral embedding) : 为图中每个顶点找到低维表示 (如用一个实数表示), 使得相邻顶点有相似的表示

例: 每个顶点用三个实数表示 (每个顶点的嵌入属于 \mathbb{R}^3)



-0.3825	0.0560	-0.1090
-0.3825	0.3675	-0.4547
-0.3825	-0.6572	0.4485
-0.2470	0.2336	0.1152
0.2470	0.2336	0.1152
0.3825	-0.4999	-0.5806
0.3825	0.2662	0.4654
0.3825	0.0000	0.0000
v_{n-1}	v_{n-2}	v_{n-3}



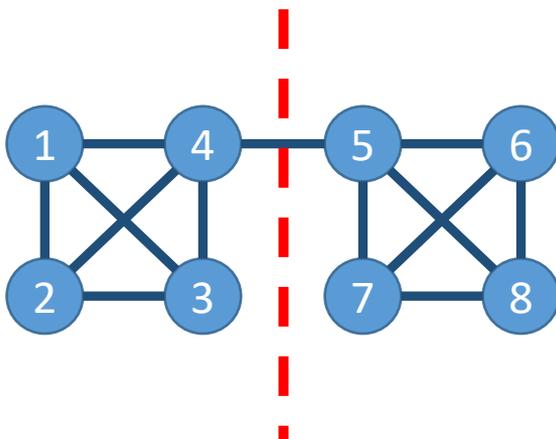
谱图论

谱聚类



谱聚类

谱聚类 (spectral clustering) : 如何根据图的谱 (如 L 的特征值) 对图中顶点聚类?



谱聚类

谱聚类 (spectral clustering) : 如何根据图的谱 (如L的特征值) 对图中顶点聚类?

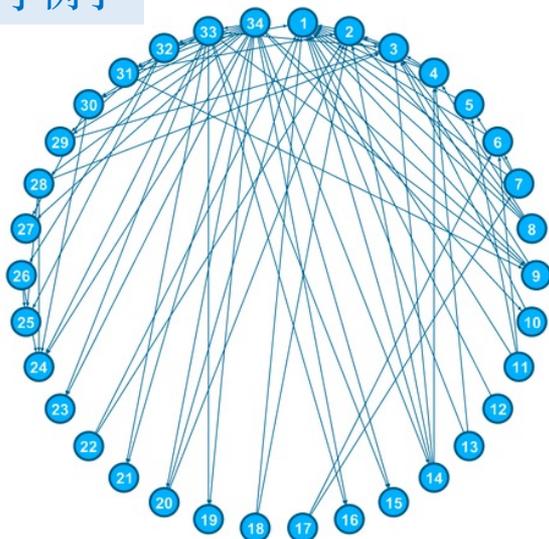
An information flow model for conflict and fission in small groups

WW Zachary - Journal of anthropological research, 1977 - journals.uchicago.edu

Data from a voluntary association are used to construct a new formal model for a traditional anthropological problem, fission in small groups. The process leading to fission is viewed as an unequal flow of sentiments and information across the ties in a social network. This flow is unequal because it is uniquely constrained by the contextual range and sensitivity of each relationship in the network. The subsequent differential sharing of sentiments leads to the formation of subgroups with more internal stability than the group as a whole, and results in ...

☆ Save 剪 Cite Cited by 5329 Related articles All 13 versions

人类学例子



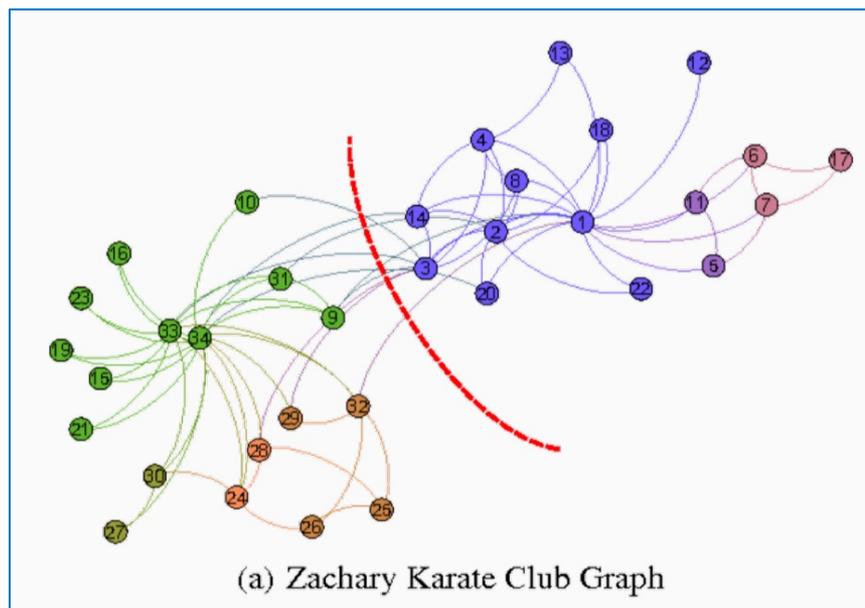
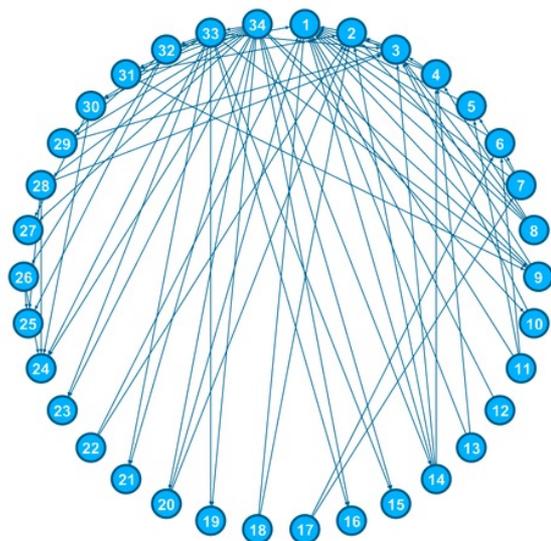
观察一个空手道俱乐部在1970~1972期间的变化

俱乐部部长与教练就是否涨价的问题分道扬镳，俱乐部一分为二

论文作者Zachary根据34个会员此前在俱乐部外的社交数据，准确预测出33个会员的去向

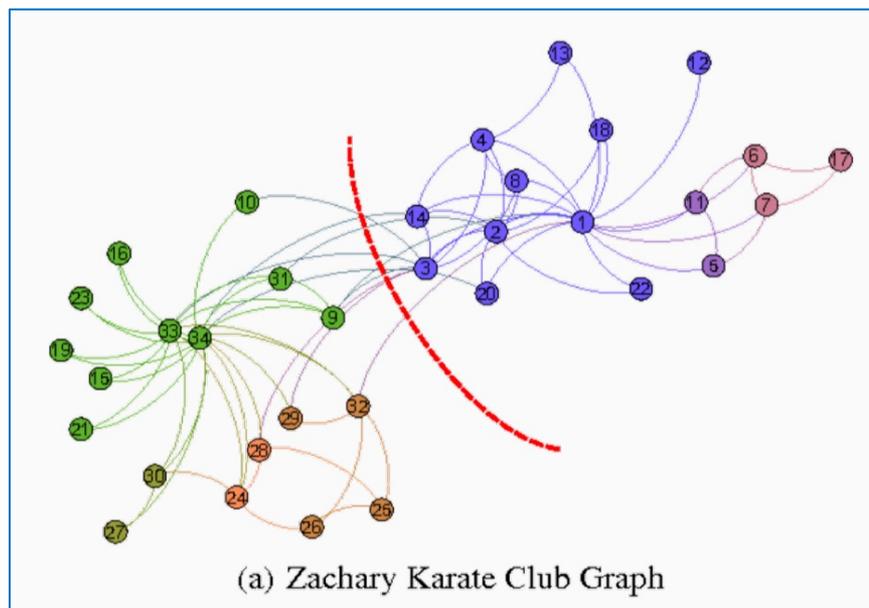
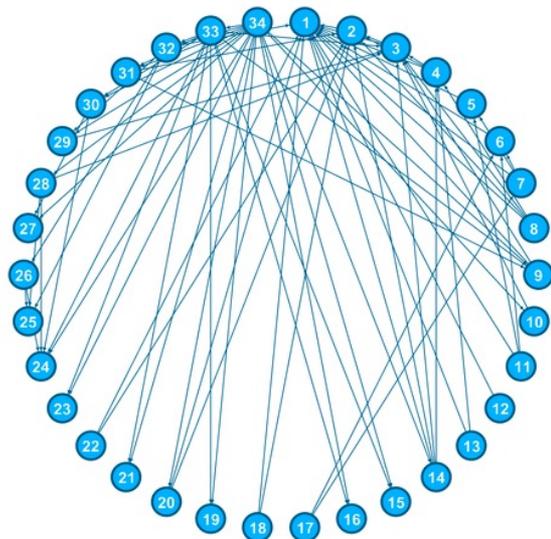
谱聚类

谱聚类 (spectral clustering) : 如何根据图的谱 (如 L 的特征值) 对图中顶点聚类?



谱聚类

谱聚类 (spectral clustering) : 如何根据图的谱 (如 L 的特征值) 对图中顶点聚类?



解决谱聚类问题有助于:

- (1) 解决社交网络中许多应用问题;
- (2) 解决无监督机器学习中非线性聚类问题

谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

* 对于如何建模图的分割问题, 有多种不同的方法

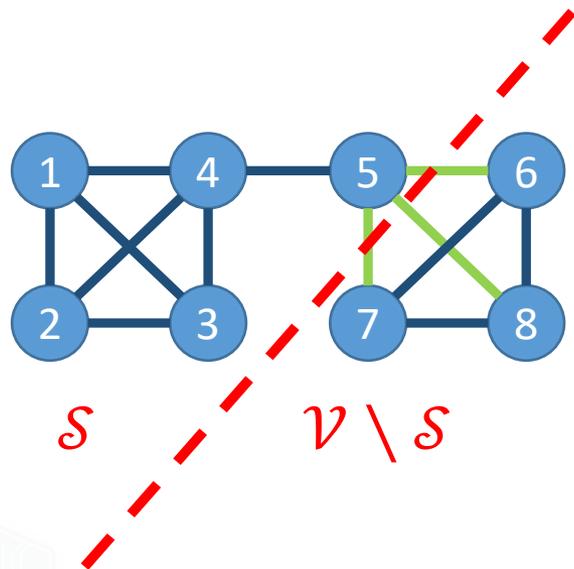
谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$ ，求解以下问题：

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}$$

其中， $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。



$$\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S}) = 3$$

$$\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\} = \min\{5, 3\} = 3$$

目标函数值为1

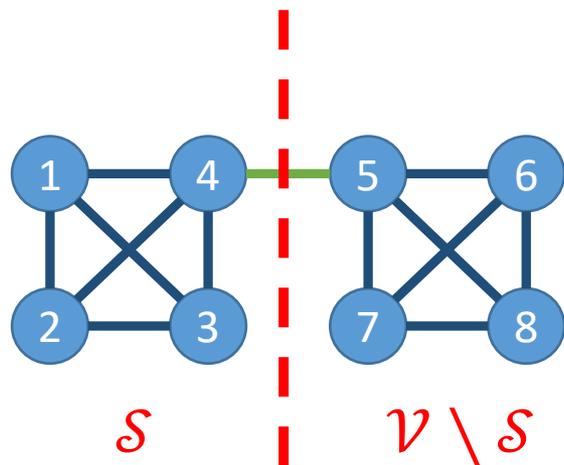
谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。



$$\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S}) = 1$$

$$\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\} = \min\{4, 4\} = 4$$

$$\text{目标函数值为 } \frac{1}{4}$$

谱聚类

稀疏割问题 (Sparsest Cut)

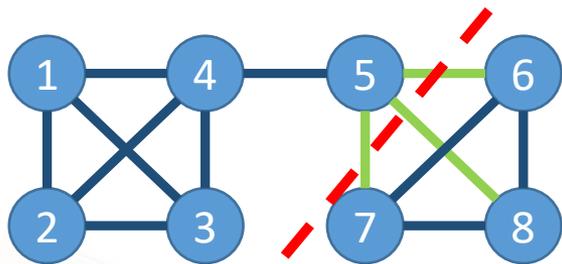
给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}$$

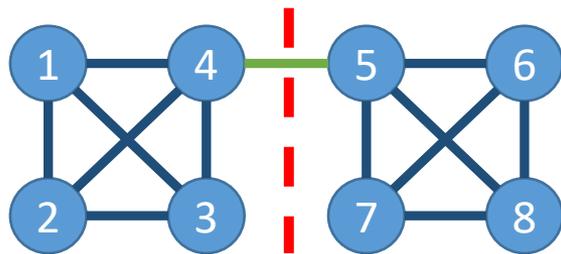
其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

尽量减小 $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$: 避免将过多有连接的点分割到不同集合

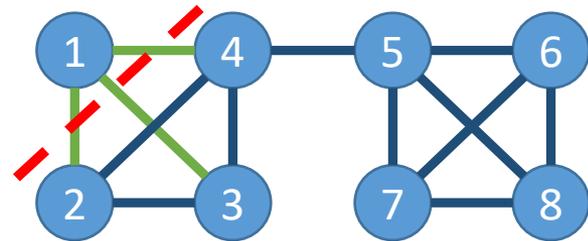
尽量增大 $\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}$: 避免分割出的两个集合大小不均衡



目标函数值为1



目标函数值为 $\frac{1}{4}$



目标函数值为3

谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

NP难问题, 一种 (近似) 解决方案: 用拉普拉斯矩阵等改写问题 (新问题是原问题的近似)



谱聚类

稀疏割问题 (Sparsest Cut)

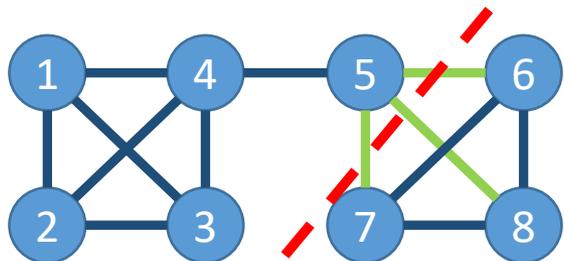
给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ (即所有 v_i 只能取 1 或 -1) 表示各顶点被划到哪个集合:

$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} , $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$



$$\mathbf{v} = [-1 \quad -1 \quad -1 \quad -1 \quad -1 \quad 1 \quad 1 \quad 1]^T$$

谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$ ，求解以下问题：

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中， $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ （即所有 v_i 只能取 1 或 -1）表示各顶点被划到哪个集合：

$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} ， $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$

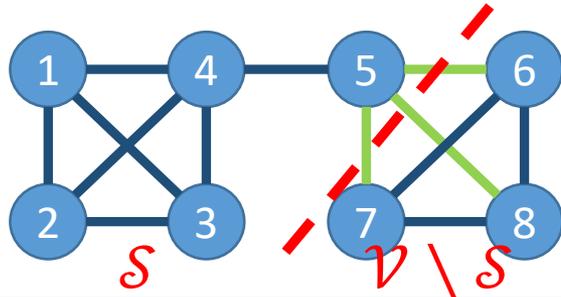
将原问题近似为：

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s. t.} & \mathbf{v}^T \mathbf{1} = 0, \\ \text{var. } & \mathbf{v} \in \{-1, 1\}^n. \end{aligned}$$

拉普拉斯矩阵



谱聚类



稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ (即所有 v_i 只能取 1 或 -1) 表示各顶点被划到哪个集合:

$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} , $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$

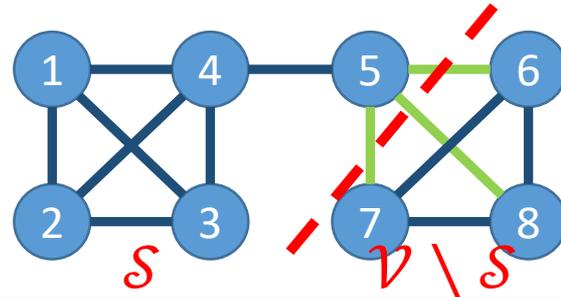
将原问题近似为:

$$\begin{aligned} & \min \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ & \text{s. t. } \mathbf{v}^T \mathbf{1} = 0, \\ & \text{var. } \mathbf{v} \in \{-1, 1\}^n. \end{aligned}$$

$$\text{回顾目标函数 } \mathbf{v}^T \mathbf{L} \mathbf{v} = \sum_{i < j: (i, j) \in \mathcal{E}} (v_i - v_j)^2$$

易证在当前 \mathbf{v} 定义下, $\mathbf{v}^T \mathbf{L} \mathbf{v} = 4 \text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$

谱聚类



稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ (即所有 v_i 只能取 1 或 -1) 表示各顶点被划到哪个集合:

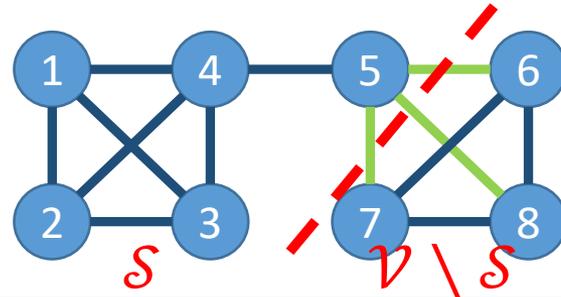
$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} , $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$

将原问题近似为:

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s. t.} & \mathbf{v}^T \mathbf{1} = 0, \\ \text{var. } & \mathbf{v} \in \{-1, 1\}^n. \end{aligned}$$

$$\begin{aligned} \mathbf{v}^T \mathbf{1} &= v_1 + v_2 + \dots + v_n \\ \text{易证 } \mathbf{v}^T \mathbf{1} &= |\mathcal{V} \setminus \mathcal{S}| - |\mathcal{S}| \end{aligned}$$

谱聚类



稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ (即所有 v_i 只能取 1 或 -1) 表示各顶点被划到哪个集合:

$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} , $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$

将原问题近似为:

$$\begin{aligned} & \min \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ & \text{s. t. } \mathbf{v}^T \mathbf{1} = 0, \\ & \text{var. } \mathbf{v} \in \{-1, 1\}^n. \end{aligned}$$

$$\begin{aligned} \mathbf{v}^T \mathbf{1} &= v_1 + v_2 + \dots + v_n \\ \text{易证 } \mathbf{v}^T \mathbf{1} &= |\mathcal{V} \setminus \mathcal{S}| - |\mathcal{S}| \end{aligned}$$

将原问题中增大 $\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}$ 的目的近似为等式约束 $\mathbf{v}^T \mathbf{1} = 0$, 即 $|\mathcal{V} \setminus \mathcal{S}| = |\mathcal{S}|$

(因为 $|\mathcal{S}| + |\mathcal{V} \setminus \mathcal{S}|$ 取值固定, 二者相等时 $\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}$ 值最大)

谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$ ，求解以下问题：

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中， $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ （即所有 v_i 只能取 1 或 -1）表示各顶点被划到哪个集合：

$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} ， $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$

近似问题1

$$\begin{aligned} & \min \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ & \text{s. t. } \mathbf{v}^T \mathbf{1} = 0, \\ & \text{var. } \mathbf{v} \in \{-1, 1\}^n. \end{aligned}$$

离散决策变量不好处理

谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

用向量 $\mathbf{v} \in \{-1, 1\}^n$ (即所有 v_i 只能取 1 或 -1) 表示各顶点被划到哪个集合:

$v_i = -1$ 对应顶点 i 被划到集合 \mathcal{S} , $v_i = 1$ 对应顶点 i 被划到集合 $\mathcal{V} \setminus \mathcal{S}$

近似问题1

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s.t.} & \mathbf{v}^T \mathbf{1} = 0, \\ \text{var.} & \mathbf{v} \in \{-1, 1\}^n. \end{aligned}$$

再近似

近似问题2

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s.t.} & \mathbf{v}^T \mathbf{1} = 0, \\ \text{var.} & \mathbf{v} \in \mathbb{R}^n. \end{aligned}$$



先考虑连续变量, 得到最优解后再近似为离散变量 (如根据正负等)

离散决策变量不好处理

谱聚类

近似问题2

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s. t.} & \mathbf{v}^T \mathbf{1} = 0, \\ & \text{var. } \mathbf{v} \in \mathbb{R}^n. \end{aligned}$$

(得到最优解后再近似为离散变量)



谱聚类

回顾：L的最小特征值对应单位特征向量

$$\mathbf{v}_n = \left[\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right]^T$$

近似问题2

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s.t.} & \mathbf{v}^T \mathbf{1} = 0, \\ \text{var.} & \mathbf{v} \in \mathbb{R}^n. \end{aligned}$$

“等价于”

近似问题2

$$\begin{aligned} \min & \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s.t.} & \mathbf{v}^T \mathbf{v}_n = 0, \\ & \|\mathbf{v}\| = 1, \\ \text{var.} & \mathbf{v} \in \mathbb{R}^n. \end{aligned}$$

(得到最优解后再近似为离散变量)



谱聚类

近似问题2

$$\begin{aligned} \min & \frac{1}{4} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s.t.} & \mathbf{v}^T \mathbf{1} = 0, \\ \text{var.} & \mathbf{v} \in \mathbb{R}^n. \end{aligned}$$

“等价于”

近似问题2

$$\begin{aligned} \min & \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \text{s.t.} & \mathbf{v}^T \mathbf{v}_n = 0, \\ & \|\mathbf{v}\| = 1, \\ \text{var.} & \mathbf{v} \in \mathbb{R}^n. \end{aligned}$$

(得到最优解后再近似为离散变量)

回顾前面介绍的
Courant-Fischer最小最大定理

$$\begin{aligned} \mathbf{v}_n &= \arg \min_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{L} \mathbf{v} \\ \mathbf{v}_{n-1} &= \arg \min_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_n} \mathbf{v}^T \mathbf{L} \mathbf{v} \end{aligned}$$

所以，近似问题2的最优解即：

拉普拉斯矩阵的第二小特征值对应的特征向量 \mathbf{v}_{n-1}

谱聚类

稀疏割问题 (Sparsest Cut)

给定 $G = (\mathcal{V}, \mathcal{E})$, 求解以下问题:

$$\min_{\mathcal{S} \subset \mathcal{V}} \frac{\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})}{\min\{|\mathcal{S}|, |\mathcal{V} \setminus \mathcal{S}|\}}.$$

其中, $\text{cut}(\mathcal{S}, \mathcal{V} \setminus \mathcal{S})$ 是横跨两个顶点集合 \mathcal{S} 与 $\mathcal{V} \setminus \mathcal{S}$ 的边的数目。

基于谱方法的近似分割

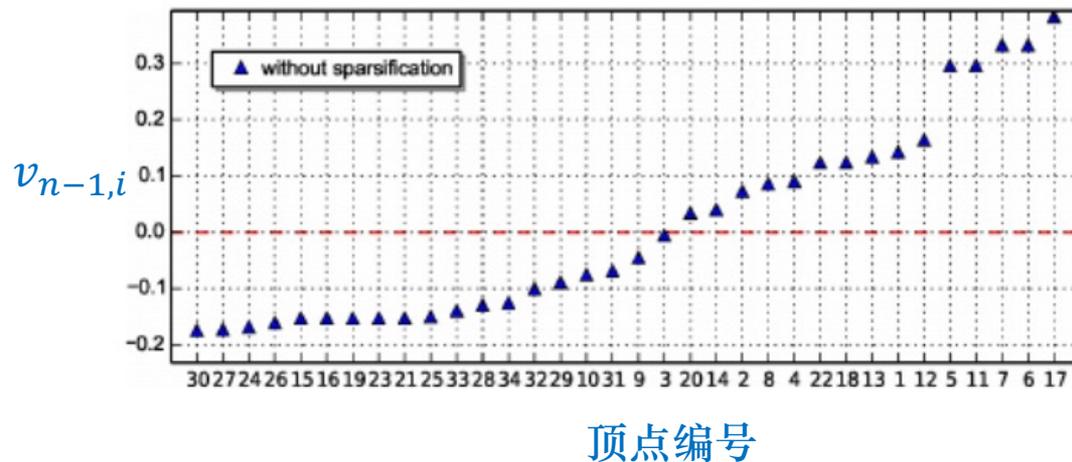
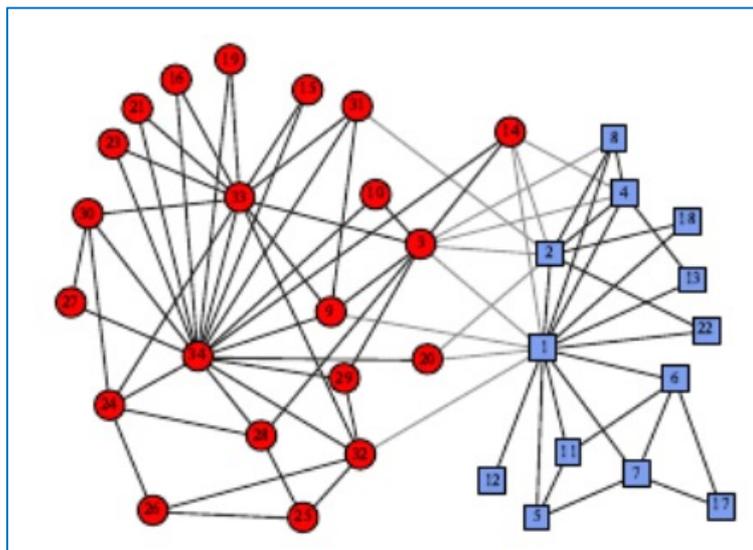
- (1) 计算图 G 的拉普拉斯矩阵 L
- (2) 计算矩阵 L 的第二小特征值对应的特征向量 v_{n-1}
- (3) 根据每个顶点 i 的值 $v_{n-1,i}$ 将其划到集合 \mathcal{S} 或集合 $\mathcal{V} \setminus \mathcal{S}$ (如以 0 或中值为界)



谱聚类

基于谱方法的近似分割

- (1) 计算图 G 的拉普拉斯矩阵 L
- (2) 计算矩阵 L 的第二小特征值对应的特征向量 v_{n-1}
- (3) 根据每个顶点 i 的值 $v_{n-1,i}$ 将其划到集合 S 或集合 $V \setminus S$ (如以0或中值为界)

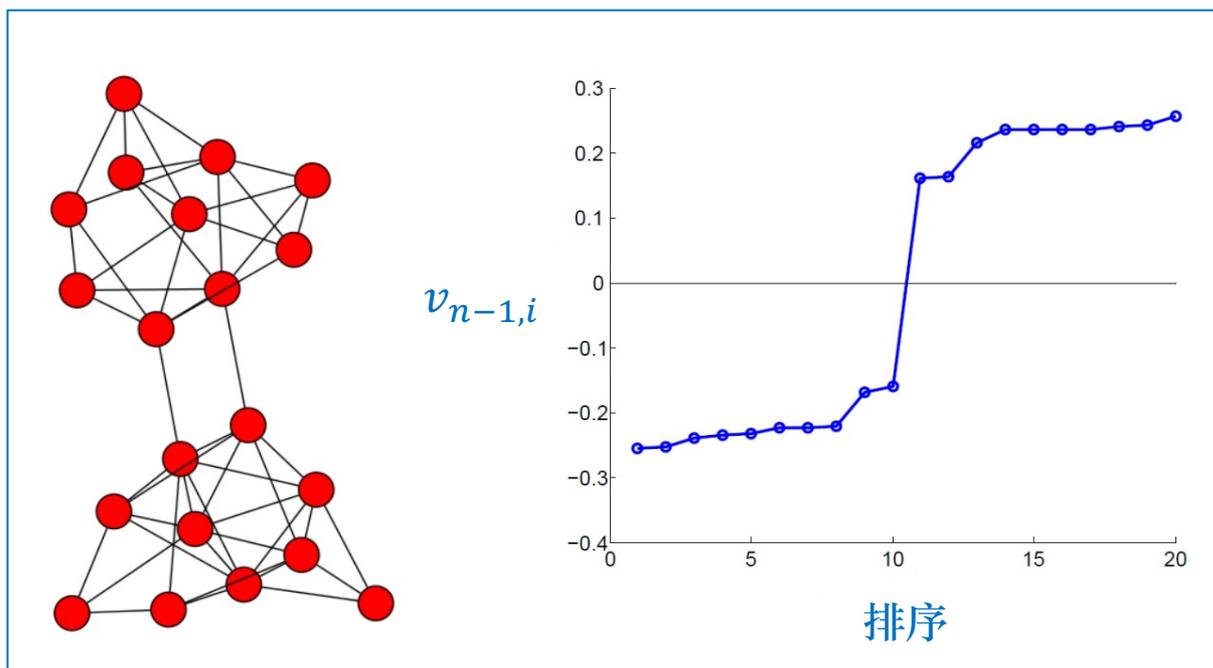


前述空手道俱乐部例子

谱聚类

基于谱方法的近似分割

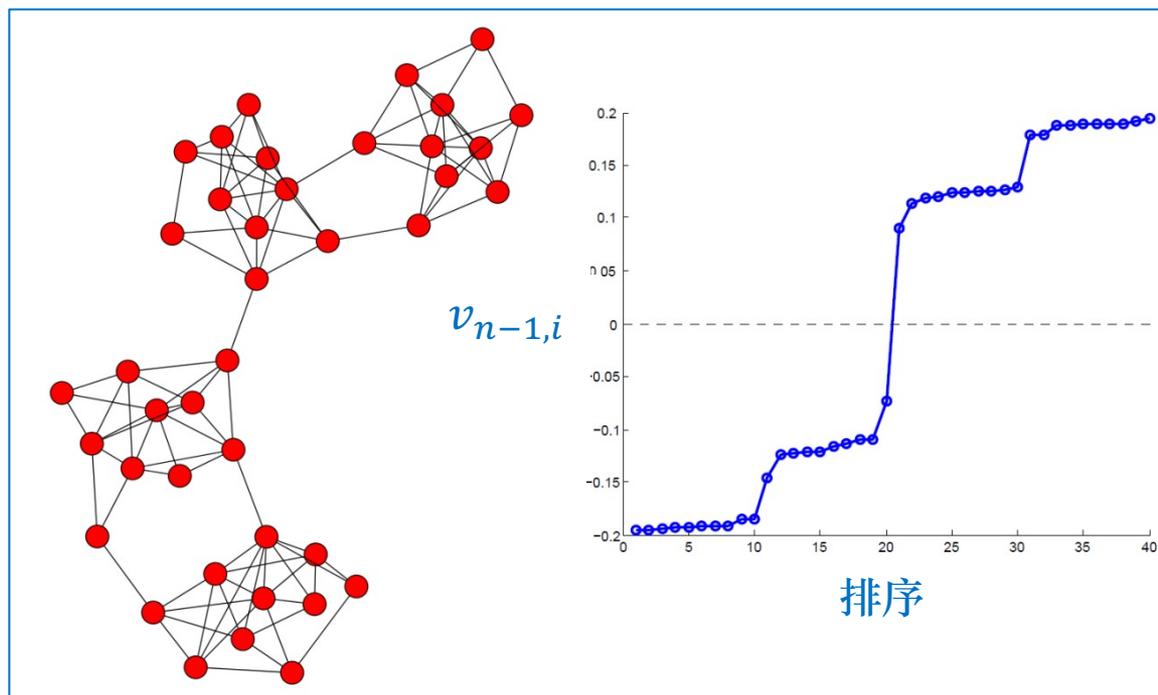
- (1) 计算图 G 的拉普拉斯矩阵 L
- (2) 计算矩阵 L 的第二小特征值对应的特征向量 v_{n-1}
- (3) 根据每个顶点 i 的值 $v_{n-1,i}$ 将其划到集合 S 或集合 $V \setminus S$ (如以0或中值为界)



谱聚类

基于谱方法的近似分割

- (1) 计算图 G 的拉普拉斯矩阵 L
- (2) 计算矩阵 L 的第二小特征值对应的特征向量 v_{n-1}
- (3) 根据每个顶点 i 的值 $v_{n-1,i}$ 将其划到集合 S 或集合 $V \setminus S$ (如以0或中值为界)



本讲小结



拉普拉斯矩阵的特征值与特征向量



谱嵌入与谱聚类的应用

主要参考资料

Tim Roughgarden and Gregory Valiant <CS 168 - The Modern Algorithmic Toolbox> Lecture Notes

Cameron Musco <COMPSCI 514 - Algorithms for Data Science> Slides

Christopher Musco <NYU CS-GY 6763 - Algorithmic Machine Learning and Data Science> Slides

Xianyi Zeng <UTEP MATH 5330 - Computational Methods of Linear Algebra> Lecture Notes

Wing-Kin Ma <CUHK ENGG 5781 - Matrix Analysis and Computations> Lecture Notes

WW Zachary <An Information Flow Model for Conflict and Fission in Small Groups> Paper

Wikipedia <Zachary's karate club> Page

Jianjun Cheng et. al, <A Divisive Spectral Method for Network Community Detection> Paper

Enhong Chen and Linli Xu <Machine Learning and Knowledge Discovery - Spectral Clustering> Slides

谢谢!

