西安电子科技大学
XIDIAN UNIVERSITY

复杂网络与群体智能

吴建设
Email: jshwu@mail.xidian.edu.cn
西安电子科技大学人工智能学院

第二讲：基本概念与结构模型

➢基本概念

（1）网络的图表示

（2）节点的聚类系数

（3）节点的介数

（4）平均度和度分布

➢结构模型

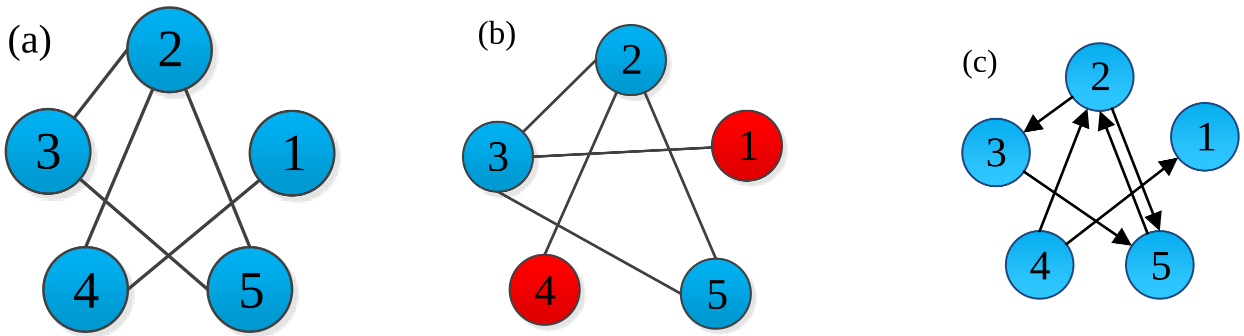
（1）小世界网络

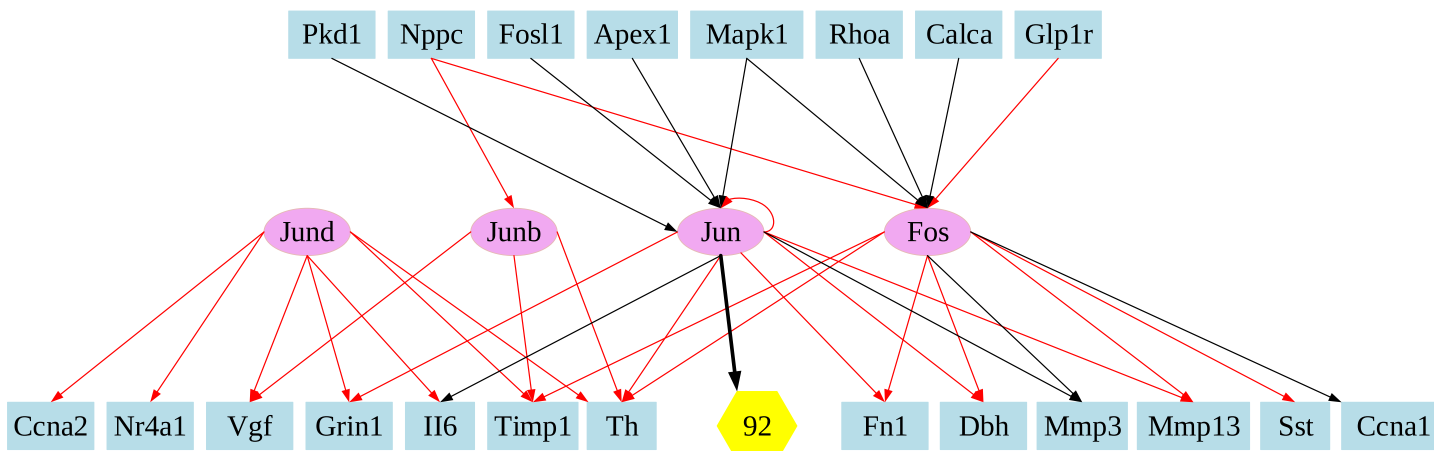
（2）无标度网络

（3）社区网络

1. 网络的图表示

无向网络与有向网络：如果网络中的边是没有方向的，e_{ij}与e_{ji}是同一条边

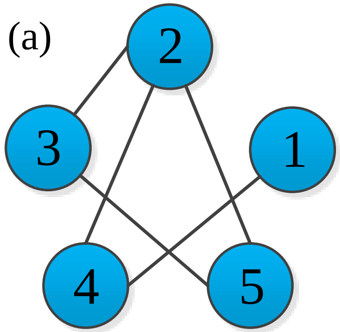


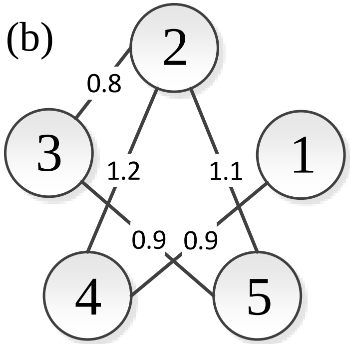


The gene regulatory network of TF family AP1 in rat.

1. 网络的图表示

无权网络与加权网络：(a) 无权网络；(b) 加权网络。



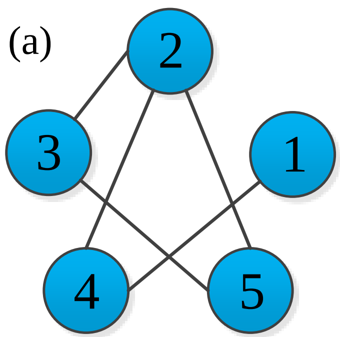


二分网络。网络中的节点也可以有不同的类型，比如描述网上销售行为的网络，一类节点是消费者（购买者），另一类节点则是商品。一个消费者购买过一种商品，则它们之间有一条连边，这样的网络称为二分网络。二分网络中同类节点之间没有连边。

1. 网络的图表示

邻接矩阵。网络的连接关系也可以用一个$N\times N$的矩阵$A=[a_{ij}]$表示，称为邻接矩阵。如果节点v_i有到节点v_j的连边，则$a_{ij}=1$，否则$a_{ij}=0$。对无向网络，邻接矩阵一定是对称的，而且其主对角线一定为零；对无向网络而言，邻接矩阵一般是非对称的。

(a)



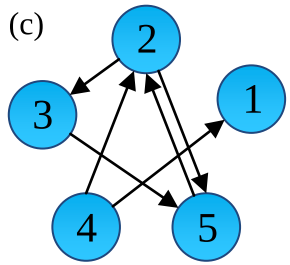
$$A=\begin{pmatrix}0&0&0&1&0\\0&0&1&1&1\\0&1&0&0&1\\1&1&0&0&0\\0&1&1&0&0\end{pmatrix}$$

1. 网络的图表示

图的连通性。在一个有向图 G 中，如果存在从节点v_i到节点v_j的路径，则称v_i和v_j是连通的。如果图中任意两点都是连通的，那么这个图是连通图。如果G是无向图，如果存在从节点v_i到节点v_j的路径，那么也必然存在从v_j到v_i的路径。

强连通图：对于有向图 G，若对于任意两个不同的节点v_i和v_j，都存在从v_i到v_j以及从v_j到v_i的路径，则称 G是强连通图。

弱连通图：对于有向图 G，将所有的有向边替换为无向边，所得到的图称为原有向图的基图。如果一个有向图的基图是连通图，则该有向图是弱连通图。



问题：左图是强连通图还是弱连通图？

➢平均路径长度L

- **两个节点之间的距离**。假定d_{ij}表示任意两个节点v_i与v_j的距离，两个节点之间通常有多条边，每条路径的距离长度通常也不同，d_{ij}指的是v_i与v_j之间最短路径的距离。
- **网络直径**。d_{ij}在整个网络的最大值称为网络的直径。
- 假定每条边的距离都是1，则距离d_{ij}就是这两个节点之间最短路径上边的个数。简单地说，要计算网络直径，可利用遍历的方法计算每个节点到其他节点的最短路径，则其最大值就是网络直径。

➢平均路径长度L

- **平均路径长度**。网络的平均路径长度是网络中两个节点之间路径长度d_{ij}对整个网络的平均，即:

$$L=\frac{\sum_{i>j}d_{ij}}{N(N-1)/2}$$

- 式中的分子是对所有节点对之间的距离求和，分母是网络中节点对的个数。

➢网络密度

节点的度：对于无向网络，一个节点的度是与其相连的边的数目

$$k_i = \sum_{j=1}^{N} a_{ij}$$

对于有向网络，还有出度和入度之分。出度：该节点指向其它节点的边数；入度：其它节点指向该节点的边数。有向网络的度是出度与入度之和。

平均度。网络中所有节点的度的平均值称为网络节点的平均度，记为$\langle k \rangle$。

网络密度。网络密度是网络中存在的边数与可能存在的最大边数的比值。对于一个节点数为N的无向网络，可能存在的最大边数为$N(N-1)/2$，已存在边数为$M=|E|$, 则网络密度为：

$$D = \frac{2M}{N(N-1)}$$

对于节点数为N的有向网络，可能存在的最大有向边数为$N(N-1)$，已存在有向边数为$M=|E|$，则网络密度为：

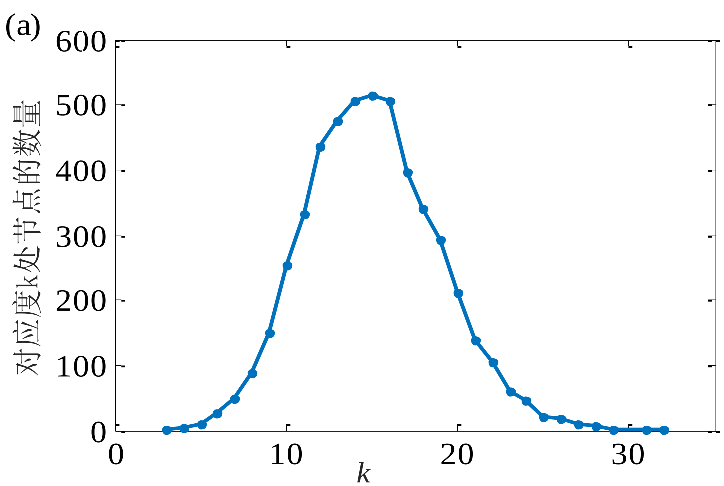
$$D = \frac{M}{N(N-1)}$$

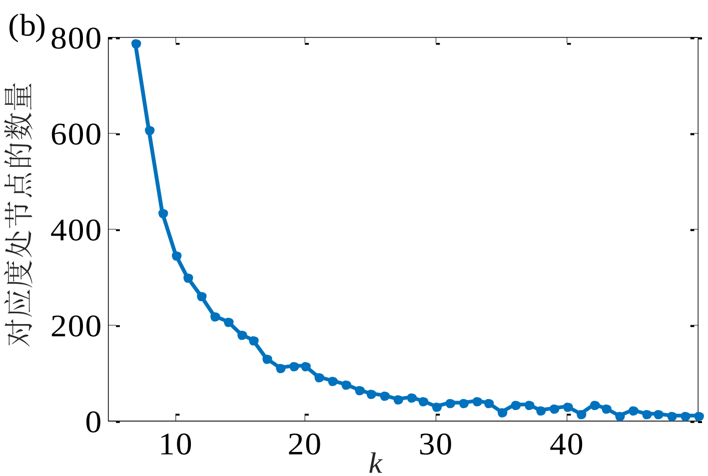
➢度分布

度分布。度分布是指网络中具有各种度值的节点数目分布情况，也可以理解成任意给定一个节点v_i, 它的度k_i等于k的概率。 度分布通常用度函数$P(k_i)$来表示。ER随机网络的度分布是泊淞(Poisson)分布，即：

$$P(k_i = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

可以看到随着度的增大，其存在的概率是指数下降的。图(a)为一个由ER随机图模型(N=5000,p=0.003)生成的服从泊淞分布(λ=Np=15)的度分布图。



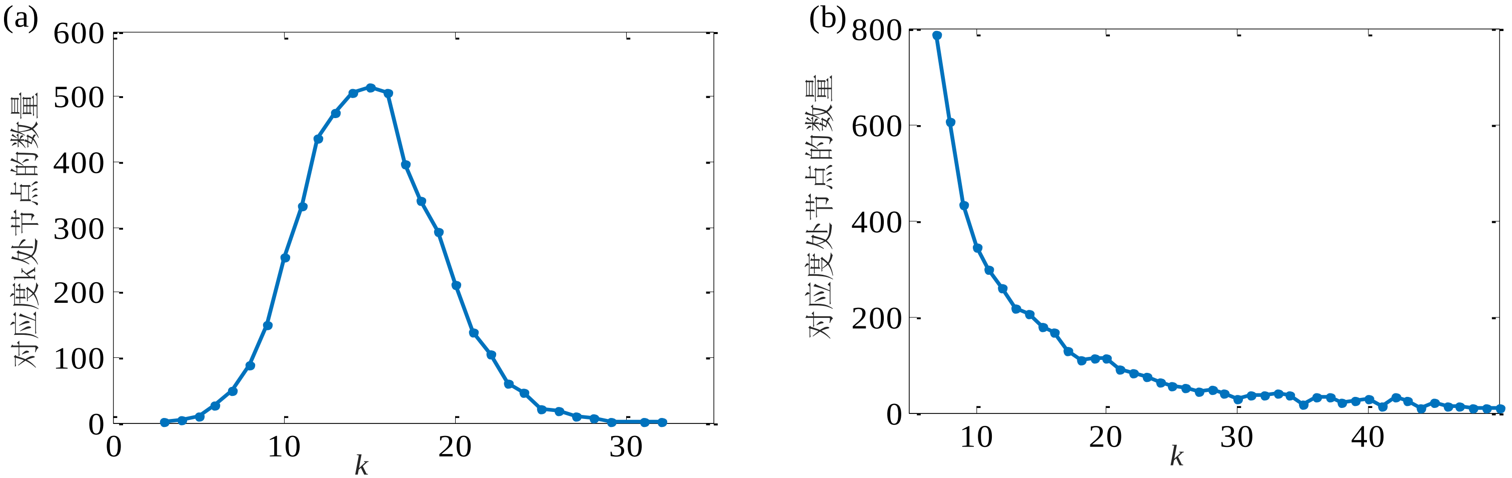


➢度分布

度分布。然而，近几年来的研究发现，在许多实际网络中随着度的增加，节点存在的概率不是指数下降的，度分布不是服从泊淞分布，而是服从幂律分布，即：

$$P(k_i = k) \propto k^{-\gamma}$$

按照幂律分布，随着度的增加其节点存在的概率下降要比指数下降缓慢的多。图(b)为一个由社区网络模型（N=5000,γ=2）生成的服从幂律分布的度分布图。



➢ 节点的聚类系数

- 节点的聚类系数：是表示一个网络中节点聚集程度的系数。
  - 在一个网络中，如果节点i的度为k, 即节点i有k个邻居，这k个邻居之间也可能相连，连边的多少体现了节点的聚集程度。
  - 在无向网络中，聚类系数（CC）的计算公式如下（以节点示i为例）：

$$C_i = \frac{e_i}{k_i(k_i-1)/2}$$

式中，分子表示节点i的所有相邻节点之间相互连接的边的个数，分母是邻居节点所有可能的连边数。

整个网络的聚集系数是所有节点的聚集系数的平均值。

➤节点的聚类系数

例如，社会网络中总是存在熟人圈或朋友圈，其中每个成员都认识其他成员。集聚程度反映网络集团化的程度；这是一种网络的内聚倾向。集团化概念反映的是一个大网络中各集聚的小网络分布和相互联系的状况。例如，它可以反映这个朋友圈与另一个朋友圈的相互关系。

➢节点中心性

中心性(Centrality)指标用来衡量一个节点的相对重要性。

度中心性（Degree Centrality）。就是直接以节点的度作为节点的中心性指标，度越大就意味着这个节点的度中心性越高，该节点在网络中就越重要。

介数中心性：顶点u的介数含义为网络中所有的最短路径之中，经过u的数量。它反映了顶点u的重要程度。

比如在交通网络中，要找到哪个节点对整个交通网络比较重要，就可以根据介数中心性找到介数最大的那个节点。

➢节点中心性

中心性(Centrality)指标用来衡量一个节点的相对重要性。

度中心性（Degree Centrality）。就是直接以节点的度作为节点的中心性指标，度越大就意味着这个节点的度中心性越高，该节点在网络中就越重要。

介数中心性：顶点u的介数含义为网络中所有的最短路径之中，经过u的数量。它反映了顶点u的重要程度。

比如在交通网络中，要找到哪个节点对整个交通网络比较重要，就可以根据介数中心性找到介数最大的那个节点。

近性中心性。用一个节点到其它节点的平均距离来衡量节点的中心性，距离越近则中心性越高，距离越远则中心性越低。

➢节点中心性

- **特征向量中心性** (Eigenvector Centrality)。度中心性完全由节点的度决定，与邻居的中心性值没有关系。实际上，即使度相同的两个节点A和B，A的邻居比B的邻居更重要，在有理由认为A比B更重要。特征向量中心性既考虑节点本身的重要性，同时也考虑其邻居的重要性，这可以通过使节点的中心性值通过边向邻居节点传播来实现，即：

$$x_j = \frac{1}{\lambda}\sum_{j=1}^{N} a_{ij}x_i$$

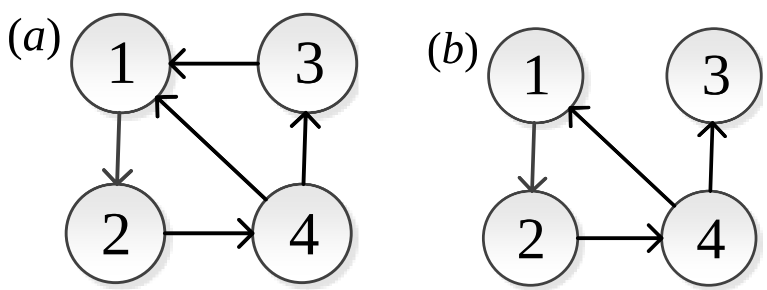
- 其中，为网络的邻接矩阵，x_i为节点v_i的中心性值，λ是一个常数。 用$X=(x_1,\ldots, x_N)^T$ 表示节点的中心性值向量，假设经过多次迭代后，各个节点的中心性值收敛且不再改变，即： $\lambda X = AX$
- 可见，收敛后的X是邻接矩阵A的特征向量，λ为最大特征值。

➢ 节点中心性

- **PageRank**. PageRank是Google用于评价网页的重要性的一种方法，目前很多的网页分析算法都是在PageRank的基础上衍生出来的。与特征向量中心性类似，PageRank既考虑了网页的入链数量，也考虑了入链网页的质量。计算上也是首先赋予各个网页初始的*PR*值，通过不断迭代更新各个网页的*PR*值。
- 为了表述方便，这里把网页看作节点，网页之间的链接关系用邻接矩阵$A=[a_{ij}]$描述，$a_{ij}=1$ 代表网页i有指向网页j的链接，否则$a_{ij}=0$。如果网页i初始*PR*值为$PR(i)$，网页i有$L(i)$条出向链接，即出度为$L(i)$，则更新后网页i将它的*PR*值平均分配给它指向的$L(i)$个网页。更新后网页i的*PR*值则是链入网页分配给它的*PR*值之和，即：
- $$PR(j)=\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)}$$

➢节点中心性

- **PageRank.** $$PR(j)=\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)}$$
- 例1：一个四节点的网络如图(a)所示，邻接矩阵为A, 各节点初始PR值都是1/4，求稳定后四个节点的PR值。



解：则四个节点的PR值更新规则为右式

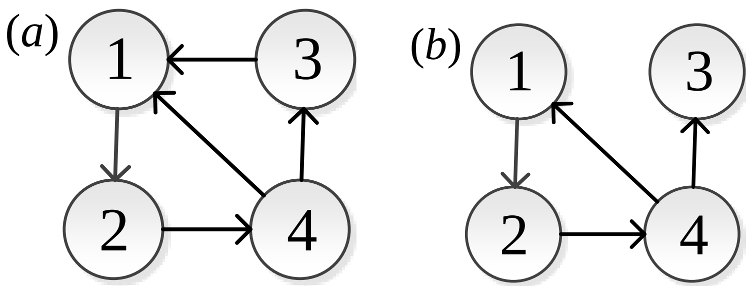
$$\begin{cases} PR(1)=\dfrac{PR(3)}{L(3)}+\dfrac{PR(4)}{L(4)} \\ PR(2)=\dfrac{PR(1)}{L(1)} \\ PR(3)=\dfrac{PR(4)}{L(4)} \\ PR(4)=\dfrac{PR(2)}{L(2)} \end{cases}$$

稳定后的各节点PR值为：
$PR(1)= PR(2)= PR(4) =2/7$，$PR(3)=1/7$

➤节点中心性

$$PR(j)=\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)} \qquad (2-14)$$

- 例1：一个四节点的网络如图(a)所示，邻接矩阵为A, 各节点初始PR值都是1/4，求稳定后四个节点的PR值。



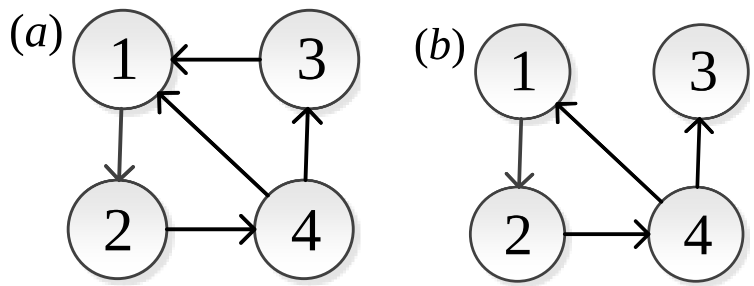
$$\begin{cases} PR(1)=\dfrac{PR(3)}{L(3)}+\dfrac{PR(4)}{L(4)} \\ PR(2)=\dfrac{PR(1)}{L(1)} \\ PR(3)=\dfrac{PR(4)}{L(4)} \\ PR(4)=\dfrac{PR(2)}{L(2)} \end{cases}$$

由图(*a*)可以看出，迭代公式给出的*PR*值更新规则只是把各个节点的*PR*值在各个节点上重新进行分配，*PR*值的总量并没有改变。这种更新规则在所有节点都有出边的情况下是可行的，如果部分节点没有出边就会有问题。如图(*b*)所示，节点3没有出边，可以验证（作业题2），无论初始*PR*值是多少，经过迭代稳定后，各个节点的*PR*值都是0。而没有出边的网页是普遍存在的，因此，必须对迭代公式（2-14）进行改进。

➢节点中心性

$$PR(j)=\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)} \tag{2-14}$$

- 例1：一个四节点的网络如图(a)所示，邻接矩阵为A, 各节点初始PR值都是1/4，求稳定后四个节点的PR值。



$$\begin{cases} PR(1)=\dfrac{PR(3)}{L(3)}+\dfrac{PR(4)}{L(4)} \\ PR(2)=\dfrac{PR(1)}{L(1)} \\ PR(3)=\dfrac{PR(4)}{L(4)} \\ PR(4)=\dfrac{PR(2)}{L(2)} \end{cases}$$

由图(*a*)可以看出，迭代公式给出的*PR*值更新规则只是把各个节点的*PR*值在各个节点上重新进行分配，*PR*值的总量并没有改变。这种更新规则在所有节点都有出边的情况下是可行的，如果部分节点没有出边就会有问题。如图(*b*)所示，节点3没有出边，可以验证（作业题2），无论初始*PR*值是多少，经过迭代稳定后，各个节点的*PR*值都是0。而没有出边的网页是普遍存在的，因此，必须对迭代公式（2-14）进行改进。

➢节点中心性

- 式(2-14)的主要问题在于对没有出边的节点，因其PR值不能输出而损失掉了，迭代后总的*PR*值有损失，而不是无损失的重新分配，稳定后各个节点的*PR*值都变成了0。实际上人们访问没有出向链接的网页，可以键入新的网址而跳转到新的网页，每个网址都有可能被键入而访问。因而，一种改进方式是对没有出向链接的网页，认为其可以1/*N*概率跳转到*N*个网页中任何一个网页。改进后的*PR*值更新规则如(2-15)所示：

$$PR(j)=\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)}+\sum_{i=1}^{N}\frac{a_{i\times}PR(i)}{N} \qquad (2-15)$$

- 式中，$a_{i\times}$=1表示节点v_i没有出边。PageRank设计的初衷是形成所有网页的一个概率分布，所以，所有网页的*PR*值之和是1。

➢节点中心性

- 实际上，人们访问每个网页，无论其有没有出向链接，都有可能键入新网址而访问任何一个网页，可以认为所有网页都有可能以概率1/N跳转到任何一个新网页，而不管其有无出向链接。这样就得到更新规则（2-16）：

$$PR(j) = d\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)} + (1-d)\sum_{\substack{i=1\\ i\neq j}}^{N}\frac{PR(i)}{N}$$
$$= d\sum_{i=1}^{N}\frac{a_{ij}PR(i)}{L(i)} + (1-d) \qquad (2-16)$$

- 式中d为阻尼系数，通常d=0.85。式(2-16)右侧第一项是考虑有链接指向节点v_j的情况，第二项是考虑除了节点v_j以外的所有节点均有可能随机跳转到v_j的情况。

➢节点中心性

- 作业题：
- 无向图有没有弱连通图和强连通图之分？说明原因。
- 如图2.4(b)所示的网页链接关系图，四个节点的初始PR值都是1，写出计算过程，验证经过公式(2-14)迭代收敛后的四个节点的PR值是多少？

西安电子科技大学
XIDIAN UNIVERSITY

谢谢