Chapter 5: 节点重要性与相似性

赵　磊

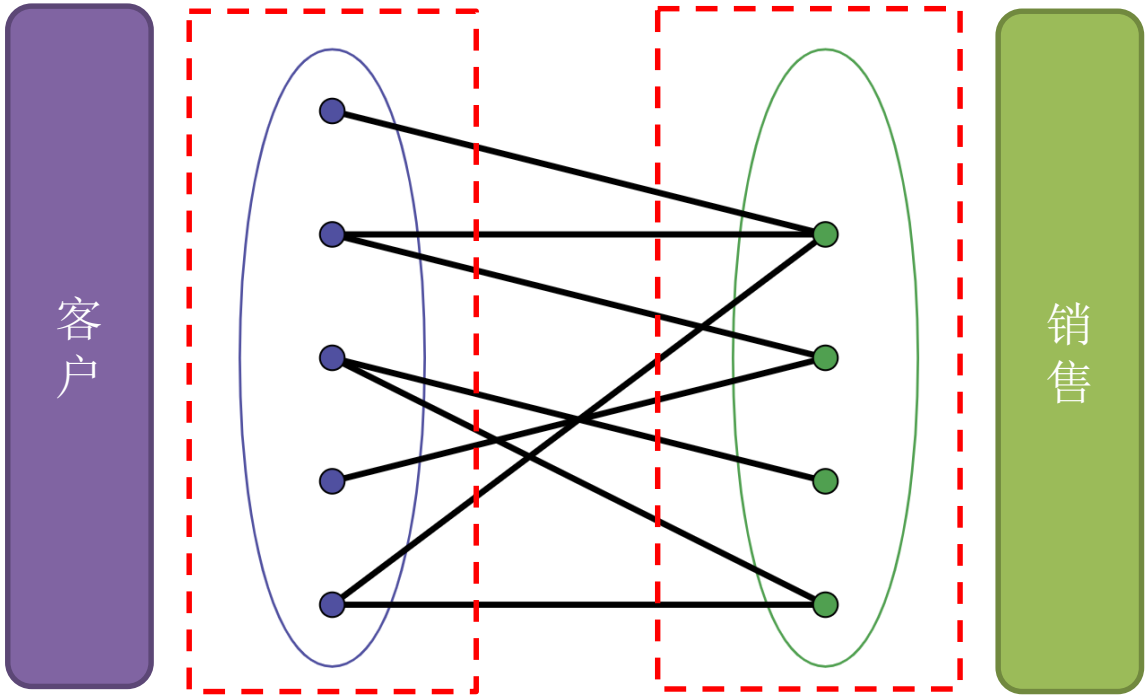
2020年11月19日

Outline

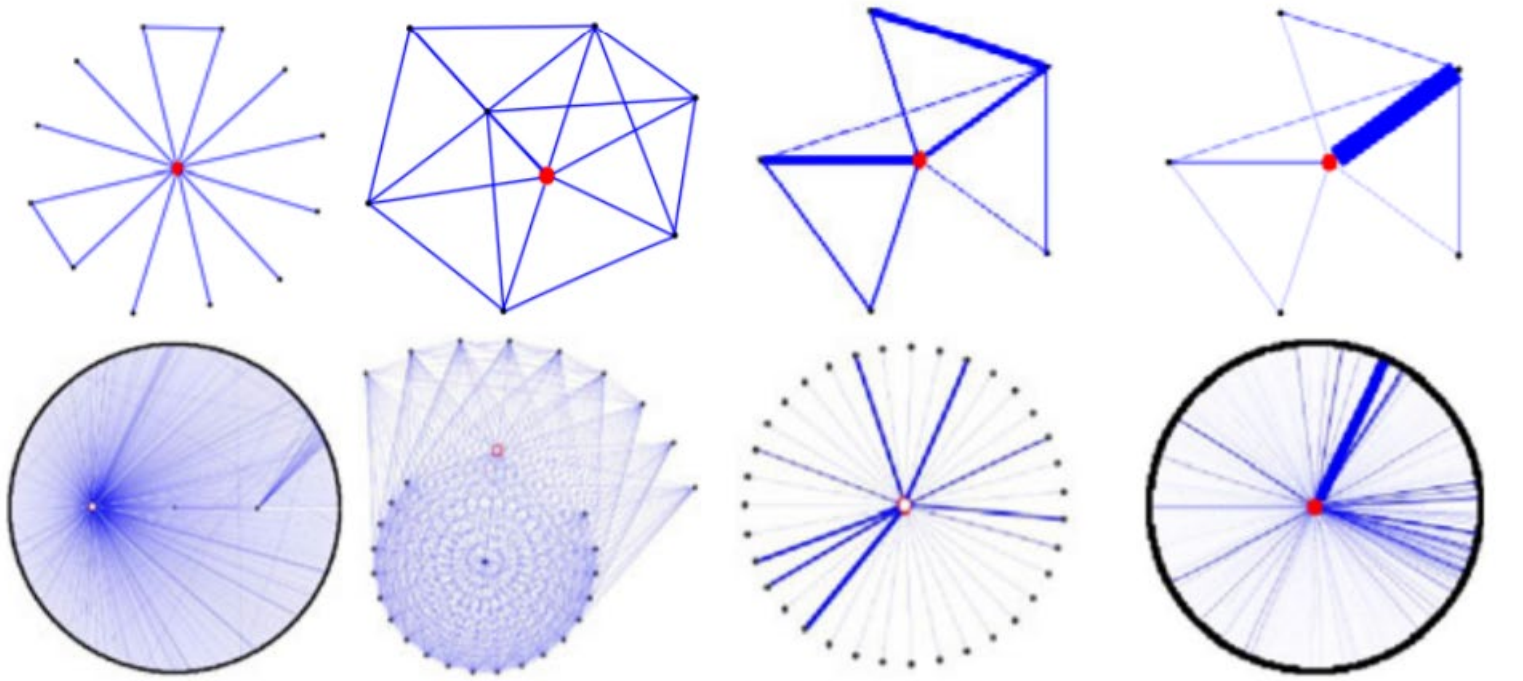
- 往期回顾
- 节点重要性刻画方法
- 节点相似性和链路预测

- **第二章：**图论概念，包括图的表示、路径和连通性、生成树和二分图匹配。

- **Problem：**



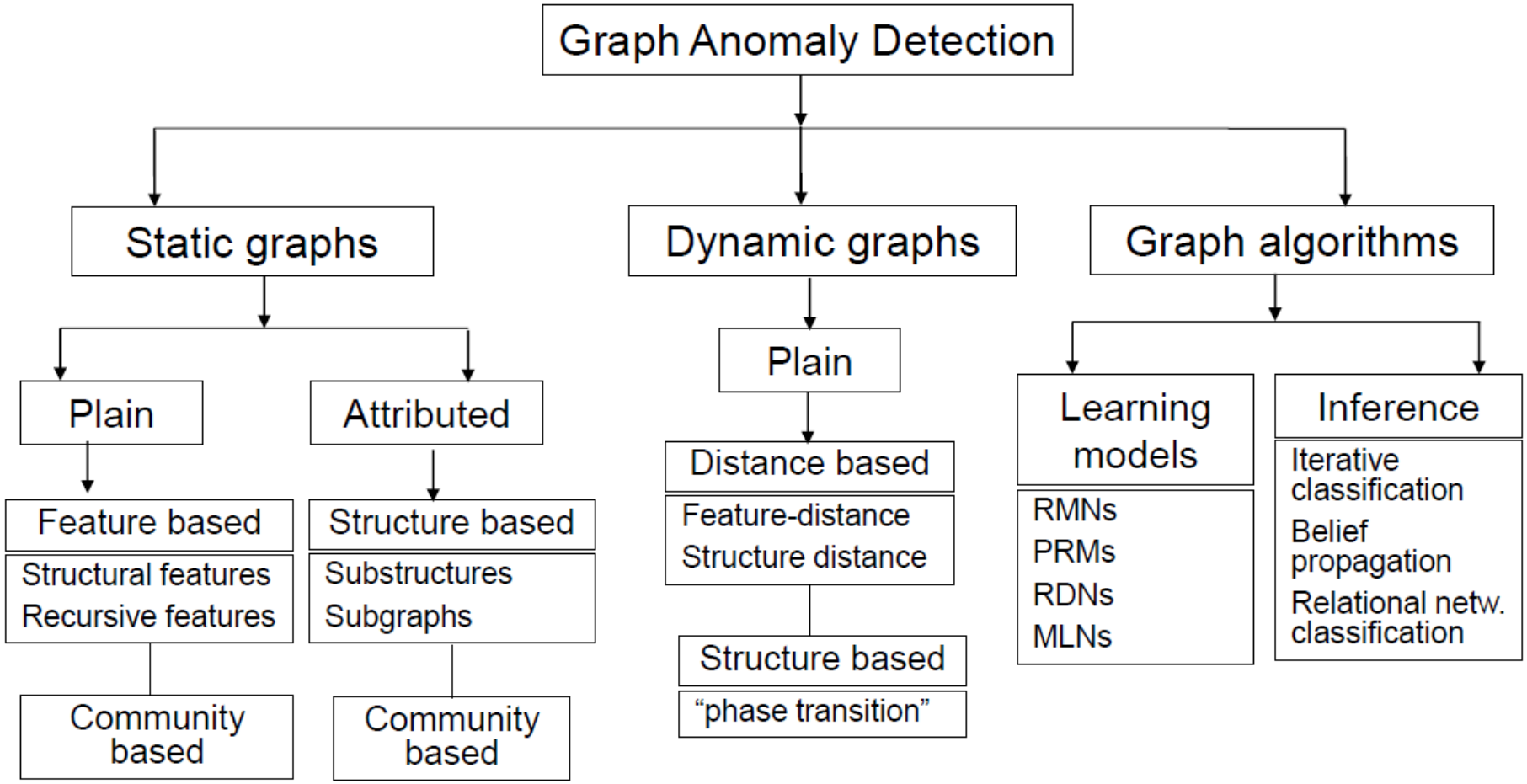
- **第三章：**网络的拓扑概念，连通性、稀疏性、路径长度、聚类系数和度分布。



(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

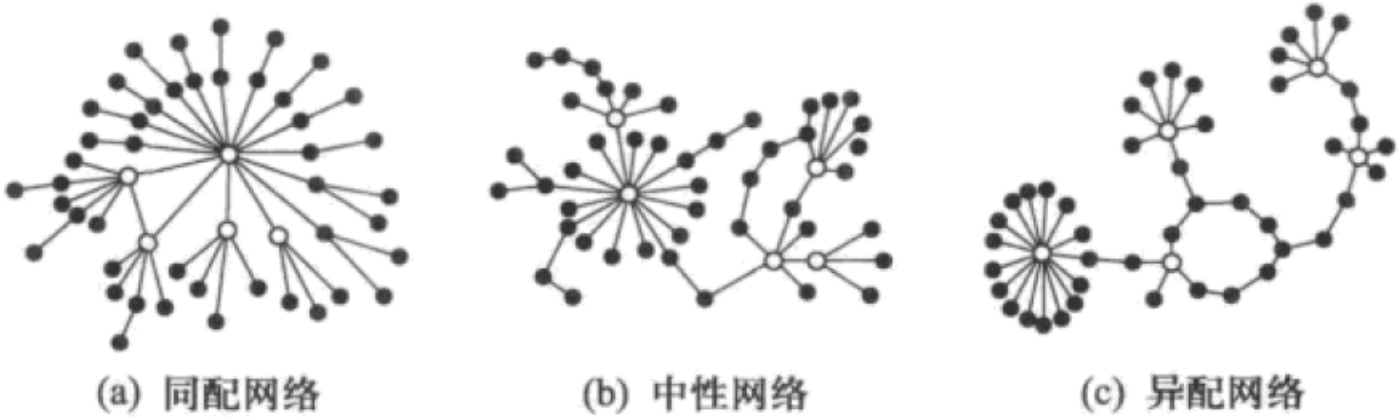
E vs N　　W vs E　　λ vs W

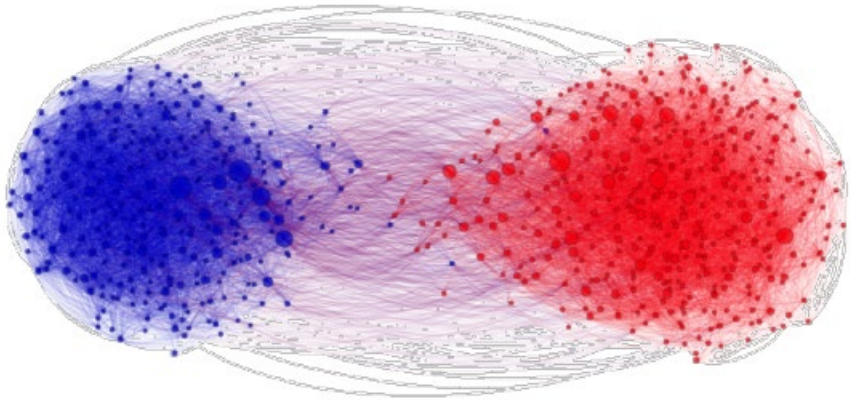
论文：Akoglu, Leman & McGlohon, Mary & Faloutsos, Christos. (2010). OddBall: Spotting Anomalies in Weighted Graphs. 410-421. 10.1007/978-3-642-13672-6_40.



- Graph-based Irregularity and Fraud Detection (ICDM'12)

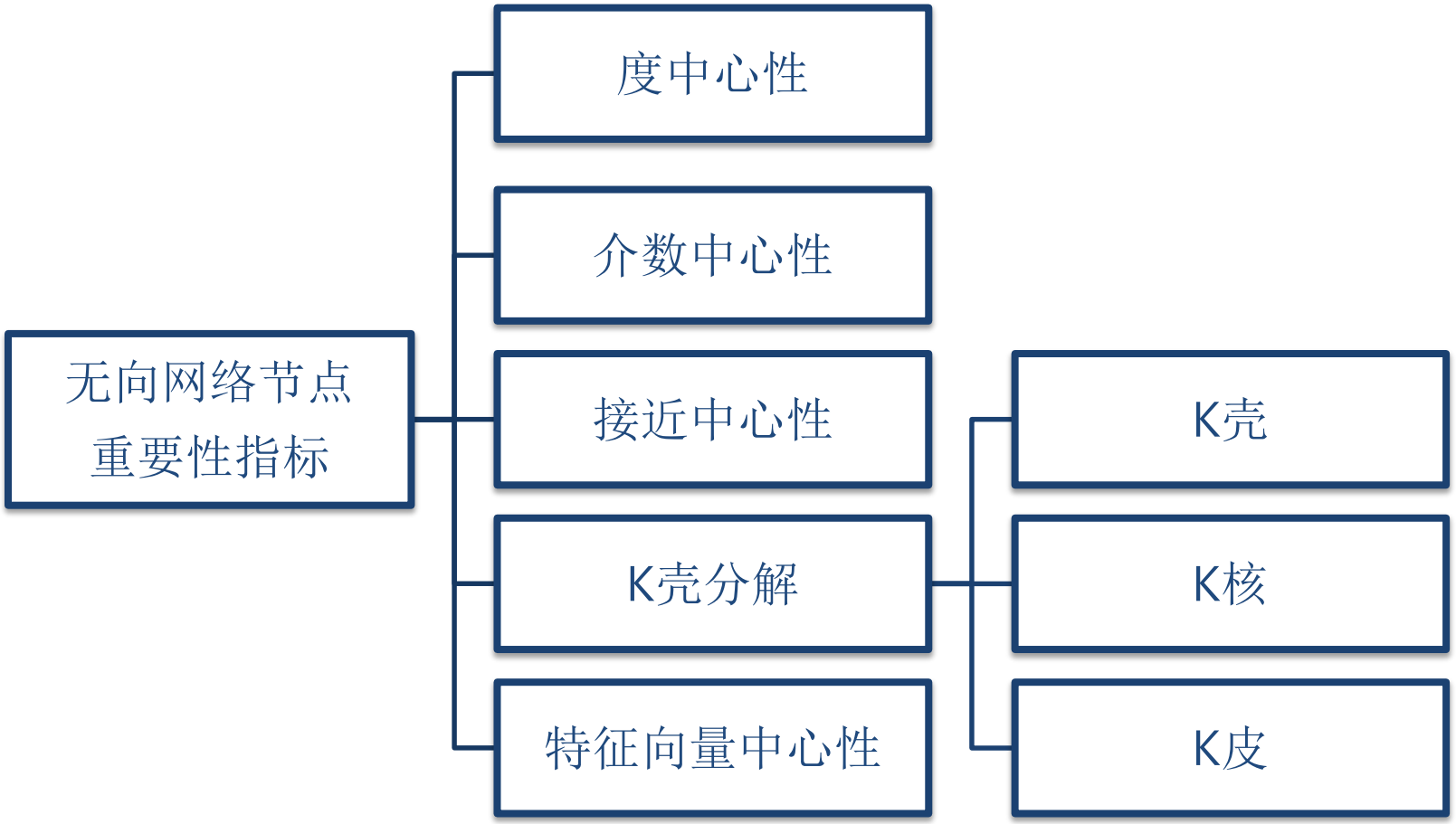
- **第四章：**进一步刻画网络的拓扑结构，引入高阶度分布、同配网络、社团结构。





- **无向网络节点重要性刻画：**度中心性、介数中心性、接近中心性、k-壳、k-核、k-皮、特征向量中心性。
- **有向网络节点重要性刻画：**HITS算法、PageRank算法。
- **节点相似性与链路预测：**基于局部信息、全局信息和随机游走的相似性指标。

无向网络节点重要性指标



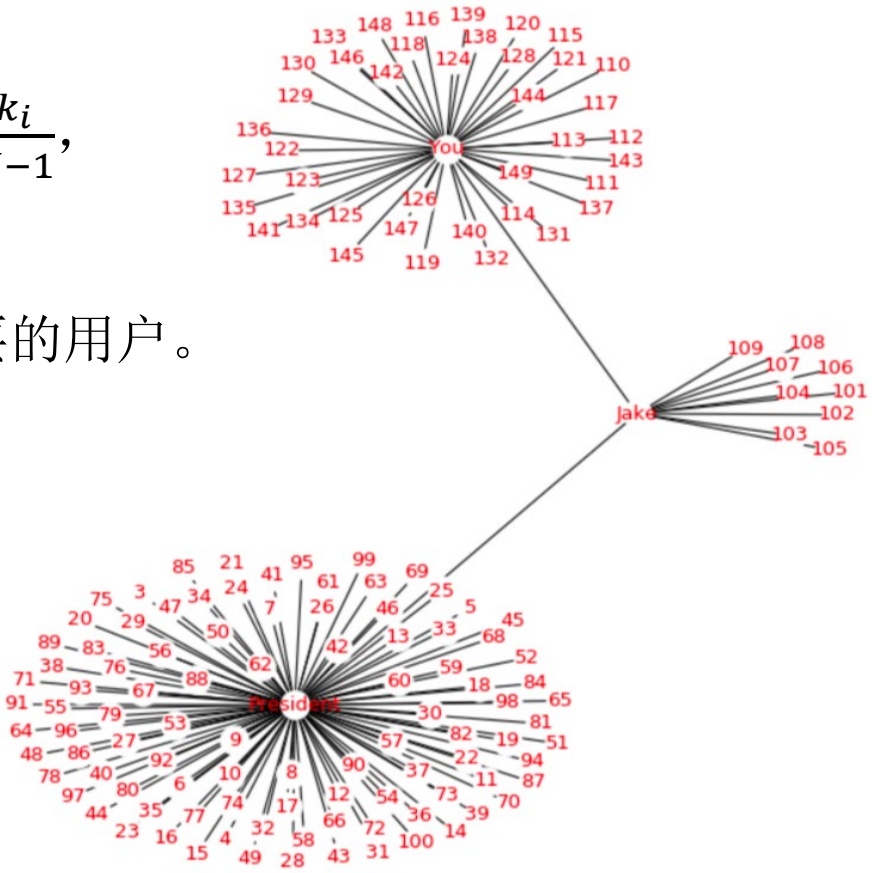
度中心性

- **度中心性**（Degree centrality）：

$$DC_i = \frac{k_i}{N-1},$$

其中，k_i为第i个节点的度。

- **Idea：**朋友数最多的用户是最重要的用户。

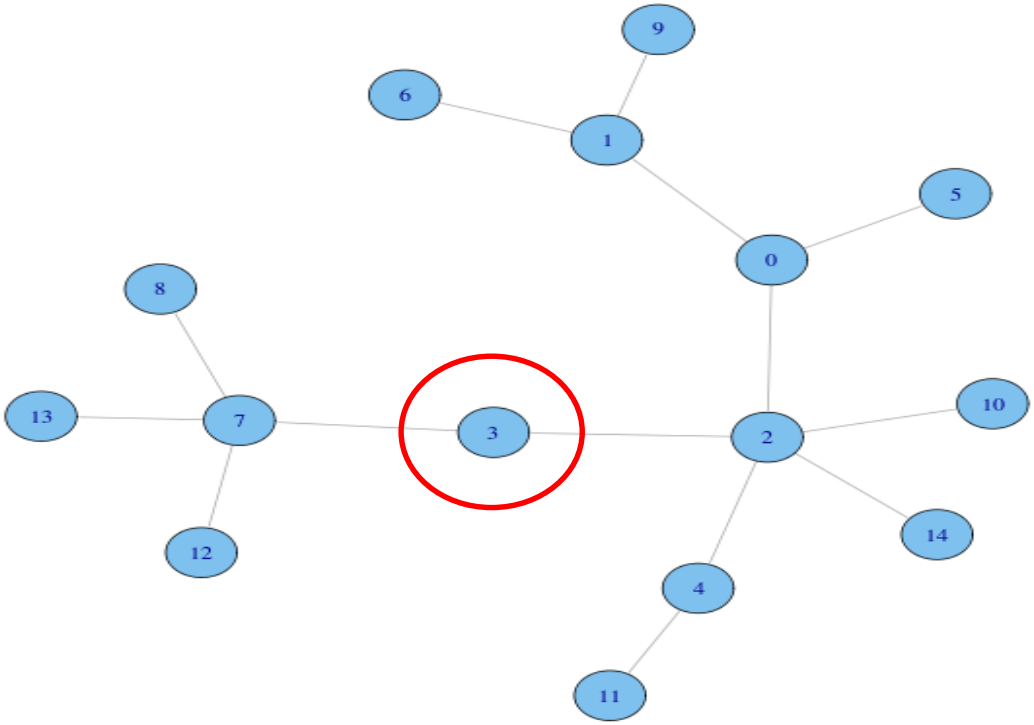


- **介数中心性**（Between centrality）：

$$BC_i = \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}},$$

其中，g_{st}为从节点s到t的最短路径数目，n_{st}^i为从节点s到节点t的g_{st}条最短路径中经过节点i的最短路径数目。

- **Idea:** 体现了对网络节点对之间沿着最短路径传输信息的控制能力。



介数中心性

- 介数中心性归一化的两种方式：

$$BC_i = \frac{1}{\binom{N-1}{2}} \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}},$$

or

$$BC_i = \frac{1}{N^2} \sum_{s,t} \frac{n_{st}^i}{g_{st}}.$$

- 尽管二者计算结果会有所不同，但是不会影响网络中节点按介数大小的排序结果，而后者通常才是我们关心的。

- **Open problem：** 节点介数的快速有效计算。

- **接近中心性**（Closeness centrality）：

$$CC_i = \frac{1}{d_i} = \frac{N}{\sum_{j=1}^{N} d_{ij}},$$

其中，d_{ij}为从节点i到j的距离，$d_i = \frac{\sum_{j=1}^{N} d_{ij}}{N}$节点$i$到网络中所有节点的距离的平均值。

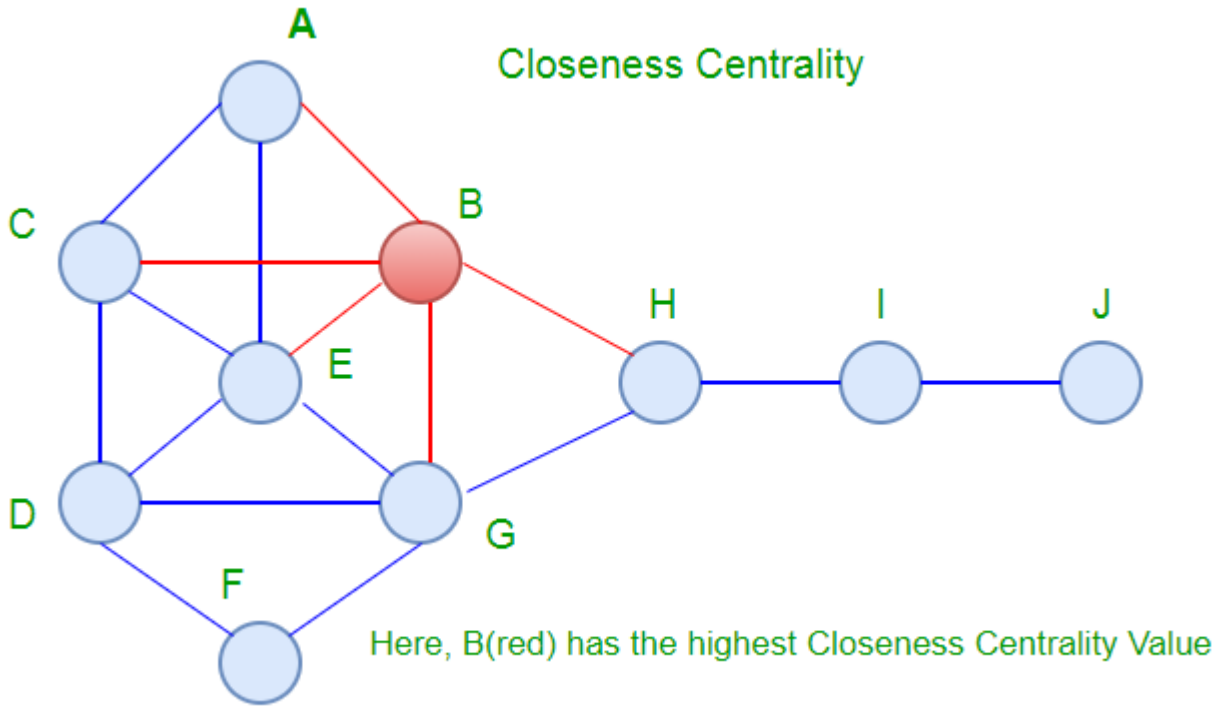
- 当网络不连通时，去掉与节点自身的距离，并将$\frac{\sum_{j=1}^{N} d_{ij}}{N}$改为调和平均数（Harmonic mean）。此时，接近中心性定义为调和平均数的倒数，此定义易受较小值影响：

$$HCC_i = \frac{1}{N-1}\sum_{j\neq i}\frac{1}{d_{ij}}.$$

- 由于$\frac{N-1}{\sum_{j\neq i}\frac{1}{d_{ij}}} \leq \frac{\sum_{j\neq i} d_{ij}}{N-1}$，即调和平均数≤等于算术平均数。据此，

$$HCC_i \geq \frac{N-1}{\sum_{j\neq i} d_{ij}} \approx CC_i.$$

- **Idea:** 介数最高的节点对于网络中信息的流动具有最大的控制力，而接近数最大的节点对于信息的流动具有最佳的观察视野。
- 如下图，介数最高的是节点H，接近数最大的是节点B和G。



- **K壳分解**（K-Shell decomposition）：

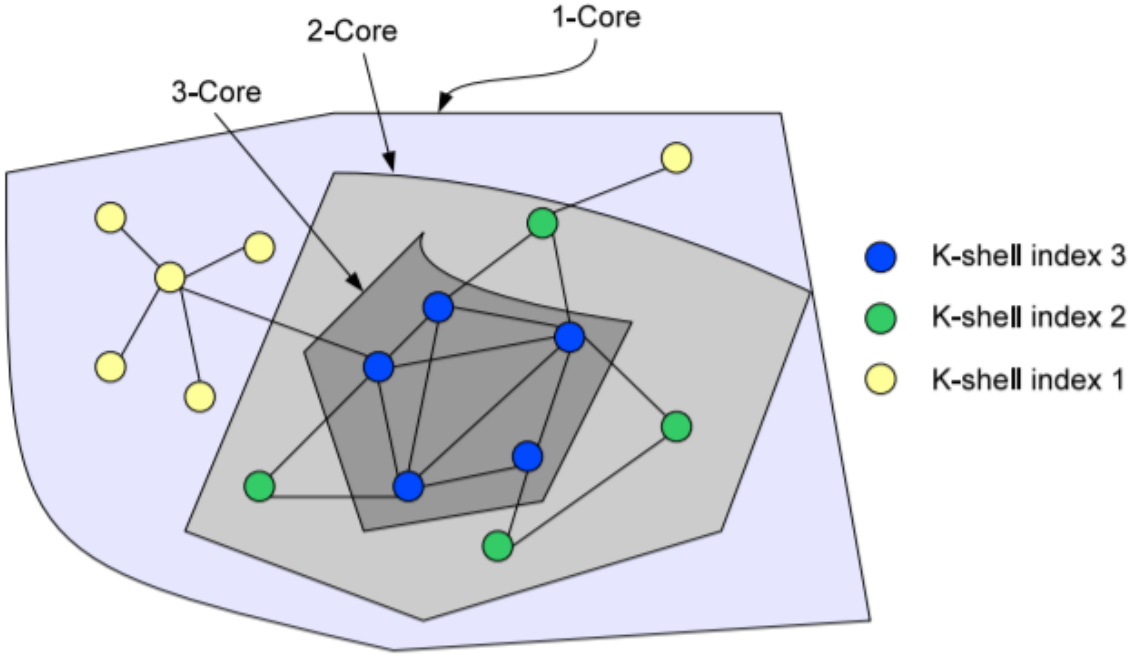
$$K_{sd}: \boldsymbol{V} \to \mathbb{N},$$

对于节点$v \in \boldsymbol{V}$，$K_{sd}(v) = k \in \mathbb{N}$。

- **1**）记$k\text{-}Shell = \{v \in \boldsymbol{V} | K_{sd}(v) = k\}$，那么有

$$\forall v \in k\text{-}Shell,\ Degree(v) \geq k.$$

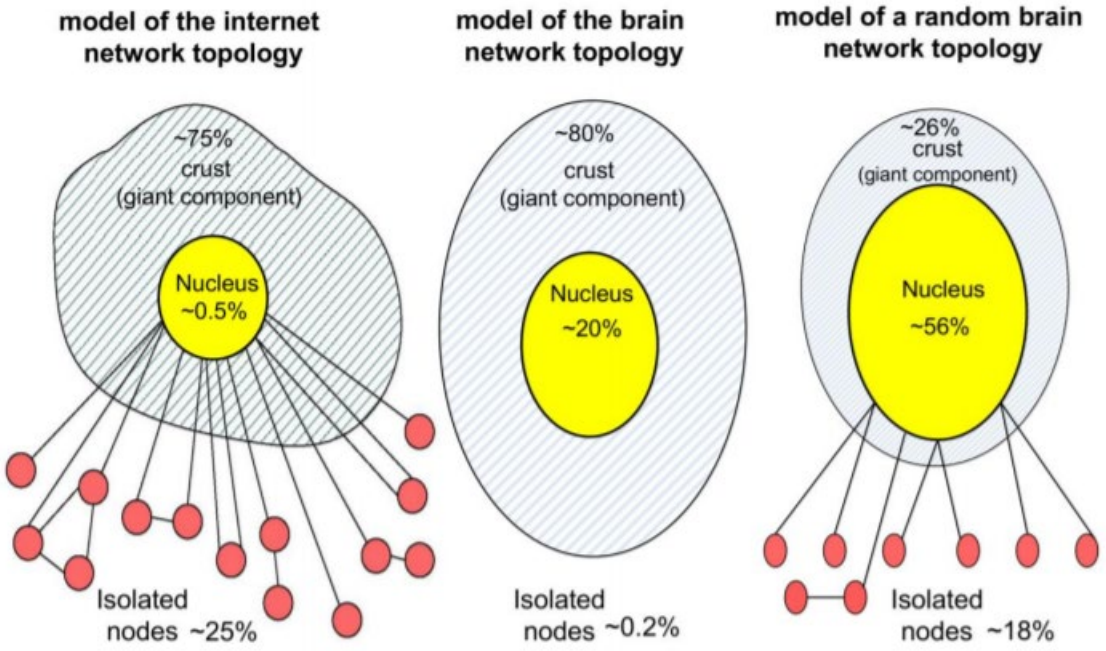
- **2**）度大的节点K_{sd}值不一定大（传染病模型）。
- **3**）$k\text{-}Core = \{v \in \boldsymbol{V} | K_{sd}(v) \geq k\},\ k\text{-}Crust = \{v \in \boldsymbol{V} | K_{sd}(v) \leq k\}.$



- **核心**（Nucleus）：$k_{max}(\boldsymbol{V})-Shell$，其中
  $$k_{max}(\boldsymbol{V}) = \max\{k \in \mathbb{N} \mid k-Shell\ of\ \boldsymbol{V}\ is\ not\ empty\}.$$

- **对等联通片**（Peer-connected component）：$[k_{max}(\boldsymbol{V})-1]-Crust$的最大联通片。

- **孤立片**（Isolated component）：$[k_{max}(\boldsymbol{V})-1]-Crust$中不属于最大联通片的节点的集合。



- **特征向量中心性**（Eigenvector centrality）：基本思想，一个节点的重要性也取决于其邻居节点的重要性。记x_i为节点i的重要性度量值，那么应该有

$$x_i = c\sum_{j=1}^{N} a_{ij}x_j,$$

其中，c为非0常数，$\boldsymbol{A} = (a_{ij})$为网络的邻接矩阵。记$\boldsymbol{x} = [x_1, \dots, x_N]^T$,可写为

$$\boldsymbol{x} = c\boldsymbol{A}\boldsymbol{x},$$

此式蕴含，$\boldsymbol{x}$为$\boldsymbol{A}$对应c^{-1}特征值的特征向量。故特征向量中心性得名。

- **Proof：** 给定初值$\boldsymbol{x}(0)$，采用迭代格式

$$\boldsymbol{x}(k) = c\boldsymbol{A}\boldsymbol{x}(k-1)，\ k = 1,2,\dots$$

分析迭代算法收敛性，由于矩阵A是对称阵，可以得到其特征值$|\lambda_1| \geq \cdots \geq |\lambda_N|$，其对应的特征向量为$\boldsymbol{v}_1,\dots,\boldsymbol{v}_N$，其构成了$\mathbb{R}^N$上的一组基，于是

$$\boldsymbol{x}(0) = \sum_{i=1}^{N} \gamma_i \boldsymbol{v}_i,$$

代入迭代格式得

$$\boldsymbol{x}(k) = c^k \sum_{i=1}^{N} \gamma_i \lambda_i^k \boldsymbol{v}_i,$$

取$c = \lambda_1^{-1}$，令$k \to \infty$，得$\boldsymbol{x}(k) = \sum_{i=1}^{N} \gamma_i \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{v}_i \to \gamma_1 \boldsymbol{v}_1.$

- **Problem：** 非0常数c可以任取吗？

- **乘幂法（求解矩阵主特征值和特征向量问题）：**
  - **假设1：**矩阵$\boldsymbol{A}$的有完备的特征向量系，记为$\boldsymbol{v_1}, \dots, \boldsymbol{v_N}$。
  - **假设2：**矩阵$\boldsymbol{A}$模最大的特征根为单根，即
$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_N|,$$
其中，λ_i对应的特征向量为$\boldsymbol{v_i}$。

【乘幂法方法分析】

- **1**）设$\boldsymbol{x}(0)$为初始化非零向量，则$\boldsymbol{x}(0)=\sum_{i=1}^{N}\gamma_i\boldsymbol{v}_i$，令

$$\boldsymbol{x}(k)=\boldsymbol{A}\boldsymbol{x}(k-1),\ k=1,2,\ldots$$

则

$$\boldsymbol{x}(k)=\boldsymbol{A}^k\boldsymbol{x}(0)=\sum_{i=1}^{N}\gamma_i\lambda_i^k\boldsymbol{v}_i。$$

- **2**）从而有

$$\boldsymbol{x}(k)_j=\sum_{i=1}^{N}\gamma_i\lambda_i^k\boldsymbol{v}_{ij},\ \frac{\boldsymbol{x}(k+1)_j}{\boldsymbol{x}(k)_j}=\frac{\sum_{i=1}^{N}\gamma_i\lambda_i^{k+1}\boldsymbol{v}_{ij}}{\sum_{i=1}^{N}\gamma_i\lambda_i^k\boldsymbol{v}_{ij}},$$

假设$\gamma_1\neq 0$，$\boldsymbol{v}_{1j}\neq 0$，则有

$$\frac{\boldsymbol{x}(k+1)_j}{\boldsymbol{x}(k)_j}=\lambda_1\frac{1+\sum_{i=2}^{N}\frac{\gamma_i\boldsymbol{v}_{ij}}{\gamma_1\boldsymbol{v}_{1j}}\left(\frac{\lambda_i}{\lambda_1}\right)^{k+1}}{1+\sum_{i=2}^{N}\frac{\gamma_i\boldsymbol{v}_{ij}}{\gamma_1\boldsymbol{v}_{1j}}\left(\frac{\lambda_i}{\lambda_1}\right)^{k}}\rightarrow\lambda_1.$$

【乘幂法方法分析】

- **3**）上述假设$\boldsymbol{v}_{1j} \neq 0$，很容易满足，因为$\boldsymbol{v}_1$是特征向量，必然存在$\boldsymbol{v}_{1j} \neq 0$，而$\gamma_1 \neq 0$不一定总是满足，事实上如果随机选取初始向量，那么$\gamma_1 = 0$的概率为0。

- **4**）观察到$\boldsymbol{x}(k) = \sum_{i=1}^{N} \gamma_i \lambda_i^k \boldsymbol{v}_i = \gamma_1 \lambda_1^k \left[\boldsymbol{v}_1 + \sum_{i=1}^{N} \frac{\gamma_i}{\gamma_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \boldsymbol{v}_i \right]$，当$k$充分大时，$\boldsymbol{x}(k) \approx \gamma_1 \lambda_1^k \boldsymbol{v}_1$。
  - 若$|\lambda_1| > 1$，当$k \to \infty$时，$\boldsymbol{x}(k)$的分量会趋于无穷；
  - 若$|\lambda_1| < 1$，当$k \to \infty$时，$\boldsymbol{x}(k)$的分量会趋于0；
  - 因此，需要在每一步将$\boldsymbol{x}(k)$规范化：

$$\begin{cases} \boldsymbol{y}(k) = \boldsymbol{A}\boldsymbol{x}(k-1) \\ \boldsymbol{x}(k) = \dfrac{\boldsymbol{y}(k)}{\|\boldsymbol{y}(k)\|_\infty} \end{cases}.$$

- **乘幂法（求解矩阵主特征值和特征向量问题）：**
  - **假设1：**矩阵$\boldsymbol{A}$的有完备的特征向量系，记为$\boldsymbol{v_1, \dots, v_N}$。
  - **假设2：**矩阵$\boldsymbol{A}$模最大的特征根为单根，即
$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_N|,$$
其中，λ_i对应的特征向量为$\boldsymbol{v_i}$。

注：假设2不满足时，向量收敛到主特征值对应的特征空间的任意一个特征向量。

有向网络节点重要性指标

- **权威性**（Authority）：
  - 复旦大学主页、高引用论文、权威专家。
  - 节点的权威性由指向它的节点的枢纽性度量。
- **枢纽性**（Hub）：
  - hao123、综述文章、市长司机。
  - 节点的枢纽性由它指向的节点的权威性度量。

- **HITS算法**（John Kleinberg1997）：
  - 网络中节点的authority值和hub值分别$\boldsymbol{x}$和$\boldsymbol{y}$，二者皆为N维列向量，即$\boldsymbol{x} = [x_1, \dots, x_N]^{\boldsymbol{T}}$，$\boldsymbol{y} = [y_1, \dots, y_N]^{\boldsymbol{T}}$。
  - 第k次迭代：

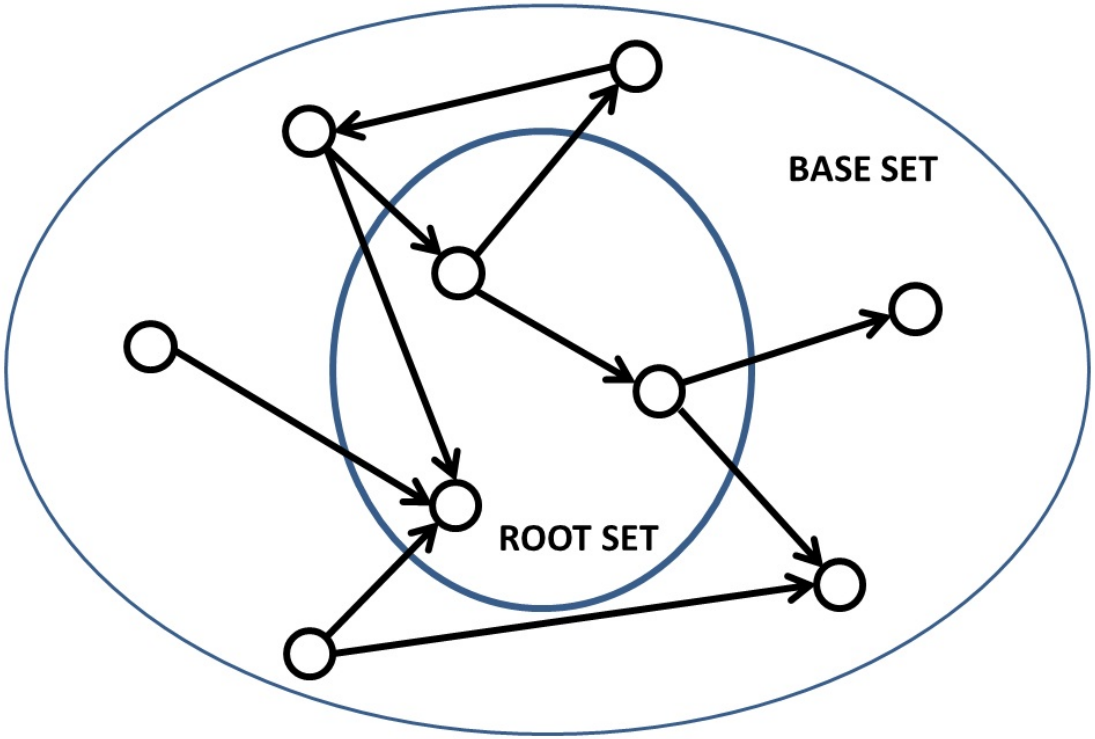
$$x_i(k) = \sum_{j=1}^{N} a_{ji} y_j(k-1)，\ y_i(k) = \sum_{j=1}^{N} a_{ij} x_j(k).$$

向量表示为：$\boldsymbol{x}(k) = \boldsymbol{A}^T \boldsymbol{y}(k-1)$，$\boldsymbol{y}(k) = \boldsymbol{A}\boldsymbol{x}(k)$，进而有

$$\boldsymbol{x}(k) = \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x}(k-1)，\ \boldsymbol{y}(k) = \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{y}(k-1).$$

  - 上述迭代过程即“乘幂法”，需要对每次迭代做向量标准化。

- **HITS算法在搜索引擎的应用：**
  - 先基于某种检索算法（如TF-IDF）找出与查询主题最相关的top-n页面作为root集合。
  - 在根集root的基础上，对root集合进行扩充形成集合base，凡与root集内网页有直接链接指向关系的网页都被扩充到集合base。
  - 在base集合内寻找authority值和hub值较高的页面。



- **HITS算法缺点：**
  - **主题漂移：**集合base中authority值和hub值较高的页面不一定与初始的查询主题有关。
  - **网页作弊：**首先人工生成一个高hub值的网页，再将该hub网页指向一个作弊网页，则作弊网页的authority值也很高。
  - **结构不稳定：**在base集合中改变少数页面/链接会使得算法排名结果产生很大改变。

- **基本思想：**
  - WWW上一个页面的重要性取决于指向它的页面的数量和质量。

- **PageRank算法**（Larry Page1996）：
  - **1**）初始步：初始化$PR_i(0)$，$i = 1, \dots, N$，$\sum_{i=1}^{N} PR_i(0) = 1$。
  - **2**）校正：$PR_i(k) = \sum_{j=1}^{N} a_{ji} \frac{PR_j(k-1)}{k_j^{out}}$，$i = 1, \dots, N$。

- 矩阵表示：设有向网络邻接矩阵$\boldsymbol{A} = (a_{ij})_{N\times N}$，定义$\overline{\boldsymbol{A}} = (\overline{a_{ij}})_{N\times N}$，

  其中，$\overline{a_{ij}} = \begin{cases} \frac{1}{k_i^{out}}, & i \to j \\ 0, & \text{否则} \end{cases}$，那么矩阵形式为：

$$\boldsymbol{PR}(k) = \overline{\boldsymbol{A}}^T \boldsymbol{PR}(k-1).$$

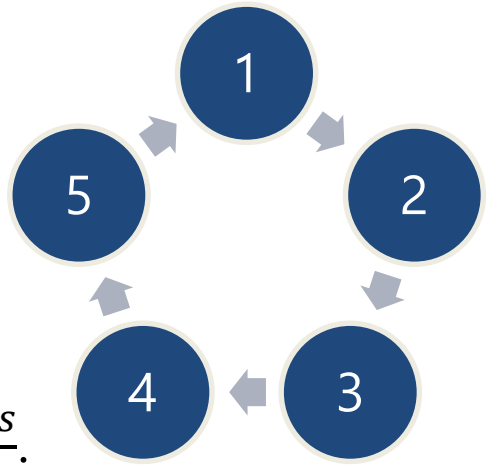
【乘幂法！！！】

- **缺陷一：**
  - 可能存在出度为0的节点，即悬挂节点（Dangling node）。
  - 解决方法：以相同概率$1/N$随机访问网络中的任一节点，即在$\overline{\boldsymbol{A}}$中的全零行替换为$1/N$。具体地，

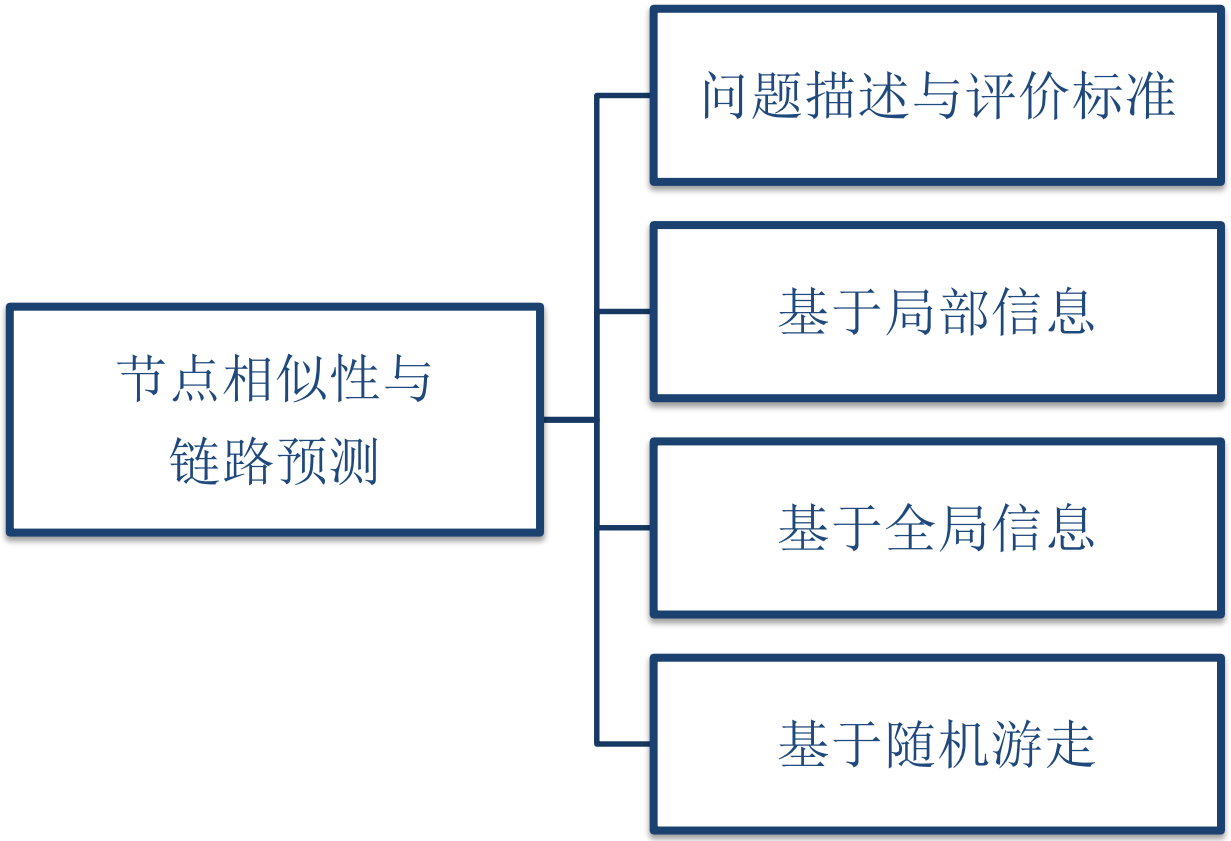
$$\overline{a_{ij}} = \begin{cases} 1/k_i^{out}, & i \to j \text{且} k_i^{out} > 0 \\ 0, & i \nrightarrow j \text{且} k_i^{out} > 0. \\ 1/N, & k_i^{out} = 0 \end{cases}$$

- **缺陷二：**
  - 可能存在周期性，即如右图。
  - 解决方法：

$$PR_i(k) = s\sum_{j=1}^{N} \bar{a}_{ji} PR_j(k-1) + \frac{1-s}{N}.$$



- **鲁棒性分析**：在保持每个节点的度值不变的情况下，通过随机重连机制生成具有相同度分布的不同网络，有如下结论：
  - 均匀的随机网络中节点的PR值排序对网络扰动较为敏感。
  - 非均匀的无标度网络中会涌现个别超稳定的PR值最大的节点，它们在按照PR值排序中的位置对于网络扰动具有很高的鲁棒性。


节点相似性与
链路预测
问题描述与评价标准
基于局部信息
基于全局信息
基于随机游走

- **问题描述：**
  - 节点的相似性，顾名思义。
  - 节点相似性分析可以用来做链路预测（Link prediction），即通过已知的各种信息预测给定网络中尚不存在连边的两个节点之间产生链接的可能性。
  - 这种预测，既包含了对未知链接（existing yet unknown link），也称丢失链接（missing link）的预测，也包含了对未来链接（future link）的预测。
  - 定义：对于任意的$x, y \in V, (x, y) \notin E$，赋予分数$S_{xy} = S\big((x, y)\big)$，并将所有未连接的节点对按照从大到小排序，分数越高，连边概率越大。

- **现实应用：**
  - 推断生物网络中，节点间是否存在相互作用关系，这些关系是通过大量实验得到的，依赖技术的进步和高额的实验成本。可以通过链路预测，指导实验从而提高实验的成功率、降低实验成本，以及纠正一些错误的虚假连边。
  - 预测演化网络中未来可能出现的链接，在社交网站中，基于当前的网络结构去预测哪些现在尚未结交的用户，并做推荐。

- **评价标准**：将连边集E分为训练集E^T和测试集E^P两部分，设E^U为不属于E的任意一对节点之间的可能连边的集合。

  - ***AUC***：从整体上衡量算法的精确度。

    1. Count = 0, FOR $i = 1:n$, do
    2. Randomly select e^P, e^U from E^P and E^U respectively
    3. IF $S(e^P) > S(e^U)$: Count += 1
    4. ELIF $S(e^P) = S(e^U)$: Count += 0.5
    5. END do
    6. Return $AUC = Count/n$

  - ***Precision***：只考虑排在前L位的边是否预测准确，即

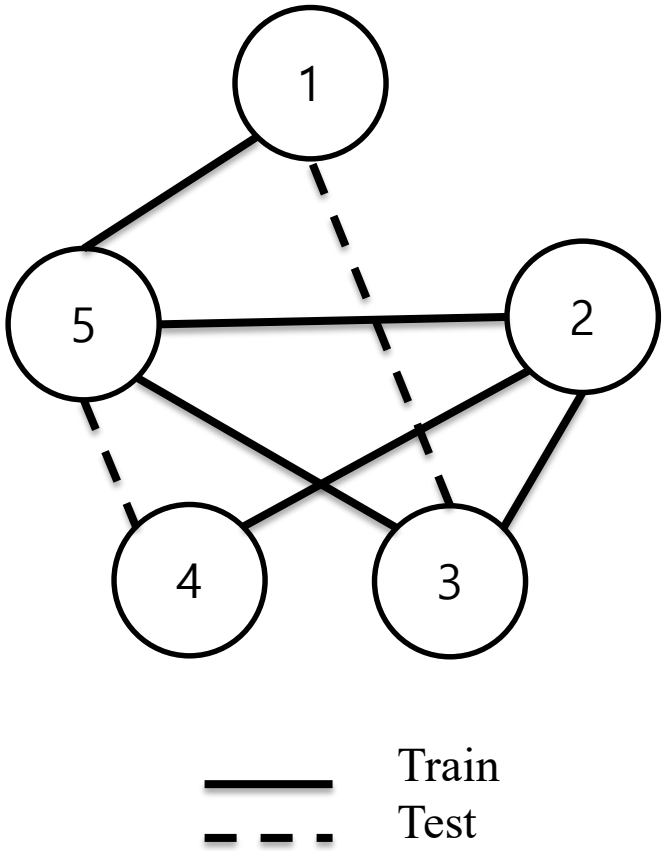
$$Precision = \frac{m}{L},$$

其中，m为排在前L位的边中出现在测试集中的个数。【*AUC*相同的情况下，*Precision*越大越好，倾向于将真正连边的节点对排在前面】

- **Example:**

$$AUC = \frac{1}{6}(1+1+1+0.5+0.5) = \frac{2}{3},\ \ Prec(L=2) = 0.5$$

连边	所属集合	S_{xy}
(1,2)	E^U	0.4
(1,3)	E^P	0.5
(1,4)	E^U	0.6
(1,5)	E^T	—
(2,3)	E^T	—
(2,4)	E^T	—
(2,5)	E^T	—
(3,4)	E^U	0.5
(3,5)	E^T	—
(4,5)	E^P	0.6



基于局部信息的节点相似性指标

- **思想：**两个节点的共同邻居（Common neighbors）的数量越多，这两个节点就越相似。

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|,$$

其中，$\Gamma(x)$为节点x的邻居节点的集合。

- 基于此，可以考虑共同邻居的相对数量，如下表，记$k(x) = |\Gamma(x)|$。

名称	定义	名称	定义
共同邻居(CN)	$S_{xy} = \|\Gamma(x) \cap \Gamma(y)\|$	Hub Depressed(HDI)	$S_{xy} = \frac{\|\Gamma(x) \cap \Gamma(y)\|}{\max\{k(x), k(y)\}}$
Salton	$S_{xy} = \frac{\|\Gamma(x) \cap \Gamma(y)\|}{\sqrt{k(x) \times k(y)}}$	LHN-I	$S_{xy} = \frac{\|\Gamma(x) \cap \Gamma(y)\|}{k(x) \times k(y)}$
Jaccard	$S_{xy} = \frac{\|\Gamma(x) \cap \Gamma(y)\|}{\|\Gamma(x) \cup \Gamma(y)\|}$	优先链接(PA)	$S_{xy} = k(x) \times k(y)$
Sorenson	$S_{xy} = \frac{2\|\Gamma(x) \cap \Gamma(y)\|}{k(x) + k(y)}$	Adamic-Adar(AA)	$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k(z)}$
Hub Promoted(HPI)	$S_{xy} = \frac{\|\Gamma(x) \cap \Gamma(y)\|}{\min\{k(x), k(y)\}}$	资源分配(RA)	$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}$

Paper: Lü, Linyuan & Zhou, Tao. (2010). Link Prediction in Complex Networks: A Survey. Physica A: Statistical Mechanics and its Applications. 390. 10.1016/j.physa.2010.11.027.

基于局部信息的节点相似性指标

- **效果比较**：将上述10种基于局部信息的相似性指标用于6个实际网络的链路预测，结果如下表。其中，RA总体表现最好，PA总体表现最差，在INT网络中*AUC*低于0.5，不如随机的预测好。

Indices	PPI	NS	Grid	PB	INT	USAir
CN	0.889	**0.933**	**0.590**	0.925	**0.559**	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	**0.933**	**0.590**	0.882	**0.559**	0.901
Sørensen	0.888	**0.933**	**0.590**	0.881	**0.559**	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	**0.933**	**0.590**	0.877	**0.559**	0.895
LHN1	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	**0.590**	0.922	**0.559**	0.925
RA	**0.890**	**0.933**	**0.590**	**0.931**	**0.559**	**0.955**

基于全局信息的节点相似性指标

- **基于全局信息的节点相似性指标：**

1）局部路径（Local path, LP）指标，考虑了三阶邻居的贡献：

$$\boldsymbol{S} = \boldsymbol{A}^2 + \alpha \boldsymbol{A}^3,$$

其中，α为可调节参数，$\boldsymbol{A}$为网络的邻接矩阵，$(A^n)_{xy}$给出了节点x和y之间长度为n的路径数。当$\alpha = 0$时，LP指标就等于CN指标。

2）Katz指标。考虑所有的路径数，且对越短的路径赋予越大的权重，定义为：

$$s_{xy} = \sum_{l=1}^{\infty} \beta^l [A^l]_{xy},$$

其中，β为权重衰减因子，那么矩阵形式为

$$\boldsymbol{S} = \sum_{l=1}^{\infty} \beta^l \boldsymbol{A}^l = (\boldsymbol{I} - \beta \boldsymbol{A})^{-1} - \boldsymbol{I}.$$

为保证收敛性，必须有$|\lambda_{max}(\beta \boldsymbol{A})| < 1$。

3）LHN-II指标。Katz指标的变种，考虑两个节点的邻居节点是相似的，那么这两个节点也是相似的，所以有

$$\boldsymbol{S} = \phi \boldsymbol{A}\boldsymbol{S} + \psi \boldsymbol{I} = \psi(\boldsymbol{I} - \phi \boldsymbol{A})^{-1} = \psi(\boldsymbol{I} + \phi \boldsymbol{A} + \phi^2 \boldsymbol{A}^2 + \cdots)$$

由于人们关注的是$\boldsymbol{S}$的相对值，我们可以令$\psi = 1$，这就和Katz指标很接近了，即

$$\boldsymbol{S} = \boldsymbol{I} + \phi \boldsymbol{A} + \phi^2 \boldsymbol{A}^2 + \cdots.$$

- What is the problem ?

  - 注意到一些顶点之间有一条甚至多条这样的路径：例如，度很高的顶点几乎肯定会有一条或多条长度为2的路径连接它们，即使顶点之间的连接只是随机建立的。

  - 简单地计算路径数量不足以建立相似性，需要知道什么时候**一对顶点之间具有相同长度的路径比我们预期的要多**。

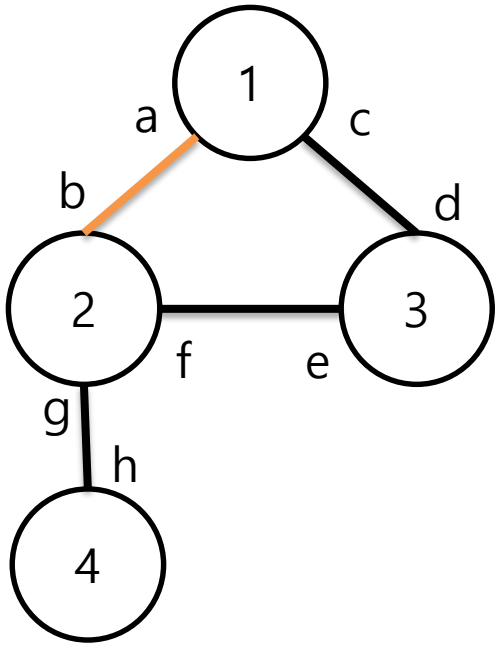
  - 说白了，就是$\boldsymbol{A}^l$前的权重系数ϕ^l设置不合理。

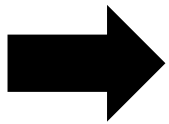
- How to do it ?
  - 独立地对每一对(i,j)分配不同的系数，即

$$S_{ij} = \sum_{l=0}^{\infty} C_l^{ij} [A^l]_{ij}.$$

Let us choose each coefficient to be equal to **1 over the expected number of paths of the corresponding length between the same pair of vertices on a network with the same degree sequence as the network under consideration**, but in which the vertices are otherwise randomly connected. Such a network is called a ***configuration model*(Chapter 6).**

- How to do it ?
  - 很容易知道$C_0^{ij} = \delta_{ij}$.
  - 对于C_1^{ij}，设节点i和节点j的度分别为k_i和k_j，考虑从节点i出发的k_i中的任意一条边，它有$k_j/2M$的可能性落到节点j的边上，那么从节点i到节点j的期望路径数为$\frac{k_i k_j}{2M}$，因此$C_1^{ij} = \frac{2M}{k_i k_j}$.





- 从a出发，有三个地方可以抵达：b、f、g.
- 从c出发，有三个地方可以抵达：b、f、g.
- 所以，从节点1到节点2的期望路径数为6/8.

- How to do it ?
  - 当$l \geq 2$时闭解计算过于复杂，直接给出近似结论：

$$C_l^{ij} = \frac{2M}{k_i k_j} \lambda_1^{-l+1},$$

其中，λ_1为矩阵$\boldsymbol{A}$的最大特征值。

  - 但是直接将$C_l^{ij} = \frac{2M}{k_i k_j} \lambda_1^{-l+1}$带入$S_{ij} = \sum_{l=0}^{\infty} C_l^{ij} [A^l]_{ij}$，可能会导致该级数发散，那么引入一个额外的参数$\alpha \in (0,1)$，得到LHN-II指标的计算公式如下：

$$\begin{aligned} S_{ij} &= \delta_{ij} + \frac{2M}{k_i k_j} \sum_{l=1}^{\infty} \alpha^l \lambda_1^{-l+1} [A^l]_{ij} \\ &= \left[1 - \frac{2M\lambda_1}{k_i k_j}\right] \delta_{ij} + \frac{2M\lambda_1}{k_i k_j} \left[\left(\boldsymbol{I} - \frac{\alpha}{\lambda_1} \boldsymbol{A}\right)^{-1}\right]_{ij}. \end{aligned}$$

- How to do it ?
  - 考虑到，

$$S_{ij} = \left[1 - \frac{2M\lambda_1}{k_i k_j}\right]\delta_{ij} + \frac{2M\lambda_1}{k_i k_j}\left[\left(\boldsymbol{I} - \frac{\alpha}{\lambda_1}\boldsymbol{A}\right)^{-1}\right]_{ij},$$

其中第一项仅仅影响节点与自己的相似度，这通常不是我们关心的，所以将第一项删去，得到

$$S_{ij} = \frac{2M\lambda_1}{k_i k_j}\left[\left(\boldsymbol{I} - \frac{\alpha}{\lambda_1}\boldsymbol{A}\right)^{-1}\right]_{ij}.$$

  - 等价地，将其写为矩阵的形式：

$$\boldsymbol{S} = 2M\lambda_1 \boldsymbol{D}^{-1}\left(\boldsymbol{I} - \frac{\alpha}{\lambda_1}\boldsymbol{A}\right)^{-1}\boldsymbol{D}^{-1},$$

其中，$\boldsymbol{D}_{ij} = k_i\delta_{ij}$为度值矩阵。

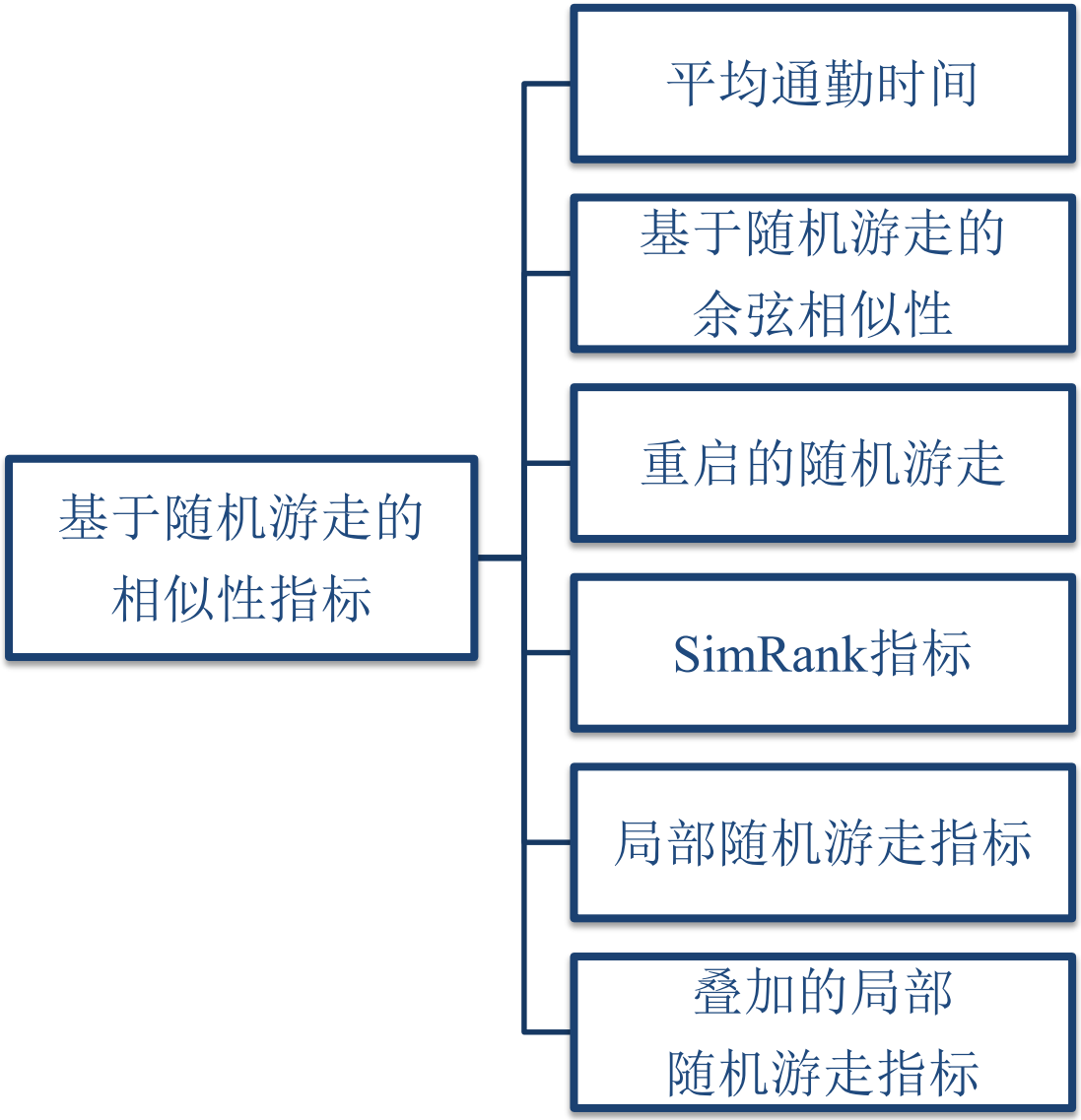
基于全局信息的节点相似性指标

- **效果比较**：将上述3种基于全局信息的相似性指标用于6个实际网络的链路预测，结果如下表。其中，从AUC看来，Katz总体表现最好，LP表现也不错，特别是在PB和USAir网络中。

AUC	PPI	NS	Grid	PB	INT	USAir
LP	0.970	**0.988**	0.697	**0.941**	0.943	**0.960**
LP*	0.970	**0.988**	0.697	0.939	0.941	0.959
Katz	**0.972**	**0.988**	**0.952**	0.936	**0.975**	0.956
LHN2	0.968	0.986	0.947	0.769	0.959	0.778
Precision	**PPI**	**NS**	**Grid**	**PB**	**INT**	**USAir**
LP	**0.734**	**0.292**	**0.132**	**0.519**	**0.557**	**0.627**
LP*	**0.734**	**0.292**	**0.132**	0.469	0.121	**0.627**
Katz	0.719	0.290	0.063	0.456	0.368	0.623
LHN2	0	0.060	0.005	0	0	0.005

注：LP*表示$\alpha = 0.01$的LP指标(USAir网络中$\alpha = -0.01$)，其他指标的参数根据数据最佳确定。对于Precision的计算，取L=100.

基于随机游走的相似性指标



- **预备知识**：考虑无向无权图。
  - Laplacian Matrix：$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$。
  - $\boldsymbol{L}$是半正定矩阵：因为，$\boldsymbol{L}$为**实对称对角占优矩阵**。
  - 特征值中0出现的次数就是图连通区域的个数。
  - 最小特征值是0，因为$\boldsymbol{L}$每一行的和均为0。
  - SVD: $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.
  - Pseudo-inverse: $\boldsymbol{L}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T$.

1）平均通勤时间（Average commute time, ACT）：

- 设$m(x,y)$为一个随机粒子从节点x到y平均需要走的步数，那么节点x和y的平均通勤时间定义为

$$n(x,y) = m(x,y) + m(y,x).$$

- 设$\boldsymbol{L}^+$为$\boldsymbol{L}$的伪逆，那么$n(x,y)$的数值解为

$$n(x,y) = M\left(l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+\right).$$

- 由此定义，ACT相似性为（忽略常数M）

$$S_{xy}^{ACT} = \frac{1}{l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+}.$$

2）基于随机游走的余弦相似性（Cos+）：

- 设$\boldsymbol{L}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T$，由于$\boldsymbol{L}$是半正定矩阵，所以有，

$$\boldsymbol{L}^+ = \boldsymbol{U}\boldsymbol{\Lambda}^+\boldsymbol{U}^T = \boldsymbol{U}(\boldsymbol{\Lambda}^+)^{1/2}(\boldsymbol{\Lambda}^+)^{1/2}\boldsymbol{U}^T = \boldsymbol{Q}\boldsymbol{Q}^T,$$

其中，$\boldsymbol{Q} = \boldsymbol{U}(\boldsymbol{\Lambda}^+)^{1/2}$，$\boldsymbol{\Lambda}$为$\boldsymbol{L}$的特征值对角阵。

- 矩阵$\boldsymbol{Q}^T$，对应一个线性变换，即，$\forall u, v \in \mathbb{R}^N$，可以得到

$$u' = \boldsymbol{Q}^T u \ , \ v' = \boldsymbol{Q}^T v.$$

进而考虑u'和v'的内积，$\langle u', v' \rangle = u^T\boldsymbol{Q}\boldsymbol{Q}^T v = u^T\boldsymbol{L}^+ v.$ 【$\boldsymbol{L}^+$ —Kernel】

- 对于节点x和y，可取$u = e_x, v = e_y$，即有

$$S_{xy}^{cos+} = \cos(x, y)^+ = \frac{l_{xy}^+}{\sqrt{l_{xx}^+ l_{yy}^+}}.$$

3）重启的随机游走（Random walk with restart, RWR）：

- 设$\boldsymbol{P}$为概率转移矩阵，某一粒子初始时刻在节点x，那么$t+1$时刻该粒子到达网络各个节点的概率为，

$$q_x(t+1) = c\boldsymbol{P}^T q_x(t) + (1-c)e_x$$

该式的稳态解为，$q_x = (1-c)(\boldsymbol{I} - c\boldsymbol{P}^T)^{-1}e_x$，其中第$y$个元素$q_{xy}$表示从节点$x$出发的粒子，最终有多少概率到达节点$y$.

- $S_{xy}^{RWR} = q_{xy} + q_{yx}$.

4）SimRank指标（SimR）：

- 基本假设：如果两节点所连接的节点相似，那么这两个节点就相似。
- $S_{xy}^{SimR} = C\dfrac{\sum_{z\in\Gamma(x)}\sum_{z'\in\Gamma(y)} S_{zz'}^{SimR}}{k_x k_y}$，其中$S_{xy}^{SimR} = 1$，$C \in (0,1)$为相似性传递时的衰减参数，该指标可以用来描述分别从节点x和y出发的粒子平均多久会相遇。

5）局部随机游走指标（Local random walk，LRW）：

- 思想：只考虑有限步数的随机游走过程。
- $\boldsymbol{\pi}_x(t+1) = \boldsymbol{P}^T\boldsymbol{\pi}_x(t)$，$t = 0,1,\ldots$，得到
$$S_{xy}^{LRW}(t) = q_x \cdot \pi_{xy}(t) + q_y \cdot \pi_{yx}(t).$$

6）叠加的局部随机游走指标（Superposed random walk，SRW）：

- 思想：与目标节点更近的节点更有可能与目标节点相连。
- 在LRW基础上将t步及其以前的结果求和便得到SRW值，即，
$$S_{xy}^{SRW}(t) = \sum_{l=1}^{t} S_{xy}^{LRW}(l) = q_x \sum_{l=1}^{t} \pi_{xy}(l) + q_y \sum_{l=1}^{t} \pi_{yx}(l).$$

基于随机游走的相似性指标

For each network, **the training set contains 90% of the known links**. Each number is obtained by averaging over **1000 implementations** with independently random divisions of training set and probe set. The parameters $\varepsilon = 0.001$ for LP (for USAir, $\varepsilon = -0.001$) and $c = 0.9$ for RWR. The numbers inside the brackets denote the optimal step of LRW and SRW indices. For example, 0.972(2) means the optimal AUC is obtained at the second step of LRW.

AUC	CN	RA	LP	ACT	RWR	LRW	SRW
USAir	0.954	0.972	0.952	0.901	0.977	0.972(2)	**0.978**(3)
NetScience	0.978	0.983	0.986	0.934	**0.993**	0.989(4)	0.992(3)
Power	0.626	0.626	0.697	0.895	0.760	0.953(16)	**0.963**(16)
Yeast	0.915	0.916	0.970	0.900	0.978	0.974(7)	**0.980**(8)
C.elegans	0.849	0.871	0.867	0.747	0.889	0.899(3)	**0.906**(3)
Precision	CN	RA	LP	ACT	RWR	LRW	SRW
USAir	0.59	0.64	0.61	0.49	0.65	0.64(3)	**0.67**(3)
NetScience	0.26	0.54	0.30	0.19	**0.55**	0.54(2)	0.54(2)
Power	0.11	0.08	**0.13**	0.08	0.09	0.08(2)	0.11(3)
Yeast	0.67	0.49	0.68	0.57	0.52	**0.86**(3)	0.73(9)
C.elegans	0.12	0.13	**0.14**	0.07	0.13	**0.14**(3)	**0.14**(3)

基于随机游走的相似性指标

- **LRW and SRW methods perform better** than other indices with their respective **optimal walking step** positively correlated with the **average shortest distance of the network.**

- Furthermore, the **computational complexity** of LRW and SRW is lower th an ACT and RWR whose time complexity in calculating inverse and pseud oinverse is approximately O(N^3), while the time complexity of n-steps LR W and SRW are approximately O(N$\langle k \rangle^n$). That is to say, when n is small LRW and SRW run much faster than other random-walk-based global similarity indices.

- The advantage of LRW and SRW for their low calculation complexity is prominent especially in the huge size (i.e. large N) and sparse (i.e. small $\langle k \rangle$) networks. For example, LRW or SRW for power grid is thousands time faster than ACT, cos+ and RWR, even for n ≃ 10.

Thank you