科学通报 2014 年 第 59 卷 第 13 期：1175 ~ 1197
特邀评述 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

网络重要节点排序方法综述

任晓龙，吕琳媛*

杭州师范大学阿里巴巴复杂科学研究中心，杭州 310036
* 联系人, E-mail: linyuan.lv@gmail.com

2013-11-21 收稿, 2014-02-25 接受, 2014-04-04 网络版发表
国家自然科学基金(11205042)、杭州师范大学科研启动基金和 CCF-腾讯科研基金资助

摘要 复杂网络的重要节点是指相比网络其他节点而言，能够在更大程度上影响网络的结构与功能的一些特殊节点. 近年来，节点重要性排序研究受到越来越广泛的关注，不仅因为其重大的理论研究意义，更因为其广泛的实际应用价值. 由于应用领域极广，且不同类型的网络中节点的重要性评价方法各有侧重，学者们从不同的实际问题出发设计出各种各样的方法. 本文系统地综述了复杂网络领域具有代表性的 30 余种重要节点挖掘方法，并将其分为四大类，详细比较各种方法的计算思路、应用场景和优缺点. 在此基础上，本文分析了重要节点排序研究现存的一些问题，并展望了若干重要的开放性问题.

关键词
复杂网络
重要节点
节点排序
节点中心性
传播模型

食物链中哪些物种对整个生态的影响最大?

“种子短信”发给哪些手机用户可以获得更多的转发?

全球经济体系中哪些国家或地区对于体系的健康发展至关重要?

当传染病来临的时候，我们应该采取何种接种免疫策略来避免其大规模爆发?

高价雇佣微博大号做新产品的推广和营销真的有用吗? 如果有用，又如何找到合适的达人?

为什么俄亥俄州克利夫兰市的几条烧断的高压线能够造成北美大停电事故，导致数百亿美元的损失?

借助网络科学的发展，对于这些问题，如今我们已经有了一些定量化的描述和解决办法. 实际上，几乎所有的复杂系统(比如社会、生物、信息、技术、交通运输系统)都可以自然地表示为网络. 其中，节点代表系统的各种构成要素，节点间的连边表示要素之间的联系. 最复杂的人类社会系统就可以用一个社会网络刻画，节点是人，人与人之间的各种关系构成社会网络中的链接. 应用复杂网络的理论和方法能够帮助我们更好地理解这些复杂系统的特征，并对其进行更好地预测和控制. 如上述的大停电事故归根到底是“网络相继攻击的脆弱性”问题. 如果我们能够事先对这个电力网络的结构有所了解，并找到关键的地区采取预防措施，就可能避免如此巨额的经济损失. 这里最核心的问题就是如何识别这些重要的节点. 所谓的重要节点是指相比网络其他节点而言能够在更大程度上影响网络的结构与功能的一些特殊节点. 这里的网络结构包括度分布、平均距离、连通性、聚类系数、度相关性等，网络功能涉及网络的抗毁性、传播、同步、控制等[1]. 重要节点一般数量非常少，但其影响却可以快速地波及到网络中大部分节点[2]. 例如，在对一个无标度网络的蓄意攻击中，少量最重要节点被攻击就会导致整个网络瓦解[3,4]; 微博中最有影响力的几个用户所发的微博很快就能传遍整个网络[5]; 仅仅 1%的公司却控制着 40%的全球经济[6]. 可见重要节点对网络的结构和功能有着巨大的影响，节点重要性的排序和重要节点的挖掘意义重大.

网络的“小世界特性”[7]和“无标度特性”[8]的发现

引用格式: 任晓龙，吕琳媛. 网络重要节点排序方法综述. 科学通报, 2014, 59: 1175–1197
Ren X L, Lü L Y. Review of ranking nodes in complex networks (in Chinese). Chin Sci Bull (Chin Ver), 2014, 59: 1175–1197, doi: 10.1360/972013-1280

掀起了网络科学持续 10 多年至今丝毫没有降温的研究热潮. 网络科学研究的热点逐渐从早期发现跨越不同网络的宏观上的普适规律转变为着眼于从中观(社团结构、群组结构)和微观层面(节点、链路)去解释不同网络所具有的不同特征[9]. 这一转变, 是因为随着研究的深入, 人们发现宏观指标不能很好表现网络结构和功能上的特征, 真正精细可靠的解释, 哪怕是针对宏观现象, 也必须立足于微观上的深入认识. 类似地, 多年以前, 一批学者就提倡关注网络结构和功能的相互影响[10], 但是早期的研究都集中在网络宏观或者中观上的一些特征与网络具体功能表现之间的关系, 所得到的一些结论, 譬如“热力学极限下无标度网络传染病(SIS 模型)没有阈值, 随机网络有阈值”[11]等, 都只是一些统计上有意义, 大多数情况下正确, 定性上可以部分解释, 定量上无法开展预测的结果. 这是因为宏观指标以及基于宏观量的运算, 已经把很多个体的特征进行了“平均化”, 而一些非常关键个体的表现被这种“平均化”淹没了. 还是回到刚才的例子, 如果我们从个体出发, 比如仔细考虑一个节点在 SIS 传播动力学中可能的自维持特性, 就会得到颠覆性的结论: “热力学极限下随机网络上的 SIS 模型也没有阈值”[12]. 可见, 基于微观层面, 即节点个体的分析, 有望揭示网络功能上精细入微的特征. 总之, 随着网络科学研究从整体宏观到个体微观的转变, 重要节点的排序和挖掘已成为近年来的研究热点[13~19].

除了在理论层面的重大意义, 重要节点的挖掘还具有非常直接的实际应用价值. 譬如说节点影响力的排序结果, 应用于 Twitter 和新浪微博等大规模的具有媒体属性的社会网络中, 可以用来给出在感兴趣的方向(domain)或者标签(tag)上最有影响力的用户排名, 从而帮助用户(特别是新用户)尽快找到有关联的信息源; 又譬如在信息传播功能背景下对于节点角色的区分和排序, 挖掘具有“信息引爆能力”的关键节点(在微博中常常是公众名人和领域专家), 可以应用于市场营销和广告投放策略的设计中. 这就是 Gladwell[20]所说的引爆流行的三大关键因素之一的“个别人物法则”.

事实上, 重要节点的挖掘算法是一个与信息科学领域有深厚渊源的研究方向, 它与链路预测和推荐构成了网络信息挖掘领域的主要核心问题[9]. 与传统挖掘方法相比, 网络科学为我们提供了新的视角和方法, 这类方法上的普适性使得信息挖掘不再仅仅是信息科学和计算机科学所关注的问题, 而成为各个学科所共同关注的问题, 这些来自各个学科的共同努力将真正推动交叉科学的发展. 关于链路预测[21~23]和推荐算法[24~27]已有一些综述性文章. 该方向受到物理科学、信息科学、管理科学等多学科的广泛关注.

到目前为止, 人们根据所研究的具体问题, 提出了多种多样的重要节点排序方法, 作了一些总结探讨[28~30], 但尚不全面, 对于这一研究领域的图景认识仍然不完整. 本文详细介绍了 30 余种具有代表性的挖掘方法. 将它们按照最合理的方式分成 4 类, 绘制出该问题目前为止最清晰的图景. 最后, 我们将主要方法及其特点总结在文章的表 S1.

为了行文方便, 我们约定: 一个网络的拓扑图记为 $G(V, E)$, 其中 $V=\{v_1, v_2,\cdots,v_n\}$是节点集合, $E=\{e_1, e_2,\cdots,e_m\}$是边的集合, n 与 m 分别是节点数和边数. 一个图的邻接矩阵记为 $A_{n\times n}=(a_{ij})$, 无向网络中 $a_{ij}=1$ 当且仅当节点 v_i 与 v_j 之间有连边, 否则 $a_{ij}=0$; 有向网络中 $a_{ij}=1$ 当且仅当存在一条从节点 v_i 指向 v_j 的有向边, 否则 $a_{ij}=0$. 特别地, 含权网络中一个图的邻接矩阵记为 $W_{n\times n}=(w_{ij})$, 如果节点 v_i 与 v_j 之间有连边则 w_{ij} 为连边上的权值, 否则 $w_{ij}=0$. 同时约定所有在网络中传播的信息、病毒、车流、人流、电流等统称为网络流(network flow). 网络中的一条路径是类似这样的一组节点和边的交替序列: $v_1, e_1, v_2, e_2, \cdots, e_{n-1}, v_n$, 其中 v_i, v_{i+1} 是 e_i 的两个端点. 如果任意一对节点之间都存在一条路径使它们相连, 就称这个网络是连通的.

1 基于节点近邻的排序方法

本类方法是最简单直观的方法, 度中心性考察节点的直接邻居数目, 半局部中心性考虑了节点 4 层邻居的信息. k-壳分解可以看作度中心性的一种扩展, 它根据节点在网络中的位置来定义其重要性, 认为越是在核心的节点越重要.

1.1 度中心性

社会网络分析中, 节点的重要性也称为“中心性”, 其主要观点是节点的重要性等价于该节点与其他节点的连接使其具有的显著性[31]. 度中心性(degree centrality)[32]认为一个节点的邻居数目越多, 影响力就越大, 这是网络中刻画节点重要性最简单的指标.

节点 v_i 的度，记为 k_i，是指与 v_i 直接相连的节点的数目，是节点最基本的静态特征．在有向网络中，根据连边的方向不同，节点的度有入度和出度之分．在含权网络中节点度又称为节点的强度(strength)，定义为与节点相连的边的权重之和．度中心性刻画的是节点的直接影响力[33]，它认为一个节点的度越大，能直接影响的邻居就越多，也就越重要．值得注意的是，不同规模的网络中有相同度值的节点有不同的影响力，为了进行比较，定义节点 v_i 的归一化度中心性指标为

$$DC(i)=\frac{k_i}{n-1}, \tag{1}$$

其中，$k_i=\sum_i a_{ij}$，a_{ij} 即网络邻接矩阵 $\boldsymbol{A}$ 中第 i 行第 j 列元素，n 为网络的节点数目，分母 n–1 为节点可能的最大度值．在有向网络中入度和出度有不同的意义(如社交网络中入度代表受欢迎程度，出度代表合群程度)，一般会分别计算入度和出度的中心性．

度中心性指标拥有简单、直观、计算复杂度低等特点．在网络鲁棒性和脆弱性研究中，针对无标度网络或指数网络，如果攻击前一次性选择若干个攻击目标，采用度中心性指标的攻击效果比介数中心性、接近中心性、特征向量中心性要好(参见 6.1 节)．度中心性指标的缺点是仅考虑了节点的最局部的信息，是对节点最直接影响力的描述，没有对节点周围的环境(例如节点所处的网络位置、更高阶邻居等)进行更深入细致地探讨，因而在很多情况下不够精确．

1.2 半局部中心性

度中心性指标计算方便简单，但实际效果欠佳．基于全局信息的方法，如在下一节中介绍的介数中心性和接近中心性指标，虽然具有较好的刻画节点重要性的能力，但计算复杂度太高，难以在大规模网络上使用．为了权衡算法的效率和效果，Chen 等人[14]提出了一种基于半局部信息的节点重要性排序方法，简称半局部中心性(semi-local centrality)．首先定义 $N(w)$为节点 v_w 的两层邻居度，其值等于从 v_w 出发 2 步内可到达的邻居的数目，然后定义

$$Q(j)=\sum_{w\in\Gamma(j)} N(w), \tag{2}$$

其中 $\Gamma(j)$ 表示节点 v_j 的一阶邻居节点的集合．最终节点 v_i 的局部中心性定义为

$$SLC(i)=\sum_{w\in\Gamma(i)} Q(w). \tag{3}$$

可见，半局部中心性涉及了节点的四阶邻居信息．文献[16]用 D-S 证据理论(参见 5.6 节)将本方法推广到了含权网络．文献[14]指出半局部中心性方法的计算复杂度随网络规模线性增长，消耗非常少的计算时间，就能够得到远好于度中心性和介数中心性的排序结果．近期，Chen 等人[34]还提出一种针对有向网络的半局部算法(ClusterRank)，该算法不仅考虑了邻居节点的数量，还考虑了聚类系数对信息传播的影响：聚类系数越大越不利于信息的广泛传播．定义 $CR(i)=f(c_i)\sum_{j\in\Gamma(i)}(k_j^{\text{out}}+1)$，其中 $\Gamma(i)$ 为节点 v_i 的邻居节点的集合，$f(c_i)$是节点 v_i 的聚类系数 c_i 的函数，c_i 越大，$f(c_i)$越小．两个数据集上的实验结果显示 ClusterRank 算法优于 PageRank 和 LeaderRank 算法，并且计算复杂度最低．

1.3 *k*-壳分解法

度中心性仅考察节点最近邻居的数量，认为度相同则重要性相同．然而，近期的一些研究表明在刻画节点重要性的时候节点在网络中的位置也是至关重要的因素．在网络中，如果一个节点处于网络的核心位置，即使度较小，往往也有较高影响力；而处在边缘的大度节点影响力往往有限．基于此，Kitsak 等人[13]提出用 *k*-壳分解法(*k*-shell decomposition)确定网络中节点的位置，将外围的节点层层剥去，处于内层的节点拥有较高的影响力．这一方法可看成是一种基于节点度的粗粒化排序方法．具体分解过程如下[35]：网络中如果存在度为 1 的节点，从度中心性的角度看它们就是最不重要的节点．如果把这些度为 1 的节点及其所连接的边都去掉，剩下的网络中会新出现一些度为 1 的节点，再将这些度为 1 的节点去掉，循环操作，直到所剩的网络中没有度为 1 的节点为止．此时，所有被去掉的节点组成一个层，称为 1-壳(记为 k_s=1)．对一个节点来说，剥掉一层之后在剩下的网络中节点的度就叫该节点的剩余度．按上述方法继续剥壳，去掉网络中剩余度为 2 的节点，重复这些操作，直到网络中没有节点为止．更广泛地，可定义初始度为 0 的孤立节点属于 0-壳，即 k_s=0．网络中的每一个节点属于唯一的一层，显然所有节点均满足 $k\geqslant k_s$．

图 1 给出一个 *k*-壳分解的示例．其中(a)为原网络，(b), (c), (d)分别表示 1-壳, 2-壳和 3-壳．可见，大度节

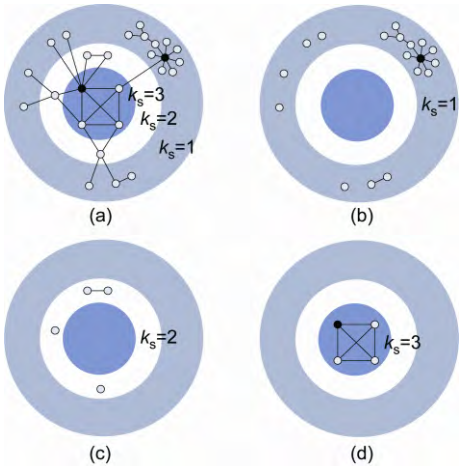

图 1 (网络版彩色)一个可分解为三层壳的简单网络[13]

点有可能因处于核心位置而拥有较大的 k_s 值(如图 1(d)中的深色节点), 也可能因为处于边缘而具有较小的 k_s 值(如图 1(b)中的深色节点). 在这个方法下, 大度节点不一定是重要节点. 文献[13]用网络上的传播实验对此进行了验证.

k-壳分解法计算复杂度低, 在分析大规模网络的层级结构等方面有很多应用. 然而, 此方法也有一定局限性. 第一, *k*-壳分解法有很多不能发挥作用的场景. 比如在树形图, 规则网络和 BA 网络[8]中, 所有(或大部分)节点都会被划分在同一层. 更极端的例子是星形图, 显然中心节点有最强的传播能力, 但是 *k*-壳分解的时候, 星形网络的所有节点会被划分在同一层(k_s=1). 第二, *k*-壳分解法的排序结果太过粗粒化, 使得节点的区分度不大. *k*-壳分解法划分的层级比度中心性方法划分的层级少很多, 很多节点处在同一层上, 它们之间的重要性难以比较. 第三, *k*-壳分解法在网络分解时仅考虑剩余度的影响, 这相当于认为同一层的节点在外层都有相同的邻居数目, 显然不合理. Zeng 等人[36]提出了在每一步剥去一部分外围节点之后, 同时考虑节点剩余的邻居数 k_i^{r} 和节点已经移除的邻居数 k_i^{e} 的方法, 定义节点 v_i 的混合度为 $k_i^{\mathrm{r}}+\lambda\times k_i^{\mathrm{e}}$, 根据新的混合度值对网络继续分层. 这种采用混合度值的 *k*-壳分解法能够很好地区分树形图以及 BA 网络中不同节点的传播能力, 并且分层的层数大大增加(甚至可超过度中心性), 提高了节点传播能力的区分度. 另外, Liu 等人[37]指出壳数相同的节点传播能力差距可能很大, 并提出了一种可以进一步区分具有相同壳数的节点的传播能力的排序方法, 从而较 Kitsak 等人[13]的方法有所进步; Hu 等人[38]将 *k*-壳分解法与社区结构相结合, 提出一种改良指标, 在 SIR 模型上的实验表明该方法较 Kitsak 等人的方法略佳.

2 基于路径的排序方法

在交通、通信、社交等网络中存在着一些度很小但是很重要的节点, 这些节点是连接几个区域的“桥节点”, 它们在交通流和信息包的传递中担任重要的角色. 此时, 刻画节点重要性就需要考察网络中节点对信息流的控制力, 这种控制力往往与网络中的路径密切相关. 基于最短路径的排序方法假设网络中的信息流只经过最短路径传输, 而真实的通信网络中必须考虑负载平衡, 容错机制, 服务水平协议(SLA)等[39]. 除了路径长度, 路径上的中间节点个数对传播也有不可忽视的影响. 一对节点的中间节点会增加这两个节点之间进行互动所需要的消耗. 第一, 中间节点越多, 一对节点之间互动所需要的时间就越长; 第二, 中间节点相当于在一对进行互动的节点之间引入了“第三方”, 这会使传递的信息失真或者延迟传递. 另一方面, 从提高网络的可靠性和抗毁性角度看, 任意节点对之间的路径数目越多, 网络的鲁棒性就越高. 此外类似于“桥节点”, 程学旗等人提出了刻画网络边重要性的指标用来寻找“桥链路”, 相关讨论参见文献[40].

2.1 离心中心性

在连通网络中, 定义 d_{ij} 为节点 v_i 与 v_j 之间的最短路径长度, 也称最短距离, 一个节点 v_i 的离心中心性 (Eccentricity)为它与网络中所有节点的距离之中的最大值[41], 即:

$$ECC(i)=\max_j\left(d_{ij}\right),\ j=(1,2,\cdots,n). \tag{4}$$

网络直径定义为网络 G 中所有节点的离心中心性中的最大值, 网络半径定义为所有节点的离心中心性值中的最小值. 显然, 网络的中心节点就是离心中心性值等于网络半径的节点, 一个节点的离心中心性与网络半径越接近就越中心. 要强调的是, 网络直径在复杂网络研究中还有多种不同的定义, 例如 Albert 等人[42]在研究万维网的时候定义网络直径为网络中所有节点对的最短路径的平均值. 离心中心性的缺点是极易受特殊值的影响, 如果一个节点与

大部分节点的距离都很小，只与极小部分节点的距离很大，这个节点的离心中心性仍然会取其中的最大值．接近中心性则采取距离平均值的方式克服了这一缺点．

2.2 接近中心性

接近中心性(closeness centrality)[33]通过计算节点与网络中其他所有节点的距离的平均值来消除特殊值的干扰．一个节点与网络中其他节点的平均距离越小，该节点的接近中心性就越大．接近中心性也可以理解为利用信息在网络中的平均传播时长来确定节点的重要性．平均来说，接近中心性最大的节点对于信息的流动具有最佳的观察视野．对于有 n 个节点的连通网络，可以计算任意一个节点 v_i 到网络中其他节点的平均最短距离:

$$d_i = \frac{1}{n-1}\sum_{j\neq i} d_{ij}, \tag{5}$$

d_i 越小意味着节点 v_i 更接近网络中的其他节点，于是把 d_i 的倒数定义为节点 v_i 的接近中心性，即:

$$CC(i) = \frac{1}{d_i} = \frac{n-1}{\sum_{j\neq i} d_{ij}}. \tag{6}$$

上面定义的缺点是仅能用于连通的网络中，文献[43]在研究网络效率时对上式进行了改进，使其能够用于非连通网络中，即:

$$EFF(i) = \sum_{j=1}^{n} \frac{1}{d_{ij}}. \tag{7}$$

如果节点 v_i 和 v_j 之间没有路径可达则定义 $d_{ij}=\infty$，即 $1/d_{ij}=0$．接近中心性利用所有节点对之间的相对距离确定节点的中心性，在研究中应用非常广泛，但时间复杂度比较高．

2.3 Katz 中心性

与接近中心性不同, Katz 中心性不仅考虑节点对之间的最短路径，还考虑它们之间的其他非最短路径[44]. Katz 中心性认为短路径比长路径更加重要，它通过一个与路径长度相关的因子对不同长度的路径加权．一个与 v_i 相距有 p 步长的节点，对 v_i 的中心性的贡献为 s^p ($s\in(0,1)$ 为一个固定参数)．设 $l_{ij}^{(p)}$ 为从节点 v_i 到 v_j 经过长度为 p 的路径的数目．显然 $\boldsymbol{A}^2 = \left(l_{ij}^{(2)}\right) = \left(\sum_k a_{ik}a_{kj}\right)$，其中元素 $l_{ij}^{(2)}$ 即从节点 v_i 到 v_j 经过的边数为 2 的路径的数目，同理我们可以得到 $\boldsymbol{A}^3, \boldsymbol{A}^4 \cdots \boldsymbol{A}^p \cdots$，将这些值赋予不同权重然后相加，便可以得到一个描述网络中任意节点对之间路径关系的矩阵:

$$\boldsymbol{K} = s\boldsymbol{A} + s^2\boldsymbol{A}^2 + \cdots + s^p\boldsymbol{A}^p + \cdots = \left(\boldsymbol{I} - s\boldsymbol{A}\right)^{-1} - \boldsymbol{I}. \tag{8}$$

其中，$\boldsymbol{I}$ 为单位矩阵．$\boldsymbol{K}$ 矩阵中第 i 行 j 列对应的元素 k_{ij} 实际上就是我们所熟知的节点 v_i 和 v_j 的 Katz 相似性[21]．为保证 $\boldsymbol{K}$ 可写成公式(8)右侧的矩阵形式，要求参数 s 小于邻接矩阵的最大特征值的倒数．由此可定义一个节点 v_j 的 Katz 中心性为矩阵 $\boldsymbol{K}$ 第 j 列元素的和:

$$Katz(j) = \sum_i k_{ij}. \tag{9}$$

Katz 中心性使用矩阵求逆的方法虽然比直接数路径数目简单，但时间复杂度依然比较高．另一方面，在考虑所有路径长度时，如果节点 v_i 与 v_j 之间存在长度为 p 的路径，在使用 $\boldsymbol{K}$ 矩阵计算节点间长度为 p 的奇数倍的路径时，这条路径会被重复计算多次．衰减因子 s 的引入正好削弱了这些由于重复计算产生的对中心性值的影响，特别是当 s 很小时，高阶路径的贡献就非常小了，使 Katz 指标的排序结果接近于局部路径指标. Katz 中心性主要用在规模不太大，环路比较少的网络中．受到 Katz 中心性指标的启发，我们还可以应用其他刻画节点间相似性的指标[21]来定义节点中心性．

2.4 信息指标

信息指标(information indices)[45]通过路径中传播的信息量来衡量节点重要性．该方法假定信息在一条边上传递的时候存在一定的噪音，路径越长噪音就越大．一条路径上的信息传输量等于该路径长度的倒数．一对节点(v_i, v_j)间能够传输的信息总量就等于它们之间所有路径传输的信息量之和，记为 q_{ij}. 值得注意的是，如果我们把网络看成一个电阻网络，每条边的电阻记为 1，则 $1/q_{ij}$ 相当于以 2 个节点 v_i 和 v_j 为两端点的电阻值(q_{ij} 相当于电导)[46]，于是我们可以通过计算矩阵 $\boldsymbol{R} = (r_{ij}) = (\boldsymbol{D} - \boldsymbol{A} + \boldsymbol{F})^{-1}$ 获得 q_{ij}，其中 $\boldsymbol{D}$ 是 n 阶对角矩阵，对角线元素都是对应节点的度值，非对角线元素为 0，$\boldsymbol{F}$ 是每个元素均为 1 的 n 阶方阵．由此可得该网络中每一对节点(v_i, v_j)间通过所有路径能够传播的信息总量为

$$q_{ij} = \left(r_{ii} + r_{jj} - 2r_{ij}\right)^{-1}. \tag{10}$$

最后, 用调和平均数的方法定义节点 v_i 的中心性指标(有时也采用算术平均数)[47]:

$$INF(i)=\left[\frac{1}{n}\sum_j\frac{1}{q_{ij}}\right]^{-1}. \tag{11}$$

信息指标考虑了所有路径, 并可通过电阻网络简化繁复的计算过程. 该方法可以很容易地扩展到含权网络, 也适用于非连通的网络.

可见, 无论是接近中心性、Katz 中心性还是信息指标, 它们的思路是一致的. 如果用一个矩阵 $\boldsymbol{M}=(m_{ij})$ 来表示网络中所有节点之间的关系, $\boldsymbol{M}$ 的每一个元素 m_{ij} 刻画了节点 v_i 和 v_j 之间的某种联系, 这个联系既可以是它们之间的距离(如接近中心性), 也可以是某种相似性, 于是一个节点 v_i 的重要性可表示为 $Centrality(i)=\sum_j m_{ij}$. 由此可见, 只要我们能够给出一种刻画节点关系的方式, 就能够基于这个方法定义一个节点的中心性.

2.5 介数中心性

通常提到的介数中心性(betweenness centrality)一般指最短路径介数中心性(shortest path BC), 它认为网络中所有节点对的最短路径中(一般情况下一对节点之间存在多条最短路径), 经过一个节点的最短路径数越多, 这个节点就越重要. 介数中心性刻画了节点对网络中沿最短路径传输的网络流的控制力. 节点 v_i 的介数定义为

$$BC(i)=\sum_{i\neq s,i\neq t,s\neq t}\frac{g_{st}^i}{g_{st}}, \tag{12}$$

其中, g_{st} 为从节点 v_s 到 v_t 的所有最短路径的数目, g_{st}^i 为从节点 v_s 到 v_t 的 g_{st} 条最短路径中经过 v_i 的最短路径的数目. 显然, 当一个节点不在任何一条最短路径上时, 这个节点的介数中心性为 0, 比如星形图的外围节点. 对于一个包含 n 个节点的连通网络, 节点度的最大可能值为 $n-1$, 节点介数的最大可能值是星形网络中心节点的介数值: 因为所有其他节点对之间的最短路径是唯一的并且都会经过该中心节点, 所以该节点的介数就是这些最短路径的数目, 即为 $(n-1)(n-2)/2$. 于是得到一个归一化的介数:

$$BC'(i)=\frac{2}{(n-1)(n-2)}\sum\frac{g_{st}^i}{g_{st}}. \tag{13}$$

此外, 文献[48]讨论了介数中心性在有向网络中的应用. 测度单个节点的介数中心性可以很容易地扩展到基于节点集的介数中心性[49,50]. 在有些情况下, 节点的度中心性和接近中心性都相等, 这时就可用介数中心性来区分节点的重要性[51].

介数中心性可用于设计网络的通信协议、优化网络部署、检测网络瓶颈等. 王延庆[52]将介数应用于负载网络, 提出用过载函数法研究网络的连接失效问题. 此外, Goh 等人[53]提出的负载中心性(traffic load centrality)采用类似网络中信息包传递的机制: 每一对节点之间沿着最短路径传输一个单位的网络流, 如果最短路径不止一条, 则在几条最短路径的分叉处将网络流平均分配到这些最短路径上. 忽略时延, 网络中所有节点对之间都互不干扰地传输一个单位的信息流时, 一个节点上传输过的网络流的数量称为该节点的负载. 一个节点的负载越大, 该节点就越重要. 介数中心性的计算时间复杂度较高, 使其在实际应用中受到限制, 相关讨论可参见文献[54,55].

2.6 流介数中心性

介数中心性仅考虑网络流通过最短路径传输. Yan 等人[56]的研究指出, 如果选择最短路径来运输网络流, 很多情况下反而会延长出行时间、降低出行效率. 把一对节点之间的每条路径看作一条单独的管道, 一条管道能够传输一个单位的网络流, 从源节点 v_s 到目标节点 v_t 的最大流量是指 v_s 与 v_t 之间所有管道可同时运输的网络流的总和(实际上, 这种假设没有实际意义, 多条路径往往有重合的部分, 重合部分的流量就会超过假设的情况). 基于这样的假设, 流介数中心性(flow betweenness centrality)[57]认为网络中所有不重复的路径中, 经过一个节点的路径的比例越大, 这个节点就越重要. 由此得到节点 v_i 的流介数中心性为

$$FBC(i)=\sum_{s<t}\frac{\tilde{g}_{st}^i}{\tilde{g}_{st}}, \tag{14}$$

其中, $\tilde{g}_{st}$ 为网络中节点 v_s 与 v_t 之间的所有路径数(不包含回路), $\tilde{g}_{st}^i$ 为节点对 v_s 与 v_t 之间经过 v_i 的路径数. 介数中心性和流介数中心性考虑的是两个极端, 前者只考虑最短路径, 后者考虑所有路径并认为每条路径作用相同, 接下来介绍两种介于两者之间的介数中心性算法.

2.7 连通介数中心性

连通介数中心性依然考虑节点对之间的所有路

径，并且赋予较长的路径较小的权值. 首先，定义节点对(v_p, v_q)之间的连通度(communicability)为[58]

$$G_{pq}=\frac{1}{s!}P_{pq}^{(s)}+\sum_{k>s}\frac{1}{k!}W_{pq}^{(k)}, \tag{15}$$

其中 $P_{pq}^{(s)}$ 为从节点 v_p 到 v_q 的最短路径的数目，s 为最短路径的长度；$W_{pq}^{(k)}$ 是从 v_p 到 v_q 的非最短路径中路径长度为 k 的路径的数目. 连通度用邻接矩阵的形式可表示为

$$G_{pq}=\sum_{k=1}^{\infty}\frac{\left(A^k\right)_{pq}}{k!}=\left(\mathrm{e}^{A}\right)_{pq}. \tag{16}$$

基于连通度的概念，可定义节点 v_r 的连通介数中心性(communicability betweenness centrality)为[59]

$$CBC(r)=\frac{1}{C}\sum\nolimits_{p}\sum\nolimits_{q}\frac{G_{prq}}{G_{pq}},\ p\neq q\neq r,$$

其中 $C=(n-1)^2-(n-1)$是归一化常数，G_{prq} 为考虑过节点 v_r 的路径得到的连通度. 定义 $A(r)$为邻接矩阵 $\boldsymbol{A}$ 中第 r 行和第 r 列上的元素均为 0 的矩阵，则上式可写为

$$CBC\left(r\right)=\frac{1}{C}\sum_{p}\sum_{q}\frac{G_{prq}}{G_{pq}}=\frac{1}{C}\sum_{p}\sum_{q}\frac{\left(\mathrm{e}^{A}\right)_{pq}-\left(\mathrm{e}^{A(r)}\right)_{pq}}{\left(\mathrm{e}^{A}\right)_{pq}}. \tag{17}$$

2.8 随机游走介数中心性

从源节点 v_s 到目标节点 v_t 的随机游走的过程中经过 v_i 的次数可表征 v_i 的重要性. 基于此，Newman[60]提出了基于随机游走的介数中心性算法(random walk betweenness centrality). 在随机游走过程中短的路径计数次数较多，相当于赋予其更高的权重. 在随机游走过程中，如果网络流不断地从一个节点来回经过无疑会提高这个节点的介数中心性，但是这样的刻画实际上是毫无意义的. 为了避免这种偏差，约定在一次随机游走中如果网络流两次分别从相反方向经过某一节点，则它们对这个节点的介数中心性的贡献相互抵消. 于是，节点 v_i 的随机游走介数中心性可表示为

$$RWBC\left(i\right)=\frac{\sum_{s<t}I_{st}^{i}}{n\left(n-1\right)/2}, \tag{18}$$

其中 I_{st}^{i} 表示从源节点 v_s 到目标节点 v_t 的随机游走过程中，经过节点 v_i 的次数. 事实上，如果我们将网络看成一个电阻网络，每条边的电阻值为 1，从节点 v_s 输入一单位电流，从节点 v_t 输出电流，那么 I_{st}^{i} 就相当于经过节点 v_i 的电流值. 定义节点 v_i 的电压为 Vol_i，根据基尔霍夫电流定律可知流入节点 v_i 的电流总和与流出的电流总和相等，于是可得

$$\begin{cases}\sum_{j}a_{ij}(Vol_i-Vol_j)=0, & i\neq s, i\neq t,\\ \sum_{j}a_{ij}(Vol_i-Vol_j)=1, & i=s,\\ \sum_{j}a_{ij}(Vol_i-Vol_j)=-1, & i=t.\end{cases}$$

上式可写成矩阵形式为 $(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{Vol}=\boldsymbol{s}$，其中，$\boldsymbol{Vol}$ 为 $n\times1$ 列向量，$\boldsymbol{s}$ 为 $n\times1$ 列向量，其中第 s 位元素等于 1，第 t 位元素等于–1，其他元素等于 0. 由于拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$ 不可直接求逆，为了求得上述方程的解我们假设网络某一个节点 v_k 的电压值为 0，从而关注其他节点相对节点 v_k 的电压的相对值，于是将矩阵 $\boldsymbol{L}$ 的第 k 行第 k 列的元素去掉，得到一个$(n-1)\times(n-1)$维矩阵 $\boldsymbol{L}_{(-k)}$，对矩阵 $\boldsymbol{L}_{(-k)}$求逆之后，在原来 k 行 k 列的位置上补齐元素 0，得到 $n\times n$ 维矩阵 $\boldsymbol{Z}$，则以节点 v_s 为电流输入，节点 v_t 为电流输出的时候节点 v_i 的电压为 $Vol_{st}^{i}=z_{is}-z_{it}$，其中 z_{is} 为矩阵 $\boldsymbol{Z}$ 第 i 行第 s 列对应元素，于是可得流经节点 v_i 的电流等于

$$I_{st}^{i}=\frac{1}{2}\sum_{j}a_{ij}\left|Vol_{st}^{i}-Vol_{st}^{j}\right|=\frac{1}{2}\sum_{j}a_{ij}\left|z_{is}-z_{it}-z_{js}+z_{jt}\right|, \tag{19}$$

当 $i=s$ 或者 t 的时候，$I_{st}^{s}=I_{st}^{t}=1$. 该方法计算复杂度较高，对于 n 个节点 m 条边的网络，其计算复杂度为 $\mathrm{O}\left(\left(m+n\right)n^2\right)\approx\mathrm{O}(n^3)$.

2.9 路由介数中心性

计算机网络中，每个路由器都有一个包含很多行记录的路由表，每行记录存储着要到达的目标地址及下一跳地址. 显然，每个路由器只记录了局部的网络结构信息. 对网络中的每一对节点(v_s, v_t)，将分布在各个路由器中的信息聚合，可形成一个关于这一对节点的有向无环图 $R(s, t)$. 定义 $p(s, u, v, t)$为有向无环图 $R(s, t)$中节点 v_u 转发给节点 v_v 一个从源节点 v_s 到目标节点 v_t 的信息包的概率. 如果 $p(s, u, v, t)>0$，则在 $R(s, t)$中存在一条从 v_u 指向 v_v 的有向边. 用 $\delta_{s,t}(u)$ 表示信息包从 v_s 到 v_t 的传递过程中，经过节点的 v_u 概率，显然 $\delta_{s,t}(s)=\delta_{s,t}(t)=1$，用 $Pred_{s,t}(v)$表示 $R(s, t)$中节点 v_v 的直接前驱的集合，那么有向无环图 $R(s, t)$中经过任意一个节点 v_v 的概率可由下式得出:

$$\delta_{s,t}(v)=\sum_{u\in Pred_{s,t}(v)}\delta_{s,t}(u)\,p(s,u,v,t). \tag{20}$$

定义 $q(s,\ t)$为从 v_s 到 v_t 传输的信息包的数目，则在所有的信息包传输过程中经过节点 v_i 的信息包数量的期望值即可反映该节点在网络中的重要性，称为路由介数中心性(routing betweenness centrality)[39]:

$$RBC(i)=\sum_{s,t\in V}\delta_{s,t}(i)q(s,t). \tag{21}$$

路由介数中心性算法采用无环路由策略，将网络用拓扑排序表示为从源节点到目标节点的有向无环图，忽略了由路由震荡产生的临时环路. 实际应用中，$q(s,\ t)$可以代表任何有意义的值，比如两个节点的传输字节数、会话数或联系的紧密程度等. 如果所有节点对对节点中心性贡献的权值相同，则 $q(s,\ t)\equiv1$. 路由介数中心性与负载中心性(参见 2.5 节)采用的方法相似，但它除了考虑最短路径，还考虑了网络流在传输过程中的非最短路径，但并非全部路径.

2.10 子图中心性

当我们考虑经过节点的路径为一个封闭环的时候，就可以定义子图中心性(subgraph centrality)[51]. 该方法从全局的视野考察了网络中所有可达的邻居对节点中心性的增强作用，并且认为增强作用会随距离的增加而衰减. 与图论中的概念有所不同，这里一个子图特指从一个节点开始到这个节点结束的一条闭环回路. 一个节点 v_i 的子图数目就是以该节点为首尾的闭环回路的个数. 子图中心性认为闭环回路的路径长度越小，回路信息交流越便利，节点之间的联系越紧密，对节点的中心性贡献越大，其定义为

$$SC(i)=\sum_{t=1}^{\infty}\frac{a_{ii}^{t}}{t!}, \tag{22}$$

其中 a_{ii}^{t} 为网络的邻接矩阵 $\boldsymbol{A}$ 的 t 次幂的第 i 个对角线元素. t=1 时，$a_{ii}^{1}\equiv0$; t=2 时，a_{ii}^{2} 为节点 v_i 的度值，即 $a_{ii}^{2}=k_i$，此时，子图中心性就等价于度中心性；$t\geqslant3$ 时，a_{ii}^{t} 表示从点 v_i 开始，经过 t 条边又回到 v_i 的路径的数目. 子图中心性赋予较短的回路较高的权重，使得节点的度在其中发挥较大作用的同时，还考虑了高阶回路[61]. 在实际应用时，根据具体计算需求，t 可以取到任意值截断. 子图中心性用邻接矩阵特征值和特征向量可表示为

$$SC(i)=\sum_{t=1}^{\infty}\frac{a_{ii}^{t}}{t!}=\sum_{j=1}^{N}\left(\xi_j^i\right)^2\mathrm{e}^{\lambda_j}, \tag{23}$$

其中，$\lambda_j\ (j=1,2,\cdots,n)$ 为邻接矩阵 $\boldsymbol{A}$ 的特征值，ξ_j 是 λ_j 所对应的特征向量，ξ_j^i 表示特征向量的第 i 个元素. 有些情况下，度中心性，接近中心性以及介数中心性都不能区分网络中某些节点谁更重要时，可用子图中心性来对这些节点进行更加细致地区分[51]. 另外，子图中心性的方法还能够应用于网络中模体的检测[51].

3 基于特征向量的排序方法

前面介绍的方法都是从邻居的数量上考虑对节点重要性的影响，基于特征向量的方法不仅考虑节点邻居数量还考虑了其质量对节点重要性的影响. 下面将详细介绍 7 种方法. 其中前两种方法，即特征向量中心性和累计提名方法一般用在无向网络中，后者收敛更快. 后面五种方法可看成特征向量中心性在有向网络中的应用. PageRank 算法和 LeaderRank 算法通过模拟用户上网浏览网页的过程，使节点的分值沿着访问路径增加，用于识别网页重要性. 实验结果显示，LeaderRank 表现好于 PageRank 算法. HITs 算法、自动信息汇集算法, SALSA 算法中考虑节点的双重角色：权威性和枢纽性，并认为两者相互影响. 本类方法在理论和商业上都受到了极大的关注，很有借鉴意义.

3.1 特征向量中心性

特征向量中心性(eigenvector centrality)[32]认为一个节点的重要性既取决于其邻居节点的数量(即该节点的度)，也取决于每个邻居节点的重要性. 记 x_i 为节点 v_i 的重要性度量值，则:

$$EC(i)=x_i=c\sum_{j=1}^{n}a_{ij}x_j, \tag{24}$$

其中 c 为一个比例常数. 记 $\boldsymbol{x}=\left[x_1,x_2,x_3,\cdots,x_n\right]^{\mathrm{T}}$，经过多次迭代到达稳态时可写成如下的矩阵形式:

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

这表示 x 是矩阵 $\boldsymbol{A}$ 的特征值 c^{-1} 对应的特征向量. 计算向量 x 的基本方法是给定初值 $x(0)$，然后采用如下迭代算法:

$$x(t)=\boldsymbol{c}\boldsymbol{A}\boldsymbol{x}(t-1),\ t=1,\ 2,\ \cdots,$$

直到归一化的 $x'(t)=x'(t-1)$ 为止. 文献[32]证明，每一步的迭代过程中，如果给 x 除以邻接矩阵 $\boldsymbol{A}$ 的主特征值 λ，这一个方程就能得到一个收敛的非零解. 即

$\boldsymbol{x}=\lambda^{-1}A\boldsymbol{x}$. 于是, 常数 $c=\lambda^{-1}$. 特征向量中心性更加强调节点所处的周围环境(节点的邻居数量和质量), 它的本质是一个节点的分值是它的邻居的分值之和, 节点可以通过连接很多其他重要的节点来提升自身的重要性, 分值比较高的节点要么和大量一般节点相连, 要么和少量其他高分值的节点相连. 从传播的角度看, 特征向量中心性适合于描述节点的长期影响力, 如在疾病传播、谣言扩散中, 一个节点的EC分值较大说明该节点距离传染源更近的可能性越大, 是需要防范的关键节点[61].

特征向量法完全用与某节点相连接的其他节点的信息来评价该节点的重要性. Bonacich等人[62]认为节点的重要性还可能受到不依赖于节点连接信息的一些来自外部的信息的影响. 例如在微博上有人喜爱转发其他人发布的信息(依赖于网络连接的内部信息), 有的人却比较热衷于发布原创信息或从其他网站转发一些信息(不依赖于网络连接的外部信息). 由此Bonacich等人提出阿尔法中心性(Alpha-centrality), 即 $x=\alpha Ax+e$, 其中α为刻画来自网络内部连接影响的内因参数, e 为刻画那些不受网络连接影响的外因参数. 不失一般性, e 可以设置为一个所有元素都等于1的向量, 此时阿尔法中心性与Katz中心性一致.

当网络中有一些度特别大的节点的时候, 特征向量中心性会出现分数局于化现象(localiztion), 即大多数分值都集中在大度节点上, 使得其他节点的分值区分度很低. 为了避免这一现象, Martin等人[63]对特征向量中心性进行改进, 提出在计算节点v_i的分值时, 求和中其邻居的分值不再考虑节点v_i的影响.

3.2 累计提名

特征向量中心性中, 一个节点的打分值完全由邻居决定, 收敛过程缓慢. 此外, 当不存在一个正的自然数 t, 使得转移矩阵的 t 次幂所有元素都是正的时, 节点打分值会出现周期性循环, 不能收敛. 为了使打分值能够收敛并且快速收敛, 累计提名(cumulative nomination)[47]方法在每次迭代过程中, 同时考虑邻居节点和自身的打分值. 设 $\tilde{p}_i^t$ 为节点 v_i 在时刻 t 时得到的提名次数, 假设 t=0 时每个节点都获得 1 次提名(即 $\tilde{p}_i^0=1$), 每个时间步每个节点从所有相邻的节点处获得新增的提名, 新增的提名数为邻居节点已有的提名数的总和. 于是定义节点在 t+1 时刻的累积提名为

$$\tilde{p}_i^{t+1}=\tilde{p}_i^t+\sum_j a_{ij}\tilde{p}_j^t, \tag{25}$$

其中 $\tilde{p}_i^t$ 为迭代 t 步之后节点 v_i 得到的提名次数, 将这个值归一化, 得到:

$$p_i^t=\frac{\tilde{p}_i^t}{\sum_j \tilde{p}_j^t}. \tag{26}$$

如果所有节点归一化后的提名次数不再变化, 则停止迭代. 稳态时每个节点的提名次数占所有节点的提名次数的比例就是其重要性权值. 特征向量中心性算法在每次迭代的时候, 一个节点 v_i 的中心性值完全等于邻居的中心性值之和, 而累计提名算法则保留了节点 v_i 上一步的中心性值, 实验结果显示累积提名相比原始的特征向量中心性收敛速度更快. 累积提名和 Alpha 中心性在数学形式上非常相似, 但 Alpha 中心性中的 e 是固定值, 即每次迭代的时候不变, 而累积提名中添加的是上一时间步的打分值, 这个打分值会随着每步更新变化.

3.3 PageRank 算法

特征向量中心性及其变体应用广泛, 例如网页排序领域中最著名的 PageRank 算法[64], 是谷歌搜索引擎的核心算法. 传统的根据关键字密度判定网页重要程度的方法容易受到“恶意关键字”行为的诱导, 使搜索结果可信度低. PageRank 算法基于网页的链接结构给网页排序, 它认为万维网中一个页面的重要性取决于指向它的其他页面的数量和质量, 如果一个页面被很多高质量页面指向, 则这个页面的质量也高. 初始时刻, 赋予每个节点(网页)相同的 PR 值, 然后进行迭代, 每一步把每个节点当前的 PR 值平分给它所指向的所有节点. 每个节点的新 PR 值为它所获得的 PR 值之和, 于是得到节点 v_i 在 t 时刻的 PR 值为

$$PR_i(t)=\sum_{j=1}^{n} a_{ji}\frac{PR_j(t-1)}{k_j^{\text{out}}}, \tag{27}$$

其中 k_j^{out} 为节点 v_j 的出度. 迭代直到每个节点的 PR 值都达到稳定时为止. 公式(27)的缺陷在于 PR 值一旦到达某个出度为零的节点(称为悬挂节点 Dangling node), 就会永远停留在该节点处而无法传递出来, 从而不断吸收 PR 值[64,65]. 为解决这一问题, PageRank 算法在上述过程基础上引入一个随机跳转概率 c. 每一步, 不管一个节点是否为悬挂节点, 其 PR 值都将

以 c 的概率均分给网络中所有节点，以 $1-c$ 的概率均分给它指向的节点. 该过程实际上是考虑到了现实中网络用户除了通过超链接访问页面之外，还可以通过直接输入网址的形式对网页进行访问的行为，从而保证了即使是没有任何入度的网页也有机会被访问到. 其实质是将有向网络变成强连通的，使邻接矩阵成为不可约矩阵，保证了特征值 1 的存在. 由此可得含参数 c 的 PageRank 算法:

$$PR_i(t)=(1-c)\sum_{j=1}^{n}a_{ji}\frac{PR_j(t-1)}{k_j^{\text{out}}}+\frac{c}{n}. \tag{28}$$

参数 c 的取值要视具体的情况而定. c 取值越大收敛越快，c=0 时回到公式(27). c 取值越大算法的有效性越低，c=1 时所有节点都有相同的 PR 值. 针对万维网的网页排序，以前的研究显示，c=0.15 是一个比较好的参数. PageRank 算法作为谷歌搜索引擎的核心算法，它在商业应用上的极大成功激发了人们深入研究 PageRank 的热忱，研究者们提出了一系列基于 PageRank 的改进算法. 例如 Kim 和 Lee[65]为了避免悬挂节点囤积 PR 值的问题，将每一步到达悬挂节点的 PR 值平均分给网络中的 n 个节点，即将概率转移矩阵中悬挂节点所在的列的 n 个元素修改为 $1/n$; PageRank 中从一个网页上的链接中挑选下一个访问目标时是等概率的，Zhang 等人[66]认为这 n 个目标网页出度越大的越有可能被点击，并提出 N-step PageRank 算法用以描述这一思想. 2012 年 Brin 和 Page[67]以相同的题目重新出版了当年提出 PageRank 算法的博士学位论文，在文中他们对这十几年的网页排序算法进行了回顾，并就如何用 PageRank 实现大规模搜索进行了深入讨论. 另外，作为有向网络节点排序最经典的算法，PageRank 及其改进算法广泛应用于其他领域，如对期刊的排序[68]、对社交网络上用户的排序[5]、对风投公司(VC)的排序[69]、对科学论文的排序[70~73]以及科学家影响力的排序[74~77]等.

3.4 LeaderRank 算法

PageRank 算法中，每一个节点的随机跳转概率都是相同的，即从任意网页出发，采用输入网址来访问其他网页的概率相等. 然而在现实中人们在内容丰富的热门网页(出度大的节点)上浏览的时候选择使用地址栏跳转页面的概率要远小于浏览信息量少的枯燥网页(出度小的节点). 另一方面，PageRank 算法中的参数 c 的选取往往需要实验获得，并且在不同的应用背景下最优参数不具有普适性[35]. LeaderRank 算法的出现很好地解决了以上两个问题. 在有向网络的随机游走过程中，通过添加一个背景节点以及该节点与网络中所有节点的双向边来代替 PageRank 算法中的跳转概率 c，从而得到一个无参数且形式上更加简单优美的算法. LeaderRank 算法在某一页面输入网址访问下一个页面的概率就相当于从这个页面访问背景节点的概率，这个概率和一个网页上的链接数负相关，链接数越多，网页的内容越丰富，越倾向于从本地的链接访问，访问背景节点的概率就越低. 注意，背景节点的存在同样保证了网络的强连通性. 初始时刻给定网络中除背景节点 v_g 以外的其他节点单位资源，即 $LR_i(0)=1,\ \forall i\neq g$； $LR_g(0)=0$. 经过以下的迭代过程直到稳态:

$$LR_i(t)=\sum_{j=1}^{n+1}\frac{a_{ji}}{k_j^{\text{out}}}LR_j(t-1). \tag{29}$$

注意，迭代过程中邻接矩阵为 n+1 阶(包含背景节点). 稳态时将背景节点的分数值 $LR_g(t_c)$平分给其他 n 个节点，于是得到节点 v_i 的最终 LeaderRank 分数值为

$$LR_i=LR_i(t_c)+\frac{LR_g(t_c)}{n}. \tag{30}$$

LeaderRank 算法在衡量社会网络中节点的影响力等方面有非常优异的表现[78]，因此得名. 实验发现 LeaderRank 比 PageRank 在很多方面表现得更好: (1) 与 PageRank 相比收敛更快[79]; (2) 能够更好地识别网络中有影响力的节点，挖掘出的重要节点能够将网络流传播的更快更广; (3) 它在抵抗垃圾用户攻击和随机干扰方面相比 PageRank 有更强的鲁棒性. 这些优点使得 LeaderRank 算法广受关注. 标准 LeaderRank 算法中背景节点和所有节点的连接都一样，Li 等人[79]对此提出改进，认为从背景节点出发访问其他节点时，入度大的节点应该有更高的概率被访问到. 如果一个节点 v_i 的入度为 k_i^{in}，则背景节点指向 v_i 的边权 $w_{gi}=(k_i^{\text{in}})^{\alpha}$，网络其他节点之间的连接的权重都等于 1，由此得到改进后的 LeaderRank 的迭代公式为

$$LR_i(t)=\sum_{j=1}^{n+1}\frac{w_{ji}}{\sum_{l}^{n+1}w_{jl}}LR_j(t-1). \tag{31}$$

这种改进更加重视网络中的大度节点，在多个数据集上的实验发现新方法比标准的 LeaderRank 的性能在多个方面均有提升. 虽然这一方法的提出最初是

为了提升 LeaderRank 算法在无权网络中的排序效果，但是这种思路也可以应用到含权网络中，关于 LeaderRank 算法在含权网络中的扩展参见 5.5 节.

3.5 HITs 算法

一个网络中不同类型的节点功能不同，每个节点的重要性往往不能由单独的一个指标给出，HITs 算法[80]赋予每个节点两个度量值：权威值(authorities)和枢纽值(hubs). 权威值衡量节点对信息的原创性，枢纽值反映了节点在信息传播中的作用. 枢纽页面是那些指向权威页面的、链接数较多的页面，反映网页上链接的价值. 节点的权威值等于所有指向该节点的网页的枢纽值之和，节点的枢纽值等于该节点指向的所有节点的权威值之和. 因而，节点若有高权威值则应被很多枢纽节点关注，节点若有高枢纽值则应指向很多权威节点. 简单地说，权威值受到枢纽值的影响，枢纽值又受到权威值的影响，最终通过迭代达到收敛.

在一个包含 n 个节点的网络中，定义 a_i^t 和 h_i^t 分别为节点 v_i 在时刻 t 的权威值和枢纽值，于是在每一时间步的迭代中:

$$a_i^t=\sum_{j=1}^{n}a_{ji}h_j^{t-1},\qquad h_i^t=\sum_{j=1}^{n}a_{ij}a_j^t. \tag{32}$$

每一时间步结束后需进行归一化处理:

$$a_i^t=\frac{a_i^t}{\left\|a^t\right\|},\qquad h_i^t=\frac{h_i^t}{\left\|h^t\right\|}. \tag{33}$$

HITs 首次用不同指标同时对网络中的节点进行排序，具有开创意义. HITs 除了可以用于确定一个节点上多个相互关联的属性，还可以处理更复杂的排序问题[81,82]，譬如在信誉评价系统中如何评价用户的信誉度以及产品的质量[83]. 这类评价系统通常包含两类节点(用户和产品)，信誉排序问题解决的是包含两类节点的各自的排序问题. 与 HITs 类似的是两类节点的分数值也是相互影响的，最终通过迭代寻优获得两类节点的排序值. 例如文献[83]利用这种思路提出一种可以有效抵抗恶意评分的排序方法，该方法认为一个商品得到的打分反映了这个商品的质量，自然地，应该给可信度高的用户更大的权重；反过来，一个用户打分的可信度，可以用他的打分和商品质量的接近程度来衡量.

需要指出的是，特殊的网络结构会影响 HITs 算法、PageRank 算法这类应用邻居之间相互传递打分值进行排序的方法的表现. 例如万维网中广泛存在紧密连接社团(tightly-knit community)，社团内节点间非常紧密的链接关系会使这些节点的权威值和枢纽值相互增强(mutual reinforcement)，从而使网页的排序结果更倾向于将社团内部的页面排在前面而偏离搜索的主题，出现主题漂移(topic draft)现象[84].

3.6 自动信息汇集算法

Kleinberg 与其合作者对 HITs 算法进行了改进，提出了自动信息汇集(automatic resource compilation, ARC)算法[85]. HITs 算法仅考虑网页之间的链接关系(即仅考虑网络结构)，ARC 算法在此基础上，还考虑了页面内容与搜索主题的相关性，给每个链接赋予不同的权值，提高页面排序的真实可靠性. 算法的具体过程如下：取一个含有搜索主题 T 的网页的增广集，这个集合中的网页抽象为节点，它们之间的链接抽象为节点之间的连边. 每个节点 v_i 都有权威值 a_i 和枢纽值 h_i，所有节点的初始权威值设为 1. 假设某一个页面上有一个指向另一个页面的链接，如果链接周围有较多关于搜索主题 T 的内容，则认为链接的权值较大. 记 t 为链接前后 B 字节范围内关于主题 T 的内容出现的次数，定义链接的权值 $w_{ij}=t+1$，在每步迭代之后进行归一化. 作者提出 ARC 算法时建议 $B=50$. 接下来，通过下面的迭代过程使权威值和枢纽值达到稳定:

$$\boldsymbol{a}=W\boldsymbol{h},\qquad \boldsymbol{h}=W^{\mathrm{T}}\boldsymbol{a}. \tag{34}$$

与此类似，文献[86]也提出一种同时考虑页面之间的链接和页面内容的排序算法，与 ARC 不同的是它对页面内容采用的是语义分析技术.

3.7 SALSA 算法

SALSA 算法[84]，即链接结构的随机分析法(stochastic approach for link structure analysis)，是 HITs 算法的另一种改进. SALSA 算法不仅考虑了用户在浏览网页时顺着网页之间的链接方向访问网页，还考虑了逆着链接方向访问原来的网页的情况. SALSA 算法用随机游走的方法，通过访问网页的马尔科夫过程来确定网页的权威值和枢纽值的大小. 万维网用有向网络 G 表示，所有入度不为零的节点构成权威集合 S_{A}，所有出度不为零的节点构成枢纽集合 S_{H}，两类节点之间的关系用无向边来表示：图 G 中从节点 v_i 指向 v_j 的边表示为边(i_{H}, j_{A})，由此将原始网络 G

转换为无向二分网络 $\tilde{G}$，图 2 给出一个示例.

用 $\tilde{G}$ 中长度为 2 的路径模拟用户上网的随机游走过程，则每一个随机游走的路径都是从集合 S_A 到集合 S_H 再到集合 S_A 或从集合 S_H 到集合 S_A 再到集合 S_H，其中每一个从集合 S_H 到集合 S_A 的路径都是沿着链接方向访问，每一个从集合 S_A 到集合 S_H 的路径都表示逆着链接方向访问. 每一随机游走过后节点上的权值都会进行重新分配. 于是可以根据枢纽值和权威值定义两个随机游走过程. 对于计算枢纽值而言，初始时刻赋予枢纽集合中的每个节点一单位初始权值，用向量 $\boldsymbol{h}^0$ 表示，权值转换的过程可表示为 $\boldsymbol{h}^t = \tilde{H}\boldsymbol{h}^{t-1}$，其中权值转换矩阵 $\tilde{H}$ 的元素 $(\tilde{H})_{ij} = \tilde{h}_{ij}$ 表示节点 v_j 将其枢纽值传给节点 v_i 的概率，即:

$$\tilde{h}_{ij} = \sum_{i,j\in S_H,\alpha\in S_A} \frac{a_{i\alpha}}{k_j}\cdot\frac{a_{j\alpha}}{k_\alpha}, \tag{35}$$

其中 $a_{i\alpha}$ 为二分网络 $\tilde{G}$ 的邻接矩阵元素，如果节点 v_i ($\in S_A$)与节点 v_α($\in S_H$)相连接则 $a_{i\alpha}=1$，否则 $a_{i\alpha}=0$. k_c 表示二分图中节点 v_c 的度，当 $v_c\in S_A$ 时 k_c 相当于节点 v_c 在图 G 中的入度，当 $v_c\in S_H$ 时 k_c 相当于节点 v_c 在图 G 中的出度. 类似地，权威值的转换过程为 $\boldsymbol{a}^t = \tilde{A}\boldsymbol{a}^{t-1}$，其中转移矩阵 $\tilde{A}$ 的元素 $(\tilde{A})_{\alpha\beta} = \tilde{a}_{\alpha\beta}$ 表示节点 v_β 将其权威值传给节点 v_α 的概率，即:

$$\tilde{a}_{\alpha\beta} = \sum_{\alpha,\beta\in S_A,\, i\in S_H} \frac{a_{i\alpha}}{k_\beta}\frac{a_{i\beta}}{k_i}. \tag{36}$$

多次迭代后每个节点上的值都达到稳定时停止迭代，于是得到节点最终的权威值和枢纽值. 由于计算枢纽值和权威值的随机过程是相互独立的，因此不会出现两者相互增强的情况，相比 HITs 算法而言，SALSA 算法能够更好地避免主题漂移的问题. SALSA 算法实际上考虑的是一个基于二部分图的随机游走过程，这一思路也被成功地应用在信息挖掘的另外两个领域中，即基于网络结构的链路预测问题[87]和个性化推荐算法[25,88]. 实际上这里介绍的 SALSA 算法和推荐算法中的物质扩散算法如出一辙[88,89]，其区别在于以下几点: (1) 推荐系统中的物质扩散算法通常只考虑扩散两步的结果，并不考虑稳态的结果; (2) 在个性化推荐中初始向量的设定根据目标用户不同而异，而 SALSA 算法不会针对某一个节点设置不同的初始向量值; (3) 在推荐算法中通常只考虑用户没有选择过的产品的排序结果，而 SALSA 考虑的是对所有节点的排序结果.

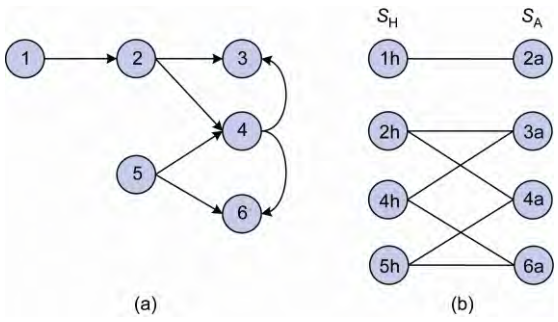

图 2 (网络版彩色)将有向网络(a)转化为一个二分网络(b)

4 基于节点移除和收缩的排序方法

节点(集)的移除和收缩方法与系统科学中确定一个系统的核心的思路暗合[90]，其最显著的特点是在重要节点排序的过程中，网络的结构会处于动态变化之中，节点的重要性往往体现在该节点被移除之后对网络的破坏性. 从衡量网络的健壮性角度看，一些节点一旦失效或移除，网络就有可能陷入瘫痪或者分化为若干个不连通的子网. 实际生活中的很多基础设施网络，如输电网、交通运输网、自来水-天然气供应网络等，都存在“一点故障，全网瘫痪”的风险. 为了预防风险，研究人员提出了很多方法来研究节点收缩或者移除之后网络的结构与功能的变化，从而为新系统的设计与建造提供依据. 比较典型的是系统的“核与核度”理论. 许进等人[91]在定义规则网络图的核概念基础上，提出了核度的测量方法，研究了网络核度与节点数、边数的关系，并根据它们之间的关系设计了规则网络构造定理; 李鹏翔等人[90]认为直接的联系往往是间接联系的必经之路，在评估节点重要性的过程中更加重要，用节点集被删除后形成的所有不直接相连的节点对之间的最短距离的倒数之和来反映节点删除对网络连通的破坏程度; 陈勇等人[92]分析了通信网络，考察去掉节点(集)及其相关边后所得到的图的生成树的数目，数目越小，表明该节点(集)越重要; 谭跃进等人[93]用收缩节点方法替代删除节点法，综合考虑了节点的度以及经过该节点的最短路径的数目，将节点收缩后网络的聚集度作为节点重要性评估的标准. 系统科学的方法给我们提供了新的视角，但由于计算复杂度较高，目前这类方法还仅限于小规模的网络实验. 此外，Restrepo 等人[94]提出通过考察网络最大特征值在移除节点后

的变化来衡量节点重要性的方法，该方法还可以应用于刻画网络连边的重要性.

4.1 节点删除的最短距离法

破坏性反映重要性. 节点删除的最短距离法[90]认为一个节点移除后的破坏性与所引起的距离变化有关: 移除一个节点(集)会引起网络分化，并形成若干个连通分支，网络中节点对之间较短距离的变化越大，被移除的节点就越重要. 该算法区别对待不同长度的路径，认为“相对直接的、近距离的联系所造成的破坏性大于相对间接的、远距离的联系所造成的破坏性”[90]. 具体地，在连通图中一个节点被删除之后，对网络的整体状况的影响体现在两个方面: 直接损失和间接损失.

直接损失是指被删除的节点与其他剩余的节点之间不再存在通路，如果连通网络中共有 n 个节点，删除一个节点后产生的不连通节点对的数目为 $n-1$. 如果删除的是节点集，直接损失还应该包括删除的节点集内节点之间的不再连接的损失. 间接损失是指删除一个节点造成剩余节点之间不连通而引发的损失: 用 N_k (k=1, 2, …, s)表示一个节点 v_i 被删除后，网络分化成的 s 个连通子图中第 k 个连通子图的节点数，则该节点被删除后所形成的不再连通的节点对的数目为 $\sum_{t=1}^{s}\sum_{r=t+1}^{s}N_tN_r$，记由于删除节点 v_i 造成的不再相连的节点对表示为集合 E (包括直接损失和间接损失两部分)，那么节点 v_i 的重要性等于集合 E 中节点对之间的最短距离的倒数之和，即:

$$DSP(i)=\sum_{(j,k)\in E}\frac{1}{d_{jk}}. \tag{37}$$

d_{jk} 为删除节点 v_i 之前 v_j 与 v_k 间的最短距离. 注意，当 j 或 k=i 的时候，相当于直接损失；当 $j\neq k\neq i$ 的时候，相当于间接损失. 节点删除的最短距离法在衡量一些节点集的重要性方面优势比较突出. 在实际的大规模网络中，仅删除一个节点时网络的拓扑图一般不会分化为几个连通子图，网络的间接损失为 0，节点删除的最短距离法效果并不明显. 而如果同时删除多个节点，则很容易使网络不再连通，这时该方法的优越性就显现出来了.

4.2 节点删除的生成树法

在通信网络中，节点删除后网络中节点对之间最短距离会发生变化，但一般对网络时延影响不大，用最短距离法不一定准确. 这时可通过考察节点删除后网络拓扑图的生成树个数来衡量节点的重要性. 在图论中，一个图的树是该图的一个连通的无环子图，一个图的生成树定义为拥有该图的所有顶点的树. 节点删除的生成树法[92]认为一个节点删除后对应的网络的生成树的数目越少，该节点越重要. 给定一个无向连通图，其邻接矩阵为 $\boldsymbol{A}$，网络拉普拉斯矩阵 $\boldsymbol{L}$=$\boldsymbol{D}$–$\boldsymbol{A}$(将矩阵 $\boldsymbol{A}$ 主对角线上的元素 a_{ii} 替换为节点 v_i 的度值，非对角线上的元素值全部乘以–1). 那么，这个连通无向图的生成树个数 t_0 为矩阵 $\boldsymbol{L}$ 的任意一个元素 l_{pq} 的余子式 M_{pq} 的行列式，即: $t_0=\left|M_{pq}\right|$. 删除任意一个节点 v_i，网络的邻接矩阵变为 A_{-i}，然后用上面的方法计算网络的生成树个数为 t_{-i}. 由此可定义节点 v_i 的中心性指标为

$$DST(i)=1-\frac{t_{-i}}{t_0}. \tag{38}$$

在节点的移除对网络的连通性影响不大的网络中，节点删除的生成树法优于最短距离法. 但节点删除的生成树法有一些缺点，例如，只能用在连通网络中. 若一个节点删除后网络变得不再连通，这些节点的重要性就难以判断了，这时可采用节点收缩法评估节点的重要性.

4.3 节点收缩法

节点收缩就是将一个节点和它的邻节点收缩成一个新节点[93]. 如果 v_i 是一个很重要的核心节点，将它收缩后整个网络将能更好地凝聚在一起. 最典型的就是星形网络的核心节点收缩后，整个网络就会凝聚为一个大节点. 从社会学的角度讲，社交网络中人员之间联系越方便(平均最短路径长度 d 越小)，人数越少(节点数 n 越小)，网络的凝聚程度就越高. 因此定义网络的凝聚度为

$$\partial[G]=\frac{1}{n\cdot d}=\frac{1}{n\cdot\dfrac{\sum_{i\neq j}d_{ij}}{n(n-1)}}=\frac{n-1}{\sum_{i\neq j}d_{ij}},$$

其中 d_{ij} 表示 v_i 与 v_j 的最短路径长度. n=1 时，令凝聚度 $\partial[G]=1$，显然 $0<\partial[G]\leqslant 1$. 节点收缩法主要考察节点收缩前后网络凝聚度的变化幅度，由此判定网络中节点的重要性，故定义节点 v_i 的重要性指标为

$$IMC(i)=1-\frac{\partial[G]}{\partial\left[G_{-v_i}\right]}, \tag{39}$$

其中 $\partial\left[G_{-v_i}\right]$ 表示将节点 v_i 收缩后所得到的网络的凝聚度. 由上式可得:

$$IMC(i)=1-\frac{n\cdot d(G)-(n-k_i)\cdot d\left(G_{-v_i}\right)}{n\cdot d(G)}. \quad (40)$$

可见, 节点收缩法中节点的重要程度由节点的邻居数量和节点在网络路径中的位置共同决定. 由于每次收缩一个节点, 都要计算一次网络的平均路径长度, 时间复杂度比较高, 不适于计算大规模网络.

4.4 残余接近中心性

为了研究网络的抗毁性, Dangalchev[95]提出了残余接近中心性(residual closeness centrality), 用来衡量节点的移除对网络带来的影响. 残余接近中心性认为若一个节点的删除使得网络变得更加脆弱, 该节点就越重要. 文献[43]对接近中心性的改进使得接近中心性应用的范围从连通图扩展到了非连通图. 该方法对接近中心性进行了改进, 分母取以 2 为底的指数, 相当于提升了短路径的影响力, 同时会使本算法更易计算和扩展(文献[43]给出了将几个图合并为一个图计算接近中心性的详细算法). 在移除一个节点 v_i 之后, 定义其残余接近中心性为

$$RCC(i)=\sum_j\sum_{k\neq j}\frac{1}{2^{d_{jk}(-i)}}, \quad (41)$$

其中 $d_{jk}(-i)$为删除节点 v_i 之后, 节点 v_j 与 v_k 的最短距离. 残余接近中心性在测度网络的脆弱性方面比图坚韧度(graph toughness)、离散数(scattering number)、节点完整度(vertex intergrity)[1)]等方法表现要好. 基于该方法可以定义出边的残余接近中心性和节点集、边集的残余接近中心性.

5 含权网络中的节点中心性

无权网络采用粗粒化的二分法来表示网络中节点间的联系(有边为 1, 无边为 0), 不考虑联系的强弱信息. 然而边的权重信息能帮助我们更加细致地理解网络的结构与功能. 如在社交网络中, 边的权值可代表情感关系的强弱、交流与服务的频次、任务执行时间的长短等. 科学家合作网络中可用两个科学家合作论文的数量刻画两个科学家的联系紧密性. 航空运输网络中可以用两个机场之间所有班次上的座位数表示这两个机场的通勤情况. 那么, 如何能够有效地利用网络的边权重信息进行重要节点的挖掘呢? 到目前为止, 大多数的研究思路都是将基于无权网络中心性指标在含权网络上进行扩展应用, 专门针对含权网络进行设计的方法鲜见.

5.1 含权的度中心性

为了更好地利用含权网络中的边权信息, 至少应当在统计层次上找到一个合适的、能表征网络的特性的量. 统计显示, 两个节点之间的连边的权重 w_{ij} 服从右偏分部(right-skewed distribution)[96], 这也揭示了网络的异质性. 以往大部分工作都是基于节点度的, 单单靠边的权重还不能为我们提供一个很好的观察网络复杂性的图景, 我们需要通过重新定义含权网络中节点的度. 为了加以区别, 含权网络中节点 v_i 的度称为强度(strength), 定义为所有与 v_i 相连的边的权值之和:

$$b_i=\sum_{j=1}^{n}w_{ij}, \quad (42)$$

其中 $W=(w_{ij})$为网络的含权邻接矩阵, w_{ij} 为连边(v_i, v_j)上的权值, 两节点之间无连边则 $w_{ij}=0$. 同理, 可定义含权网络中的出度为与 v_i 相连的出边的权值之和, 即 b_i^{out}, 含权网络中的入度为与 v_i 相连的入边的权值之和, 即 b_i^{in}. 进一步定义含权网络中节点 v_i 的度中心性为

$$WDC(i)=\frac{b_i}{\sum_{j=1}^{n}b_j}. \quad (43)$$

含权网络中, 公式(43)在计算节点中心性时比较通用, 即由边权的累和确定节点的中心性. 受此启发, Gao 等人[17]模拟阿米巴虫寻找食物来确定网络中每条路径上的流量, 同样的思想, 用边上流量的累和确定节点的中心性. 需要特别指出的是, 含权网络中节点强度的定义并不统一, 还有其他多种定义方式, 比较有代表性的是 Garas 等人[97]提出的用节点的邻居数量和节点周围的边权共同表示含权网络中节点的度值(参见公式(47)).

1) 定义一个图 G 的节点集为 V, 连通片的个数为 $c(G)$, 其中最大联通片中的节点数目为 $m(G)$. 定义图 G 的一个分割集(separator) S 为: $S\subset V$ 并且 $c(G[V\setminus S])>1$, 则图坚韧度(graph toughness)定义为: $GT(G)=\min\{|S|/c(G[V\setminus S])\}$; 离散数(scattering number)定义为: $SN(G)=\max\{c(G[V\setminus S])-|S|\}$; 节点完整度(vertex intergrity)定义为: $VI(G)=\min\{|S|+m(G[V\setminus S])\}$

5.2 *H*-度中心性

节点的强度(strength)经常会被视为节点的度在含权网络中的扩展，以此为前提将很多不含权网络中的性质向含权网络中推广[96]. 但此前提有着极大的隐患. 节点 v_i 和 v_j 的度相等时说明节点 v_i 与 v_j 有相同个邻节点，而在含权网络中，当节点 v_i 和 v_j 的有相同的强度时并不能说明节点 v_i 与 v_j 是否有相同个邻节点，甚至当节点 v_i 的强度大于 v_j 的时候，也不能说明 v_i 有着比 v_j 多的邻居节点. 一个简单的例子如图 3 所示. 显然，与节点的度相比，强度在表征节点性质时有一定的信息丢失. 信息计量学中用 H 指数[98]衡量学者的贡献：如果一个人在其所有学术文章中最多有 n 篇论文分别被引用了至少 n 次，他的 H 指数就是 n. Korn 等人[99]受此启发，提出用 Lobby 指数衡量无权网络中的节点的重要性：一个节点的 Lobby 指数为 n 如果该节点至多有 n 个邻居且这些邻居的度至少为 n. Zhao 等人[100]将 Lobby 指数扩展到含权网络中，用 H-度中心性来更好地衡量含权网络中节点的重要性：一个节点的 H-度中心性为 n，如果至多有 n 个权重不小于 n 的边与之相连. 一定程度上，H-度中心性可看成单纯用节点强度或度衡量中心性的一种折中. 不过这种方法有一个巨大缺陷，就是边权值的取值或者数量级需要在合适的范围，否则无法得到有效的排序结果.

5.3 含权的介数中心性和接近中心性

无权网络中，节点的介数中心性是所有节点对的最短路径中，通过该节点的数目的比例. 含权的网络中，节点之间的路径长短由边权决定. 如在传染病传播过程中，病毒更有可能在两个接触密切的人(边权大)之间快速传播. 于是，采用边权的倒数来衡量

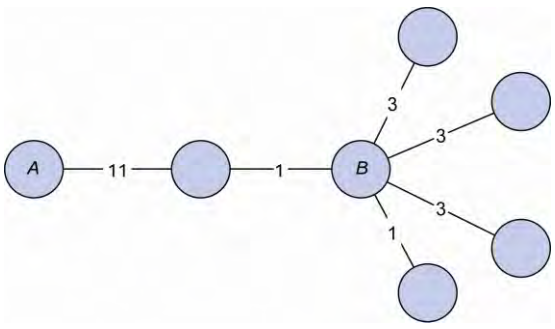

图 3 (网络版彩色)节点 *A* 和 *B* 有相同强度却有不同性质[100]

距离的长短，比如一条边的权重是另一条边的权重的两倍，那么前者的距离就是后者距离的一半. 此时，两个节点 v_i 和 v_j 之间的距离表示为

$$d_{ij}^{w}=\min\left(\frac{1}{w_{ih}}+\cdots+\frac{1}{w_{hj}}\right),$$

其中节点 v_h 是从 v_i 到 v_j 的路径上的中间节点. 实际上，上面的式子忽略了一条路径中中间节点的数目对评估路径长度的影响，这对于万维网、输电网、通信网等路径上节点数目的多少对网络流传输的干扰和时延不太敏感的网络是适用的. 但在很多网络中，中间节点的数目对网络的传输效率会产生不可忽略的影响，例如有些情况下中间节点越多，传播过程中的时延就越大，有些情况却相反. 为了体现中间节点的数目对路径的影响，需要添加一个可调节的参数[101]，于是，节点 v_i 和 v_j 之间的距离表示为

$$d_{ij}^{w\alpha}=\min\left(\frac{1}{\left(w_{ih}\right)^{\alpha}}+\cdots+\frac{1}{\left(w_{hj}\right)^{\alpha}}\right), \tag{44}$$

其中，

$$\begin{cases}\alpha=0, & \text{(不含权网络)},\\ 0<\alpha<1, & \text{(含权网络, 考虑边权重, 同时中间节点越少, 距离越短)},\\ \alpha=1, & \text{(含权网络, 考虑边权重, 不考虑中间节点的数目)},\\ \alpha>1, & \text{(含权网络, 考虑边权重, 同时中间节点越多, 距离越短)}.\end{cases}$$

有了含权网络中节点距离的定义，就很容易推导出含权网络中接近中心性和介数中心性的定义：

$$WCC(i)=\left[\sum_{j}^{n}d_{ij}^{w\alpha}\right]^{-1}, \tag{45}$$

$$WBC(i)=\sum_{i\neq j\neq k}\frac{g_{jk}^{w\alpha}(i)}{g_{jk}^{w\alpha}}, \tag{46}$$

其中，$g_{jk}^{w\alpha}$ 为从节点 v_j 到 v_k 的所有最短路径的数目，$g_{jk}^{w\alpha}(i)$ 为从节点 v_j 到 v_k 的 $g_{jk}^{w\alpha}$ 条最短路径中经过 v_i 的最短路径的数目.

5.4 含权的 *k*-壳分解法

不含权的 k-壳分解法中，用迭代法层层剥去外围节点，以确定节点在网络中的位置，依靠位置信息来确定节点的重要性. 含权网络中，重新定义了节点

的度: 一个节点的强度等于与该节点相连的所有边的权值之和, 这种定义的缺点显而易见, 完全忽视了节点的连接状况, 经常出现有很多邻居而节点的强度却很小的情况. Garas 等人[97]提出了另一种节点强度的定义方式, 从而将经典的 k-壳分解法扩展到了含权网络中, 在这里, 重新定义节点 v_i 的度值为

$$k_i' = \left[k_i^{\alpha} \left(\sum_j^{k_i} w_{ij} \right)^{\beta} \right]^{\frac{1}{\alpha+\beta}}, \tag{47}$$

其中 k_i 为节点 v_i 邻居的数目, w_{ij} 为节点 v_i 与其邻居 v_j 之间连边的权值, α 和 β 为参数, 用以调节前两者的比重. 文献[97]中采用最简单的形式, 将这两个指标同等对待: $\alpha=\beta=1$. 于是 $k_i' = \sqrt{k_i \left(\sum_j^{k_i} w_{ij} \right)}$, 其中 $\sum_j^{k_i} w_{ij}$ 即为 5.1 节中介绍的节点的强度 b_i. 当 $\alpha=1, \beta=0$ 的时候, 公式(47)就变成无权网络的情形. 该方法具有很强的可扩展性. 实验发现, 含权重的 k-壳分解法更加全面地考虑了节点所处的周围环境, 对节点的属性刻画更为细致, 因此划分的层级数目远远多于不含权重的 k-壳分解法, 能更加准确地找到重要节点.

5.5 含权的 LeaderRank 算法

标准的 LeaderRank 算法是针对无权有向网络的一种简单, 能够快速收敛, 重要节点识别准确, 抗攻击和干扰能力强的高效算法. 在 3.4 节介绍的 Li 等人[79]对其进行的改进算法同样适用于含权网络. 下面给出含权 LeaderRank 算法的一般表达形式

$$WLR_i(t) = \sum_{j=1}^{n+1} w_{ji} \frac{WLR_j(t-1)}{b_j^{\text{out}}}, \tag{48}$$

其中 $b_j^{\text{out}} = \sum_{i=1}^{n+1} w_{ji}$ 为含权有向网络中节点 v_j 的出强度. 在计算时可以假设与背景节点相连的所有边的权重是 1. 当然, 如果需要也可以对这个权重进行调整, 用以调节背景节点的作用. 不同网络里边权在节点的重要性排序中对节点作用不同, 为此, 我们提出了一个含参数的 LeaderRank 方法, 具体定义为

$$WLR_i(t) = \sum_{j=1}^{n+1} w_{ji}^{\alpha} \frac{WLR_j(t-1)}{b_j^{\text{out}}}, \tag{49}$$

其中 $b_j^{\text{out}} = \sum_{i=1}^{n+1} w_{ji}^{a}$, α为一个自由参数用来调节权重. 显然, 当α=0 的时候, 算法回到原始的无权情况; 当α=1 的时候就回到公式(48)所述的无参含权 LeaderRank 算法.

5.6 基于 D-S 证据理论的节点中心性

在含权网络中, 节点的中心性程度与节点的邻居数、节点的强度关系密切. 直观地看, 一个节点的邻居数越多或者强度越大就越重要. 用来处理不确定性推理问题的 D-S 证据理论, 为我们将这两个因素整合在一起提供了依据. 我们先简单介绍一下 D-S 证据理论.

D-S 证据理论比贝叶斯定理需要的条件要弱 (不必满足概率的可加性), D-S 证据理论中的基本概率分配(basic probability assignment, BPA)具有直接表达“不确定”和“不知道”两种状态的能力[102]. 对于某个假设, 可分别计算出关于该假设的信任函数 $Bel(x)$和似然函数 $Pl(x)$组成信任区间[$Bel(x)$, $Pl(x)$], 用以表示对该假设的确认程度.

D-S 证据理论涉及的理论比较多, 这里只简要介绍本方法中用到的一些概念. 用 k_M 和 k_m 分别表示网络中节点邻居数的最大值和最小值, 用 w_M 和 w_m 分别表示网络中节点强度的最大值和最小值. 在度或强度指标下, 对一个节点是“重要节点”和该节点是“不重要节点”的信任程度分别用 *high* 和 *low* 表示, 由此得到一个识别框架 $\theta=(high, low)$, 也称假设空间. 基本概率分配函数 $m_{di}(h)$ 和 $m_{di}(l)$ 分别表示用节点度来衡量重要性时, 对 v_i 是重要节点的信任程度和对 v_i 不是重要节点的信任程度. 同理 $m_{wi}(h)$ 和 $m_{wi}(l)$ 分别表示用节点的强度来衡量重要性时, 对 v_i 是重要节点的信任程度和对 v_i 不是重要节点的信任程度, 其取值为

$$\begin{aligned} m_{di}(h) &= \frac{|k_i - k_m|}{k_M - k_m + 2\mu}, \ m_{di}(l) = \frac{|k_i - k_M|}{k_M - k_m + 2\mu}, \\ m_{wi}(h) &= \frac{|w_i - w_m|}{w_M - w_m + 2\varepsilon}, \ m_{wi}(l) = \frac{|w_i - w_M|}{w_M - w_m + 2\varepsilon}, \end{aligned} \tag{50}$$

其中 $0<\mu,\varepsilon<1$ 是两个参数(文献[15]指出这两个参数对节点最终的排序没有影响, 可取 $\mu=\varepsilon=0.15$). 于是, 我们可以得到节点 v_i 的两个基本概率分配函数:

$$\begin{aligned} M_d(i) &= \left(m_{di}(h), m_{di}(l), m_{di}(\theta)\right), \\ M_w(i) &= \left(m_{wi}(h), m_{wi}(l), m_{wi}(\theta)\right), \end{aligned} \tag{51}$$

其中, $m_{di}(\theta)=1-m_{di}(h)-m_{di}(l)$, $m_{wi}(\theta)=1-m_{wi}(h)-m_{wi}(l)$, 表示在两种度量下, 不知道节点重要不重要的信任程度. 节点的度和节点的强度可看成两种重要性指

标，引入 Dempster 证据合成规则，将节点 v_i 的度和强度融合在一起，形成一个新的指标 $M(i)$：

$$M(i)=\left(m_i(h),m_i(l),m_i(\theta)\right), \tag{52}$$

其中，

$$\begin{aligned}
m_i(h)&=\left[m_{di}(h)\cdot m_{wi}(h)+m_{di}(h)\cdot m_{wi}(\theta)\right.\\
&\quad\left.+m_{wi}(h)\cdot m_{di}(\theta)\right]\Big/(1-K),\\
m_i(l)&=\left[m_{di}(l)\cdot m_{wi}(l)+m_{di}(l)\cdot m_{wi}(\theta)\right.\\
&\quad\left.+m_{wi}(l)\cdot m_{di}(\theta)\right]\Big/(1-K),\\
m_i(\theta)&=\left[m_{di}(\theta)\cdot m_{wi}(\theta)\right]/(1-K),\\
K&=m_{di}(h)\cdot m_{wi}(l)+m_{di}(l)\cdot m_{wi}(h).
\end{aligned} \tag{53}$$

通常情况下，将 $m_i(\theta)$的值平分给 $m_i(h)$和 $m_i(l)$. 即

$$\begin{aligned}
M_i(h)&=m_i(h)+\frac{1}{2m_i(\theta)},\\
M_i(l)&=m_i(l)+\frac{1}{2m_i(\theta)}.
\end{aligned} \tag{54}$$

$M_i(h)$和 $M_i(l)$分别代表节点是重要节点和节点是不重要节点的信任程度，对一个节点来说，$M_i(h)$越大，同时 $M_i(l)$越小，节点的重要性就越大. 于是，定义节点的证据重要性 EVC 为

$$\mathrm{EVC}(i)=M_i(h)-M_i(l)=m_i(h)-m_i(l). \tag{55}$$

EVC 方法在使用 D-S 证据理论时默认节点的度是均匀分配，文献[16]对此进行了改进，使其能够用在度值为幂律分布等的网络中. 另外，文献[16]还用证据理论将半局部中心性的使用范围扩展到了含权网络. 用 D-S 证据理论将多个影响因素有效结合，有较强的可扩展性.

6 节点重要性排序方法的评价标准

网络科学研究的早期，所关注的网络中节点数目较少，典型的有同性恋接触网络[45,90]、女生用餐伙伴选择网络[103]、空手道俱乐部网络[45]等，对于这些小规模网络，可以通过调查问卷等方式对每个节点的重要性进行打分，然后将实际的调查结果作为标准与其他算法结果进行比较，分析各种方法的表现和优劣. 随着科技的发展和进步，大数据时代已经来临，现在我们所面对的网络规模迅速增长，想要得到一个对所有节点的重要性的较为客观的评价标准极为困难. 目前评价各种排序算法优劣的主要思路是：将排序算法得出的重要节点作为研究对象，通过考察这些节点对网络某种结构和功能及对其他节点状态的影响程度来判断排序是否恰当. 例如，如果一个排序算法得出节点 v_i 比 v_j 更重要，单独考察 v_i 比 v_j 发现前者对网络的结构功能或对其他节点的影响程度更大，就说明这种排序算法比较符合实际. 常用来评价各排序算法的方法有基于网络的鲁棒性和脆弱性方法以及基于网络的传播动力学模型的方法.下面分别对这两类方法进行简单的介绍.

6.1 用网络的鲁棒性和脆弱性评价排序算法

本类方法着重考察网络中一部分节点移除后网络结构和功能的变化，变化越大移除的节点越重要. 用某一种重要节点挖掘方法将网络中所有节点按重要性进行排序，然后按重要性从大到小的顺序，将一部分节点从网络中移除，用$\sigma(i/n)$表示移除 i/n 比例的节点后，网络中属于巨片(giant component)[104]的节点数目的比例，网络的鲁棒性(robustness)可用 R-指标刻画[105]:

$$R=\frac{1}{n}\sum_{i=1}^{n}\sigma(i/n). \tag{56}$$

显然，不论对何种算法，星形图中，R 取最小值$(1/n-1/n^2)$，完全图中 R 取最大值$(1-1/n)/2$，当 n 比较大时 $R\in(0,1/2)$. 可定义 $V=1/2-R$ 来表示网络对于所实施的移除方法的脆弱性(vulnerability)，可见，V-指标越大表示采用该方法进行攻击的效果越好. V-指标和 R-指标可从整体上反应各种重要节点挖掘方法的有效性. 另外也可画出 i/n 与$\sigma(i/n)$在二维坐标上的曲线，对节点移除的影响进行详细分析. 例如文献[106]中考察了在无标度网络中使用 4 种排序方法移除节点后对网络最大连通集的影响，这 4 种方法包括度中心性、介数中心性、接近中心性和特征向量中心性，并和随机移除节点的方法进行比较. 用于实验的无标度网络节点数为 n=10000，平均度为 4 (图 4(a)) 和 6(图 4(b))，移除节点时采用同时移除的方法.

6.2 用传播动力学模型评价排序算法

复杂网络上传播研究的对象极广[107,108]，比如通信网络中的病毒传播[109]、社会网络中的信息传播[110]、电力网络中的相继故障[111]、经济网络中的危机扩散等[35]. 在评价各种节点重要性挖掘方法时广泛采用的是传染病模型，主要包括 SIS 模型[13]和 SIR 模型[32,78]. 在 SIS 模型中一个节点的传播能力被定义为稳态下该节点被感染的概率；在 SIR 模型中，一个

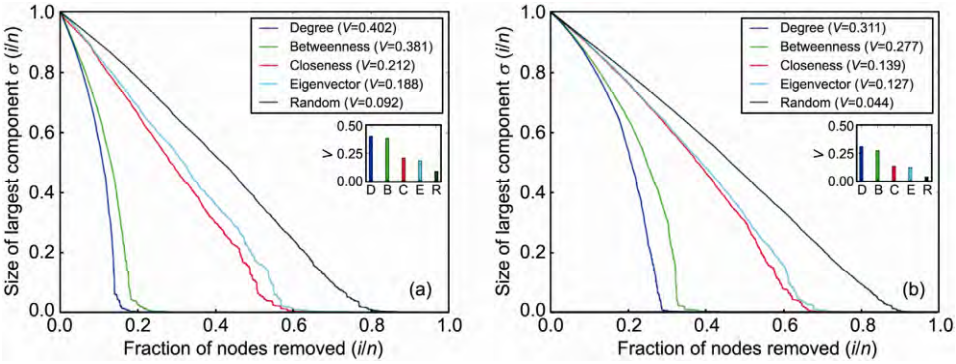

图 4 无标度网络中移除节点数目和网络最大连通集规模的关系[106]

(a) 平均度为 4 的无标度网络; (b) 平均度为 6 的无标度网络

节点的传播能力被定义为该节点的平均传播范围.

下面简要介绍 SIR 模型及一个应用的例子. SIR 假设网络中的节点有三个状态: 易染态 S (susceptible, 可被处于感染态的邻节点感染), 感染态 I (infected, 处于 I 态的节点一定时间后会变为免疫态), 免疫态 R (recovered, 免疫态的节点不会被感染, 也不会传播病毒). SIR 模型有单点接触和全接触两种[112], 前者指在每一时间步内, 处于 I 态的节点感染其邻居的时候将随机选择一个 S 态的邻居, 然后以概率 p 使其由 S 态变为 I 态; 后者指处于 I 状态的节点感染邻居的时候选择的是所有 S 态的邻居, 每个 S 状态的邻居都有机会以概率 p 转变为 I 态. 设置一个(组)节点为初始感染节点(即处于 I 态), 观察每一时间步网络中感染过的节点数目和最终稳定态时(没有 I 态的节点时)感染过的节点数目, 可通过病毒的传播速度和范围两个方面来考察节点的真实影响力. 要对比两种重要节点挖掘方法的优劣, 可分别用这两种方法对网络中的节点按重要性进行排序, 取相同数目的最重要的节点设为初始感染态, 用 SIR 模型在网络上进行实验, 如果一个排序方法的结果使得网络流传播地又快又广, 则说明该重要节点排序方法优于其他方法. 例如文献[78]中应用 SIR 模型比较了 LeaderRank 算法和 PageRank 算法的排序结果. 图 5 显示了使用两种方法获得的前 20 个(图 5(a))最重要的节点中, 以不同的节点为初始感染源进行 SIR 传播的过程. 可见, 以 LeaderRank 获得的节点为初始感染源的传播又快又广, 说明 LeaderRank 算法比 PageRank 算法更能够识别网络中传播影响力高的节点. 图 5(b)为考虑前 50 个节点的情况.

需要注意的是, 网络中信息传播和病毒传播有很大的不同. 文献[110]深入比较了信息传播与病毒传播的不同, 提出了网络中的信息传播模型. 文中还全面总结了影响网络流在网络中传播速度和快慢的 7 种因素, 比如边的强度、信息内容、传播者的角色、记忆效应、时间延迟效应等. 因此, 在评价节点信息传播影响力的时候, 例如社交网站上意见领袖挖掘, 应该考虑更加符合实际传播方式的模型.

7 总结与展望

大自然的发展演化并不是尊崇简单的随机规则,

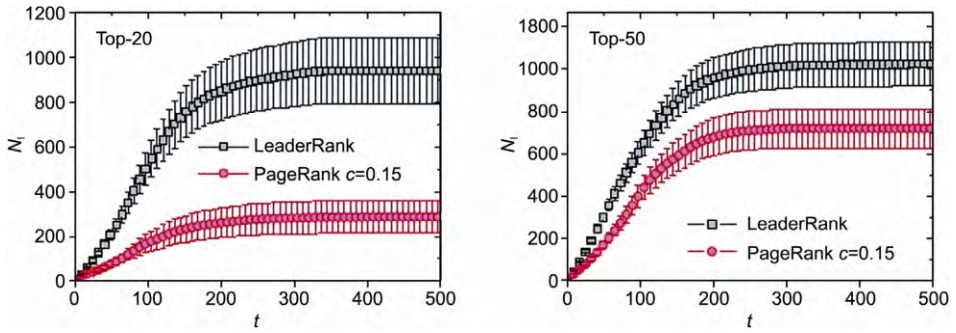

图 5 用 LeaderRank 和 PageRank 算法获得的最重要节点的传播影响力比较[78]

而是受到一些潜在规则的支配，对这些规则的探索吸引着无数研究人员的热情. 最近十几年异军突起的复杂网络科学为我们提供了一系列全新的理论与方法来探索大自然的奥秘. 复杂网络理论能够将形形色色的各种复杂系统表征为网络的结构，网络中节点的属性与功能对网络的生长演化影响重大，复杂网络最重要的模型之一——Barabási-Albert 模型[8]就是建立在与重要节点优先连边的基础上的. 重要节点挖掘是网络攻击和网络流传播及控制等领域中最核心的问题之一，这些少数的节点对网络的控制力和影响力超乎想象. 我们已经明白造成网络节点角色差异的根本原因是网络结构的异质性[113]，但困难的是如何找出哪些节点才是最重要的节点，研究人员已经提出非常多的重要节点挖掘方法，这些方法站在各自的立场，从不同角度为我们探索不同背景下节点的重要性提供可选的方案. 纵观各种方法，可看出它们基本上遵从以下几个思路:

(1) 从节点的局部环境考虑. 节点的局部环境包括直接邻居、间接邻居及连边、节点的聚类系数等. ClusterRank 同时考虑节点的度和聚类系数. 度中心性考虑节点直接邻居的数量，半局部中心性考虑节点四层邻居的信息. 在考虑数量的同时，还有一些挖掘方法从邻居节点的重要性相互增强角度进行了探索，这主要是指基于特征向量的系列方法. 另外，接近中心性、Katz 中心性和信息指标从节点与全局范围内所有节点的联系强弱角度评估节点的重要性.

(2) 从节点所处的位置考虑. 这里包括节点在路径中的位置和节点在网络中的位置两方面. 前者主要包括图中心性、介数中心性及其他基于路径的挖掘方法，而 *k*-壳分解法则是后者的典型代表.

(3) 从节点对网络功能的影响考虑. 这类方法主要考察将节点移除后网络结构和功能的变化，例如，节点删除的最短距离法和残余接近中心性主要关注网络中平均最短距离的变化，生成树法关注节点删除后网络生成树的变化，节点收缩法关注节点删除后网络凝聚度的变化.

除了上面提到的几种评价算法外，混合指标的综合评价方法在实际应用中往往能取得比较好的效果. 混合指标的方法[61]主要有综合加权法、模糊综合评价法、灰色系统评价法、层次分析法、多元统计分析法(主成分分析法、因子分析法、聚类分析法和判别分析等)、TOPSIS 法、神经网络法等.

综上所述，一方面，复杂网络的理论和方法为原本属于信息挖掘领域的重要节点识别问题给出了全新的视角，另一方面，重要节点的挖掘对于网络科学的理论研究和应用拓广也有重要的意义和价值. 现有的工作只是起点，还有很多需要进一步研究和完善的地方. 我们认为主要存在的问题包括以下 3 个方面: 一是理论研究跟不上应用需求，二是结构重要性和功能重要性脱轨，三是理论和实验脱轨. 我们总结了在该领域未来可能成为研究热点的八大问题，希望在不久的将来，这些问题能得到完美解决.

(1) 针对大规模动态变化的网络设计快速有效的重要节点挖掘方法. 目前的算法存在“简单而不准确，准确但太复杂”的缺点. 基于全局信息的方法相对准确但难以在大规模网络中应用. 如何在网络规模达到上千万甚至过亿节点的情况下快速而准确的识别出重要节点，以及在网络结构不断变化情况下的增量计算问题视为挑战一.

(2) 重要节点对网络特定结构和功能的影响. 所谓的“重要”只有结合网络的结构和功能才有意义. 因此，在衡量算法有效性的时候必须结合具体网络的结构和某一功能目标进行. 例如，考察分析各种方法挖掘出来的重要节点对于网络结构稳定性和传播的影响. 如何比较不同的排序方法对于目标功能的适用程度，最终给出各种有代表性的排序方法在不同类型的网络及不同目标功能上表现情况的完整图景视为挑战二.

(3) 在相互关联的网络中如何确定一个节点的重要性. 不同类型的社交网络往往迎合不同的用户需求，尽管一个人可能在多个网络上都注册过，但每类社交网络都有其特定的活跃用户群. 例如，针对一批在腾讯 QQ、微博、微信都有用户名而在不同网络上活跃度不一的用户，当如何确定其影响力，这方面的研究鲜见报道. 如何通过考察一个用户在不同网络上的行为来评估用户的影响力，即可以是某一个领域或小圈子的影响力，也可以是整个社交圈子的影响力，是前沿的研究热点，视为挑战三.

(4) 利用功能表现挖掘重要节点的反问题研究. 在不知道网络结构或只知道部分结构的情况下，通过分析网络的一些功能表现判断节点的重要性，并分析这种方法得到的结果和通过结构信息得到的结果之间的区别和联系，并尝试将两种方式结合，设计更加精确的预测、排序算法. 这类反问题研究极具理

论价值和应用意义，无疑将成为未来研究的一大热点，视为挑战四.

(5) 含时网络中用户的影响力评价方法及预测. 现实的网络大都是随着时间不断演化的，这类网络被称为含时网络. 含时网络的研究已成为网络科学研究的热点之一[114,115]. 在这类网络中，随着网络的演化节点的重要性也在不断变化. 而目前的大多数研究方法都是针对静态网络或者动态网络某一时刻的快照进行的. 在含时网络中如何评估节点的影响力以及预测节点未来影响力，特别是在网络结构变化之前的预测更具意义，此视为挑战五.

(6) 重要节点在网络演化过程中的作用以及对网络控制的意义. 对于现实系统的认识、理解和预测的最终目的是能够实现有效的控制. 如何找到最佳的控制点，从而减少成本，提高效率以达到某种既定目标是最核心的问题. 例如，众所周知的“富者愈富”机制被认为是无标度网络形成的两大机制之一，然而这种倾向富者的增长方式对于系统多样性的发展却是不利的. 研究不同节点在网络演化中的角色以更好地引导系统健康的发展成为具有创新意义之挑战六.

(7) 用真实网络验证各种算法得出的重要节点的真实影响力. 目前，大部分的研究都是使用网络模型对各种算法进行效果评估. 例如，使用 SIS/SIR 模型计算节点的传播影响力，或者采用攻击的方法确定节点的重要性大小. 需要注意的是，这些模型和方法本身只是对现实世界的模拟和高度抽象，虽然具有一定的指导意义，但是结果往往和真实情况差异很大. 这是由于真实系统中的个体的行为具有很大的不确定性，从而很难通过模型进行精确地刻画. 在此情况下，利用真实系统做实验来验证节点重要性不仅将提高评价可信度，还可能会产生不同于以往模型评价的颠覆性结果，有望催生出全新的节点重要性评价体系. 但是真实实验的方法往往难以重复且成本较高. 如何设计真实系统中可行的实验以更加客观有效准确地评估算法优劣视为挑战七.

(8) 基于已有成果开发实际产品并在真实平台上验证实际应用效果. 在充分进行小规模真实实验后，可应用于开发实际产品，收集数据并在更大范围内开展有监督的实验. 例如，可基于微博平台找到影响力高端用户，并将结果应用于电子商务网站中，进行产品营销和广告投放策略的设计等. 与此同时，通过观察和收集用户的真实反馈情况，帮助提高算法的性能，从而进一步改善产品质量和服务. 基于重要节点挖掘理论和技术的相关应用产品开发，以及应用领域探索是最终研究的应用价值所在，在此视其为挑战八.

最后，网络科学已经成为最大的交叉科学和最具活力的交叉科学. 网络科学的发展，有望推动甚至解决若干具有重大理论和实践价值的大科学工程问题，包括解释和预测金融经济危机的发生和发展[116]，理解人际关系的变动和社会结构的形成[117]，优化互联网时代海量信息的组织、共享、导航和推荐[25]，预测和控制全球传染病的流行[118]，揭示生物基本功能得以实现的内在机制[119]等. 我们相信，在这些问题上的任何一点进步，都会带给人类从知识创造到社会经济再到生活健康的巨大价值.

致谢 感谢电子科技大学张千明博士对于本文算法复杂性的讨论.

参考文献

1 Newman M E J. Networks: An Introduction. Oxford: Oxford University Press, 2010
2 Albert R, Jeong H, Barabási A L. Error and attack tolerance of complex networks. Nature, 2000, 406: 378–382
3 Callaway D S, Newman M E J, Strogatz S H, et al. Network robustness and fragility: Percolation on random graphs. Phys Rev Lett, 2000, 85: 5468
4 Cohen R, Erez K, Ben-Avraham D, et al. Breakdown of the Internet under intentional attack. Phys Rev Lett, 2001, 86: 3682
5 Weng J, Lim E P, Jiang J, et al. Twitterrank: Finding topic-sensitive influential twitterers. Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010. 261–270
6 Vitali S, Glattfelder J B, Battiston S. The network of global corporate control. PLoS One, 2011, 6: e25995
7 Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ networks. Nature, 1998, 393: 440–442
8 Barabási A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286: 509–512

9 吕琳媛, 陆君安, 张子柯, 等. 复杂网络观察. 复杂系统与复杂性科学, 2010, 7: 173–186
10 Newman M E J. The structure and function of complex networks. SIAM Rev, 2003, 45: 167–256
11 Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys Rev Lett, 2001, 86: 3200
12 Castellano C, Pastor-Satorras R. Thresholds for epidemic spreading in networks. Phys Rev Lett, 2010, 105: 218701
13 Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. Nat Phys, 2010, 6: 888–893
14 Chen D B, Lü L, Shang M S, et al. Identifying influential nodes in complex networks. Physica A, 2012, 391: 1777–1787
15 Wei D, Deng X, Zhang X, et al. Identifying influential nodes in weighted networks based on evidence theory. Physica A, 2013, 392: 2564–2575
16 Gao C, Wei D, Hu Y, et al. A modified evidential methodology of identifying influential nodes in weighted networks. Physica A, 2013, 392: 5490–5500
17 Gao C, Lan X, Zhang X, et al. A bio-inspired methodology of identifying influential nodes in complex networks. PLoS One, 2013, 8: e66732
18 任卓明, 邵凤, 刘建国, 等. 基于度与集聚系数的网络节点重要性度量方法研究. 物理学报, 2013, 62: 128901
19 任卓明, 刘建国, 邵凤, 等. 复杂网络中最小 K-核节点的传播能力分析. 物理学报, 2013, 62: 108902
20 Gladwell M. The tipping point: How little things can make a big difference. Hachette Digital, Inc., 2006
21 Lü L, Zhou T. Link prediction in complex networks: A survey. Physica A, 2011, 390: 1150–1170
22 吕琳媛. 复杂网络链路预测. 电子科技大学学报, 2010, 39: 652
23 吕琳媛, 周涛. 链路预测. 北京: 高等教育出版社, 2013
24 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展. 自然科学进展, 2009, 19: 1–15
25 Lü L, Medo M, Yeung C H, et al. Recommender systems. Phys Rep, 2012, 519: 1–49
26 朱郁筱, 吕琳媛. 推荐系统评价指标综述. 电子科技大学学报, 2012, 41: 163–175
27 Zhang Z K, Zhou T, Zhang Y C. Tag-aware recommender systems: A state-of-the-art survey. J Comput Sci Tech, 2011, 26: 767–777
28 赫南, 李德毅, 淦文燕, 等. 复杂网络中重要性节点发掘综述. 计算机科学, 2007, 34: 1–5
29 孙睿, 罗万伯. 网络舆论中节点重要性评估方法综述. 计算机应用研究, 2012, 29: 3606–3608
30 刘建国, 任卓明, 郭强, 等. 复杂网络中节点重要性排序的研究进展. 物理学报, 2013, 62: 179801
31 Burt R S, Minor M J, Alba R D. Applied network analysis: A methodological introduction. Sage Publications Beverly Hills, 1983
32 Bonacich P. Factoring and weighting approaches to status scores and clique identification. J Math Sociol, 1972, 2: 113–120
33 Freeman L C. Centrality in social networks conceptual clarification. Soc Netw, 1979, 1: 215–239
34 Chen D B, Gao H, Lü L, et al. Identifying influential nodes in large-scale directed networks: The role of clustering. PLoS One, 2013, 8: e77455
35 汪小帆, 李翔, 陈关荣. 网络科学导论. 北京: 高等教育出版社, 2012
36 Zeng A, Zhang C J. Ranking spreaders by decomposing complex networks. Phys Lett A, 2013, 377: 1031–1035
37 Liu J G, Ren Z M, Guo Q. Ranking the spreading influence in complex networks. Physica A, 2013, 392: 4154–4159
38 Hu Q, Gao Y, Ma P, et al. A new approach to identify influential spreaders in complex networks. In: Web-Age Information Management 2013. New York: Springer, 2013. 99–104
39 Dolev S, Elovici Y, Puzis R. Routing betweenness centrality. J ACM, 2010, 57: 25
40 Cheng X Q, Ren F X, Shen H W, et al. Bridgeness: A local index on edge significance in maintaining global connectivity. J Stat Mech-Theory E, 2010, 2010: P10011
41 Hage P, Harary F. Eccentricity and centrality in networks. Soc Netw, 1995, 17: 57–63
42 Albert R, Jeong H, Barabási A L. Internet: Diameter of the world-wide web. Nature, 1999, 401: 130–131
43 Latora V, Marchiori M. Efficient behavior of small-world networks. Phys Rev Lett, 2001, 87: 198701
44 Katz L. A new status index derived from sociometric analysis. Psychometrika, 1953, 18: 39–43
45 Stephenson K, Zelen M. Rethinking centrality: Methods and examples. Soc Netw, 1989, 11: 1–37
46 Altmann M. Reinterpreting network measures for models of disease transmission. Soc Netw, 1993, 15:1–17
47 Poulin R, Boily M C, Mâsse B. Dynamical systems to define centrality in social networks. Soc Netw, 2000, 22: 187–220
48 Freeman L C. A set of measures of centrality based on betweenness. Sociometry, 1977, 40: 35–41
49 Everett M G, Borgatti S P. The centrality of groups and classes. J Math Sociol, 1999, 23: 181–201
50 Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E, 2001, 64: 016132
51 Estrada E, Rodriguez-Velazquez J A. Subgraph centrality in complex networks. Phys Rev E, 2005, 71: 056103
52 王延庆. 基于接连失效的复杂网络节点重要性评估. 网络安全技术与应用, 2008, 3: 026
53 Goh K I, Kahng B, Kim D. Universal behavior of load distribution in scale-free networks. Phys Rev Lett, 2001, 87: 278701

54 Brandes U. A faster algorithm for betweenness centrality. J Math Sociol, 2001, 25: 163–177
55 Zhou T, Liu J G, Wang B H. Notes on the algorithm for calculating betweenness. Chin Phys Lett, 2006, 23: 2327
56 Yan G, Zhou T, Hu B, et al. Efficient routing on complex networks. Phys Rev E, 2006, 73: 046108
57 Freeman L C, Borgatti S P, White D R. Centrality in valued graphs: A measure of betweenness based on network flow. Soc Netw, 1991, 13: 141–154
58 Estrada E, Hatano N. Communicability in complex networks. Phys Rev E, 2008, 77: 036111
59 Estrada E, Higham D J, Hatano N. Communicability betweenness in complex networks. Physica A, 2009, 388: 764–774
60 Newman M E J. A measure of betweenness centrality based on random walks. Soc Netw, 2005, 27: 39–54
61 郭世泽, 陆哲明. 复杂网络基础理论. 北京: 科学出版社, 2012
62 Bonacich P, Lloyd P. Eigenvector-like measures of centrality for asymmetric relations. Soc Netw, 2001, 23: 191–201
63 Martin T, Zhang X, Newman M E J. Localization and centrality in networks. 2014, arXiv:14015093
64 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Sys, 1998, 30: 107–117
65 Kim S J, Lee S H. An improved computation of the pagerank algorithm. Adv Infor Retr, 2002, 2291: 73–85
66 Zhang L, Qin T, Liu T Y, et al. N-step PageRank for web search. Adv Infor Retr, 2007, 4425: 653–660
67 Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine. Comput Netw, 2012, 56: 3825–3833
68 Bollen J, Rodriquez M A, Van de Sompel H. Journal status. Scientometrics, 2006, 69: 669–687
69 Bhat H S, Sims B. InvestorRank and an inverse problem for PageRank. Doctor Dissertation. Merced: University of California, 2012
70 Chen P, Xie H, Maslov S, et al. Finding scientific gems with Google's PageRank algorithm. J Informetr, 2007, 1: 8–15
71 Walker D, Xie H, Yan K K, et al. Ranking scientific publications using a model of network traffic. J Stat Mech Theor Exp, 2007, 6: P06010
72 Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Inform Process Manag, 2008, 44: 800–810
73 Jomsri P, Sanguansintukul S, Choochaiwattana W. CiteRank: Combination similarity and static ranking with research paper searching. Int J Int Technol Secur Tran, 2011, 3: 161–177
74 Ding Y, Yan E, Frazho A, et al. PageRank for ranking authors in co-citation networks. J Assn Inf Sci Technol, 2009, 60: 2229–2243
75 Petersen A M, Wang F, Stanley H E. Methods for measuring the citations and productivity of scientists across time and discipline. Phys Rev E, 2010, 81: 036114
76 Zhou Y B, Lü L, Li M. Quantifying the influence of scientists and their publications: Distinguishing between prestige and popularity. New J Phys, 2012, 14: 033033
77 Radicchi F, Fortunato S, Markines B, et al. Diffusion of scientific credits and the ranking of scientists. Phys Rev E, 2009, 80: 056103
78 Lü L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case. PLoS One, 2011, 6: e21202
79 Li Q, Zhou T, Lü L, et al. Identifying Influential Spreaders by Weighted LeaderRank. 2013, Physica A, 2014, 404: 47–55
80 Kleinberg J M. Authoritative sources in a hyperlinked environment. JACM, 1999, 46: 604–632
81 Laureti P, Moret L, Zhang Y C, et al. Information filtering via iterative refinement. Europhys Lett, 2006, 75: 1006
82 Jiang L L, Medo M, Wakeling J R, et al. Building reputation systems for better ranking. 2010, arXiv:10012186
83 Zhou Y B, Lei T, Zhou T. A robust ranking algorithm to spamming. Europhys Lett, 2011, 94: 48002
84 Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Comput Netw, 2000, 33: 387–401
85 Chakrabarti S, Dom B, Raghavan P, et al. Automatic resource compilation by analyzing hyperlink structure and associated text. Comput Netw ISDN Sys, 1998, 30: 65–74
86 Weiss R, Vélez B, Sheldon M A. HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering: Proceedings of the seventh ACM conference on Hypertext. Washington, DC: ACM Press, 1996. 180–193
87 Zhou T, Lü L, Zhang Y C. Predicting missing links via local information. Eur Phys J B, 2009, 71: 623–630
88 Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation. Phys Rev E, 2007, 76: 046115
89 Zhou T, Jiang L L, Su R, et al. Effect of initial configuration on network-based recommendation. Europhys Lett, 2008, 81: 58004
90 李鹏翔, 任玉晴, 席酉民. 网络节点(集)重要性的一种度量指标. 系统工程, 2004, 22: 13–20
91 许进, 席酉民, 汪应洛. 系统的核与核度理论. 系统工程学报, 1999, 14: 243–257
92 陈勇, 胡爱群, 胡啸. 通信网中节点重要性的评价方法. 通信学报, 2004, 25: 129–134
93 谭跃进, 吴俊, 邓宏钟. 复杂网络中节点重要度评估的节点收缩方法. 系统工程理论与实践, 2006, 26: 79–83
94 Restrepo J G, Ott E, Hunt B R. Characterizing the dynamical importance of network nodes and links. Phys Rev Lett, 2006, 97: 094102
95 Dangalchev C. Residual closeness in networks. Physica A, 2006, 365: 556–564
96 Barrat A, Barthelemy M, Pastor-Satorras R, et al. The architecture of complex weighted networks. Proc Natl Acad Sci USA, 2004, 101: 3747–3752

97 Garas A, Schweitzer F, Havlin S. A k-shell decomposition method for weighted networks. New J Phys, 2012, 14: 083030
98 Hirsch J E. An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA, 2005, 102: 16569
99 Korn A, Schubert A, Telcs A. Lobby index in networks. Physica A, 2009, 388: 2221–2226
100 Zhao S X, Rousseau R, Ye F Y. h-Degree as a basic measure in weighted networks. J Informetr, 2011, 5: 668–677
101 Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. Soc Netw, 2010, 32: 245–251
102 Shafer G. A Mathematical Theory of Evidence. Princeton: Princeton university Press, 1976
103 Moreno J L. The Sociometry Reader. New York, NY, US: Free Press, 1960
104 Dereich S, Mörters P. Random networks with sublinear preferential attachment: The giant component. Ann Prob, 2013, 41: 329–384
105 Schneider C M, Moreira A A, Andrade J S, et al. Mitigation of malicious attacks on networks. Proc Natl Acad Sci USA, 2011, 108: 3838–3841
106 Iyer S, Killingback T, Sundaram B, et al. Attack robustness and centrality of complex networks. PLoS One, 2013, 8: e59613
107 Zhou T, Fu Z Q, Wang B H. Epidemic dynamics on complex networks. Prog Nat Sci, 2006, 16: 452–457
108 周涛, 傅忠谦, 牛永伟, 等. 复杂网络上传播动力学研究综述. 自然科学进展, 2005, 15: 513–518
109 Peng X L, Xu X J, Fu X, et al. Vaccination intervention on epidemic dynamics in networks. Phys Rev E, 2013, 87: 022813
110 Lü L, Chen D B, Zhou T. The small world yields the most effective information spreading. New J Phys, 2011, 13: 123005
111 Brummitt C D, D'Souza R M, Leicht E. Suppressing cascades of load in interdependent networks. Proc Natl Acad Sci USA, 2012, 109: E680–E689
112 Yang R, Wang B H, Ren J, et al. Epidemic spreading on heterogeneous networks with identical infectivity. Phys Lett A, 2007, 364: 189–193
113 Moreno Y, Pastor-Satorras R, Vespignani A. Epidemic outbreaks in complex heterogeneous networks. Eur Phys J B, 2002, 26: 521–529
114 Holme P, Saramäki J. Temporal networks. Phys Rep, 2012, 519: 97–125
115 Holme P, Saramäki J. Temporal Networks. New York: Springer, 2013
116 Schweitzer F, Fagiolo G, Sornette D, et al. Economic networks: The new challenges. Science, 2009, 325: 422
117 Borgatti S P, Mehra A, Brass D J, et al. Network analysis in the social sciences. Science, 2009, 323: 892–895
118 Colizza V, Barrat A, Barthélemy M, et al. The modeling of global epidemics: Stochastic dynamics and predictability. B Math Biol, 2006, 68: 1893–1921
119 Barabási A L, Oltvai Z N. Network biology: Understanding the cell's functional organization. Nat Rev Genet, 2004, 5: 101–113

Review of ranking nodes in complex networks

REN XiaoLong & LÜ LinYuan

Alibaba Research Center for Complexity Sciences, Alibaba Business College, Hangzhou Normal University, Hangzhou 310036, China

The important nodes in complex networks are the extraordinary nodes which play more significant role than other nodes on the structure and function of the networks. In recent years, the reaserch on indentifying inflential nodes in complex networks has attracted much attention, because of its great theoretical significance as well as the wide range of applications. Aiming at different types of networks and motivated by different problems and applications, researchers have proposed groups of methods. This article systematically reviews more than 30 representative methods which are classified into four categories, and detailedly compares them from the aspects of computing ideas and application scenarios, and futher analyzes the strongness and weakness of each method. On this basis, this article summarizes the existing problems and outlines eight open issues as main challenges in the near future.

complex networks, important nodes, node ranking, node centrality, spreading model

doi: 10.1360/972013-1280

补充材料

表 S1　各算法特性总结

本文的以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.