第三章 网络理论基础

3.1 网络与图

3.1.1 图的表示与分类

3.1.2 路径与连通性

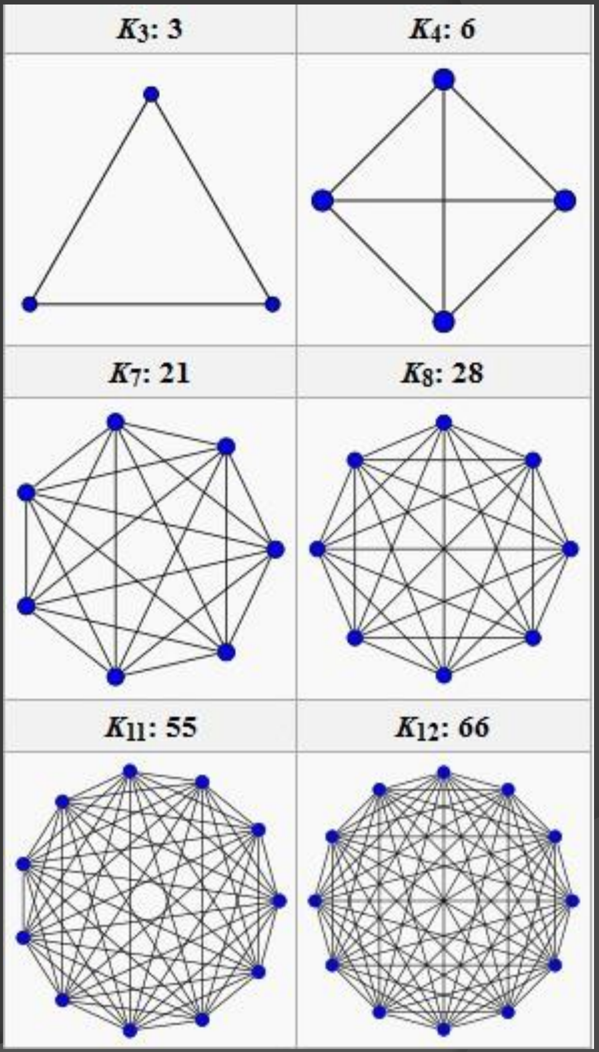
3.2 网络参数及其特性

3.2.1 主要参数中心性

3.2.2 网络的结构特性

3.2.3 网络的传输特性

3.3 大规模网络结构特征



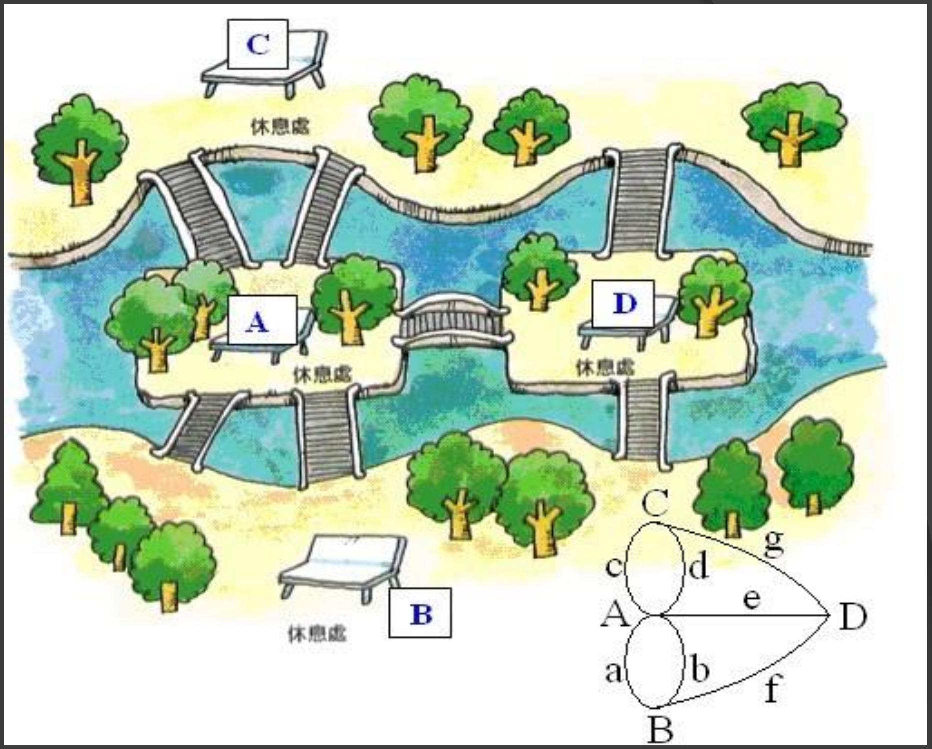
3.1 网络与图

图： 由点集与边集组成的二元组，记为$G = (V, E)$，

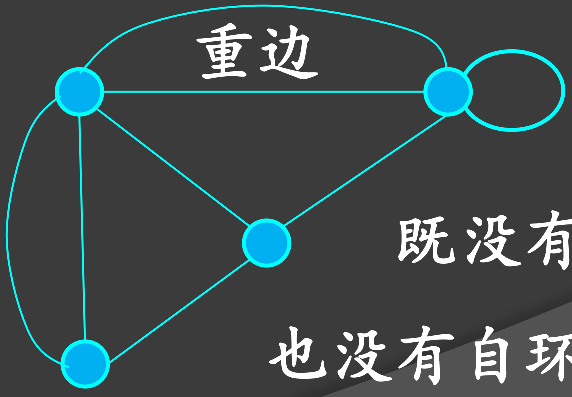
$V = (v_1, v_2, \cdots, v_n)$，

$E = (e_1, e_2, \cdots, e_m)$。

对任意一条边$e_k \in E$，必有一对$(v_i, v_j) \in V$的邻居节点与之对应。

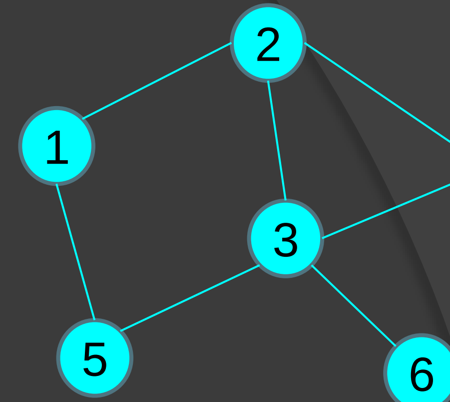


拓扑学： 研究与几何对象的大小、位置、形状、功能等因素无关，并且在几何对象连续变形下还能保持固有性质的学科。



既没有重边，也没有自环的图为**简单图**。

1、图的表示与分类



边列表： n =6，(1, 2)、(1, 5)、(2, 3)、(2, 4)、(3, 4)、(3, 5)、(3, 6)。

邻接矩阵： 图G表示为$A=(a_{ij})_{N\times N}$，其中：

$$a_{ij}=\begin{cases}1，\text{如果点}i\text{和点}j\text{之间有一条边；}\\0，\text{如果点}i\text{和点}j\text{之间没有边；}\end{cases}$$

N为点的个数。

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

- 简单图的邻接矩阵对角线的元素均为零；
- 简单图的邻接矩阵是对称的。

设图G是一个点个数为N、边个数为M的简单图，则有$0\le M\le N(N-1)\div 2$。实际大规模网络中，$M\ll N$，也称为稀疏网络，其邻接矩阵为稀疏矩阵。

加权图： 图G表示为$A=(a_{ij})_{N\times N}$，其中：

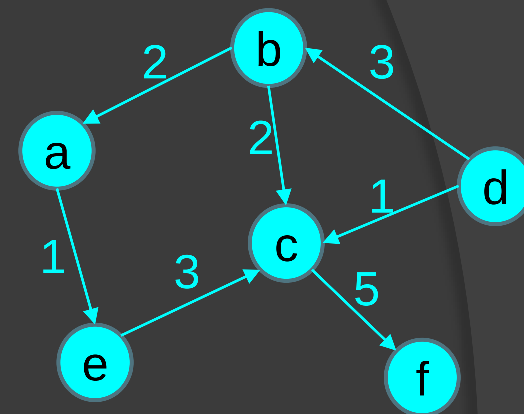
$$a_{ij}=\begin{cases} w_{ij}, & \text{如果有从点}i\text{指向点}j\text{的一条权值为}w_{ij}\text{的边;} \\ 0, & \text{如果点}i\text{和点}j\text{之间没有边;} \end{cases}$$

N为点的个数。

有向图： 图G表示为$A=(a_{ij})_{N\times N}$，其中：

$$a_{ij}=\begin{cases} 1, & \text{如果有从点}i\text{指向点}j\text{的一条边;} \\ 0, & \text{如果点}i\text{和点}j\text{之间没有边;} \end{cases}$$

N为点的个数。

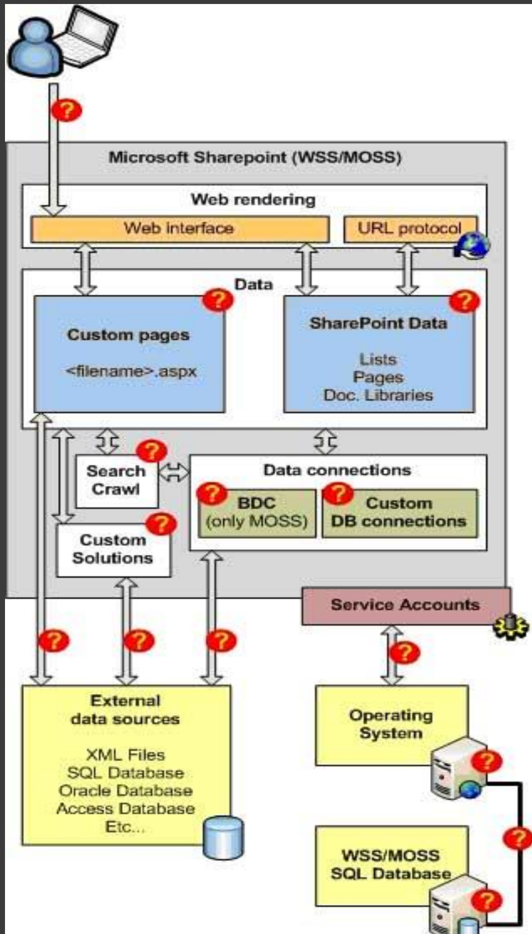


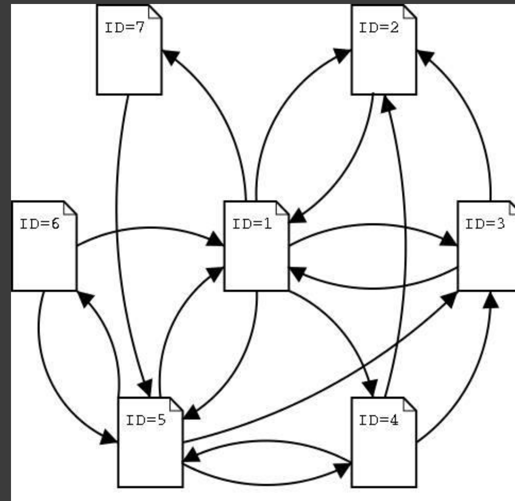
$$A=\begin{bmatrix} 0&0&0&0&1&0 \\ 2&0&2&0&0&0 \\ 0&0&0&0&0&5 \\ 0&3&1&0&0&0 \\ 0&0&3&0&0&0 \\ 0&0&0&0&0&0 \end{bmatrix}$$

有向图的邻接矩阵是非对称的。

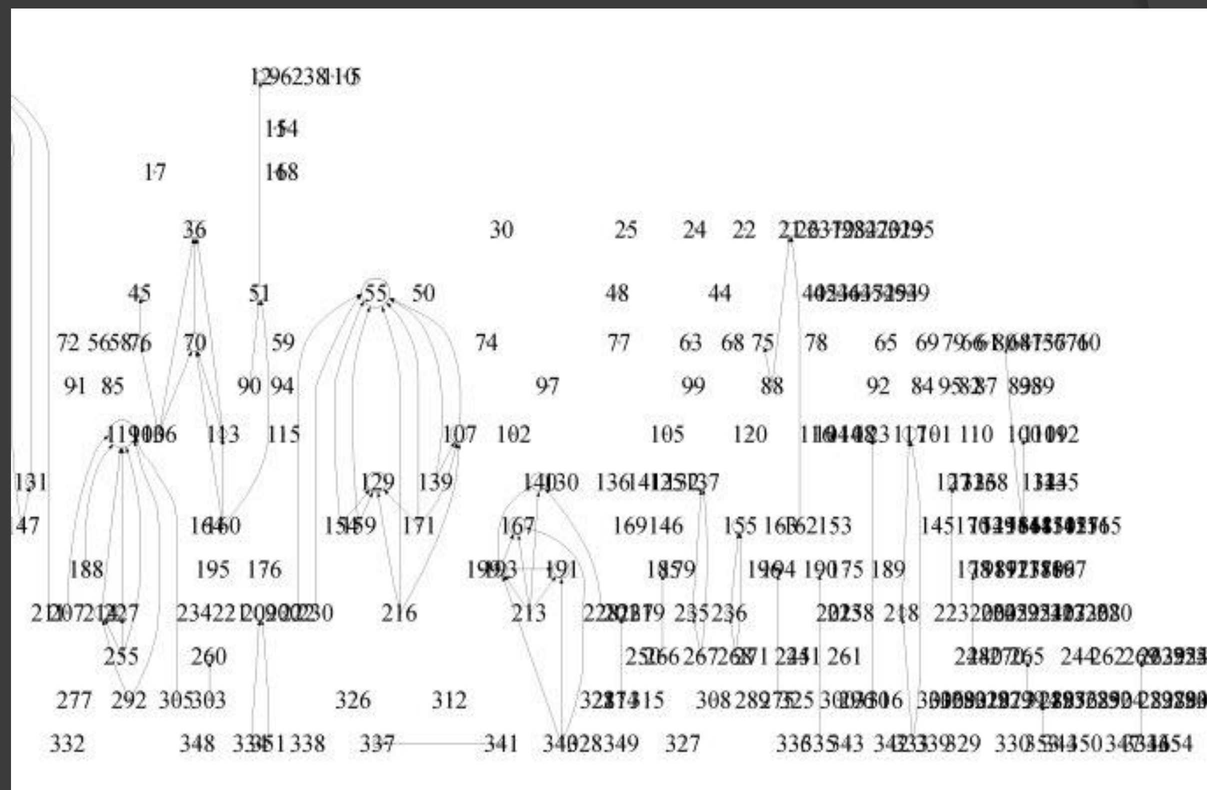
循环有向图与非循环有向图：

有向边可以构成闭合回路的图为循环有向图；否则，图为非循环有向图，图中至少有一个点没有转出方向的边。





Web网页链接关系构成一个复杂网络结构，是包括了回路的循环有向图。



引文网络是基于时间基准的非循环有向图。

共引值： 有向图G中，如果k同时有指向点i和点j的边，责令

$a_{ki}a_{kj}=1$，否则为0；那么点i和点j的共引值c_{ij}，

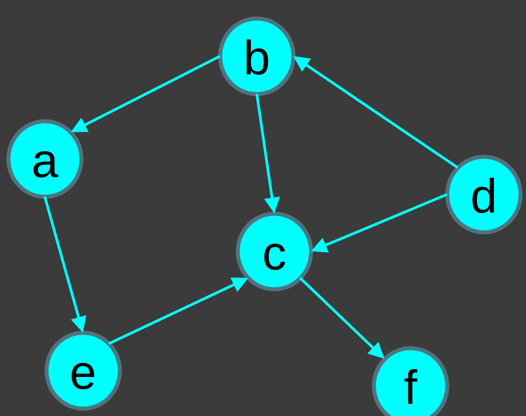
有$c_{ij}=\sum_{k=1}^{N}a_{ki}a_{kj}=\sum_{k=1}^{N}a_{ki}(a_{jk})^{T}$。

共引矩阵： $\boldsymbol{C}=\left(c_{ij}\right)_{N\times N}=\boldsymbol{A}^{T}\boldsymbol{A}$，对角元$c_{ii}=\sum_{k=1}^{N}(a_{ki})^{2}$，

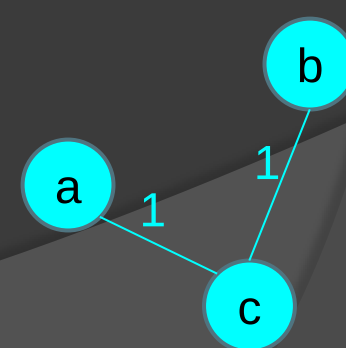
c_{ii}也就是所有指向i点的其它点的个数。

共引网络： 如果$c_{ij}>0$，$i\neq j$，则在点i和点j之间连接一条边，

边的权值为c_{ij}，故构成一个加权无向图的共引网络。



$$\Rightarrow\quad\begin{vmatrix}1&0&1&0&0&0\\0&1&1&0&0&0\\1&1&3&0&0&0\\0&0&0&0&0&0\\0&0&0&0&1&0\\0&0&0&0&0&0\end{vmatrix}\quad\Rightarrow$$

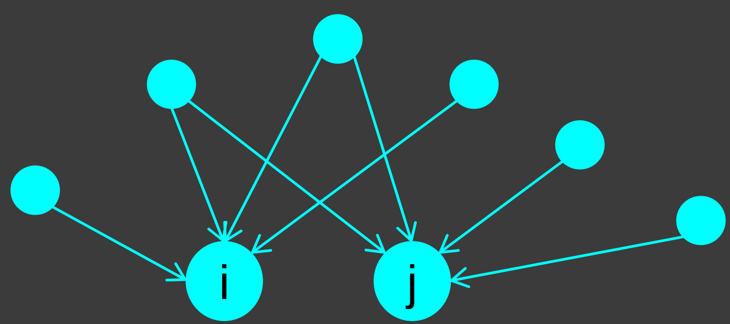


文献耦合数： 有向图G中，如果点i和点j都有指向k的边，责令

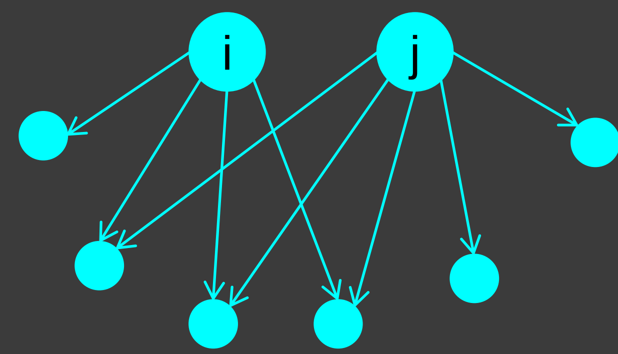
$a_{ik}a_{jk}=1$，否则为0；点i和点j的文献耦合数b_{ij}，

有$b_{ij}=\sum_{k=1}^{N} a_{ik}\,a_{jk}=\sum_{k=1}^{N} a_{ik}\,(a_{kj})^{T}$。

耦合矩阵和耦合网络： 与共引矩阵和共引网络的定义方式类似。



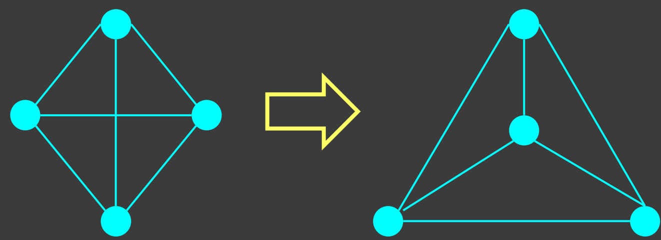
共引值=2

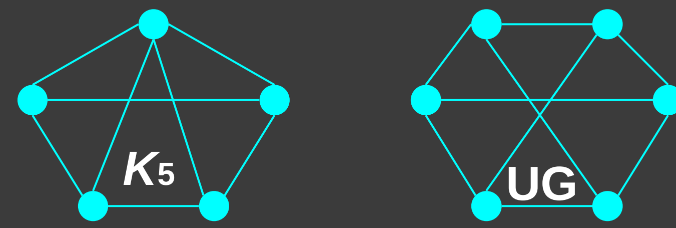


文献耦合数=3

共引与文献耦合在数学处理上相似，但许多实际应用中基于文献耦合的关联方式更为通用。因为，共引的统计有滞后性，而文献耦合的统计与文献发布同步。

平面网络： 可以画在一个平面，且边不交叉的网络。





Kuratowshi定理：任何一个非平面网络至少包含一个K_5或UG的子图或扩展子图。

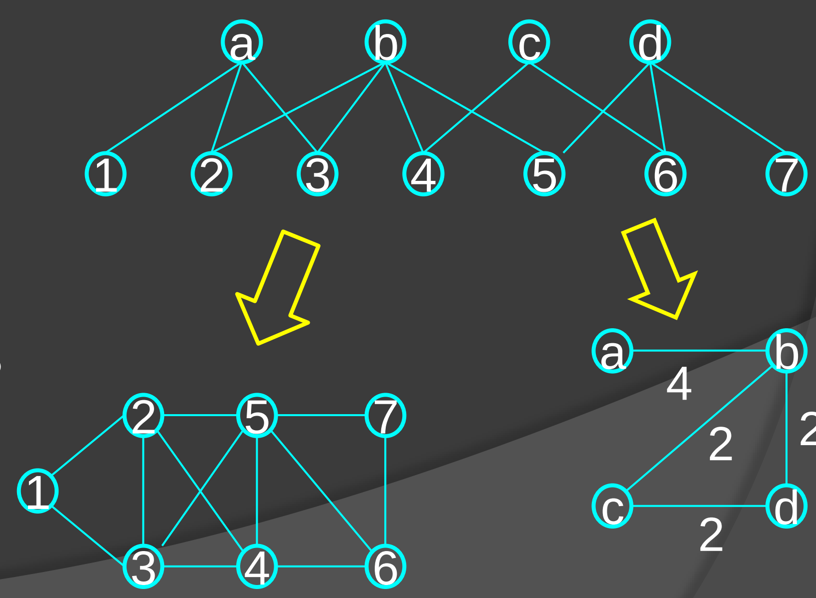
二分网络： 一类点代表原始点，另一类点表示原始点所属的群。

n代表成员数，g代表群数，

关联矩阵B是$g \times n$的矩阵，

$$B_{ij} = \begin{cases} 1, & \text{如果点}j\text{属于群}i \\ 0, & \text{其它} \end{cases}$$ 。

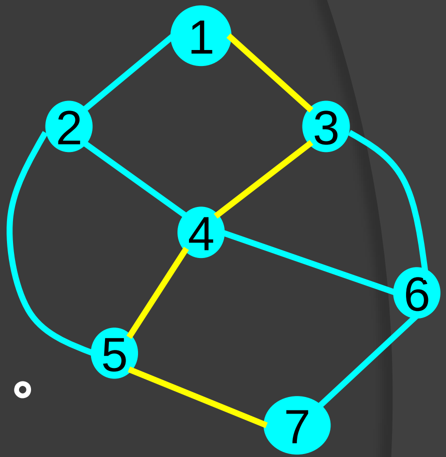
通过对二分网络**单模投影**可推出同类点之间的联系。



2、路径与连通性

路径： 无向图G中一条路径是指一个点的序列$P = v_1 v_2 \cdots v_n$，其中每对相邻的点v_i和v_j之间有一条边e_{ij}。

环路： 起点与终点重合的路径。



简单路径： 所有点和边都不相同的路径。

例如：$v_1 e_{13} v_3 e_{34} v_4 e_{45} v_5 e_{57} v_7$

路径长度： 所经过的边的个数，也称为“跳数”。

最短路径： 在指定的两个点之间的多条路径中，长度最短的简单路径即为两点之间的最短路径。

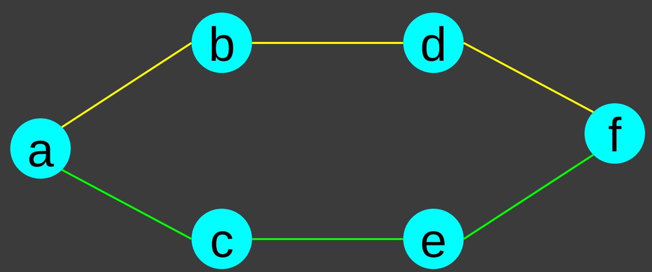
例如：$v_1 e_{13} v_3 e_{36} v_6 e_{67} v_7 / v_1 e_{12} v_2 e_{25} v_5 e_{57} v_7$

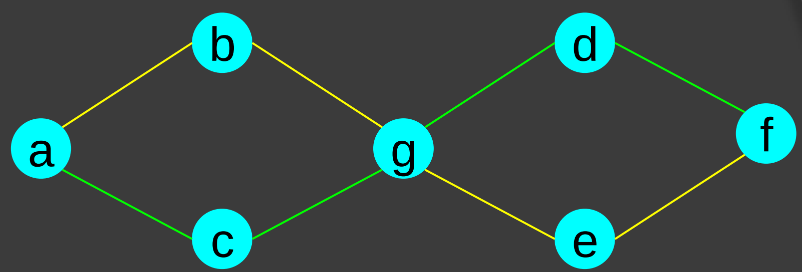
欧拉路径： 经过图中的所有边且每条边只经过一次的路经。

哈密顿路径： 经过图中的所有点且每个点只经过一次的路经。

路径独立性： 源与目的间任意两条路径是否有共用的点或边。

- 如果没有共用的点，则为点独立或**点不相交路径**。
- 如果没有共用的边，则为边独立或**边不相交路径**。





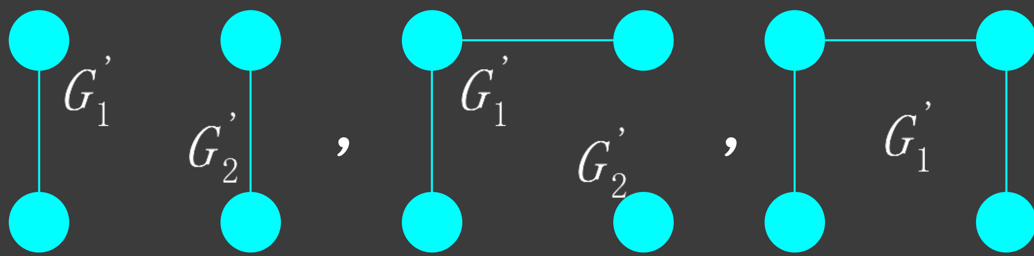
连通图： 图G的任意两点之间至少有一条路径。

- 如果点i和j经过点k有一条长度为2的路径，令$a_{ik}a_{kj}=1$，否则为0。设点i和j之间长度为2的路径总数为$(n_{ij})^2$，有
  $(n_{ij})^2=\sum_{k=1}^{N} a_{ik}\, a_{kj}=[A^2]_{ij}$ ；
- 依次类推，i和j之间长度为r的路径总数为$(n_{ij})^r=|A^r|_{ij}$。

一个图是连通的，当且仅当$I+A+[A^2]+\cdots+[A^{n-1}]$是正矩阵。

（矩阵所有元素都不为0）

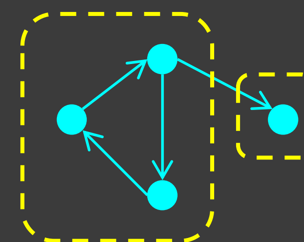
分支：图G的点子集，该子集构成连通图G'，任何一个点都属于且只属于一个分支。



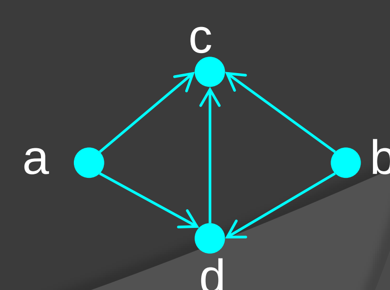
$$G = \sum G'$$

连通图有且只有一个分支，该分支构成的图为**生成子图**。

强连通分支是任意两点之间存在相互可达的有向路径的最大点子集。



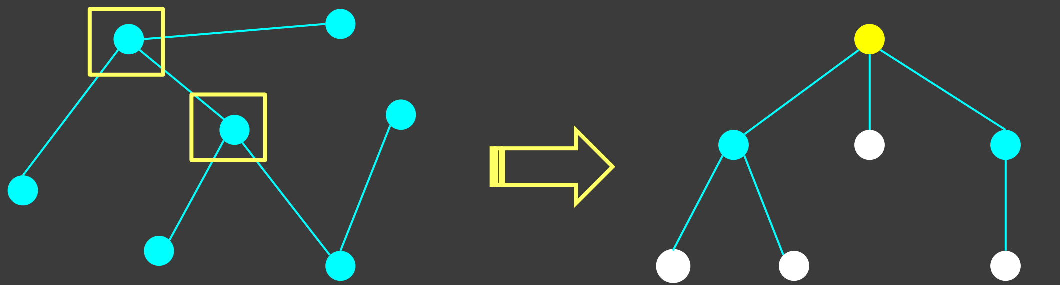
外向分支是从特定点出发，沿着有向边可以到达的所有点的集合。

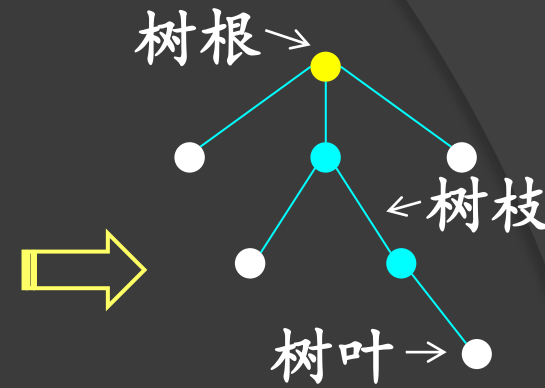


内向分支是从其它点出发，沿着有向边可以到达特定点的所有点的集合。

a的外向分支<a, c, d>；
d的内向分支<a, b, d>。

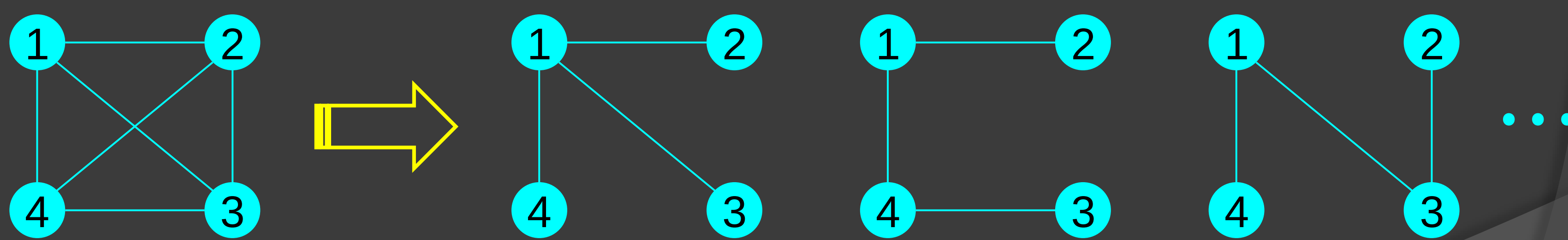
树：不包含重边和环路的无向连通图。





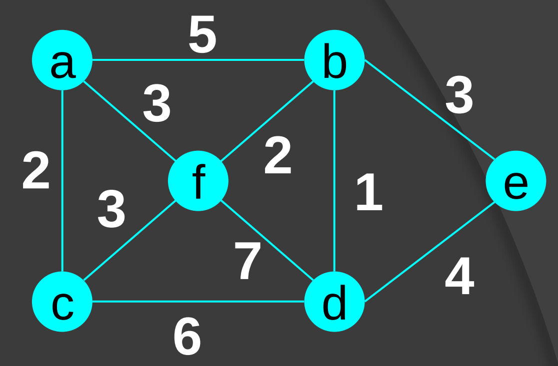
- 一棵树的任意两个点之间有且仅有一条路径。
- 一棵n个点的树有且仅有$n-1$条边。

生成树：一个连通图的树形生成子图（分支）。

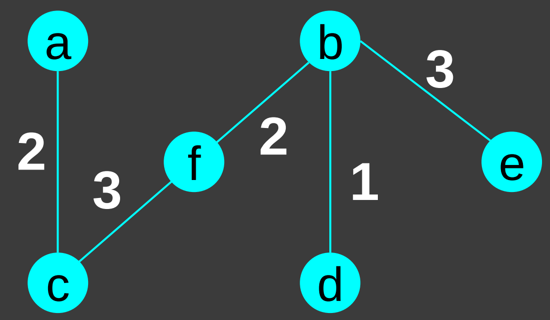


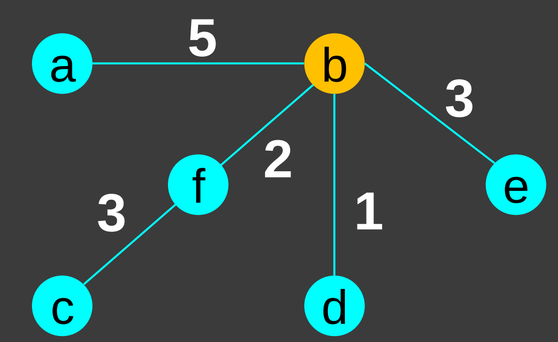
一个有n个点的完全图K_n的生成树个数$\tau(K_n)=n^{n-2}$。

加权最小生成树：一个无向的加权连通图$G=(V,E,w)$的一个权值之和最小的生成树。



$$\sum w=11$$





$$\sum w_{ba}=5 \quad \sum w_{be}=3$$

$$\sum w_{bc}=5 \quad \sum w_{bf}=2$$

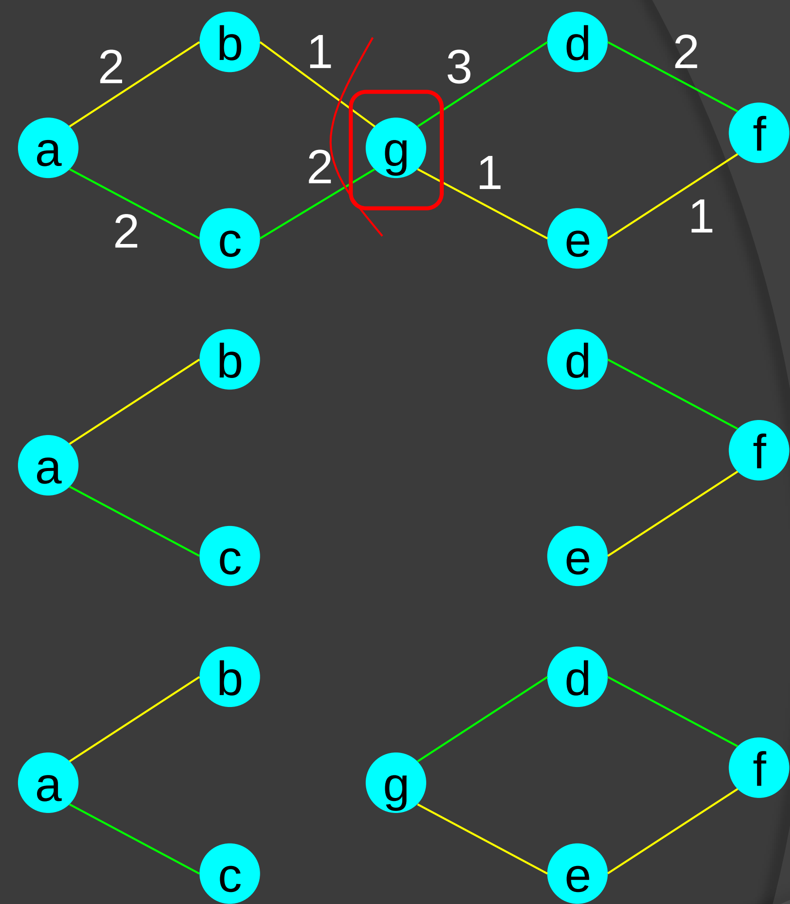
$$\sum w_{bd}=1$$

加权最短路径树：一个无向加权连通图$G=(V,E,w)$，某一源点到某一目标点的多条路径中，权值之和最小的路径为一条加权最短路径。由某一源点所有的加权最短路径构成的一棵树，即为该点的加权最短路径树。

割集：使连通图变为非连通图所需去除的点或边的集合。

- **最小点割集**，点数量最少的点割集，如$\langle v_g \rangle$。
- **最小边割集**，边数量最少的边割集，如$\langle e_{bg}、e_{cg} \rangle$。

Menger定理：如果给定两点之间不存在规模小于n的最小割集，则两点之间至少存在n条（点/边）独立路径。



加权网络最小边割集：边的权重之和（而非边数之和）最小的割集。

上例中，a与f间存在两条边不相交路径；它们之间的加权最小边割集为$\langle e_{df}、e_{ef} \rangle$。

3.2 网络参数与特性

1、主要参数中心性

度：无向图中连接某一个点i的边的个数k_i。

设图中有n个点m条边，则$k_i=\sum_{j=1}^{n}a_{ij}=\sum_{j=1}^{n}a_{ji}$；

平均度数 $\bar{k}=\frac{1}{n}\sum_{i=1}^{n}k_i=\frac{2m}{n}$，$2m=\sum_{i=1}^{n}k_i$。

有向图中点i的**出度**$k_i^{out}=\sum_{j=1}^{n}a_{ij}$，**入度**$k_i^{in}=\sum_{j=1}^{n}a_{ji}$。

因为图中总的出边与入边相等，故 $\bar{k}^{out}=\bar{k}^{in}=\frac{m}{n}$ 。

网络密度：实际边数与最大可能边数之比 $\rho=\frac{2m}{n(n-1)}$ 。

- $n\to\infty$，$\rho\to 0$，则网络是稀疏的；
- $n\to\infty$，$\rho\to$常数，则网络是稠密的。

大多数实际网络的$\bar{k}$为常数，有$\rho\sim\frac{\bar{k}}{n}\to 0$，故为稀疏网络。

度中心性$DC_i = \frac{k_i}{n-1}$，即度越大的节点越重要。

设点i的中心性为x_i，令$x_i = c\sum_{j=1}^{n} a_{ij}x_j$，c为常数选项，

设$x = [x_1, x_2, \cdots, x_n]^T$，故上式的矩阵形式为 $\boldsymbol{x} = c\boldsymbol{A}\boldsymbol{x}$ 。

因$\boldsymbol{x}$是与矩阵$\boldsymbol{A}$和$\boldsymbol{A}$的特征值c^{-1}对应的矩阵特征向量，因此，

也称这种节点重要性评估方法为特征向量中心性。

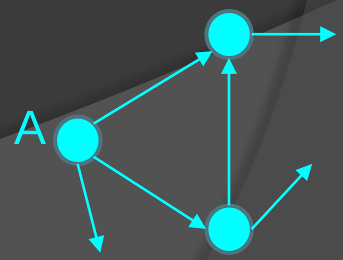
经过k步迭代后，中心性$x(k) = c\boldsymbol{A}\boldsymbol{x}(k-1)$，$k = 1,2,\cdots$。

设矩阵$\boldsymbol{A}$的特征值为θ_i，对应特征向量为v_i，ε_i为常数选项，

经矩阵运算推导出$x(k) = {\theta_1}^k \sum_i \varepsilon_i \left[\frac{\theta_i}{\theta_1}\right]^k v_i, \theta_1 > \theta_2 \cdots > 0$。

当$\boldsymbol{k} \to \infty$时，$x(k) \approx {\theta_1}^k \varepsilon_1 v_1$；$Ax = \theta_1 x$。

中心性取决于邻居的数量和邻居的中心性大小。



但在有向网络中可能出现中心性=0。

基于诱导的超链接主题搜索——HITS算法

HITS算法给点$\boldsymbol{i}$赋予一个权威中心性x_i和一个核心中心性y_i。

$x_i = \alpha \sum_j a_{ji} y_j$与指向点$\boldsymbol{i}$的邻居点的核心中心性之和成正比；

$y_i = \beta \sum_j a_{ij} x_j$与点$\boldsymbol{i}$指向的邻居点的权威中心性之和成正比；

α和β均为常数，迭代计算$x(K)$和$y(K)$。

利用矩阵表达方式，上述两个公式为：$\boldsymbol{x} = \alpha \boldsymbol{A}\boldsymbol{y}$，$\boldsymbol{y} = \beta \boldsymbol{A}^T \boldsymbol{x}$。

将$\boldsymbol{y}$代入有$\tau \boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{x}$；将$\boldsymbol{x}$代入有$\tau \boldsymbol{y} = \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{y}$，$\tau = (\alpha\beta)^{-1}$。

可见，权威中心性和核心中心性分别由具有相同特征值$\boldsymbol{\tau}$的矩阵$\boldsymbol{A}\boldsymbol{A}^T$和矩阵$\boldsymbol{A}^T\boldsymbol{A}$所对应的特征向量决定。

注意：$\boldsymbol{A}\boldsymbol{A}^T$就是共引矩阵；$\boldsymbol{A}^T\boldsymbol{A}$就是文献耦合矩阵。这样，一个节点即便没有被其它节点所指向，它的权威中心性为零，但它仍然可能有非零的核心中心性。

避免中心性为零的一个简单措施就是赋予点非零的初始值。

定义$x_i = \alpha \sum_j a_{ji} x_j + \beta_i$，$\boldsymbol{\beta}$为$\beta_i$的向量，$\mathbf{1}$为向量(1, ⋯），

上式的矩阵表示为$\boldsymbol{x} = \alpha \boldsymbol{A} \boldsymbol{x} + \beta \cdot \mathbf{1}$，也称为**Katz中心性**。

PageRank算法

基本思想：节点中心性被其指向的所有邻居平分，且$\sum x_i = 1$。

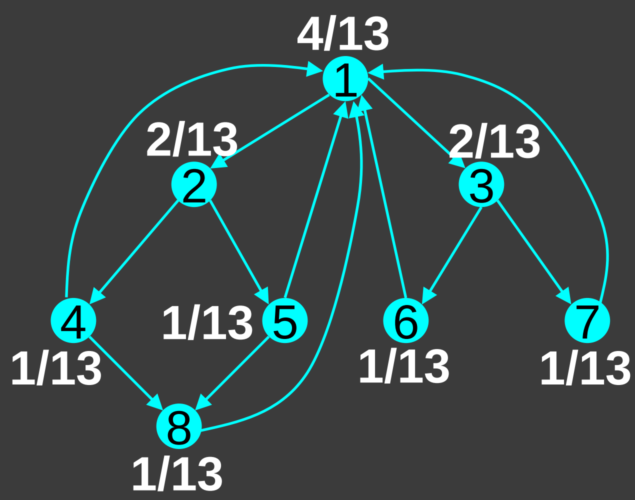
定义**Google**矩阵 $\widehat{A} = (\hat{a}_{ij})_{n \times n}$ ，$\hat{a}_{ij} = \begin{cases} {}^{1}/_{k_i{}^{out}}, & \text{有} i \text{指向} j \text{的边；} \\ 0, & \text{其它。} \end{cases}$

中心性 $\mathbf{x}(K) = \widehat{A}^T \mathbf{x}(K-1)$。

- 初始化：$x_i(0)$赋初值，且$\sum_i x_i = 1$；（一般$x_i(0) = {}^{1}/_{n}$）
- 迭代（步数为t）：$x_i(t) = \sum_{j=1}^{n} a_{ji} \frac{x_j(t-1)}{k_j{}^{out}} = \sum_j \hat{a}_{ji} x_j(t-1)$。

从算法计算过程可以看出，每个点的*PR*值只与它的入度有关，

与出度无关；但点出度的值却影响着出边对应点的*PR*值。



$$X(1)=\hat{A}^{T}X(0)=\begin{vmatrix} 0&0&0&\frac{1}{2}&\frac{1}{2}&1&1&1\\ \frac{1}{2}&0&0&0&0&0&0&0\\ \frac{1}{2}&0&0&0&0&0&0&0\\ 0&\frac{1}{2}&0&0&0&0&0&0\\ 0&\frac{1}{2}&0&0&0&0&0&0\\ 0&0&\frac{1}{2}&0&0&0&0&0\\ 0&0&\frac{1}{2}&0&0&0&0&0\\ 0&0&0&\frac{1}{2}&\frac{1}{2}&0&0&0 \end{vmatrix}\cdot\begin{vmatrix} 1/8\\1/8\\1/8\\1/8\\1/8\\1/8\\1/8\\1/8 \end{vmatrix}$$ ；…X(K) …

✓ 随机冲浪：从一个随机选择的网页开始，随机点击一个链接进入下一个网页。随机冲浪t步后到达网页X的概率，就等于应用基本PageRank算法迭代t步后网页X得到的$PR(t)$值。

悬挂节点问题：一旦到达某个出度为零的节点，会永远停留在该节点无法走出来。

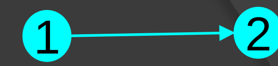
$$\hat{A}^{T}=\begin{pmatrix}0&0\\1&0\end{pmatrix}\quad PR(0)=\begin{pmatrix}1/2\\1/2\end{pmatrix}$$

$$PR(1)=\begin{pmatrix}0&0\\1&0\end{pmatrix}\begin{pmatrix}1/2\\1/2\end{pmatrix}=\begin{pmatrix}0\\1/2\end{pmatrix}$$

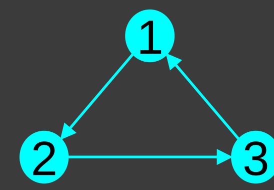
$$PR(2)=\begin{pmatrix}0&0\\1&0\end{pmatrix}\begin{pmatrix}0\\1/2\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

修正Google矩阵：

$$\hat{a}_{ij}=\begin{cases} 1/k_i^{out}, & k_i^{out}>0\text{且有}i\text{指向}j\text{的边。} \\ 0, & k_i^{out}>0\text{且无}i\text{指向}j\text{的边。} \\ 1/n, & k_i^{out}=0\text{ 。} \end{cases}$$



$$\hat{A}=\begin{vmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{vmatrix} \quad x^*=\begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix}^T$$



$$\hat{A}^T=\begin{vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix}$$

算法收敛问题：对一些特殊网络，存在迭代计算多次后中心性又回到初始值的现象。

$x(0)$	$x(1)$	$x(2)$	$x(3)$
1	0	0	1
0	1	0	0
0	0	1	0

修整规则：随机选择一个初始节点i。如果$k_i^{out}>0$，以概率α随机转到k_i^{out}指向的一个节点，以$(1-\alpha)$概率随机转到网络中任意一点；如果$k_i^{out}=0$，随机转到网络中任意一点。

$$x_i(K)=\alpha\sum_{j=1}^{n}\hat{a}_{ji}x_j(K-1)+(1-\alpha)\frac{1}{n}\text{ ，}\alpha=0.85\text{。}$$

平均路径长度

设点i到点j的最短路径长度（跳数）为l_{ij}，点i的平均路径长度有$\overline{l_i}=\frac{1}{n-1}\sum_{j\neq i}l_{ij}$，$n$为网络中的节点个数。

接近中心性：$C_i=\frac{1}{\overline{l_i}}=\frac{n-1}{\sum_{j\neq i}l_{ij}}$。

改进的计量公式：$\widehat{C_i}=\frac{1}{n-1}\sum_{j\neq i}\frac{1}{l_{ij}}$。

- 排除了处于不同分支中的点之间的影响因素；
- 突出了距离较近的那些路径的影响因素。

网络平均路经长度：$L=\frac{1}{n}\sum_i\overline{l_i}$。

改进的计量公式：$\frac{1}{\hat{L}}=\frac{1}{n(n-1)}\sum_{i\neq j}\frac{1}{l_{ij}}=\frac{1}{n}\sum_i\widehat{C_i}$，$\hat{L}=\frac{n}{\sum\widehat{C_i}}$。

网络直径：$D=\max_{i,j}l_{ij}$。如果跳数在d内的连通节点对数量占网络节点数量的90%以上，则d为网络的有效直径。

介数： 经过某个点或边的最短路径的数量。

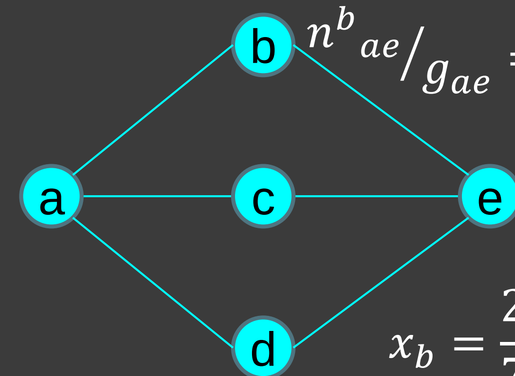
$$n^i{}_{sd}=\begin{cases}1，点 i 为源点 s 到目的点 d 的最短路径上的一个点；\\ 0，其它。\end{cases}$$

介数中心性 $x_i=\sum_{s,d} n^i{}_{sd}$ ，在有向网络中按实际路径处理。

如果点s和d之间存在g_{sd}条最短路径，则$x_i=\sum_{s,d} {n^i{}_{sd}}/{g_{sd}}$。

介数值分布在很大的范围，最大值与最小值比值可达$\frac{n}{2}$，故可归一化处理。

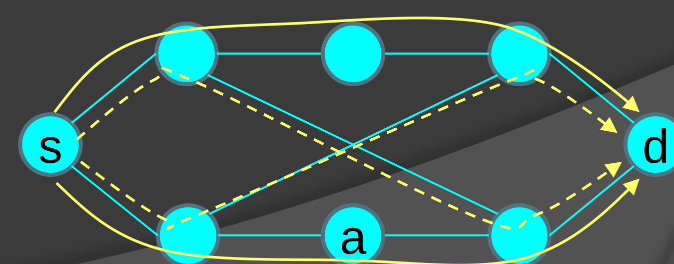
$x_i=\frac{1}{n^2}\sum_{s,d} {n^i{}_{sd}}/{g_{sd}}$，取值为[0，1]。



$n^b{}_{ae}/g_{ae}=\frac{1}{3}$

$x_b=\frac{20}{75}\approx 0.27$

流介数： $n^i{}_{sd}$为源点s到目的点d之间经过点i的独立路径数量。不再以最短路径为基准。



$n^a{}_{sd}=1$

2、网络的结构特性

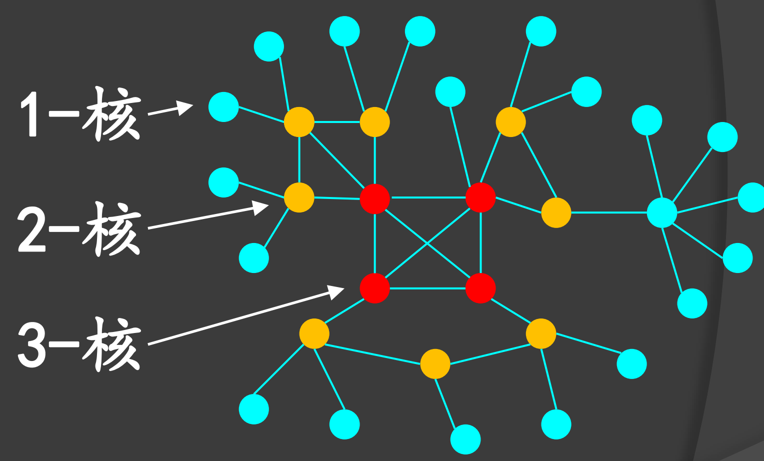
群组（社团）性

团：任何两个点之间都直接相连的最大点子集。

***k*−团:**任意两点之间边的距离不超过*k*跳的最大点子集。

***k*−核：**每个点至少与子集中其它*k*个点直接相连的最大点子集。

令$k = 1, 2, \cdots$，删除度数为k的点及其连边，直到不再存在度数为k的点，这个过程称为***k*−核分解**。该方法可一定程度定义节点连通性和重要性，注意：并非度数高的节点一定很重要。



***k*−(连通)分支:**任意两点之间至少有*k*条独立路径的最大点子集。

k−分支具有层层包含关系，$k \geq 3$的*k*−分支可以是不邻接的。

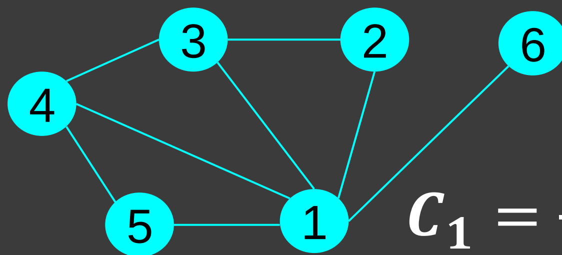
传递性

关系X，若有aXb和bXc，可推出aXc，称X具有传递性。

在网络中X为点之间的“边连接”关系，也称网络的传递性。

点i的度为k_i，聚类系数$C_i=\frac{2E_i}{k_i(k_i-1)}$，$E_i$是$i$的邻居之间的边数。

用邻接矩阵表示，$C_i=\frac{\sum a_{ij}a_{jk}a_{ki}}{\sum a_{ij}a_{ik}}$。



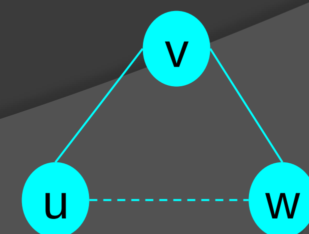
$C_1=\frac{3}{10}$。

有时邻居点之间的预期连接并不存在，

这些消失的连接称为结构洞，聚类系数与结构洞数量成反比关系。

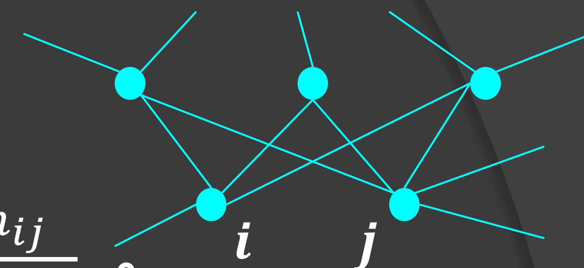
结构洞的存在对于网络路由不利，可选路经减少，但对信息传播的集中控制有利，如星形。因此，度数高而聚类系数低的节点在网络中往往具有更强的影响力和重要性。

通用聚类系数表示法，$C=\frac{\text{三角形数}\times 3}{\text{连通三元组数}}$



相似性： 利用网络结构中包含的信息确定节点之间的相似性。

余弦相似性： 如果点i和j的共享邻居较多，则它们相似度较高，反之，相似度较低。



相似性参数为σ_{ij}，$\sigma_{ij}=\frac{\sum_l a_{il}a_{lj}}{\sqrt{\sum_l a_{il}{}^2 \sum_l a_{lj}{}^2}}=\frac{n_{ij}}{\sqrt{k_i k_j}}$。

$$\sigma_{ij}=\frac{3}{\sqrt{4\times 5}}=0.671\cdots$$

σ_{ij}为1，两个节点有完全相同的邻居；

为0则无任何共享邻居。

皮尔逊相关系数： $r_{ij}=\frac{\sum_k(a_{ik}-\langle a_i\rangle)(a_{jk}-\langle a_j\rangle)}{\sqrt{\sum_k(a_{ik}-\langle a_i\rangle)^2}\sqrt{\sum_k(a_{jk}-\langle a_j\rangle)^2}}$。

其中，$\langle a_i\rangle$为随机连接产生的邻接矩阵中第i行元素的均值。

$\sum_k a_{ik}a_{jk}-\frac{k_i k_j}{n}=\sum_k(a_{ik}-\langle a_i\rangle)(a_{jk}-\langle a_j\rangle)$，行$i$和$j$的协方差。

$r_{ij}>0$，两个节点之间具有相似性，值越大相似度越高；

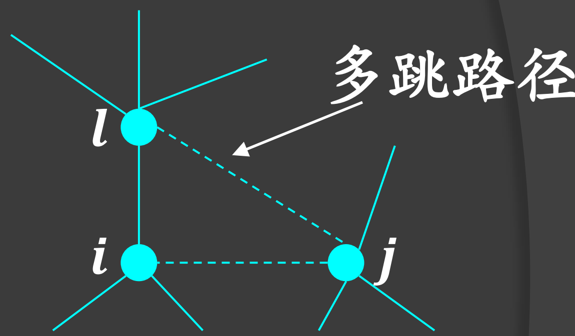
$r_{ij}\leq 0$，两个节点之间不具有相似性。

欧几里得（汉明）距离： $d_{ij}=\sum_k\left(a_{ik}-a_{jk}\right)^2$ ，

归一化有： $\delta_{ij}=1-2\times\frac{n_{ij}}{k_i+k_j}$ 。

δ_{ij}为两个节点之间的差异度，δ_{ij}越大差异也越大。

Katz相似性： 如果点$\boldsymbol{i}$的邻居$\boldsymbol{l}$与点$\boldsymbol{j}$相似度较高，则点$\boldsymbol{i}$与点$\boldsymbol{j}$的相似度也较高。有



$\sigma_{ij}=\alpha\sum_l a_{il}\sigma_{lj}+\delta_{ij}$，$\delta_{ij}$为对角元。

矩阵形式：$\boldsymbol{\sigma}=\alpha\boldsymbol{A}\boldsymbol{\sigma}+\boldsymbol{I}$ 。

$\boldsymbol{\sigma}^{(0)}=\mathbf{0}$；$\boldsymbol{\sigma}^{(1)}=\boldsymbol{I}$；

$\boldsymbol{\sigma}^{(2)}=\alpha\boldsymbol{A}+\boldsymbol{I}$；

$\boldsymbol{\sigma}^{(3)}=(\alpha\boldsymbol{A})^2+\alpha\boldsymbol{A}+\boldsymbol{I}$；

迭代次数$\rightarrow\infty$，有

$\boldsymbol{\sigma}=\sum_{m=1}^{\infty}(\alpha\boldsymbol{A})^{m-1}+\boldsymbol{I}$ 。

链路预测： 预测网络中尚不存在连边的两节点建立连接的可能性，两点的相似度越高，新连接产生的概率越大。

3、网络的传输特性

连通度： = min {最小割点集的点数，最小割边集的边数} 。

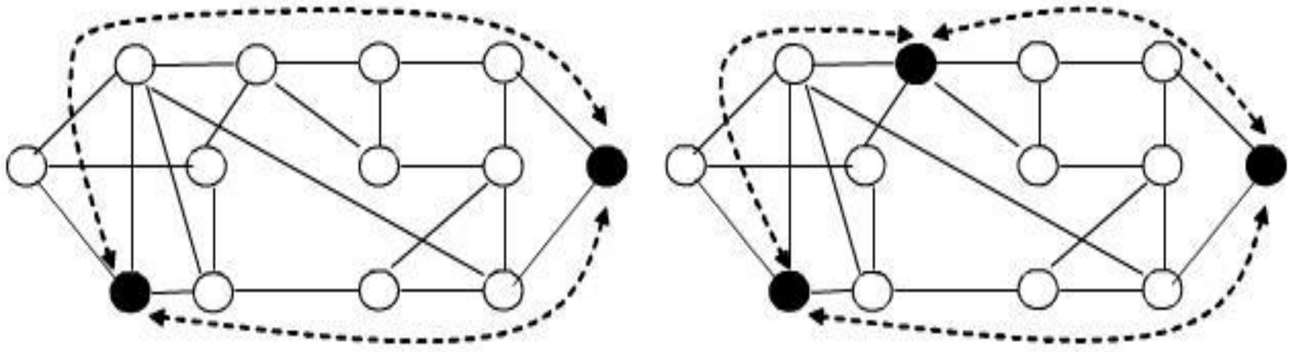
可靠度： 网络可靠度是指网络在规定条件下，按照规定要求，将信息数据完整、正确地传输的能力。

$$\text{平均无故障时间} = \frac{\text{有效运行时间}}{\text{总的运行时间}}$$ 。单个设备

$$\text{丢包率} = \frac{\text{成功接收的数据包个数}}{\text{总的传输的数据包个数}}$$ 。 单条链路

终端可靠度=节点保持连通的概率=保证可达性的概率之和。

K-终端可靠度： 在一个连通网络的概率图中，对指定$|K|$个端节点所构成的集合K，任意两个端节点之间均有一条可以通信的路径的概率p，记为$Rel_K(G)$。



$|K|=2$　　　　$|K|=3$

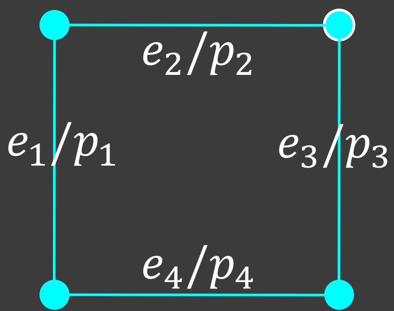
$|K|=2$，为两终端可靠度；$|K|=n$，为全终端可靠度。

设在一个连通简单网络的概率图G中，指定的K个节点之间的全部K-树有m个，记为$A_1, A_2, \cdots, A_m$，则该网络的K-终端可靠度为：$Rel_K(G) = p(A_1 \uplus A_2 \uplus \cdots \uplus A_{m-1} \uplus A_m)$

$$= p(A_1) + p(\bar{A}_1 A_2) + \cdots + p(\bar{A}_1 \bar{A}_2 \cdots \bar{A}_{m-1} A_m)。$$

- A_i、A_j无共同元素，$\bar{A}_i A_j$用摩尔定律展开。
- A_i、A_j有共同元素，$\bar{A}_i A_j = \overline{(A_i - A_j)} A_j$，类似地，$\left(\prod_{i=1}^{j-1} \bar{A}_i\right) A_j$，$j = 2,3,\cdots,m$。

不交和计算法



设网络图G的边为e_i，边正常传输的概率为p_i（不正常的概率可以设为丢包率），计算网络G的全终端可靠度Rel_{all}。

网络图G的生成树集：$\langle e_1,e_2,e_3\rangle$，$\langle e_1,e_2,e_4\rangle$，$\langle e_1,e_3,e_4\rangle$，$\langle e_2,e_3,e_4\rangle$。

不交和：$F(e_1e_2e_3 \uplus e_1e_2e_4 \uplus e_1e_3e_4 \uplus e_2e_3e_4)$

$$= e_1e_2e_3 + \overline{e_1e_2e_3}\cdot e_1e_2e_4 + \overline{e_1e_2e_3}\cdot\overline{e_1e_2e_4}\cdot e_1e_3e_4 + \overline{e_1e_2e_3}\cdot\overline{e_1e_2e_4}\cdot\overline{e_1e_3e_4}\cdot e_2e_3e_4$$

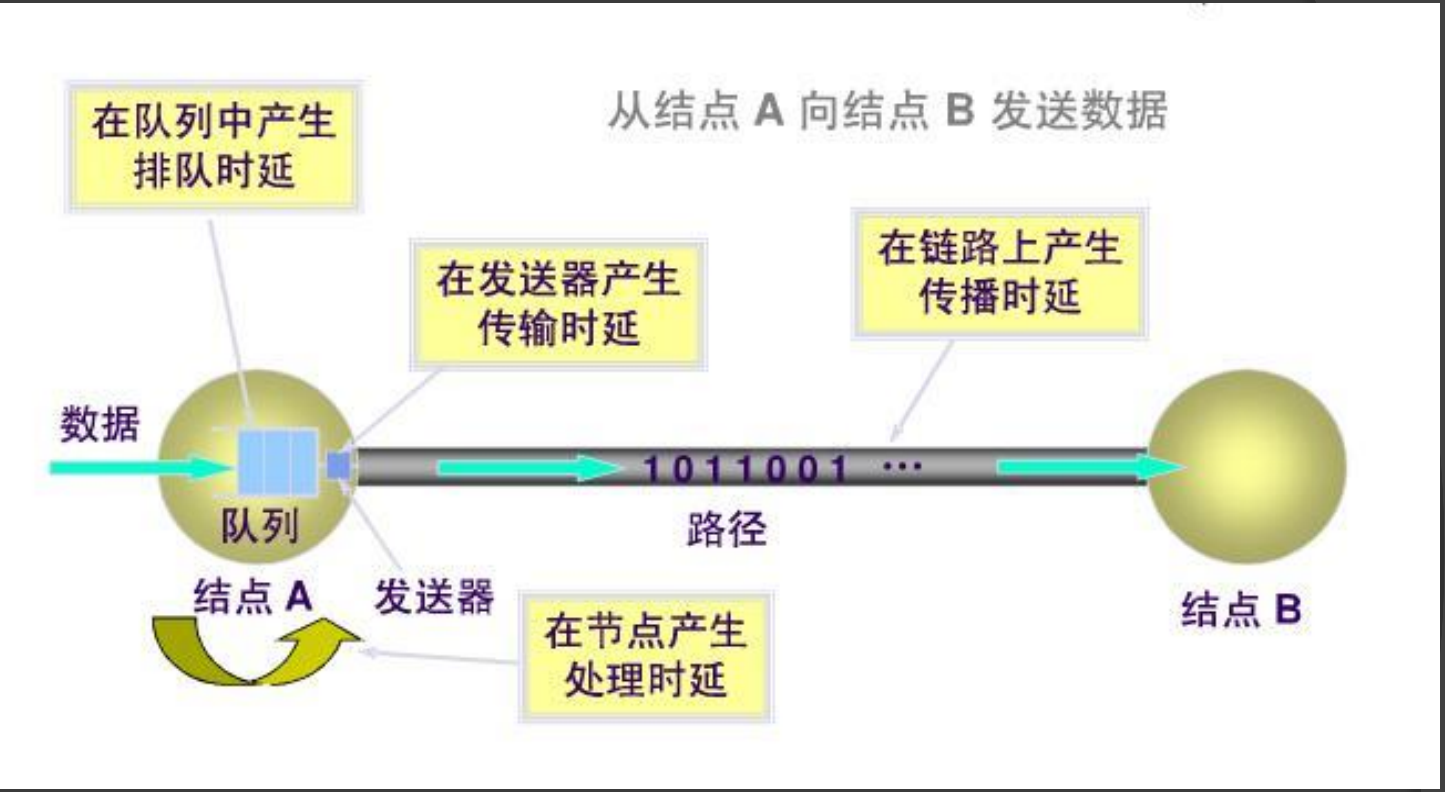
$$= e_1e_2e_3 + \bar{e}_3e_1e_2e_4 + \bar{e}_2e_1e_3e_4 + \bar{e}_1e_2e_3e_4\text{。}$$

$$Rel_{all}(G) = p(F) = p_1p_2p_3 + (1-p_3)p_1p_2p_4 + (1-p_2)p_1p_3p_4 + (1-p_1)p_2p_3p_4\text{。}$$

路径时延： 数据报文从网络一端传递到另一端所用的时间。

单跳时延的构成：

排队时延具有随机不确定性，是时延抖动变化的根本原因。



路径时延的累加性：

链路长度　报文长度　排队时间

路径L的时延 $D_L = \sum_{h=0}^{H} \frac{d_h}{v} + \sum_{h=0}^{H} \frac{l}{b_h} + \sum_{h=1}^{H} {t^p}_h + \sum_{h=1}^{H} {t^q}_h$ ，

L的长度为H跳。

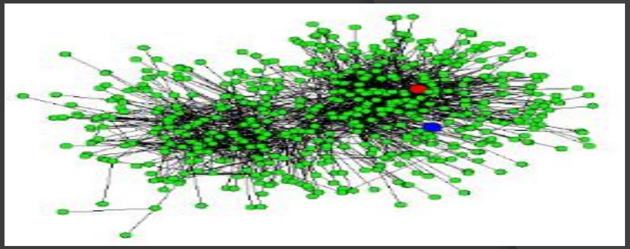
传播速度　链路带宽　处理报文时间

3.3 大规模网络结构特征

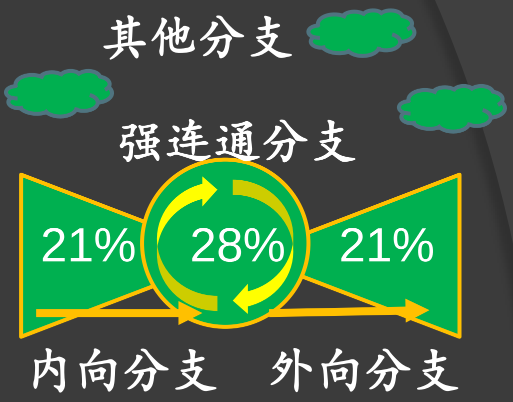
网络	类型	点数 n	边数 m	平均度数 $\bar{k}$	平均最短路径 $\bar{L}$	聚类系数 C	最大分支比例 S	幂律指数 α
电话呼叫图	无向	47000000	80000000	3.16				2.1
电子邮件	有向	59821	86300	1.44	4.95		0.95	1.5
电邮地址薄	有向	16811	57029	3.38	5.22	0.17	0.59	-
电影演员	无向	449913	25516482	113.43	3.48	0.2	0.98	2.3
www.nd.edu	有向	269504	1497135	5.55	11.27	0.11	1	2.1
Web网*	有向	203549046	1466000000	7.2	16.18		0.91	2.1
引文网络	有向	783339	6716198	8.57				3.0
Internet*	无向	10697	31992	5.98	3.31	0.012	1	2.1
电力网络	无向	4941	6594	2.67	18.99	0.1	1	-
火车线路	无向	587	19603	66.67	2.16		1	-
对等网络	无向	880	1296	1.47	4.28	0.035	0.805	2.5

“*”表示部分统计结果；“-”表示不遵循幂律；“ ”表示没有获取的数据。

分支



无向网络由一个"巨分支"和大量小分支组成。Internet只有一个分支。



有向网络以一个强连通分支及其外向和内向分支为主导。例如，Web网络中的蝴蝶结分支。

最短路径与直径

平均最短路径 $\bar{L}_{min} \sim \ln\ln n / \ln n$；

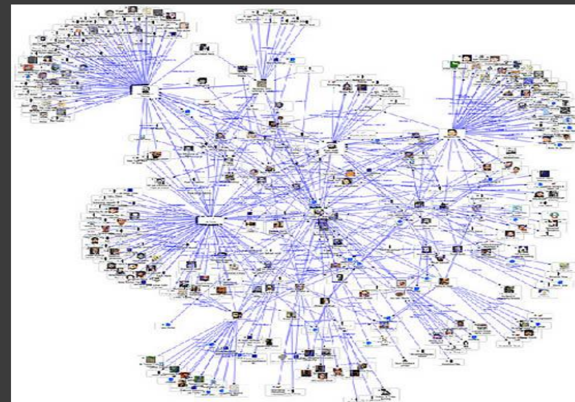
网络直径 $D \sim \log n$ ，n为节点个数。

漏斗效应：大多数节点在很少几跳内的邻居中就有一个或两个高介数节点。

Internet的AS间，约49%的最短路径经过度>=5的路由器。

度分布： p_k为网络中度数=k的节点个数所占节点总数比例。

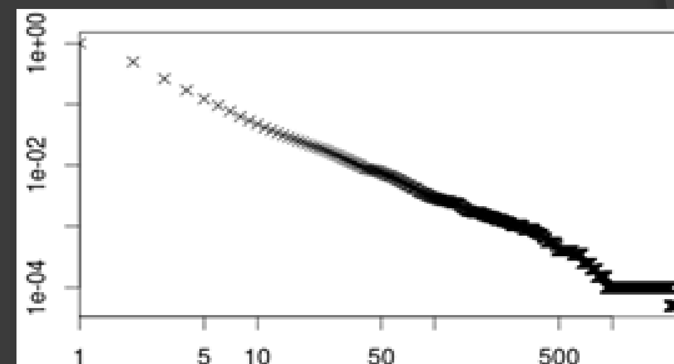
幂律分布： 绝大多数节点度数很小，个别“核心”节点度数很高。



$$\ln p_k = -\alpha \ln k + c \rightarrow p_k = e^c k^{-\alpha}$$

α和c都是常数，α为幂律指数。

右图为按度数降序排列的分布图。



累积分布函数： $P_k = \sum_{k'=k}^{\infty} p_{k'}$ ，节点度数$\geq k$的比例。

$P_k \approx e^c \int_k^{\infty} k'^{-\alpha} dk' = \frac{e^c}{\alpha-1} k^{-(\alpha-1)}$。设$k \geq k_{min}$时为幂律分布，

因$e^c \sum_{k_{min}}^{\infty} k^{-\alpha} = 1$，有$e^c \approx \frac{1}{\int_{k_{min}}^{\infty} k^{-\alpha} dk} = (\alpha - 1)(k_{min})^{\alpha-1}$ ，

故 $P_k \approx \left(\frac{k}{k_{min}}\right)^{-(\alpha-1)}$ 。

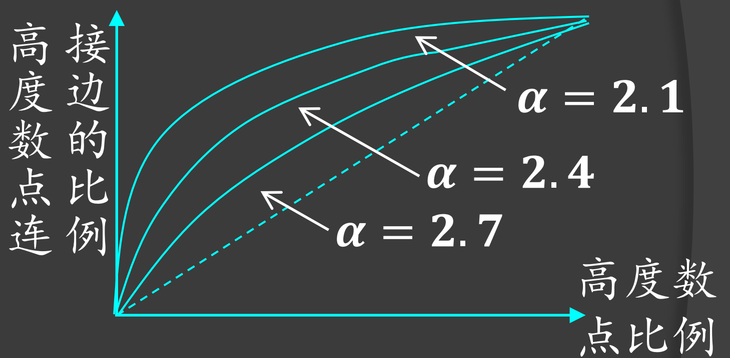
度分布的m阶矩： $\langle k^m \rangle = \sum_{k=0}^{\infty} k^m p_k$ 。

当$k \ge k_{min}$时，$\langle k^m \rangle = \sum_{k=0}^{k_{min}-1} k^m p_k + e^c \sum_{k=k_{min}}^{\infty} k^{m-\alpha}$。

$$\langle k^m \rangle \sim e^c \int_{k_{min}}^{\infty} k^{m-\alpha} dk = \frac{e^c}{m-\alpha+1} \left| k^{m-\alpha+1} \right|_{k_{min}}^{\infty} \sim n^{m-\alpha+1}。$$

Internet、Web网等，$2 \le \alpha \le 3$，所以，有有限均值，但无有限方差，故也称为无标度网络。

头重分布：设网络中与高度数节点连接的边所占总边数的比例为W，有 $W = (P_{k_h})^{\alpha-2/\alpha-1}$，$k \ge k_h$的为高度数节点。



- Web网50%的超链接都指向了1.5%的高度数网页；
- 引文网8.3%的被引用最多的论文占据了引用关系的50%；
- Internet中3.3%核心节点占据了“对等”连接关系的50%。

聚类系数

Internet聚类系数小，远低于局部聚类系数C_i的算术平均值。

设$C' = \frac{1}{n}\sum_{i=1}^{n} C_i$，有$C'_{internet} = 0.39$，而$C_{internet} = 0.012$。

Internet中存在大量的“结构洞”和一些“统治”节点。

局部聚类系数随节点度数递减，$C_i \sim k^{-\beta}$，$0 < \beta \leq 1$。

可能一些节点组成某个群组或社团，群组内的彼此连接比较多。

Internet的其他中心性

- 特征向量中心性的累积分布函数大致服从幂律分布；
- 介数中心性的累积分布函数也大致服从幂律分布；
- 接近中心性的累积分布函数不服从幂律分布。

25、1）写出图1的邻接矩阵；

2）写出基于邻接矩阵计算节点度值k_i的向量K的表达式；

3）写出网络边数的计算表达式。

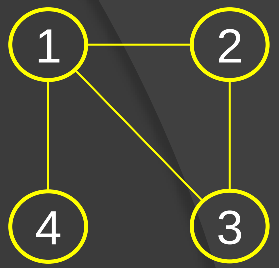


图1

26、写出图1中的最小割端集和最小割边集。

27、利用连通图的数学定义证明图1是连通图，写出计算过程。

28、分析图1中哪个点的接近中心性最好？写出分析过程。

29、图1有几棵最小生成树？举一例说明图1的最小生成树和最短路径树可以是同一棵树。

30、1）写出图2对应的共引矩阵和共引网络，

以及文献耦合矩阵和文献耦合网络；

2）写出图2的google矩阵$\hat{A}$；

3）写出$PR(1)$的计算过程（$\alpha = 0.85$），

迭代过程大约多少步可以收敛？

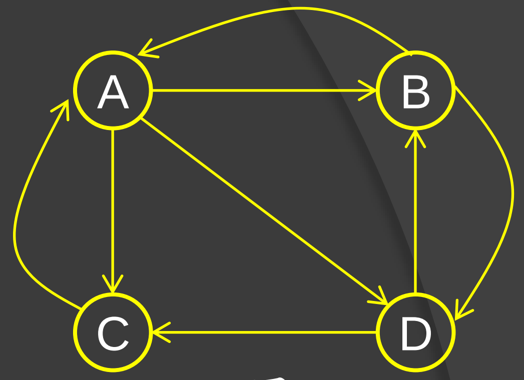


图2

31、1）说明在图3中标记出一个“3-核”的生成过程；

2）分别计算两个实心点的聚类系数；

3）计算两个实心点的余弦相似性；

4）采用皮尔逊相关系数的优点是什么？

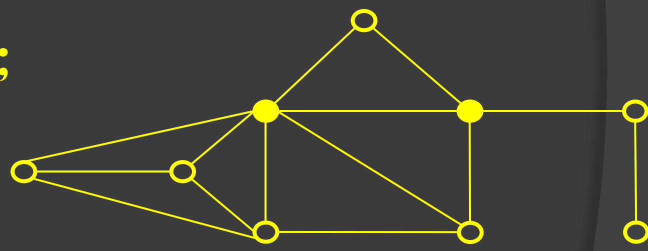


图3

32、1）说明连通度与可靠度的相同与不同之处；

2）如何使图3的连通度不小于3？画图说明。

3）图3中边可用概率为p，写出实心点间的Rel_2计算过程。