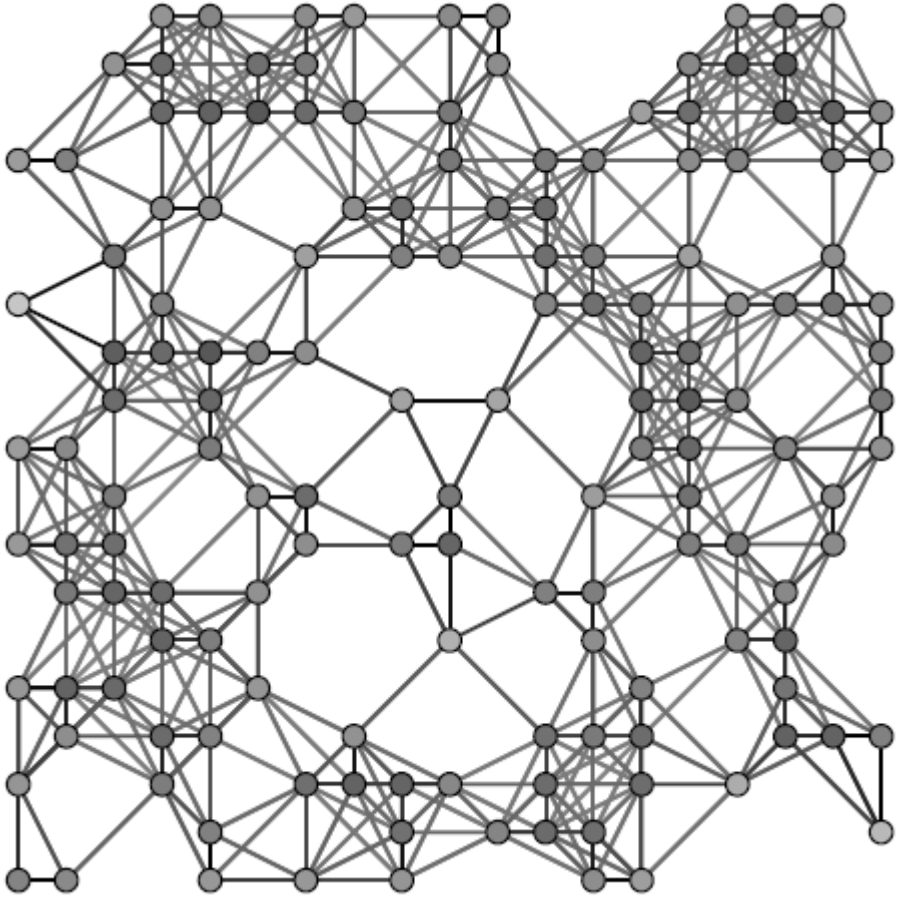
图聚类(Graph Clustering)

图聚类不是图像聚类或者图片聚类，也不是对图像中的像素点进行聚类！

图聚类的必要性

- 数据并不总是能以向量的形式表示。
- 对人的聚类，如果我们希望按照身高、体重、收入等特征进行聚类，可以将一个人表示成向量的形式。
- 但如果我们希望根据社会关系进行聚类，就不行了。
- 比如张三，是王五的好朋友，刚认识李四，对赵六很是反感，李四又跟钱七关系比较好，跟刘八是亲戚等等。这些信息就是一个社交网络。如果以图像的形式表示出来应该是如下形式。

社交网

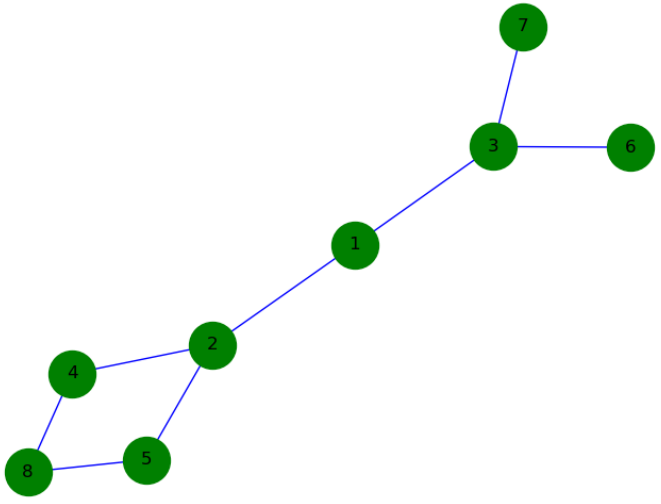


图中并没有坐标，顶点的位置和线的长短也不代表任何信息(拓扑)，两个顶点之间有线相连就说明有关，反之表示无关。

第一部分 图数据

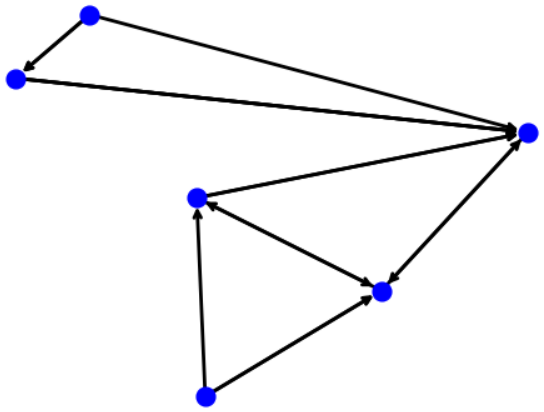
有关定义

- 图(Graph)是一种数学结构，一般用G表示。
- G=(V,E),V表示vertices/nodes，E表示edges，两个顶点(也叫节点)v_i,v_j之间的边用e=(v_i,v_j)表示，称e附着于v_i和v_j， 称v_i和v_j是邻接(adjacent)的，或者称其为邻居。
- 顶点到其自身的边(v_i,v_i)称作自环(loop)，没有自环的无向图称为简单图(simple graph)。
- 一个图顶点的数目称为阶 (order)。
- 边的数目称为图的尺寸 (size)。



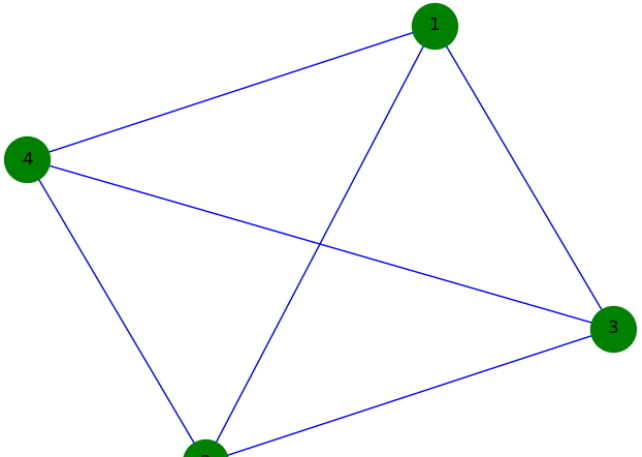
有关定义

- 有序顶点对构成的图称为有向图(directed graph)。一条有向边(v_i,v_j)也称作一条从v_i到v_j的弧(arc)。
- 称v_i为弧的尾顶点(tail)， v_j为弧的头顶点(head)。
- 如果图中每条边的重要程度有差异，则称为 带权图(weighed graph)，没有差异的图可以看成权值均为1的带权图。注意：权值与边的长短无关。



有关定义

- 如果一个图中所有的顶点对之间都有边，则称其为完全图(complete graph)或团(clique)。
- 一个顶点的度(degree) 是与之相连的边的数目，表示为$d(v_i)$或者d_i。
- 一个图的度序列(degree sequence)是所有顶点的度以非增的顺序排列的列表。



有关定义

- 图的度频率分布(degree frequency distribution)为

  $(N_0, N_1, N_2, \cdots, N_t)$

  其中，t表示图中各顶点的最大度数，N_k 表示度为k的节点数目
- 图的度频率分布给出了各个度数出现的次数，其概率可以用图的度数分布(degree distribution) 表示。图的度数分布表示为：

  $(f(0), f(1), \cdots, f(t))$

  其中，$f(k)=\frac{N_k}{n}$，n表示总的顶点 数目(即图的阶)

有关概念

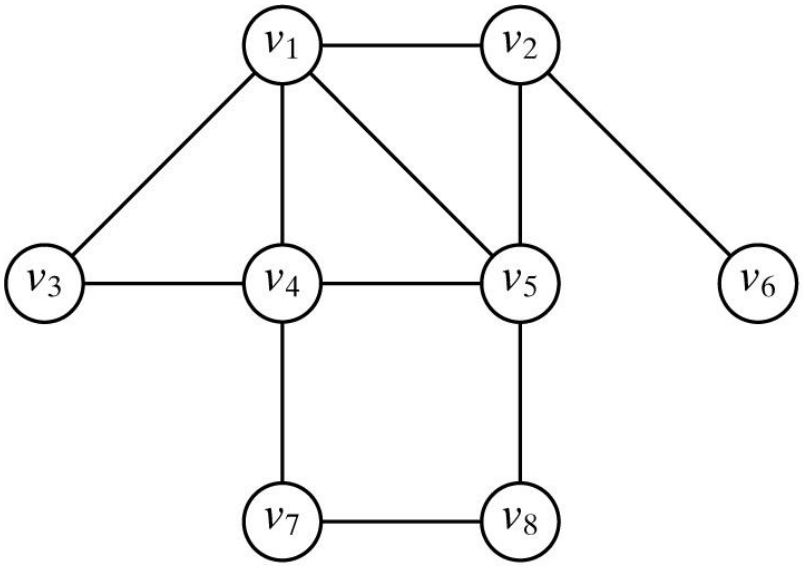
- 在有向图中，以顶点v_i为头顶点的边的数目，称为该点的入度(indegree)，表示为$id(v_i)$，以顶点v_i为尾顶点的边的数目，称为该点的出度(outdegree)，表示为$od(v_i)$。
- 两个顶点之间的最短路径称为最短路(shortest path)，最短路的长度称作这两个顶点之间的距离，记作$d(v_i,v_j)$，如果两个顶点之间不存在路径，则其距离为正无穷。

有关概念

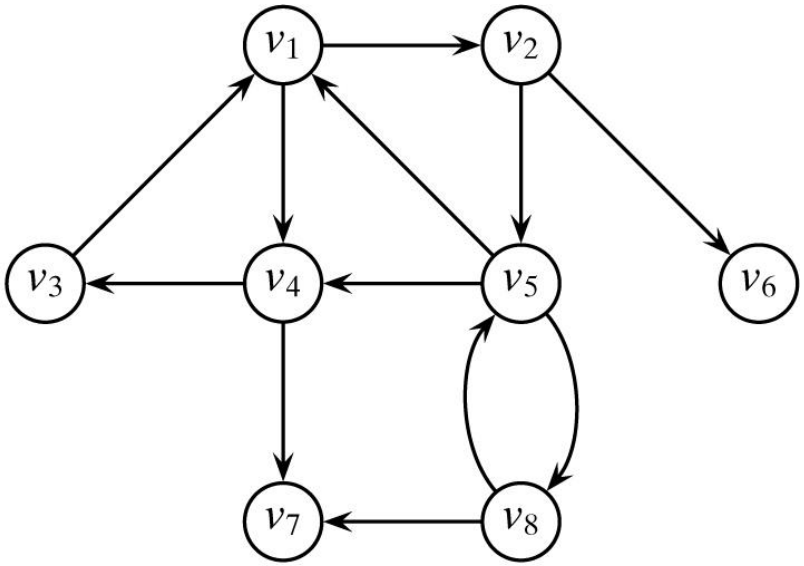
- 两个顶点之间有一条路径，则称这两个顶点是连通的(connected)，如果一个无向图的所有顶点之间都有路径，则称这个图为连通的。
- 对于有向图，如果所有有序顶点对之间都存在一条有向路径，则称之为强连通(strongly connected)的，如果只有将边看作是无向时才存在路径，则称该图是弱连通的(weakly connected)

例 1

- 图(a)的阶为8，尺寸为11，与v_1相连的边共有4条，所以v_1的度为4，整个图的度序列为(4,4,4,3,2,2,2,1)。度频率分布为(0,1,3,1,3)。度分布为(0, 0.125, 0.375, 0.125, 0.375)。



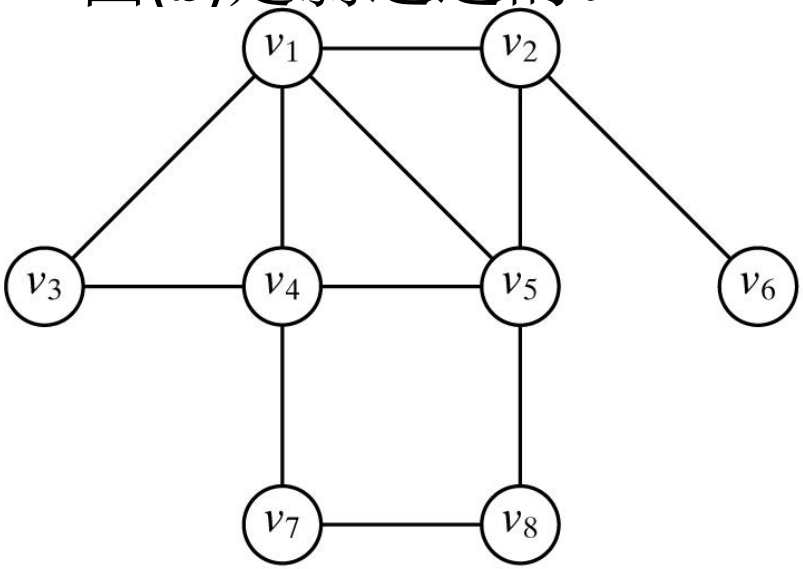
(a)



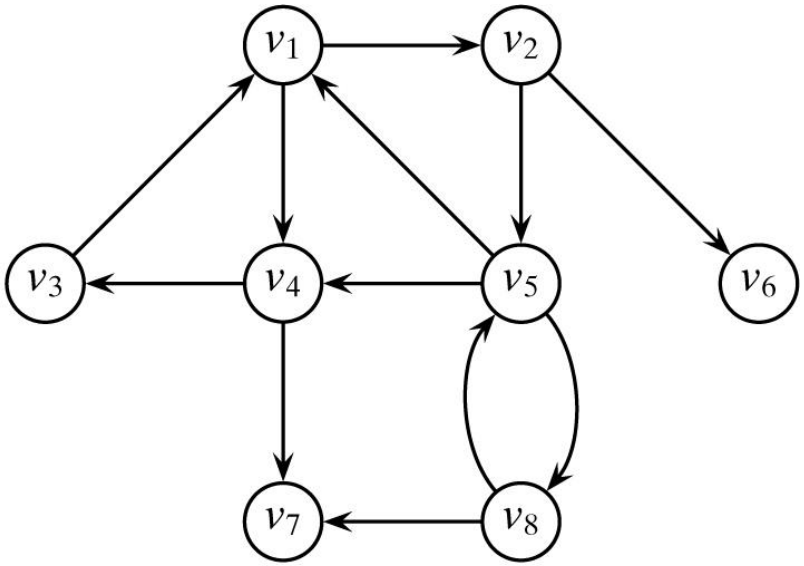
(b)

例 1

- 图(b)是一有向图。 v_7的出度为0，入度为2， (v_5,v_8) (v_8,v_5)这两条边是不同的。
- V_5到v_6的距离为3。
- V_7到v_6的距离为正无穷。
- 图(b)是弱连通的。



(a)



(b)

邻接矩阵

- 一个包含n个顶点的图可以表示为一个n×n的二值矩阵，该矩阵称为邻接矩阵(adjacency matrix)，一般以A表示。其中

$$A(i,j)=\begin{cases}1 & \text{若}v_i\text{和}v_j\text{邻接}\\ 0 & \text{否则}\end{cases}$$

- 如果图是无向的，则该矩阵为对称矩阵，如果图是有向的，则该矩阵一般为非对称矩阵。

邻接矩阵

- 如果图是带权图，则该矩阵可以表示为：

$$A(i,j)=\begin{cases} w_{ij} & \text{若}v_i\text{和}v_j\text{邻接} \\ 0 & \text{否则} \end{cases}$$

其中w_{ij}表示这两个顶点之间的边的权值

- 通过设置一个阈值，超过该阈值时，权值为1，否则为0，可以将任何一个带权邻接矩阵转化为一个二值矩阵。

向量数据转换为图数据

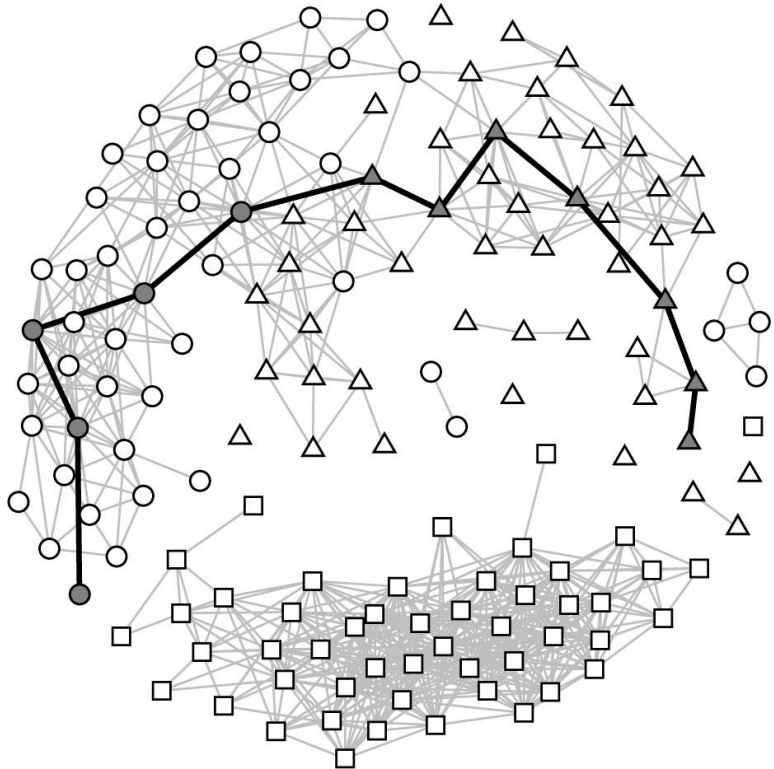
- 可以将以向量形式出现的数据集转换为图的形式。
- 对于d维空间中的n个数据点组成的数据集D，可以定义一个带权图G=(V,E)，该图中每个顶点对应一个原来的数据点，数据点之间的边表示两个数据点之间的相似度。相似度可以通过以下方式计算：

$$w_{ij} = sim\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$$

σ是一个可以调节的参数，具体内容参考SVM

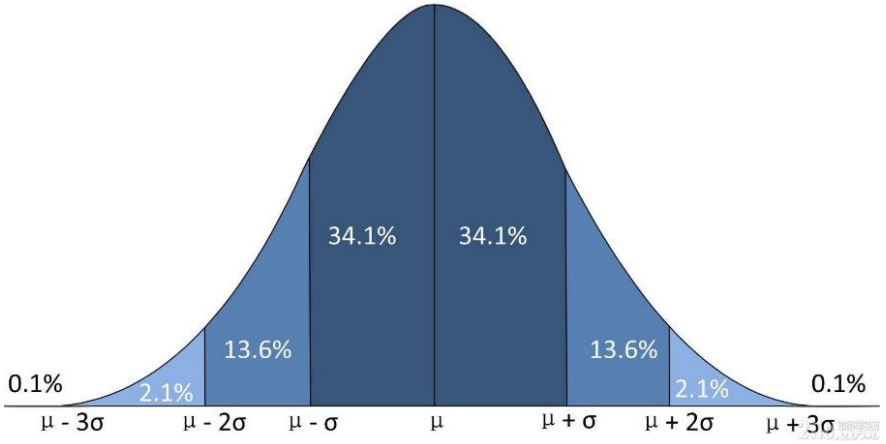
鸢尾花数据集的图表示

- 令$\sigma = 1/\sqrt{2}$，通过计算可以得到数据集各点间相似度的均值为0.197，标准差为0.290，将阈值设置为0.197+2×0.290=0.777，使得带权邻接矩阵转换为一个二值矩阵，任意两个相似度大于0.777的顶点之间都有一条边，最终形成的相似度图如图所示。



两倍标准差的意义

- 对于正态分布，处于正负两个标准差之外的数据大约占总数据的5%，处于正负三个标准差之外的数据大约占总数据的0.3%，所以依据《计数抽样检验程序》(GB2828)、《正态样本异常值的判断和处理》(GB4883)，处于两个标准差之外的值称为异常值，三个标准差之外的称为高度异常值。



聚类系数(clustering coefficient)

- 一个顶点的聚类系数是对该顶点相邻边数目的度量。其定义为：

  令$G_i=(V_i,E_i)$是由顶点v_i的邻居导出的子图，如果v_i的邻居的数目为n_i，v_i邻居所附着的边的数目(即G_i中边的数目)为m_i，则聚类系数可以表示为：

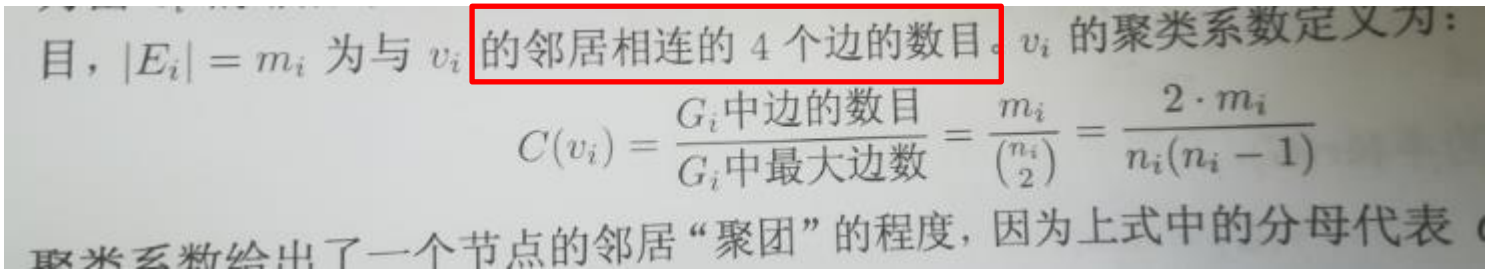
$$C(v_i)=\frac{G_i\text{中边的数目}}{G_i\text{中可以有的最大边的数目}}=\frac{m_i}{C_{n_i}^2}=\frac{2}{n_i(n_i-1)}$$

注意

- 在图书馆借阅《数据挖据与分析：概念与算法》中文版的，此书中文翻译有一定的错误，一定要结合英文版去看(英文版在qq群文件中有)。

Clustering Coefficient The *clustering coefficient* of a node v_i is a measure of the density of edges in the neighborhood of v_i. Let $G_i = (V_i, E_i)$ be the subgraph induced by the neighbors of vertex v_i. Note that $v_i \notin V_i$, since we assume that G is simple. Let $|V_i| = n_i$ be the number of neighbors of v_i, and $|E_i| = m_i$ be the number of edges among the neighbors of v_i. The clustering coefficient of v_i is defined as

$$C(v_i) = \frac{\#\text{ of edges in } G_i}{\text{maximum number of edges in } G_i} = \frac{m_i}{\binom{n_i}{2}} = \frac{2 \cdot m_i}{n_i(n_i - 1)}$$



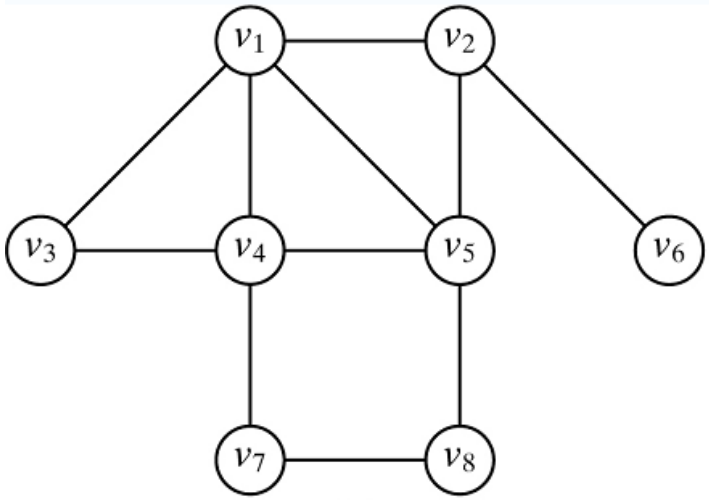
聚类系数

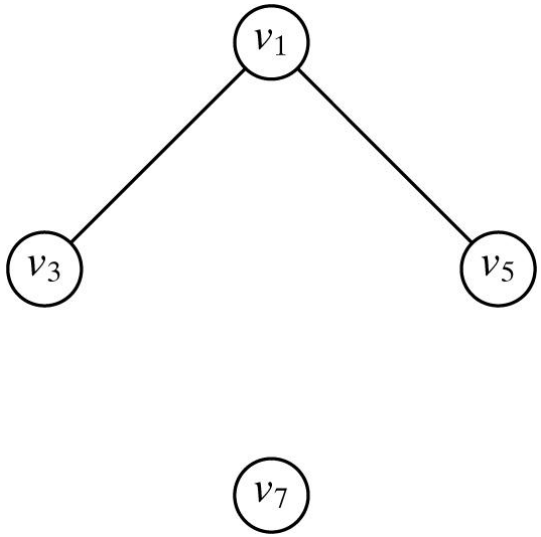
- 整个图的聚类系数是该图中所有点的聚类系数的平均值。
- 聚类系数对于度数大于等于2的顶点才有意义，小于2的顶点聚类系数为0。
- 一个顶点对的效率(efficiency)为距离的倒数，如果这两个顶点不连通，则距离为无穷大，效率为0。节点间的距离越小，其通信越有效，效率越高。
- 整个图的效率是所有节点对的效率的平均值。

例 2

- 下面左图v_4中的邻居导出的子图如右图所示。
- v_4的聚类系数为 $C(v_4)=\frac{2}{C_4^2}=0.33$
- 整个图上的聚类系数为：

$$C(G)=\frac{1}{8}\left(\frac{1}{2}+\frac{1}{3}+1+\frac{1}{3}+\frac{1}{3}+0+0+0\right)=0.3125$$





聚类系数

- 一个顶点的聚类系数是对该顶点相邻边数目的度量。这与聚类有什么关系？
- 如果一个子图由(v_i,v_j) 和(v_i,v_k)构成，则称其为以v_i为中心的连通三元组(connected triple);如果该三元组包含边(v_j,v_k) ，则称其为三角形(triangle)。
- 如果一个顶点有n个邻居，则以该顶点为中心的连通三元组共有 C_n^2 个，且包含该顶点的三角形的数目等于由该顶点导出的子图中边的数目。

聚类系数

- 所以聚类系数也可以表示为：

$$C(v_i) = \frac{\text{包含}v_i\text{的三角形的数目}}{\text{以}v_i\text{为中心的连通三元组的数目}}$$

- 以社交网络为例，如果一个人的聚类系数高说明这个人的朋友之间也互为朋友的概率比较高，那这些人就可以归为一个簇。

中心度分析

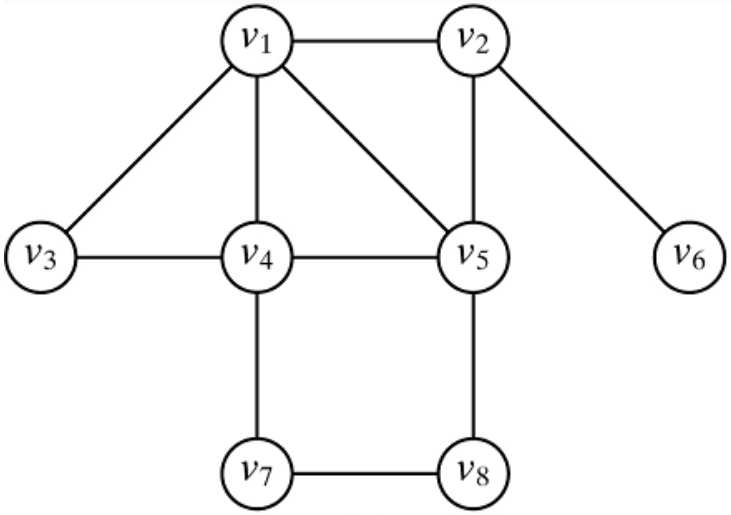
- 聚类时各样本的重要程度是不一样的，比如K-medoids算法中密集处的样本更容易成为质心点。
- 在图中，各个顶点的重要程度也是不一样的，可以用中心度(centrality)来表示顶点的居中度或者重要性。
- 基本中心度包括度数中心度、离心率中心度、接近中心度和介数中心度等。

基本中心度

- 度数中心度(degree centrality)是中心度中最简单的一种，该点的度数就是它的度数中心度。一个点的度数越高，说明与该点有关系的顶点越多，也就是说这个点越重要。如果是有向图，可以进一步考虑其入度中心度和出度中心度。
- 离心率中心度(eccentricity centrality)指一个顶点的离心率越小，该顶点越重要，离心率中心度可以定义为离心率的倒数。

离心率

- 一个顶点的离心率指从该顶点到整个图中所有顶点的最大距离(不包含非连通点)。比如下图中v_4的离心率是3。离心率最小的点称为中心节点(center node)，离心率最大的点称为边缘节点(periphery node)。



接近中心度(closeness centrality)

- 接近中心度用一个顶点的所有距离值的和的倒数来表示该顶点的重要性。总距离最小的点称为中位点(median node)。

$$c(v_1) = \frac{1}{\sum_j d(v_i, v_j)}$$

- 接近中心度为设施选址问题提出了一个目标函数，对于那些不关心离群点的选址问题(比如超市选址问题)就可以选择离所有人距离最近，即接近中心度最高的点。

介数中心度(betweenness centrality)

- 某个顶点的介数中心度衡量所有顶点对之间的最短路径中包含该顶点的比率。其计算方法为：

  令η_{jk}表示顶点v_j和v_k之间的最短路径的数量，

  $\eta_{jk}(v_i)$表示这些路径中包含v_i的数量。则有

  $$\gamma_{jk}(v_i)=\frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

  如果顶点v_j和v_k不连通，则$\gamma_{jk}(v_i)=0$

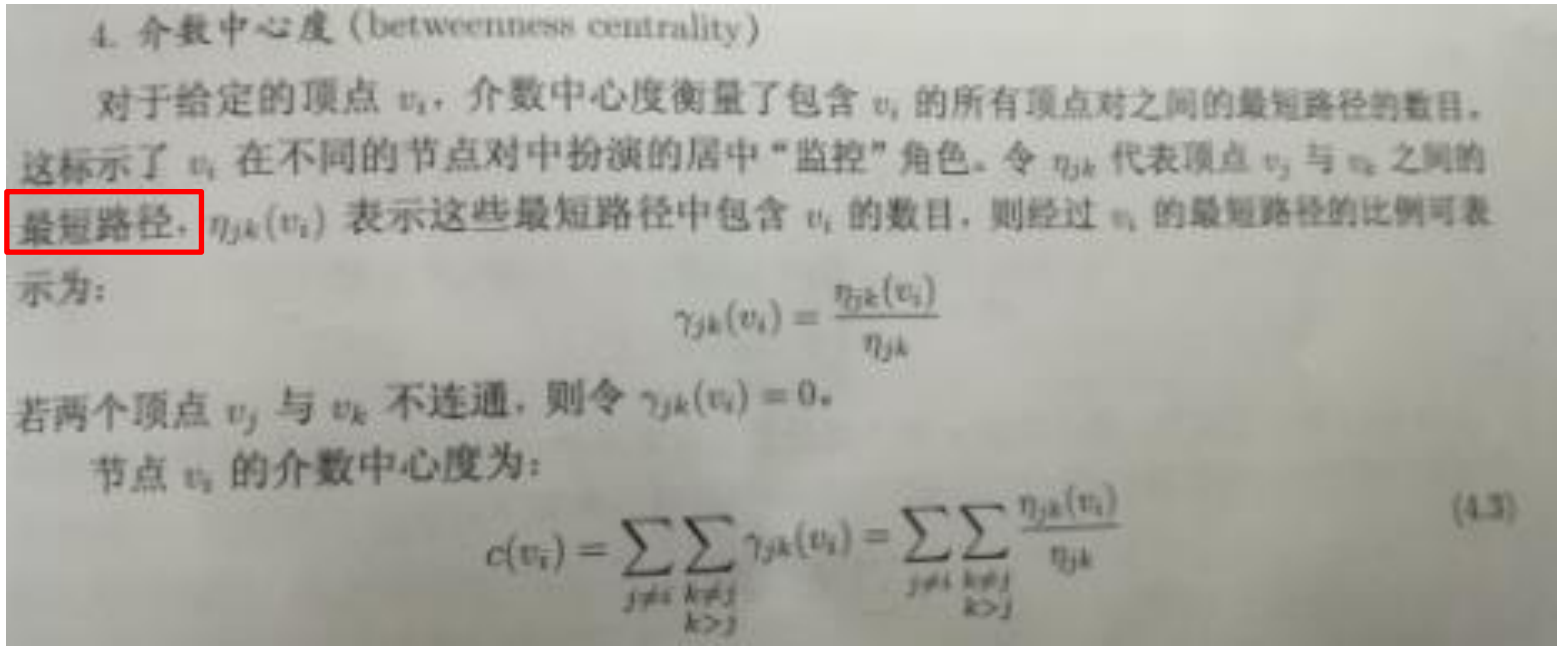
  节点v_i的介数中心度为：

  $$c(v_i)=\sum_{\substack{j\neq i\\ k\neq i\\ k>j}}\sum\gamma_{jk}(v_i)=\sum_{\substack{j\neq i\\ k\neq i\\ k>j}}\sum\frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

注意

- 翻译错误

Let η_{jk} denote the number of shortest paths between vertices v_j and v_k, and let $\eta_{jk}(v_i)$ denote the number of such paths that include or contain v_i, then the fraction of paths through v_i is denoted as



4. 介数中心度 (betweenness centrality)

对于给定的顶点 v_i，介数中心度衡量了包含 v_i 的所有顶点对之间的最短路径的数目。这标示了 v_i 在不同的节点对中扮演的居中“监控”角色。令 η_{jk} 代表顶点 v_j 与 v_k 之间的最短路径，$\eta_{jk}(v_i)$ 表示这些最短路径中包含 v_i 的数目，则经过 v_i 的最短路径的比例可表示为：

$$\gamma_{jk}(v_i) = \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

若两个顶点 v_j 与 v_k 不连通，则令 $\gamma_{jk}(v_i) = 0$。

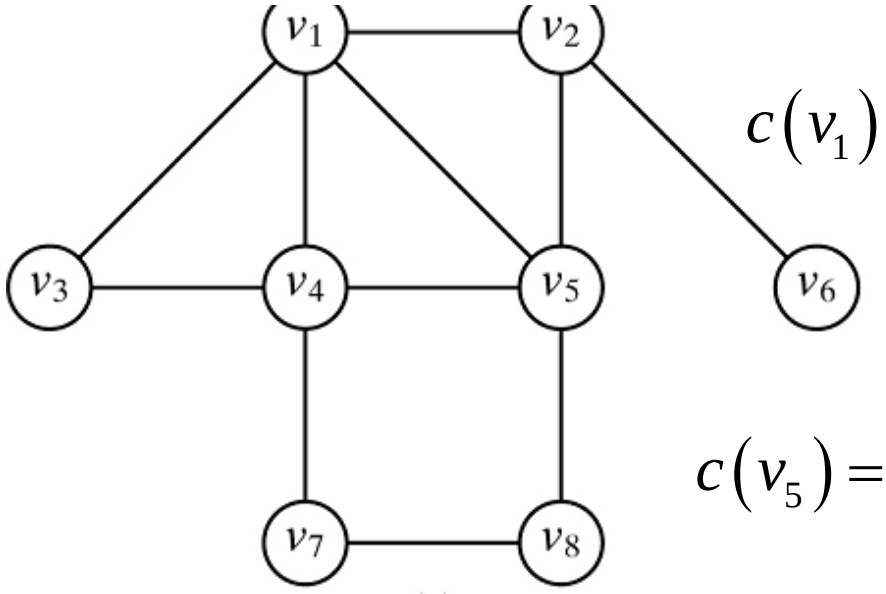
节点 v_i 的介数中心度为：

$$c(v_i) = \sum_{j \neq i} \sum_{\substack{k \neq j \\ k > j}} \gamma_{jk}(v_i) = \sum_{j \neq i} \sum_{\substack{k \neq j \\ k > j}} \frac{\eta_{jk}(v_i)}{\eta_{jk}} \qquad (4.3)$$

例 3

- 对下图求各种中心度可以得到：

Centrality	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Degree	4	3	2	4	4	1	2	2
Eccentricity	0.5	0.33	0.33	0.33	0.5	0.25	0.25	0.33
$e(v_i)$	2	3	3	3	2	4	4	3
Closeness	0.100	0.083	0.071	0.091	0.100	0.056	0.067	0.071
$\sum_j d(v_i, v_j)$	10	12	14	11	10	18	15	14
Betweenness	4.5	6	0	5	6.5	0	0.83	1.17



$$c(v_1) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{2} + 1 + \frac{1}{3} + \frac{1}{2} + \frac{1}{3} = 4.5$$

$$c(v_5) = 1 + \frac{1}{2} + \frac{2}{3} + 1 + \frac{2}{3} + \frac{1}{2} + \frac{1}{2} + \frac{2}{3} + 1 = 6.5$$

Web中心度

- 以上中心度主要针对无向图，对于有向图，尤其是在Web场景中，例如超链接文档包含指向其它文档的有向链接，科学文献的引用网络包含从一篇论文到被引用论文的有向边等，这些问题需要适合Web规模的图的中心度来表示。
- Web中心度包括声望、网页排名、hub和权威分数等。

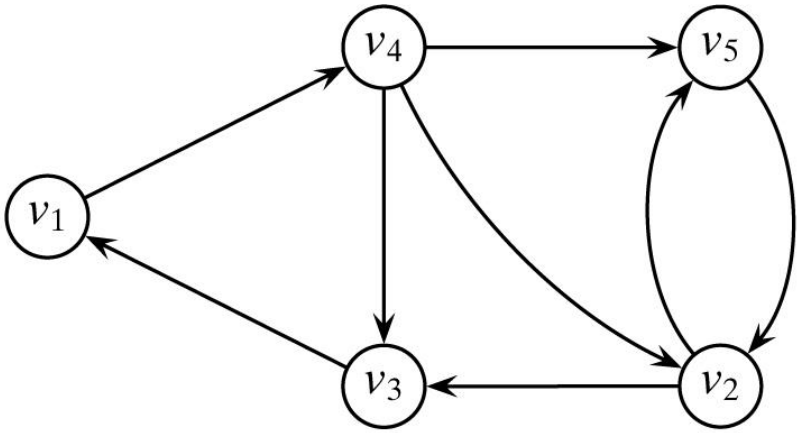
声望(Prestige)

- 度数中心度通过一个顶点的度数来衡量其重要性。在有向图中，度数又要分为入度和出度，如果考虑入度的话，则一个顶点的入度越高，声望越高。但一个顶点的重要程度不仅与有多少条边指向该点有关，还与指向它的顶点的声望有关(即与指向它的顶点的入度有关，这一般是一个递归的过程)。
- 令G=(V,E)为一个有向图，该图有n个顶点，则该图的邻接矩阵是一个n×n的不对称矩阵。

$$A(u,v)=\begin{cases}1 & (u,v)\in E\\ 0 & (u,v)\notin E\end{cases}$$

声望

- 对于下图，有



(a)

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

(b)

$$\mathbf{A}^T = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

(c)

声望

- 将顶点u的声望分数记为p(u)，一个节点的声望与指向它的节点的声望有关，所以可以定义一个节点v的声望分数为：

$$p(v)=\sum_{u} A(u,v)\bullet p(u)=\sum_{u} A^{T}(v,u)\bullet p(u)$$

例如，节点v_3的声望分数为

$$p(v_3)=\sum_{u} A(u,v_3)\bullet p(u)=A(v_2,v_3)\bullet p(v_2)+A(v_4,v_3)\bullet p(v_4$$

$$=A^{T}(v_3,v_2)\bullet p(v_2)+A^{T}(v_3,v_4)\bullet p(v_4)$$

$$=\sum_{u} A^{T}(v_3,u)\bullet p(u)$$

声望

$$p(v_3)=\sum_u A^T(v_3,u)\bullet p(u)$$

$$即 p(v_3)=A_3^T\bullet\begin{pmatrix}p(v_1)\\p(v_2)\\p(v_3)\\p(v_4)\\p(v_5)\end{pmatrix}$$

其中，A_3^T表示A^T的第三行

声望

一个顶点的声望分数会与指向其的顶点的声望分数有关，而指向其的顶点的声望由于更前一层的顶点的声望分数有关，所以一个顶点的声望分数最终会递归的收到自己的声望分数的影响，所以上式中，等号两边都有$p(v_3)$，为了方便将左侧v_3的声望分数表示为$p'(v_3)$,则

$$\begin{pmatrix} p'(v_1) \\ p'(v_2) \\ p'(v_3) \\ p'(v_4) \\ p'(v_5) \end{pmatrix} = \begin{pmatrix} A_1^T \\ A_2^T \\ A_3^T \\ A_4^T \\ A_5^T \end{pmatrix} \begin{pmatrix} p(v_1) \\ p(v_2) \\ p(v_3) \\ p(v_4) \\ p(v_5) \end{pmatrix} = A^T \begin{pmatrix} p(v_1) \\ p(v_2) \\ p(v_3) \\ p(v_4) \\ p(v_5) \end{pmatrix}$$

声望

- 上一页的公式可以简写为：

  $\overrightarrow{p}' = A^T \overrightarrow{p}$

  $\overrightarrow{p}'$和$\overrightarrow{p}$都是一个列向量，对应每个顶点的声望分数。

- 或者可以写成：

$$\overrightarrow{p_k} = A^T \overrightarrow{p_{k-1}} = A^T \left(A^T \overrightarrow{p_{k-2}} \right) = \cdots = \left(A^T \right)^k \overrightarrow{p_0}$$

  其中p_0表示初始声望向量。

思考

- 初始声望向量怎么确定？

幂方法

设n 阶矩阵B 的特征值按绝对值从大到小排列分别为$\lambda_1,\lambda_2,\cdots,\lambda_n$,对应的线性无关的特征向量分别为$x_1,x_2,\cdots,x_n$,则这些向量可以构成$n$维空间的一组基,该空间的任一向量$p_0$可以表示为

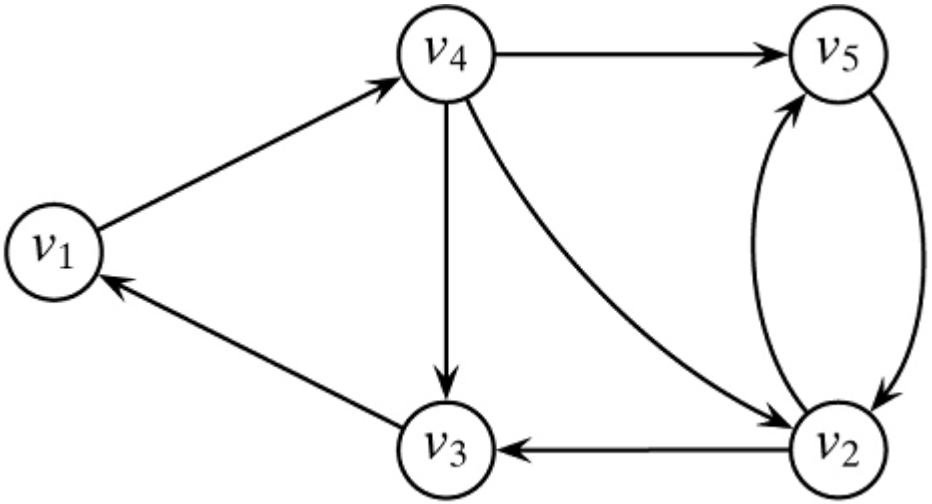
$$p_0 = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

$$p_k = Bp_{k-1}B^k p_0 = \lambda_1^k\left(a_1x_1 + \left(\frac{\lambda_2}{\lambda_1}\right)a_2x_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)a_nx_n\right) \approx \lambda_1^k a_1x_1$$

即经过多次迭代之后，无论原来的p_0是什么，p_{k-1}总能收敛到矩阵B 的主特征向量，即p_k总能收敛到矩阵B 的主特征向量。

例 4

- 如图所示，假设初始声望向量为$(1,1,1,1,1)^{\mathrm{T}}$，$A^T p_0 = (1,2,2,1,2)^T$，为防止数值溢出，将该向量的每个分量除以2 (向量中的最大分量) ，得到$p_1 = (0.5,1,1,0.5,1)^T$，令A^T乘以p_1得到p_2，然后$p_3, \cdots$



例 4

- 经过10次迭代之后，向量收敛，规范化后得到主特征向量为$(0.356, 0.521, 0.521, 0.243, 0.521)^T$。
- 从中可以看出v_2,v_3,v_5顶点的声望分数最高，v_4顶点的声望分数最低。
- 在原图中第v_2,v_3,v_5顶点入度均为2， v_1和v_4顶点虽然入度都为1，但指向v_4的顶点为v_1 ，指向v_1的顶点为v_3， v_3的声望高于v_1，即指向v_1的顶点声望高于指向v_4的顶点的声望，所以v_1的声望分数较高。

网页排名(PageRank)

- 网页排名是一种在Web环境下计算声望或节点中心度的方法，也叫网页级别、Google左侧排名或佩奇排名。Google的创始人拉里·佩奇和谢尔盖·布林于1998年在斯坦福大学发明了这项技术。
- PageRank通过网络浩瀚的超链接关系来确定一个页面的等级。Google把从A页面到B页面的链接解释为A页面给B页面投票，Google根据投票来源（甚至来源的来源，即链接到A页面的页面）和投票目标的等级来决定新的等级。简单的说，一个高等级的页面可以使其他低等级页面的等级提升。

网页排名

- 网页排名与声望类似，所不同的地方在于其加入了一个称作“随机冲浪”的机制。
- “随机冲浪(random surfing)”假设一个人在浏览网页的时候会随机的选择一个向外的链接，或者以很小的概率随机跳转到Web图中任意的其他页面。比如一个网页A中有三个超链接B,C,D，用户除了点击这三个超链接外，还有可能在地址栏中输入一个链接跳转到任意的其他界面。

网页排名

- 假设每个顶点的出度最少为1，则一个节点的出度

$$od(u) = \sum_{v} A(u,v)$$

- 如果有一条从u指向v的边，则从u访问v的概率为：

$$\frac{1}{od(u)}$$

- 每个节点的初始概率要满足：

$$\sum_{u} p_0(u) = 1$$

网页排名

- 与声望的更新方式类似：

$$p(v)=\sum_{u}\frac{A(u,v)}{od(u)}\bullet p(u)$$

$$=\sum_{u}N(u,v)\bullet p(u)$$

$$=\sum_{u}N^{T}(v,u)\bullet p(u)$$

- 因为要满足概率和为1，所以对邻接矩阵进行了规范化：

$$N(u,v)=\begin{cases}\dfrac{1}{od(u)} & (u,v)\in E\\ 0 & (u,v)\notin E\end{cases}$$

网页排名

- 对于所有节点，网页排名的表示为(参考声望一节)：

  $\overrightarrow{p}'=N^T\overrightarrow{p}$

  $\overrightarrow{p}'$和$\overrightarrow{p}$都是一个列向量，对应指向该顶点的每个顶点的网页排名。
- 在随机冲浪中，会以一个较小的概率从一个节点跳转到图中的任意其他节点，即使它们之间并不存在边，可以将Web图想象成一个全连接的图，其邻近矩阵为(假设共有n个页面)：

网页排名

$$A_r = 1_{n\times n} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

- 每个顶点的出度为n，从一个顶点跳转到另一个顶点的概率为：

$$\frac{1}{od(u)} = \frac{1}{n}$$

网页排名

- 如果不允许点击网页内的链接，只能随机的从一个页面跳转到另一个页面，则网页排名的计算方法不变：

$$p(v)=\sum_{u}\frac{A_r(u,v)}{od(u)}\bullet p(u)$$

$$=\sum_{u}N_r(u,v)\bullet p(u)$$

$$=\sum_{u}N_r^T(v,u)\bullet p(u)$$

这里的p(v)和p(u)都是数字，不是向量。

- 其中

$$N_r=\frac{1}{n}\times 1_{n\times n}$$

网页排名

- 如果将以上两者都考虑进来，假设从一个页面随机调整到另一个页面的概率为 α，则完整的网页排名计算方式应为：

$$\begin{aligned}\overrightarrow{p}' &= (1-\alpha) N^T p + \alpha N_r^T p \\ &= \left((1-\alpha) N^T + \alpha N_r^T\right) p \\ &= M^T \overrightarrow{p}\end{aligned}$$

- 可以通过声望计算中的幂迭代方法计算 $\overrightarrow{p}'$。

网页排名

- 一个页面可能没有出度，即该页面没有向外的链接，在这种情况下：

$\alpha = 1$, 该点的出度$od(u) = 0$, N矩阵中的每一行代表一个顶点没有向外的链接情况，此时该行全为0 ，令M_u 表示M矩阵的某一行，则

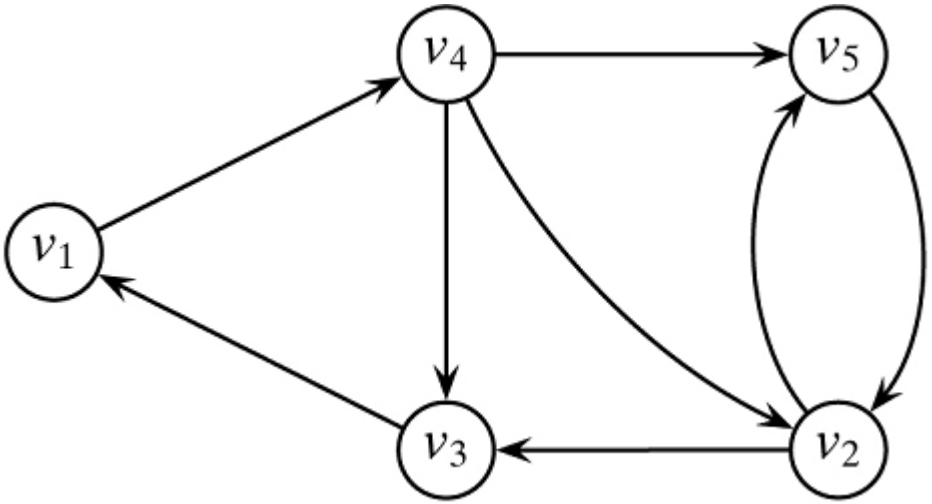
$$M_u = \begin{cases} M_u & od(u) > 0 \\ \frac{1}{n} \times 1_n^T & \mathrm{od}(u) = 0 \end{cases}$$

1_n^T表示全为1 的行向量

例 5

- 考虑如图，其规范化邻接矩阵为：

$$N = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0.33 & 0.33 & 0 & 0.33 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$



例 5

- 规范化的随机跳转矩阵为：

$$N_r = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

- 假设随机跳转概率为0.1，则合成的规范化邻接矩阵为

例 5

$$M = 0.9N + 0.1N_r = \begin{pmatrix} 0.02 & 0.02 & 0.02 & 0.92 & 0.02 \\ 0.02 & 0.02 & 0.47 & 0.02 & 0.47 \\ 0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.32 & 0.32 & 0.02 & 0.32 \\ 0.02 & 0.92 & 0.02 & 0.02 & 0.02 \end{pmatrix}$$

计算主特征向量可以得到：

$$p = \begin{pmatrix} 0.419 \\ 0.546 \\ 0.417 \\ 0.422 \\ 0.417 \end{pmatrix}$$

节点v_2的网页排名值最高。