前沿领域综述

复杂网络中节点重要性排序的研究进展*

刘建国† 任卓明 郭强 汪秉宏

(上海理工大学复杂系统科学研究中心, 上海 200093)

(2013 年 4 月 11 日收到; 2013 年 5 月 8 日收到修改稿)

如何用定量分析的方法识别超大规模网络中哪些节点最重要, 或者评价某个节点相对于其他一个或多个节点的重要程度, 这是复杂网络研究中亟待解决的重要问题之一. 本文分别从网络结构和传播动力学的角度, 对现有的复杂网络中节点重要性排序方法进行了系统的回顾, 总结了节点重要性排序方法的最新研究进展, 并对不同的节点重要性排序指标的优缺点以及适用环境进行了分析, 最后指出了这一领域中几个有待解决的问题及可能的发展方向.

关键词: 复杂网络, 节点重要性, 网络结构, 传播动力学

PACS: 89.75.Hc, 89.75.Fb **DOI:** 10.7498/aps.62.178901

1 引 言

伴随着信息技术的迅猛发展, 人类的社会活动日趋网络化. 我们的生活被各种网络包围着 [1−3], 例如与他人交流的在线社交网络、通信网络、科研合作网络; 与生活密切相关的因特网、交通网络、电力网络; 与人自身相关的新陈代谢网络、神经网络、基因调控网络等等. 研究者们通过收集不同复杂系统的数据, 分析网络的统计特征, 进一步认识网络的动力学行为 [4−6]. 然而这些网络的结构非常复杂, 而且网络数据规模也越来越庞大 [7−9]. 例如 Facebook 拥有超过 10 亿用户, 腾讯即时通讯工具 QQ 的注册用户超过 10 亿, 活跃用户超过 7 亿, 因特网具有超过万亿的统一资源定位符 (URL), 大脑神经元网络有数百亿节点. 如何用定量分析的方法度量大规模网络中节点重要程度是复杂网络研究中亟待解决的重要问题之一 [10−12].

随着网络科学的蓬勃发展, 节点重要性的研究进一步受到人们的关注, 王林 [13]、赫南 [14] 等介绍了复杂网络中几种常用的网络中心性指标, 指出了中心性指标的特点及应用场合. 孙睿等 [15] 介绍了国内外关于网络舆论中节点重要性评估的研究现状, 从基于网络拓扑结构和基于节点属性两个方面总结了现有评价节点重要性的模型和方法. 然而近几年有不少学者已从新的视角研究网络节点重要性排序, 例如 Kitsak 等人 [16] 于 2010 年首次提出了节点重要性依赖于其在整个网络中的位置的思想, 并且利用 K-核分解获得了比度、介数更为准确的节点重要性排序指标. 在短短两年半内该文的 Google Scholar 的引用次数已高达 170 余次. 因此有必要从不同角度总结目前的研究进展, 并对未来可能的研究方向进行探讨.

本文简要地回顾了网络节点重要性排序的研究进展. 我们首先介绍了基于网络结构的节点重要性排序度量指标, 这类指标主要从网络的局部属性、全局属性、网络的位置和随机游走等四个方面展开, 同时对这些方法的优缺点及适用范围进行了分析; 然后介绍了传播动力学与节点重要性度量指标的关系; 最后在总结和展望部分指出了当前面临的问题和可能的发展方向.

2 问题描述与评价方法

假设网络 $G=(V,E)$ 是由 $|V|=N$ 个节点和 $|E|=M$ 条边连接所组成的一个无向网络. 网络的

* 国家自然科学基金 (批准号: 71071098, 71171136, 91024026)、上海市教委科研创新项目 (批准号: 11ZZ135, 11YZ110)、教育部科学技术研究重点项目 (批准号: 211057) 和上海市一流学科 (系统科学) 建设项目 (批准号: XTKX2012) 资助的课题.

† 通讯作者. E-mail: liujg004@ustc.edu.cn

$A = a_{ij}$, $a_{ij} = 1$ 表示节点 i 与节点 $j(i \neq j)$ 之间直接连接, 否则 $a_{ij} = 0$. 网络中节点重要性排序方法的准确性常用传播动力学进行度量, 一般以网络节点为传播源, 利用传播动力学模型仿真, 通过计算网络中目标节点的影响范围来度量节点在传播过程中的影响力. 另一种方法是考虑节点删除前后图的连通状况的变化情况 [17−20], 将节点的重要性等价为该节点被删除后对网络的破坏性. 假设在一个网络中, 某个节点被删除, 则同时移走了与该节点相连的所有边, 从而可能使得网络的连通性变差. 节点被删去后网络连通性变得越差, 则表明该节点越重要. 经过网络抗毁性实验得出的节点重要性排序与先前的节点重要性排序方法的结果越相似, 则认为该排序方法越准确.

3 基于网络结构的节点重要性排序方法

复杂网络中节点重要性可以是节点的影响力, 地位或者其他因素的综合. 从网络拓扑结构入手是研究这一问题常用的方法之一. 最早对这一问题进行研究的是社会学家 [21−30], 随后其他领域的学者们也开始研究这一问题, 提出了一系列的评估指标. 本小节从网络的局部属性、全局属性、网络的位置以及随机游走等四个角度出发, 介绍了基于网络结构的节点重要性排序的不同指标.

3.1 基于网络局部属性的指标

基于网络局部属性的节点重要性排序指标主要考虑节点自身信息和其邻居信息, 这些指标计算简单, 时间复杂度低, 可以用于大型网络.

节点 i 的度 (Degree) 定义为该节点的邻居数目. 具体表示为

$$k(i) = \sum_{j \in G} a_{ij}. \tag{1}$$

度指标直接反映的是一个节点对于网络其他节点的直接影响力. 例如在一个社交网络中, 有大量的邻居数目的节点可能有更大的影响力, 更多的途径获取信息, 或有更高的声望. 又如在引文网络中, 利用文章的引用次数来评价科学论文的影响力 [21].

王建伟等 [22] 认为网络中节点的重要性不但与自身的信息具有一定的关系, 而且与该节点邻居节点的度也存在一定的关联, 即节点的度及其邻居节点的度越大, 节点就越重要.

Chen 等 [23] 考虑节点最近邻居和次近邻居的度信息, 定义了一个多级邻居信息指标 (local centrality), 来对网络中节点的重要性排序, 其具体定义如下:

$$L_{\mathrm{C}}(i) = \sum_{j \in \Gamma(i)} \sum_{u \in \Gamma(j)} N(u), \tag{2}$$

其中 $\Gamma(i)$ 为节点 i 最近邻居集合, $\Gamma(j)$ 为节点 j 最近邻居集合, $N(u)$ 为节点 u 最近邻居数和次近邻居数之和.

任卓明等 [24] 综合考虑节点的邻居个数, 以及其邻居之间的连接紧密程度, 提出了一种基于邻居信息与集聚系数的节点重要性评价方法. 具体表示为

$$P(i) = \frac{f_i}{\sqrt{\sum_{j=1}^{N} f_j^2}} + \frac{g_i}{\sqrt{\sum_{j=1}^{N} g_j^2}}, \tag{3}$$

其中 f_i 为节点 i 自身度与其邻居度之和, 即 $f_i = k(i) + \sum_{u \in \Gamma(i)} k(u)$, 其中 $k(u)$ 表示节点 u 的度, $u \in \Gamma(i)$ 表示节点 i 的邻居节点集合. g_i 表示为

$$g_i = \frac{\max\limits_{j=1}^{N} \left\{ \dfrac{c_j}{f_j} \right\} - \dfrac{c_i}{f_i}}{\max\limits_{j=1}^{N} \left\{ \dfrac{c_j}{f_j} \right\} - \min\limits_{j=1}^{N} \left\{ \dfrac{c_j}{f_j} \right\}}, \tag{4}$$

其中 c_i 为节点 i 的集聚系数. 该方法只需要考虑网络局部信息, 适合于对大规模网络的节点重要性进行有效分析. Centol[25] 研究在线社会网络的行为传播, 发现传播行为在高集聚类网络传播的更快, 节点的传播重要性与该节点的集聚性有关. Goel 等 [26] 通过研究 Facebook 系统中朋友关系演化特性发现邻居节点的绝对数目不是影响节点重要性的决定性因素, 起决定作用的是邻居节点之间形成的联通子图的数目.

3.2 基于网络全局属性的指标

基于网络全局属性的节点重要性排序指标主要考虑网络全局信息, 这些指标一般准确性比较高, 但时间复杂度高, 不适用于大型网络.

特征向量 (eigenvector centrality)[27,28] 是评估网络节点重要性的一个重要指标. 度指标把周围相邻节点视为同等重要, 而实际上节点之间是不平等的, 必须考虑到邻居对该节点的重要性有一定的影响. 如果一个节点的邻居很重要, 这个节点重要性很可能高; 如果邻居重要性不是很高的话, 那么即使该

节点的邻居众多, 也不一定很重要. 通常称这种情况为邻居节点的重要性反馈. 特征向量指标是网络邻接矩阵对应的最大特征值的特征向量. 具体定义如下:

$$C_e(i) = \lambda^{-1} \sum_{j=1}^{N} a_{ij} e_j, \tag{5}$$

其中 λ 为邻接矩阵 A 的最大特征值; $e = (e_1, e_2, \cdots, e_n)^{\mathrm{T}}$ 为邻接矩阵 A 对应最大特征值 λ 对应的特征向量. 特征向量指标是从网络中节点的地位或声望角度考虑将单个节点的声望看成是所有其他节点声望的线性组合, 从而得到一个线性方程组. 该方程组的最大特征值所对应的特征向量就是各个节点的重要性.

Poulin 等人 [29] 在求解特征向量映射迭代方法的基础上提出累计提名 (cumulated nomination centrality) 的方法, 该方法计算网络中的其他节点对目标节点的提名值总和. 累计提名值越高的节点其重要性就越高. 累计提名方法计算量较少、收敛速度较快, 而且适用于大型和多分支网络.

Katz 指标 [30] 同特征向量一样可以区分不同的邻居对节点的不同影响力. 不同的是 Katz 指标给邻居赋予不同的权重, 对于短路径赋予较大的权重, 而长路径赋予较小的权重. 具体定义为

$$\begin{aligned} S &= \beta A + \beta^2 A^2 + \beta^3 A^3 + \cdots \\ &= (I - \beta A)^{-1} - I, \end{aligned} \tag{6}$$

其中 I 为单位矩阵, A 为网络的邻接矩阵, β 为权重衰减因子. 为了保证数列的收敛性, β 的取值须小于邻接矩阵 A 最大特征值的倒数, 然而该方法权重衰减因子的最优值只能通过大量的实验验证获得, 因此具有一定的局限性.

紧密度 (closeness centrality)[31] 用来度量网络中的节点通过网络对其他节点施加影响的能力. 节点的紧密度越大, 表明该节点越居于网络的中心, 在网络中就越重要. 紧密度具体定义如下:

$$C_{\mathrm{c}}(i) = \frac{N-1}{\sum\limits_{j=1}^{N} d_{ij}}, \tag{7}$$

其中 d_{ij} 表示节点 i 到节点 j 的最短距离. 紧密度依赖于网络的拓扑结构, 对类似于星形结构的网络, 它可以准确地发现中心节点, 但是对于随机网络则不适合, 而且该方法的计算时间复杂度为 $O(N^3)$.

Zhang 等 [32] 考虑节点的影响范围, 定义了 Kernel 函数法, 具体定义如下:

$$U(i) = \sum_{j=1}^{N} \mathrm{e}^{-\frac{d_{ij}^2}{2h^2}}, \tag{8}$$

其中 d_{ij} 表示节点 i 到节点 j 的最短距离, h 表示 Kernel 函数的宽度, h 越大此函数越平滑, 节点影响范围越大, 反之亦然. 考虑到非最短路径的信息, Kernel 函数法另一表述为

$$U(i) = \sum_{j=1}^{N} \mathrm{e}^{-\frac{d_{ij}^2}{2h^2}} + \sum_{j=1}^{N} \mathrm{e}^{-\frac{L(p)^2}{2h^2}}, \tag{9}$$

其中 p 表示节点 i 到其他所有节点的非最短距离路线, $L(p)$ 表示这些非最短路线的长度. 虽然 Kernel 函数法较紧密度更准确, 但时间复杂度依然没有降低, 不适用于大型网络.

Huang 等 [33] 分析了美国 1996—2006 年间公司董事网络结构, 该网络中节点是由公司中的董事构成, 两位董事在同一个公司任职则表示他们有连接关系. 作者认为公司董事的影响力取决于该董事手中掌握多少获取公司信息的渠道, 提出一种识别公司董事影响力的方法. 其方法记为

$$I(i) = \frac{\sum\limits_{j=1}^{N} w_j r_{j1} r_{j2} \cdots r_{jd_j}}{\sum\limits_{j=1}^{N} w_j}, \tag{10}$$

其中 w_j 表示董事 j 所在公司拥有的信息量, 即该公司的市值. d_j 是表示董事 i 与董事 j 之间的最短路径, r_j 是信息在传递过程中的衰减率.

Freeman 于 1977 年在研究社会网络时提出介数指标 (betweenness centrality)[34,35], 该指标用于衡量个体社会地位的参数. 节点 i 的介数含义为网络中所有的最短路径之中经过节点 i 的数量, 记为

$$C_{\mathrm{c}}(i) = \sum_{s<t} \frac{n_{st}^i}{g_{st}}, \tag{11}$$

其中 g_{st} 表示节点 s 到节点 t 之间的最短路径数; n_{st}^i 表示节点 s 和节点 t 之间经过节点 i 的最短路径数. 节点的介数值越高, 这个节点就越有影响力, 即这个节点也就越重要. 例如判断社交网络中某人的重要程度, 某个人在关系网络中类似于 “交际花”, 长袖善舞能够使其与各色人群打交道, 拥有人脉越广泛, 则其影响范围越大, 则其他人与此人也就越密切相关, 因此该人也就越重要.

Travencolo 等 [36] 提出了节点可达性指标 (accessibility). 可达性指标是描述节点在自避随机游走的前提下, 行驶 h 步长之后该节点能够访问多少不同目标节点的可能性, 具体定义为

$$E(\Omega,i)=-\sum_{j=1}^{N}\begin{cases}0, & p_h(j,i)=0;\\ p_h(j,i)\log(p_h(j,i)), & p_h(j,i)\neq 0.\end{cases} \tag{12}$$

$$E(i,\Omega)=-\sum_{j=1}^{N}\begin{cases}0, & p_h(j,i)=0;\\ \left(\dfrac{p_h(j,i)}{N-1}\right)\left(\log\dfrac{p_h(j,i)}{N-1}\right), & p_h(j,i)\neq 0.\end{cases} \tag{13}$$

其中 $p_h(j,i)$ 表示从 i 点出发到 j 点的可能性, h 表示步长, $p_h(j,i)$ 即从 i 点到 j 点行驶 h 步的不同路径数与总的得到的不同路径数之比. 这里 Ω 是指除 i 以外的所有节点. 除此之外, 当随机游走遇到以下三种情况时将会停止: 1) 游走达到所定义的最大步长 H; 2) 游走达到一个点, 而该点的度数为 1, 即无法再行走下去; 3) 游走无法再进行下去, 因为所有与该点相邻的点都已经被访问过了. 作者为了完善多样性的概念, 提出了对外可达性和对内可达性两个指标, 分别记为

$$OA_h(i)=\frac{\exp(E(\Omega,i))}{N-1}, \tag{14}$$

$$IA_h(i)=\frac{\exp(E(i,\Omega))}{N-1}. \tag{15}$$

前者指在行走 h 步之后, 起始点 i 达到所有剩下点的可能性, 后者是指从每个点出发行走 h 步后, 能够到达点 i 的可能性, 也可理解为到达频率. Travencolo 等 [36] 的实验结果显示, 处于中心区域的节点有较高的对外可达性, 可以被近似看成是现实中的“交流区”, 而处于网络边缘的节点对外可达性较低.

Comin 等人 [37] 考虑介数与度的关系, 定义了一个节点重要性排序的指标, 具体定义为

$$\hat{B}(i)=\frac{B(i)}{[k(i)]^{\lambda}}, \tag{16}$$

其中 $B(i)$ 为节点 i 的介数值, $k(i)$ 为节点 i 的度, λ 为的最优参数. 该指标值越大, 则认为该节点越重要. 虽然该方法较介数和度指标的准确性要高, 但时间复杂度并没有降低, 而且引入了参数, 使得其实用性不强.

其他的基于网络路径的全局方法如李鹏翔等 [38] 提出用节点被删除后形成的所有不连通节点之间的距离 (最短路) 的倒数之和来度量所删节点的重要性. 谭跃进等 [39] 定义了网络的凝聚度, 在此基础上提出了一种评估复杂网络节点重要度的节点收缩方法, 认为会在最重要的节点会在该节点收缩后使网络的凝聚度最大. 该方法综合考虑了节点的连接度以及经过该节点最短路径的数目. 余新等 [40] 通过计算网络中的节点被移除时网络直径和网络连通度变化梯度来评估网络中节点的重要性, 利用该算法对美国 ARPA 网络的节点重要程度进行了分析. 饶育萍等 [41] 提出了一种基于全网平均等效最短路径数的网络抗毁评价模型, 认为全网平均等效最短路越多, 网络的抗毁能力越强. 并在此基础上, 提出一种节点重要性评价方法. 如果节点失效后网络抗毁度下降越多, 则该节点在网络中的重要性越大. 程克勤等 [42] 根据有权网络中边的权值计算节点的边权值, 并依据边的权值计算全网平均最短路径, 以此度量节点重要性.

3.3 基于网络位置属性的指标

Kitsak 等人 [16] 于 2010 年首次提出了节点重要性依赖于其在整个网络中的位置的思想, 并且利用 K-核分解获得了节点重要性排序指标 (k-shell), 该指标时间复杂度低, 适用于大型网络, 而且比度、介数更能准确识别在疾病传播中最有影响力的节点. 近几年不少学者受到这种思想的启发, 对 K-核进行了扩展和改进, 使其应用范围更广, 准确性更好.

K-核分解方法 [16,43] 通过递归地移去网络中所有度值小于或等于 k 的节点. K-核的定义如下: 由集合推导出的子网络 $H=(C,E|C)$, 满足 C 中的任意节点 V, 其度值均大于 k 的最大子网络被称为 K-核, 其中满足 K-核值等于 k 小于 $k+1$ 的那部分节点称为 k-shell, 简称 Ks. Ks 分解示意图如图 1, 该网络被划分为 3 个不同的层.

一些学者都认为网络中的 Hubs 节点或者高介数的节点是传播中最有影响力的节点. 这是因为 Hubs 节点拥有更多的人际关系, 而高介数的节点有更多的最短路径通过, 于是疾病控制首先要确定这样的节点. 但是, Kitsak 等人 [16] 调查了社交网络、邮件网络、病人接触网络、演员合作网络等实证网络并且通过传播动力学的建模分析指出, 对

于单个传播源情形, Hubs 节点或者高介数的节点不一定是最有影响力的节点, 而通过 K-核分解分析确定的网络核心节点 (即 Ks 值大的节点) 才是最有影响力的节点. 当然文章也指出, 当初始存在多个传播源的时候, 传播的规模很大程度依赖于初始传播源之间的距离. 在存在多个传播源的情况下, 度大的 Hub 节点往往比 Ks 大的节点具有更高传播效率. 然而 Ks 指标赋予大量节点相同的值, 例如 Barabasi-Albert (BA) 网络模型的所有节点的 Ks 值都相等, 从而导致 Ks 指标无法衡量其节点的重要性.

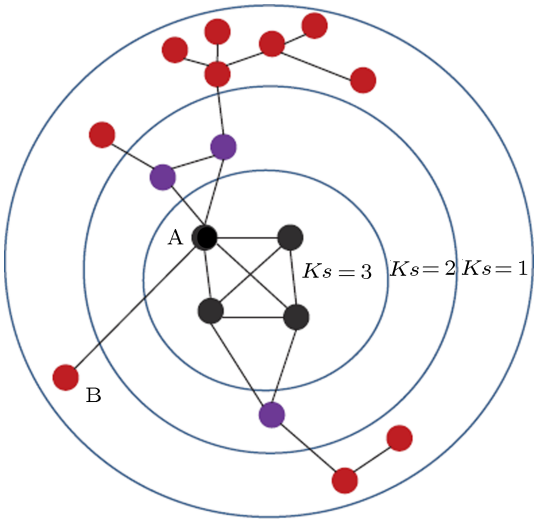

图 1 Ks 分解示意图 [16]

Zeng 等人 [44] 考虑节点的 Ks 信息和经过 Ks 分解后被移除节点的信息, 提出了混合度分解方法 (MDD):

$$k(i)_m = k(i)_r + \lambda k(i)_e, \tag{17}$$

其中 $k(i)_m$ 表示经过 Ks 分解后节点的度信息, 即节点的 Ks 值, $k(i)_e$ 表示经过 Ks 分解后, 被移除节点的度信息. 当 $\lambda = 0$, 表示节点 i 的 Ks 值, 当 $\lambda = 1$ 时, 表示节点的度信息. MDD 的分解流程为:

1) 初始状态下, 网络中节点的 $k(i)_m$ 值与 $k(i)_r$ 相等.

2) 移去网络中 k_m 值最小的节点, 这些节点赋值为 M-shell.

3) 采用公式 $k(i)_m = k(i)_r + \lambda k(i)_e$ 更新网络剩余节点的 k_m, 并通过递归移除网络中 k_m 值小于或等于 M 的节点, 将这类节点赋值为 M-shell, 直至网络中所有剩余节点的 k_m 值大于 M.

随着 M 值的变大重复步骤 2 和 3, 直到网络中所有节点都能赋予其对应的 M 值, 具体分解方法如图 2.

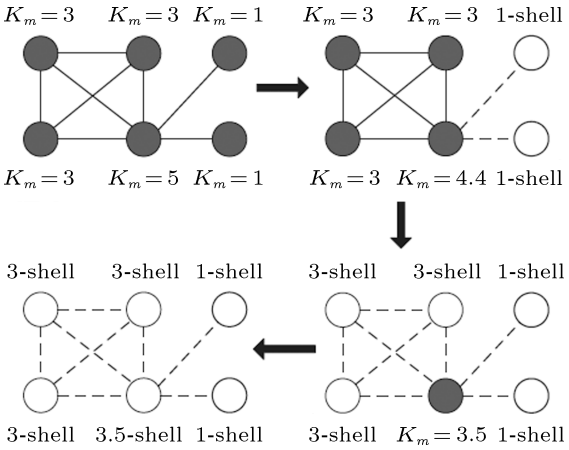

图 2 混合度分解方法示意图 [43], 其中 $\lambda = 0.7$

Garas 等 [45] 先将加权网络转变成无权网络, 再进行经典的 Ks 分解, 加权网络的 Ks 分解具体计算如下:

$$k(i)_w = \left[k(i)^{\alpha} \left(\sum_{j}^{k(i)} w_{ij} \right)^{\beta} \right]^{\frac{1}{\alpha+\beta}}, \tag{18}$$

其中 $k(i)$ 为节点 i 的度, $\sum_{j}^{k(i)} w_{ij}$ 为节点 i 的边权值之和, 其中 α, β 为可调参数. 如图 1 中, 假设边 AB 的权重为 3, 其他所有的边的权重为 1, 其中 $\alpha = \beta = 1$, 通过公式计算可得, $k(A)_w = 2$, $k(B)_w = 2$, 此时 $Ks(B) = 2$.

Liu 等 [46] 综合考虑目标节点自身 K-核的信息和与网络最大 K-核的距离, 提出了新的度量节点重要性的指标. 该指标解决了 Ks 指标赋予网络中大量节点相同的值导致其无法准确衡量其节点重要性的缺陷. 具体定义如下:

$$\theta(i|k_s) = (k_s^{\max} - k_s + 1) \sum_{j \in J} d_{ij}, \quad i \in S_{k_s}, \tag{19}$$

其中 $k_s^{\max}$ 为网络最大 K-核值, d_{ij} 表示节点 i 到节点 j 的最短距离. 节点集合 J 定义为具有网络的最大 K-核值的节点, S_{k_s} 定义为节点的 K-核值为 Ks 的节点集合.

由于最小 K-核节点的 Ks 值是相同的, 而且根据 K-核分解原理, 度为 1 以及大部分介数为 0 的节点属于最小 K-核节点, 因此仅仅依靠节点自身

K-核、度或介数信息不能很好区分这类节点的传播能力. 任卓明等 [47] 提出了基于最小 K-核节点邻居集合中最大 Ks 值的深度指标 $H(i)$, 该指标依靠最小 K-核节点与网络的其他层级节点的连接关系, 判断最小 K-核节点的重要性. 表示为

$$H(i)=\max\left\{Ks_j\right\},\quad j\in J(i), \tag{20}$$

其中 $J(i)$ 为节点 i 的邻居集合, Ks_j 为节点 j 的 Ks 值.

Hou 等 [48] 考虑度、介数、K-核三个不同的指标的对节点重要性的影响, 采用欧拉距离公式, 计算度、介数、Ks 等三个不同的指标的综合作用, 该指标记为

$$D(i)=\sqrt{k^2(i)+C_b^2(i)+K_s^2(i)}, \tag{21}$$

其中 $k(i)$, $C_b(i)$, $K_s(i)$ 表示节点 i 的度、介数和 Ks 值.

3.4 基于随机游走的节点重要性排序

基于随机游走的节点重要性排序方法主要是基于网页之间的链接关系的网页排序技术, 由于网页之间的链接关系可以解释为网页之间的相互关联和相互支持, 从而判断出网页的重要程度. 这类典型的方法有 PageRank[49,50], LeaderRank[51] 和 HITS 算法 [52].

PageRank[49,50]: 当网页 A 有一个链接指向网页 B 时, 就认为网页 B 获得了一定的分数, 该分值的多少取决于网页 A 的重要程度, 即网页 A 的重要性越大, 网页 B 获得的分数就越高. 由于网页上链接相互指向非常复杂, 该分值的计算是一个迭代过程, 最终网页将依照所得的分数对其进行排序并将检索结果送交用户, 这个量化了的分数就是 PageRank 值. 其计算公式如下:

初始步: 给定所有节点的初始 PageRank 值 (简称 PR) $\mathrm{PR}_i(0)$, $i=1,2,3,\cdots,N$, 满足 $\sum_{i=1}^{N}\mathrm{PR}_i(0)=1$.

PageRank 校正规则: 给定一个标度常数 $s\in(0,1)$. 首先按照基本的 PageRank 校正规则计算各个节点的 PR 值, 然后把每个节点的 PR 值通过比例因子 s 进行缩减. 这样, 所有节点的 PR 值之和也就缩减为 s, 再把 $1-s$ 平均分给每个节点 PR 值, 以保持网络总的 PR 值为 1. 即为

$$\mathrm{PR}_i(k)=s\sum_{j=1}^{N}\bar{a}_{ij}P. \tag{22}$$

用户在每一步都有一个较小概率 $1-s$ 随机访问互联网上的任何一个网站, 同时保持概率 s 访问当前网页所提供的链接. 同时加入最后一项可以保证算法可以走出 “悬挂节点 (dangling node)” 以及避免死循环. PageRank 算法能够根据用户查询的匹配程度在网络中准确定位节点的重要程度, 而且计算复杂度不高, 为 $O(MI)$, 其中 M 为网络中边的数目, I 为算法达到收敛所需的迭代次数.

当网络中存在孤立节点或社团时, 采用 PageRank 算法对网络中节点进行排序会出现排序不唯一的问题. Lü等 [51] 提出的 LeaderRank 弥补了这一缺陷, 具体方法是在已有节点外, 另加一个节点 (Ground node), 并且将它与已有的所有节点双向连接, 于是得到 $N+1$ 个节点的网络 (如图 3). 这个新的网络是一个强连通的网络, 再按 LeaderRank 算法对该网络的节点进行排序, 结果表明 LeaderRank 算法比 PageRank 算法排序更精准, 而且对网络噪音 (节点随机加边或删边) 有更好的容忍性. 具体计算如下:

初始步: 每个节点的初始分数 $s_i(0)=1$, 但其中 Ground node 的分数为 $s_g(0)=0$.

第二步: 在时间步 t, 节点 i 分数为 $s_i(t)$, 如果节点 i 指向连接节点 j, 则 $a_{ij}=1$, 否则为 0. k_j^{out} 为节点 j 的出度, 则

$$s_i(t+1)=\sum_{j=1}^{N+1}\frac{a_{ij}}{k_j^{\mathrm{out}}}s_i(t). \tag{23}$$

第三步: 不断重复第二步, 直到时间步 t_c 时, 节点 i 的分数不变, 收敛于一个定值 $s_i(t_c)$, 则节点 (Ground node) 的分数记为 $s_g(t_c)$.

因此, 经过上面的步骤, 节点 i 的影响力值记为

$$s_i=s_i(t_c)+\frac{s_g(t_c)}{N}. \tag{24}$$

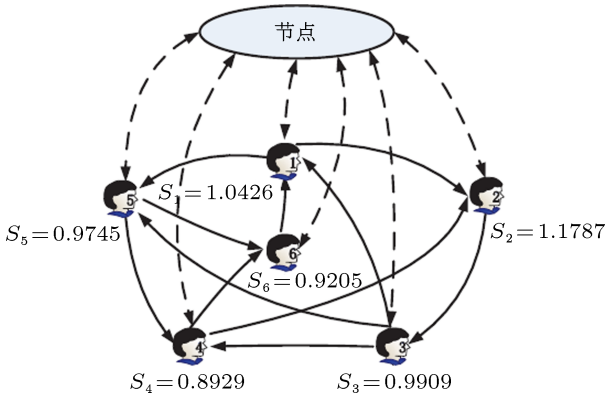

图 3 LeaderRank 算法示意图 [48]

Radicchi 等 [53,54] 提出一种扩散算法分析了有向加权网络的节点重要性排名, 并给出了科学家的科研影响力和职业运动员的影响力排名的实例分析, 该排名算法类似 PageRank 的方法. 具体计算如下:

$$P(i) = (1-q)\sum_{j=1}^{N} P(j)\frac{w_{ij}}{s_j^{\text{out}}} + \frac{q}{N} + \frac{1-q}{N}\sum_{j=1}^{N} P(j)\delta\left(s_j^{\text{out}}\right), \tag{25}$$

其中初始状态下 $P(i) = 1/N$, N 为网络节点数, $q \in [0,1]$ 为可调参数. w_{ij} 表示节点 j 到节点 i 的权重,$\delta(x)$ 函数, 其中 $x = 0$ 时, $\delta(x) = 1$; 当 $x = 1$ 时, $\delta(x) = 0$.

Masuda 等 [55] 对基于拉普拉斯算子的节点中心性度量方法进行了扩展, 扩展后的方法与 PageRank 类似. 该方法不仅适用于强连通的有向网络, 也适用于有孤立社团的网络, 不同的是, 前者是连续时间的简单随机游走, 而后者为离散时间的简单随机游走.

Kleinberg 于 1998 年提出 Hypertext-Induced Topic Search(HITS) 算法 [52]. 他将网页分为两类, 即表达某一特定主题的 Authorities 和把 Authorities 串连起来的 Hubs. Authorities 为具有较高价值的网页, 依赖于指向它的页面; 而 Hubs 为指向较多 Authorities 的网页, 依赖于它所指向的页面, 每个节点也引入了两个权值: Authority 权值和 Hub 权值. HITS 算法的目标就是通过一定的迭代计算方法得到针对某个检索提问的最具价值的网页, 即 Authority 权值排名最高的网页. HITS 算法在学术界应用较为广泛, 其计算复杂度为 $O(NI)$, 其中 N 为网络中节点的数目, I 为算法达到收敛所需的迭代次数. 然而 HITS 算法不能识别非正常目的的网页引用, 导致计算结果与实际结果有偏差.

3.5 其他方法

除以上四类方法外, 有些方法分别从网络的连通性、节点删除法、边权值、节点效率等视角度量节点重要性. 例如陈勇等 [56] 提出了一种对通信网中节点重要性进行评价的方法, 通过比较生成树的数目, 可以判断图中任意数目的两组节点的相对重要性. 从图中去掉节点以及相关联的链路后, 所得到的图对应的生成树数目越少, 则该组节点越重要. 安世虎等 [57] 利用节点删除的研究思想, 提出节点赋权网络中节点重要性的综合测度法, 并给出了该方法在知识共享网络中的应用. 吴俊等 [58] 提出了一个基于负载重分配的复杂负载网络级联失效模型的网络节点重要度评估方法. 该方法有助于发现网络中一些潜在的 “关键节点”. 陈静等 [59] 提出了一种基于节点接近度和节点在其邻域中的关键度的方法评估复杂网络中节点的重要性. 该方法综合了节点的全局和局部重要性, 即在复杂网络中, 节点的接近度越大, 而且该节点越居于网络的中心, 则其在网络中就越重要; 节点在其邻域中的关键度越大, 该节点对其邻域就越重要. 肖连杰等 [60] 用学术期刊论文的作者信息构建了作者科研合作网络, 在此基础上, 通过计算网络中节点的权值来评价作者的学术贡献, 通过计算与该节点相连的边的权值来评价作者的科研产出能力, 最后通过对节点和边的综合考察来判断节点的重要性. 叶春森等 [61] 提出了基于节点的度和凝聚度线性加权的节点的重要性指标; 并用供应链网络案例说明了该方法的有效性和优越性. 周漩等 [62] 通过定义节点效率和节点重要度评价矩阵, 综合考虑节点效率、节点度值和相邻节点重要度贡献, 用节点度值和效率值表征其对相邻节点的重要度贡献. 并且提出了一种利用重要度评价矩阵来确定复杂网络重要节点的方法. 该方法较节点删除法、节点收缩法、介数法等方法的时间复杂度低.

4 基于传播动力学的节点重要性排序方法

严刚等 [63] 在加权无标度网络上的病毒传播实验表明不仅网络拓扑结构会影响疾病传播过程, 网络中边权也会影响传播过程, 强度越大的节点越易被感染. Borge 等 [64] 研究了在线社会网络的信息扩散机制, 发现网络中只有少量的度非常大的节点时, 度指标能准确识别网络最有影响力的节点, 而在信息以 “级联效应” 爆发扩散到整个网络的情况下, Ks 指标更适合识别网络最有影响力的节点. Borge 等 [65] 也认为在真实网络的谣言传播过程中, 节点的重要性不是由该节点的 Ks 位置决定的, 而是由谣言传播扩散机制决定的. Klemm 等 [66] 提出了集群动力学中节点的重要性是由网络拓扑结构和集群动力学机制决定的观点. Aral 等 [67] 研究了脸谱 (Facebook) 网络上的 130 万用户传播行为, 发

现用户的影响力受年龄、性别、婚姻等因素左右, 在研究节点影响力与网络拓扑结构的关系时, 有影响力的节点具有集聚性, 这样易于用户网络行为的传播. 综上可以看出节点重要排序不仅仅由网络结构决定, 还受网络行为传播机制以及节点自身特性的影响 [68].

表 1 基于网络结构的节点重要性排序指标的特点, 其中 N 表示网络节点数, M 表示网络边数, $\langle k\rangle$ 表示平均度, I 为算法达到收敛所需的迭代次数

指标	优点	缺点	时间复杂度
Degree	简单直观	只反映了节点局部特征	$O(N)$
Local	比度更准确, 时间复杂度低, 适合大型网络	没有考虑邻居之间的紧密程度	$O(N^2\langle k\rangle)$
Eigenvector	考虑了节点邻居的重要性	简单的将各节点的拓扑特性线性叠加	$O(N^2)$
Katz	区分不同邻居对节点的影响力	不易获得权重衰减因子的最优值	$O(N^2)$
Closeness	通过网络对部分节点产生全局影响	不适合随机网络和大型网络	$O(N^3)$
Kernel	通过网络对所有节点产生全局影响	不适合大型网络	$O(N^3)$
Betweenness	考虑了节点的信息负载能力	不适合大型网络	$O(N^3)$
Accessibility	计算节点到达目的节点的可能性	不适合大型网络	$O(N^3)$
Ks	考虑节点在网络中位置的全局特性	不适用树状网络、BA 网络	$O(N)$
MDD	对于树状网络、BA 网络等也适用	不易确定最佳权重因子	$O(N)$
PageRank	考虑网络的全局拓扑特性	忽略了一些实际因素, 排序不唯一	$O(MI)$
LeaderRank	排序唯一, 对网络噪音有更好的容忍性	不适合无向网络	$O(MI)$
HITS	时间复杂度低	网页的非正常链接导致结果不准确	$O(NI)$

5 结论与展望

综上所述, 节点重要性排序的指标在涉及网络的结构信息时, 都是从某一个角度对于网络的某一方面的结构特点进行刻画, 如果目标网络的结构在该方面特征显著, 即可得到较好的效果; 或在复杂网络环境下, 通过节点的网络传播行为的影响力与网络结构关系判断节点的重要性. 复杂网络节点重要性问题的研究方兴未艾, 还有非常多的问题没有解决. 下面我们列出其中部分, 作为本文的结束语.

1. 节点重要性的定义. 节点的重要性含义不同, 评价节点重要性排名的结果也不同. 例如 2012 年, 美国《福布斯》全球影响力人物排行榜, 美国总统奥巴马成为 2012 年度全球最具影响力人物, 排名依据是看一个人物是否能影响一群人, 看所在国家的人口, 企业家的雇员规模, 媒体受众人数, 拥有的财富等. 而 2012 年, 美国《时代》周刊评选全球最具影响力人物, 美国 NBA 篮球运动员纽约尼克斯球队控卫林书豪位居榜首. 而时代周刊评选规则是最具有影响力的人物不一定是全球最有权力或最有钱的人, 而是一群使用想法、洞察力和行动, 对民众产生实际影响力的代表.

2. 各种指标间的内在联系. 各种节点重要性排序的方法层出不穷, 这些指标从不同视角评价节点重要性. 这些指标在不同拓扑结构的网络的准确性又是怎样呢? 例如 Silva 等 [69] 对随机网络, 小世界网络和随机集合网络等网络模型以及美国航空网络进行 SIR 传播仿真实验, 采用皮尔逊系数, 讨论了节点的拓扑性质, 例如度、可达性、节点强度 (strength)、介数、Ks 等指标与该节点传播能力的相关程度.

3. 网络结构和网络行为是如何影响节点重要性评价, 特别是对研究社会影响力非常有帮助. Robert 等人 [70] 以 2010 年美国大选为实例研究社会影响力, 发现 Facebook 用户的社会影响力与网络结构和网络行为传播机制两者都相关.

4. 时变网络中, 网络结构是变化的, 节点的各种指标具有动态性 [71], 也许此刻某个节点的重要性排在某个名次, 下一个时刻又可能是另一个名次. 此时节点重要性指标的稳定性和准确性, 计算复杂度如何, 就变得特别重要 [72,73]. 例如淘宝网每天交易量达数千万笔; 新浪微博平台平均每天发布超过 1 亿条微博. 那么如何在这种具有大数据特征的时变网络中对节点重要性排名, 这将是一个极具有挑战性的课题.

感谢上海理工大学管理学院王芳杰的交流与讨论.

[1] Albert R, Barabasi A L 2002 *Rev. Mod. Phys.* **74** 47

[2] Newman M E J 2003 *SIAM Rev.* **45** 167

[3] L*ü* L Y, Medo M, Yeung C H, Zhang Y C, Zhang Z K, Zhou T 2012 *Phys. Rep.* **519** 1

[4] Wang B H, Zhou T, Wang W X, Yang H J, Liu J G, Zhao M, Yin C Y, Han X P, Xie Y B 2008 *Complex System and Complex Science* **5** 21 (in Chinese) [汪秉宏, 周涛, 王文旭, 杨会杰, 刘建国, 赵明, 殷传洋, 韩筱璞, 谢彦波 2008 复杂系统与复杂性科学 **5** 21]

[5] Li X, Liu Z H, Wang B 2010 *Complx System and Complex Science* **7** 34 (in Chinese) [李翔, 刘宗华, 汪秉宏 2010 复杂系统与复杂性科学 **7** 34]

[6] Rong Z H, Tang M, Wang X F, Wu Z X, Yan G, Zhou T 2012 *Journal of University of Electronic Science and Technology of China* **41** 801 (in Chinese) [荣智海, 唐明, 汪小帆, 吴枝喜, 严钢, 周涛 2012 电子科技大学学报 **41** 801]

[7] Dorogovtsev S N, Mendes J F F, Samukhin A N 2000 *Phys. Rev. Lett.* **85** 4633

[8] Eagle N, Macy M, Claxton R 2010 *Science* **328** 1029

[9] Papadopoulos F, Kitsak M, Serrano M A, Boguna M, Krioukov D 2012 *Nature* **489** 537

[10] Pinto P C, Thiran P, Vetterli M 2012 *Phys. Rev. Lett.* **109** 068702

[11] Ghoshal G, Barabasi A L 2011 *Nat. Commun.* **2** 394

[12] Goltsev A V, Dorogovtsev S N, Oliveira J G, Mendes J F F 2012 *Phys. Rev. Lett.* **109** 128702

[13] Wang L, Zhang J J 2006 *Complex System and Complex Science* **3** 13 (in Chinese) [王林, 张婧婧 2006 复杂系统与复杂性科学 **3** 13]

[14] He N, Li D, Gan W Y, Zhu X 2007 *Computer Science* **34** 1 (in Chinese) [赫南, 李德, 淦文燕, 朱熙 2007 计算机科学 **34** 1]

[15] Sun R, Luo W B 2012 *Application Research of Computers* **29** 3606 (in Chinese) [孙睿, 罗万伯 2012 计算机应用研究 **29** 3606]

[16] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nat. Phys.* **6** 888

[17] Tan Y J, Yu J, Deng H Z, Zhu D Z 2006 *Systems Engineering* **24** 1 (in Chinese) [谭跃进, 吴俊, 邓宏钟, 朱大智 2006 系统工程 **24** 1]

[18] Liu J, Wang Z, Dang Y 2006 *Mod. Phys. Lett.* B **20** 815

[19] Vragovic I, Louis E, Diaz-Guilera A 2005 *Phys. Rev.* E **71** 036122

[20] Latora V, Marchiori M A 2007 *New J. Phys.* **9** 188

[21] Newman M E J 2010 *Networks An Introduction* (New York: Oxford University Press) p 168–169

[22] Wang J W, Rong L L, Guo T Z 2010 *Journal of Dalian University of Technology* **50** 822 (in Chinese) [王建伟, 荣莉莉, 郭天柱 2010 大连理工大学学报 **50** 822]

[23] Chen D B, Lü L Y, Shang M S, Zhang Y C, Zhou T 2012 *Physica* A **391** 1777

[24] Ren Z M, Shao F, Liu J G, Guo Q, Wang B H 2013 *Acta Phys. Sin.* **62** 128901 (in Chinese) [任卓明, 邵凤, 刘建国, 郭强, 汪秉宏 2013 物理学报 **62** 128901]

[25] Centola D 2010 *Science* **329** 1194

[26] Ugander J, Backstrom L, Marlow C, Kleinberg J 2012 *Proc Natl. Acad. Sci. USA* **109** 5962

[27] Stephenson K, Zelen M 1989 *Soc. Netw.* **1** 11

[28] Borgatti S P 2005 *Soc. Netw.* **27** 55

[29] Poulin R, Boily M C, Massea B R 2000 *Soc. Netw.* **22** 187

[30] Katz L 1953 *Psychometrika* **18** 39

[31] Sabidussi G 1966 *Psyehometrika* **31** 581

[32] Zhang J, Xu X K, Li P, Zhang K, Small M 2011 *Chaos* **21** 016107

[33] Huang X Q, Vodenska I, Wang F Z, Havlin S, Stanley H E 2011 *Phys. Rev.* E **84** 046101

[34] Freeman L 1977 *Sociometry* **40** 35

[35] Zhou T, Liu J, Wang B H 2006 *Chin. Phys. Lett.* **23** 2327

[36] Travencolo B A N, Costa L D 2008 *Phys. Lett.* A **373** 89

[37] Comin C H, Costa L D 2011 *Phys. Rev.* E **84** 056105

[38] Li P X, Ren Y Q, Xi Y M 2004 *Systems Engineering* **22** 13 (in Chinese) [李鹏翔, 任玉晴, 席酉民 2004 系统工程 **22** 13]

[39] Tan Y J, Wu J, Deng H Z 2006 *Sytems Engineering-Theory and Practice* **788** 79 (in Chinese) [谭跃进, 吴俊, 邓宏钟 2006 系统工程理论与实践 **788** 79]

[40] Yu X, Li Y H, Zheng X P, Zhang H Y, Guo Y L 2008 *Journal of Tsinghua University (Science and Technology)* **48** 541 (in Chinese) [余新, 李艳和, 郑小平, 张汉一, 郭奕理 2008 清华大学学报 (自然科学版) **48** 541]

[41] Rao Y P, Lin J Y, Yue D F 2009 *Computer Engineering* **35** 14 (in Chinese) [饶育萍, 林竞焉, 月东方 2009 计算机工程 **35** 14]

[42] Chen K Q, Li S W, Zhou J 2010 *Computer Engineering and Applications* **46** 95 (in Chinese) [程克勤, 李世伟, 周健 2010 计算机工程与应用 **46** 95]

[43] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E 2007 *Proc. Natl. Acad. Sci.* USA **104** 11150

[44] Zeng A, Zhang C J 2013 *Phys. Lett.* A **377** 1031

[45] Garas A, Schweitzer F, Havlin S 2012 *New J. Phys.* **14** 083030

[46] Liu J G, Ren Z M, Guo Q 2013 *Physica* A **392** 4154

[47] Ren Z M, Liu J G, Shao F, Guo Q 2013 *Acta Phys. Sin.* **62** 108902 (in Chinese) [任卓明, 刘建国, 邵凤, 郭强, 汪秉宏 2013 物理学报 **62** 108902]

[48] Hou B N, Yao Y P, Liao D S 2012 *Physica* A **391** 4012

[49] Bryan K, Leise T 2006 *SIAM Rev.* **48** 569

[50] Berkhin P 2005 *Int. Math.* **2** 73

[51] Lü L Y, Zhang Y C, Yeung C H, Zhou T 2011 *PLoS One* **6** e21202

[52] Kleinberg J M 1999 *ACM* **46** 604

[53] Radicchi F, Fortunato S, Markines B, Vespignani A 2009 *Phys. Rev.* E **80** 056103

[54] Radicchi F 2011 *PLoS One* **6** e17249

[55] Masuda N, Kori H 2010 *Phys. Rev.* E **82** 056107

[56] Chen Y, Hu A Q, Hu X 2004 *Journal of China Institute of Communications* **25** 129 (in Chinese) [陈勇, 胡爱群, 胡啸 2004 通信学报 **25** 129]

[57] An S H, Nie P R, He G G 2006 *Journal of Management Sciences in China* **9** 37 (in Chinese) [安世虎, 聂培尧, 贺国光 2006 管理科学学报 **9** 37]

[58] Wu J, Tan Y J, Deng H Z, Chi Y 2007 *Journal of Chinese Computer Systems* **28** 627 (in Chinese) [吴俊, 谭跃进, 邓宏钟, 迟妍 2007 小型微型计算机系统 **28** 627]

[59] Chen J, Sun L F 2009 *Journal of Southwest Jiaotong University* **44** 426 (in Chinese) [陈静, 孙林夫 2009 西南交通大学学报 **44** 426]

[60] Xiao L J, Wu J N, Xuan Z G 2010 *Science of Science and Management of S&T* **241** 12 (in Chinese) [肖连杰, 吴江宁, 宣照国 2010 科学学与科学技术管理 **241** 12]

[61] Ye C S, Wang C L, Liu H W 2010 *Statistics and Decision* **301** 22 (in Chinese) [叶春森, 汪传雷, 刘宏伟 2010 统计与决策 **301** 22]

[62] Zhou X, Zhang F M, Li K W, Hui X B, Wu H S 2012 *Acta Phys. Sin.* **61** 050201 (in Chinese) [周漩, 张凤鸣, 李克武, 惠晓滨, 吴虎胜 2012 物理学报 **61** 050201]

[63] Yan G, Zhou T, Wang J, Fu Z Q, Wang B H 2005 *Chinese Phys. Lett.* **22** 510

[64] Borge-Holthoefer J, Rivero A, Moreno Y 2012 *Phys. Rev.* E **85** 066123

[65] Borge-Holthoefer J, Moreno Y 2012 *Phys. Rev.* E **85** 026116

[66] Klemm K, Serrano M A, Eguiluz V M, San Miguel M 2012 *Sci. Rep.* **2** 292

[67] Aral S, Walker D 2012 *Science* **337** 337

[68] Liu J, Wu Z X, Wang F 2007 *Int. J. Mod. Phys.* C **18** 1087

[69] Silva R A P, Viana M P, Costa L 2012 *J. Stat. Mech.* **7** P07005

[70] Bond R M, Fariss C J, Jones J J, Kramer A D, Marlow C, Settle J E, Fowler J H 2012 *Nature* **489** 295

[71] Holme P, Saramaki J 2012 *Phys. Rep.* **519** 97

[72] Takaguchi T, Sato N, Yano K, Masuda N 2012 *New J. Phys.* **14** 093003 [73] Kim H, Anderson R 2012 *Phys. Rev.* E **85** 026107

Comprehensive Survey for the Frontier Disciplines

Node importance ranking of complex networks*

Liu Jian-Guo† Ren Zhuo-Ming Guo Qiang Wang Bing-Hong

(*Complex Systems Science Research Center, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

(Received 11 April 2013; revised manuscript received 8 May 2013)

Abstract

Identifying the most important nodes, or ranking the node importance by using the method of quantitative analysis in large scale networks are important problems in the complex networks. In this article, the metrics for node importance ranking in complex networks are reviewed and the latest progresses in this field are summarized from two prospects: the network structure and the spreading dynamics. The merits, weaknesses and applicable conditions of different node importance ranking metrics are analyzed. Finally, several important open problems are outlined as possible future directions.

Keywords: complex networks, node importance, network structure, spreading dynamics

PACS: 89.75.Hc, 89.75.Fb **DOI:** 10.7498/aps.62.178901

* Project supported by the National Natural Science Foundation of China (Grant Nos. 71071098, 71171136, 91024026), the Innovation Program of Shanghai Municipal Education Commission, China (Grant Nos. 11ZZ135, 11YZ110), the Key Project of Chinese Ministry of Education, China (Grant No. 211057), and the Shanghai Leading Academic Discipline Project of China (Grant No. XTKX2012).

† Corresponding author. E-mail: liujg004@ustc.edu.cn