图谱论导引(第十二期)

> 后两至三期文章将着重介绍图谱与网络. 实际上, 读者不必寻找图谱论导引原书, 笔者一开始便没有循规蹈矩地直接参考之.

$\lambda_{\min} \geq -2$图拾遗

上期文章证明了所有$\lambda_{\min} \geq -2$图之表示: 若$\lambda_{\min}(G) \geq -2$, 则$G$为广义线图(有$D_n$表示或$A_n$表示)或例外图(有$E_8$表示). 本小节大致探讨$\lambda_{\min} = -2$之重数, 以作扫尾.

对$\lambda_{\min} = -2$之线图$L(H)$, x为其特征向量若且仅若$Bx = 0$, 其中$B_{|V|\times|E|}$为描述点边相连关系的导出矩阵(incidence matrix). 证明容易, 注意到

$$A(L(H))x = -2x \Leftrightarrow B^T Bx = \mathbf{0} \Leftrightarrow \|Bx\| = 0.$$

即可. 欲研究$\ker B$, $\operatorname{rank}(B)$不得不求. 不妨设G连通, 记$B_1, B_2, \ldots, B_n$为B之横行, 若存在一组非零数组$(c_1, \ldots, c_n)$使得$\sum_i c_i B_i = \mathbf{0}$, 则$i \sim j$时$c_i + c_j = 0$. 从而当$G$为二部图时, $\operatorname{rank}(B) = n - 1$, 反之$B$满秩. 综上, $L(H)$特征值-2之重数为

$$m_{L(H)}(-2) = \begin{cases} |E| - |V| + 1, & \hat{H} \text{ is bipartite,} \\ |E| - |V|, & \hat{H} \text{ is non-bipartite.} \end{cases}.$$

据此有推论: $L(H) > 2$若且仅若H为树或由一个奇圈在顶点上添树而导出的图. 其最小特征值满足

$$-2 \leq \lambda_{\min}(L(H)) \leq \lambda_{\min}(P_d) = -2\cos\frac{\pi}{d+1}.$$

其中, 直径(diameter)d表示图中两点距离的最大值.

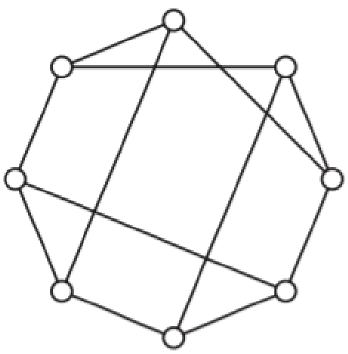
对广义线图$H(a_1, \ldots, a_n)$可同理求得$\operatorname{rank}(B) = n + \sum_{i=1}^n a_i$, 其中要求$a_i$不全为零. 故$m_{L(\hat{H})}(-2) = m - n + \sum_{i=1}^n a_i$.

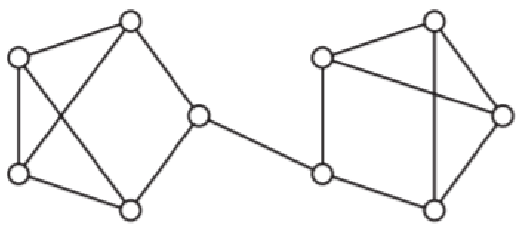
经上述分析, 一切满足$\lambda_{\min} > -2$之图无非以下两类:

- $L(H)$形式. 其中H或为树, 或为添上一片花瓣的树, 或为一个奇圈上添加树所得的图.
- 573种例外图(E_8表示).

λ_2随记

有如此一个有趣现象: 当正则图的顶点数$n \leq 14$时(即n较小时), 对度为3的正则图取λ_2, 则当λ_2较小时图较为通达(well-connected, 即直径较小, 割边更少), λ_2较大时图较为狭长(即直径较大, 割边较少). 如下图所示, 左图之λ_2为1, 右图之λ_2约为2.7785.


8.


9.

下文将着重研究λ_2与图结构之关系.

λ_2与$\bar{d}$

设正则图G顶点数为n, 度为r, 则记其删点图的导出子图为H, 则

$$\bar{d} := \frac{2|E(H)|}{|V(H)|} \le r\frac{\lambda_2^2 + \lambda_2(n-r)}{\lambda_2(n-1)+r}.$$

现选定顶点v以构造删点图: 先将顶点分为三类, 分别为v, v之邻点, 及其余者. 由于邻接矩阵可自然依照3×3之规制分块, 现将每小分块中每行总和之平均值作为元素, 则

$$A \implies B = \begin{pmatrix} 0 & r & 0 \\ 1 & r-\nu-1 & \nu \\ 0 & r-\bar{d} & \bar{d} \end{pmatrix}.$$

其中, ν为由各个v邻点向非v零点引出边数量之平均值, 即

$$\frac{\operatorname{cut}(N(v), |V| - (N(v)\cup\{v\}))}{|N(v)|}.$$

B有特征多项式

$$P_B(x) = (x-r)(x^2 - (\bar{d}-\nu-1)x - \bar{d}).$$

回顾先前所介绍的特征值插值定理: 若$Q_{m\times n}$满足$Q^TQ = I$, A之特征值$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_m$, 则$Q^TAQ$之第$k$大特征值$\mu_k$满足

$$\lambda_{n-m+k} \le \mu_k \le \lambda_k.$$

从而$P_B(\lambda_2) \le 0$, 即

$$\lambda_2^2 - (\bar{d}-\nu-1)\lambda_2 - \bar{d} \le 0.$$

由于$|N(v)|\nu = r\nu = (n-1-r)(r-\bar{d})$, 化简得

$$\bar{d} \le r\frac{\lambda_2^2 + \lambda_2(n-r)}{\lambda_2(n-1)+r}.$$

$\bar{d}$随λ_2之减小而减小.

δ上界估算

以简便故, 下统一记λ'为$1-\lambda_2$. 定义正则图的悠哉游哉(lazy ramdom walk)矩阵为$M := \dfrac{A}{2r} + \dfrac{I}{2}$, 即次走动有$\dfrac{1}{2}$的概率保持原位(谓之悠哉), 有$\dfrac{1}{2}$的概率等概率移动至邻点(谓之游哉). 记$u$, v为图中相聚较远的两个点, 设$p_t(v)$为u处出发点在时间段t后到达v之概率. 回顾Markov链模型, 正则图之稳态对应

$$\pi(v) = \frac{\deg v}{\sum_u \deg u} = 1/n.$$

若$|p_t(v) - \pi(v)| < 1/n$, 则$p_t(v) > 0$, 从而$\delta \le t$. 下将着重探究t与λ之关联.

定义l^2范数$\|p_t - \pi\|_2 := \sqrt{\sum_u [p_t(u) - \pi(u)]^2}$. 如上文所记, p_0为初始分布函数, p_t为t时间段后分布函数. 设v_i为矩阵M中μ_i对应的特征向量, 则

$$M^k \cdot p_0 = \sum_i v_i^T p_0 \cdot \mu_i^k \cdot v_i = \frac{v_1}{n} + \sum_{i\ge 2} v_i^T p_0 \cdot \mu_i^k \cdot v_i$$

从而

$$\begin{aligned}\|p_t-\pi\|_2^2 &= \sum_u [p_t(u)-\pi(u)]^2 \\ &= \|\sum_{i\geq 2} v_i^T p_0 \cdot \mu_i^t \cdot v_i\|_2^2 \\ &= \sum_{i\geq 2} (v_i^T p_0)^2 \mu_i^{2t} \\ &\leq \mu_2^t \sum_{i=1}^n (v_i^T p_0)^2 \\ &\leq \mu_2^t = (1-\lambda)^t\end{aligned}.$$

令$t=\dfrac{\ln n}{\lambda}$, 则

$$\begin{aligned}|p_t(v)-\pi(v)| \leq& (1-\lambda)^t \sqrt{\frac{\deg v}{\min_u \deg u}} \\ =&(1-\lambda)^{\ln n/\lambda} \\ <&(1/e)^{\ln n} \\ <&\frac{1}{n}\end{aligned}.$$

从而$\delta \leq \dfrac{\ln n}{\lambda}$.

偏个题: 图第二大特征值与聚类算法

倘若一类对象的彼此关系是对称的, 则其间关系可以用加权的简单图表示, 不妨设为$G(V,E,W)$, 其中W为权重. 若将图中结点分为A与B两部分, 记权重和$W(A,B):=\sum_{i\in A, j\in B} w_{ij}$为其间的总关联度.

聚类的本质较为通俗: 将节点分为若干部分, 使得各部分间的关联度较小. 不妨记

$$\mathrm{cut}(A_1,\ldots,A_n) := \frac{1}{2}\sum_{i\neq j} W(A_i,A_j) = \frac{1}{2}\sum_i W(A_i,\overline{A_i})$$

为关联度总和, 聚类的目的之一是杜绝$\mathrm{cut}(A_1,\ldots,A_n)$过大.

虽说在V中找到度最小之点即可保证$\mathrm{cut}(A_1,\ldots,A_n)$最小, 但实际上, 如此的划分方式并未给聚类带来实质性的帮助; 假若兼顾每一划分所得部分的权重和或结点数量, 应当能构造出较好的聚类指标. 常用的指标包括数目-划分比与权重-划分比, 分别定义作

$$\begin{aligned}\mathrm{RatioCut}(A_1,\ldots,A_k) :=& \frac{1}{2}\sum_{i=1}^k \frac{W(A_i,\overline{A_i})}{|A_i|}, \qquad |A_i| \text{ is the number of vertices in } A_i, \\ \mathrm{NCut}(A_1,\ldots,A_n) :=& \frac{1}{2}\sum_{i=1}^k \frac{W(A_i,\overline{A_i})}{\mathrm{vol}(A_i)}, \qquad \mathrm{vol}(A_i) \text{ is the weight of } A_i.\end{aligned}$$

记Laplace矩阵$L:=A-D$. 下仅讨论数目-划分比对应之情形. 记$n:=|V|$维向量$h^1,\ldots,h^i,\ldots,h^k$分别由$A_1,\ldots,A_i,\ldots,A_k$决定, 其中$h^i=(h_1^i,\ldots,h_n^i)$,

$$h_j^i = \begin{cases}\dfrac{1}{\sqrt{|A_i|}}, \text{ if } j\in A_i, \\ 0, \text{ else.}\end{cases}, \quad i=1,\ldots,k; j=1,\ldots,n.$$

从而

$$(h^i)^T L h^i = -\frac{1}{2}\sum_{s,t=1}^{n} w_{st}(h_s^i - h_t^i)^2 = -\frac{1}{2}\frac{W(A_i, \overline{A_i})}{|A_i|}.$$

聚类等价于求解优化问题

$$\min_{H\in\mathbb{R}^{k\times n}} -\mathrm{tr}(HLH'), \quad \text{subject to } HH' = I_k.$$

根据Rayleigh-Ritz's定理, $\{h^1,,\ldots,h^k\}$应尽量靠近L的前n个特征向量, 其中h^1为$\frac{1}{\sqrt{n}}\mathbf{1}$应予舍去. 特殊地, $k=2$时, 按照λ$_2$对应的特征向量x中各分量的正负关系对结点进行分类即可.