网络科学中相对重要节点挖掘方法综述

朱军芳[1,2]，陈端兵[1]，周 涛[1]，张千明[1*]，罗咏劼[1]

(1. 电子科技大学计算机科学与工程学院 成都 610054；2. 西南科技大学理学院 四川 绵阳 621010)

【摘要】网络科学中相对重要节点的挖掘具有重要的实际应用价值。设计衡量节点相对重要程度的指标和方法，是准确有效地识别复杂网络中相对重要节点的关键。本文对近二十多年来网络科学领域中提出的相对重要节点衡量指标和方法进行了系统性地综述，利用数值模拟方法对这些衡量指标和方法的准确度进行了分析和比较，并讨论了相对重要节点挖掘的一些开放问题和发展趋势。

关 键 词 复杂网络； 马尔科夫链； 网络结构； 相对重要性； 随机过程

中图分类号 TP301; N940 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.04.018

A Survey on Mining Relatively Important Nodes in Network Science

ZHU Jun-fang[1,2], CHEN Duan-bing[1], ZHOU Tao[1], ZHANG Qian-ming[1*], and LUO Yong-jie[1]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054;
2. School of Science, Southwest University of Science and Technology Mianyang Sichuan 621010)

Abstract Mining the relatively important nodes in network science has important practical application. Designing indicators and methods to measure the relative importance of nodes is the key to accurately and effectively identify the relative important nodes. This paper presents a systematic review on the indicators and methods proposed in the field of network science for measuring the relatively importance of nodes, analyzes and compares the accuracy of these indicators and methods by numerical simulation, and discusses some open issues and research trends on mining relatively important nodes.

Key words complex network; Markov chains; network structure; relative importance; stochastic process

近年来，网络科学的迅猛发展，使得人们对自然和社会现象的认识已经从宏观层面深入到微观层面。节点作为微观存在单元，在大多数现实网络中具有不同的作用[1-2]，因此，度量网络中节点的重要性吸引了越来越多学者的关注。目前，大部分的研究都是通过对网络中所有节点的重要性做整体排序，以查找网络中的重要节点[3-16]。然而，除了重要节点，其他相对重要的节点也具有重要的应用价值。如在案件侦查中通过已知罪犯查找其余罪犯[17, 19-21]，通过已知恐怖分子挖掘其余恐怖分子[17, 21]，通过已知致病基因查找未知致病基因[22]，或通过已知染病节点查找或预测风险节点[23]等。上面几个重要应用场景，都是通过已知部分重要节点和网络结构，挖掘可能隐匿在网络中的其他重要节点。一种典型的办法就是先通过量化一个节点相对于一个已知重要节点的重要性(称为相对重要性，有时也称为接近性或者相似性)，再计算一个节点相对已知的重要节点集有多么重要，从而找到相对重要节点，这类研究叫做相对重要节点挖掘[17-18]。

对 N 个节点构成的网络 $G(V,E)$，其中有 N_1 个节点构成重要节点集 V_1，N_2 个节点构成非重要节点集 V_2。其中重要节点集 V_1 由已知重要节点集 R 和未知重要节点集 U 组成。目的是计算未知重要节点集 U 和非重要节点集 V_2 构成的目标节点集 T 中任一节点 t 的相对重要性，最终找到目标节点集 T 中 top-k 个相对重要节点，并对其结果进行评价。

相对重要节点的查找大致可以按照如下框架依次展开[17]：

1) 根据一定的节点相对重要性评估指标或方法计算节点 $t(t\in T)$ 对已知重要节点 $r(r\in R)$ 的相对

收稿日期：2018 – 12 – 20；修回日期：2019 – 03 – 19
基金项目：国家自然科学基金(61433014，61673085，61703074)
作者简介：朱军芳(1980 –)，女，博士，主要从事复杂网络方面的研究.
通信作者：张千明，E-mail: qmzhangpa@gmail.com

重要性 $I(t|r)$，它一般是一个非负的量。

2) 计算 t 对已知重要节点集的相对重要性 $I(t|R)$，它一般定义为 $\{I(t|r):r\in R\}$ 的函数，如平均重要性 $I(t|R)=\frac{1}{|R|}\sum_{r\in R}I(t|r)$；当 t 对节点集 R 中所有节点相对重要性都很高时，可以用 $I(t|R)=\min\{I(t|r):r\in R\}$ 或其他合适的函数[17]。

3) 最后对 T 中所有节点根据相对重要性排序，找到重要性排名 top-k 的节点，利用准确性评价指标对相对重要性指标或方法进行定量的评估。

上述框架中，其核心在于相对重要性的计算。文献[17]指出，节点相对重要性计算指标和方法并非全局重要性识别方法在网络子图中的简单局域化，而是对给定已知重要节点的偏好设计。偏好是节点相对重要性方法和指标设计的核心，是相对重要性的根本体现。目前，根据偏好设计，学者们已经提出了一些相对重要节点挖掘的指标或方法。这些指标和方法主要分为两类：1) 基于网络的结构特征量的指标和方法，如NN指标(nearest neighbors)[24]、RD指标(reciprocal distance)[25-30]、Katz指标[31-33]、BTW指标[20, 34-35] (Betweenness)等；2) 基于随机游走的指标和方法，如MarC指标(Markov centrality)[36-39]、PPR方法(personalized PageRank)[40-45]、PHITS方法(HITS with prior)[42]、CHITS方法(customized HITS)[46]等。这些指标或方法已经能够找到部分相对重要节点，然而它们出现在不同的应用领域，需要归纳和梳理，同时也缺乏科学的评价和比较。

因此，本文对近20年复杂网络中出现的相对重要节点挖掘的指标和方法进行总结，然后针对其准确度进行对比分析，最后讨论了该方向的发展趋势和存在的开放性问题。

1 相对重要性计算指标和方法

相对重要节点的指标和方法包括基于结构特征和基于随机游走的指标和方法。

1.1 基于结构特征的指标

该类指标在给定已知重要节点集的情况下，根据其他节点与已知重要节点之间的网络结构特征量设计相对重要性指标。这些结构特征量包括节点之间的连边或路径等信息，能够衡量节点的相对重要性。

1) NN指标[24]

根据该指标，节点 t 的相对重要性表示为：$I(t|R)=\sum_{r\in R}w_{tr}$。其中，$w_{tr}$ 表示 t 与 r 的连边权重。如果 t 与 r 直接相连，$w_{tr}>0$；如果不相连，则 $w_{tr}=0$。在无权网络中，t 与 r 直接相连，则 $w_{tr}=1$。因此，无权网络中节点 t 的相对重要性为已知重要节点集 R 中与 t 直接相连的节点的个数。

2) RD指标[25-27]

该指标定义为：

$$I(t|R)=\sum_{r\in R}\frac{1}{d_{tr}+1} \tag{1}$$

式中，d_{tr} 表示节点 t 与 r 之间的最短距离。该指标可以找出与已知重要节点集不直接相连但关联程度很大的相对重要节点。此外，还可以采用最短距离本身[28]或最短距离的其他函数[29-30]作为相对重要性衡量指标，如最短距离的 p 范数的倒数[29]、最短距离的sigmoid函数或线性函数[25]等。

3) WSP指标[17]

该指标认为距离已知重要节点的路径越长，节点相对重要性越小，将节点 t 对已知重要节点 r 的相对重要性 $I(t|r)$ 定义为：

$$I(t|r)=\sum_{i=1}^{|P(r,t)|}\lambda^{-|P_i|} \tag{2}$$

式中，$P(r,t)$ 为从 r 到 t 的最短路径集合；P_i 是最短路径集合中第 i 条路径；$|P_i|$ 为第 i 条最短路径的路径长度，即跳数；$1\leqslant\lambda<\infty$ 是一个标量参数。除最短路径外，$P(r,t)$ 也可以选择 K 短路径集合或 K 短节点不相交路径集合[17]。此外，该指标中 λ 可以选取某种变量，如路径上的权重之和[19]等。

4) Katz指标[31-33]

Katz指标是一种基于路径的指标，该指标考虑了节点之间的所有路径，并对较短的路径赋予更大权重。定义为：

$$T=(\boldsymbol{I}-\varphi\boldsymbol{A})^{-1}-\boldsymbol{I}=\varphi\boldsymbol{A}+\varphi^2\boldsymbol{A}^2+\varphi^3\boldsymbol{A}^3+\cdots \tag{3}$$

式中，$\boldsymbol{A}$ 为邻接矩阵；$\varphi\in(0,1)$ 为衰减因子，为保证数列收敛性，φ 的取值应当小于 $\boldsymbol{A}$ 的最大特征值的倒数。在此基础上，节点相对重要性为：$I(t|R)=\sum_{r\in R}T_{rt}$。

5) BTW指标[20,34]

介数指标将节点 v 对 r 的相对重要性表示为从 r 出发经过 v 到 e 的最短路径条数 $\sigma_{re}(v)$ 与从 r 出发到 t 的所有最短路径条数 σ_{re} 之比的和，如：

$$\delta(v|r)=\sum_{e\in V}\frac{\sigma_{re}(v)}{\sigma_{re}} \tag{4}$$

它满足如下递推式：

$$\delta(v|r)=\sum_{w:v\in P_r(w)}\frac{\sigma_{rv}}{\sigma_{rw}}(1+\delta(w|r)) \tag{5}$$

式中，$w:v\in P_r(w)$表示节点v在r到w的最短路径上，且v为w的最近邻的先驱节点。最终，节点t的相对重要性为$I(t|R)=\sum_{r\in R}\delta_r(v)$（或$I(t|R)=\sum_{r\in R}\delta(t|r)$）。

除此之外，几种修改版的介数指标[20-21,35]也可以用来计算节点的相对重要性。

1.2 基于随机游走的指标和方法

给网络中所有节点赋予一定的初值，假设在一个无限长的时间内，这些初值以随机游走方式遍历整个网络，每一时步节点拥有的值按照一定随机概率传递给邻居节点，这个概率被称为转移概率，与当前节点的结构性质有关。当随机游走趋于稳定时，被更多的节点指向的点得到更大的值，最终每个节点的终值与主特征矢成正比。可以认为每个节点的终值正比于这个节点相对于所有其他节点的重要性。如果在这个随机游走过程中，加强从已知重要节点出发的随机游走，使得节点终值更大程度上受已知重要节点影响，这样得到的节点终值可以衡量节点相对于已知重要节点的相对重要性。

1) MarC指标[17,38]

该方法将节点t相对于已知重要节点集R的相对重要性定义为平均第一通过时间的平均值[36-39]的倒数：

$$I(t|R)=\frac{1}{\frac{1}{|R|}\sum_{r\in R}m_{rt}} \tag{6}$$

式中，m_{rt}是从r到t的平均第一通过时间，即从r出发第一次到达t的期望步数。

2) PPR方法[40-45]

为了描述网页的相对重要性，文献[40-41]提出了PageRank with Priors方法。该方法在初始时将节点集R中每个节点r赋值$\mathrm{PPR}(r)=1/|R|$，节点集T中每个节点t赋值$\mathrm{PPR}(t)=0$，然后进行迭代。节点在每个时间步k的PPR值由以下迭代公式得到：$\mathrm{PPR}^k=P\cdot\mathrm{PPR}^{k-1}$，$P$为转移概率矩阵，其中$p_{ji}$为从节点$j$到$i$的跳转概率：

$$p_{ji}=s\frac{a_{ji}}{k_j^{\mathrm{out}}}+(1-s)\frac{v}{|R|} \tag{7}$$

式中，$a_{ji}=1$表示节点j指向节点i，$a_{ji}=0$表示节点j没有指向节点i；如果$i\in R$，则$v=1$，否则$v=0$；s为一参数，在文献[40]中取0.75。与PageRank方法相比，由于PPR方法中，节点在每一步都以$(1-s)v/|R|$的概率跳回已知重要节点，从已知重要节点出发重新开始新的随机游走，因此，得到的值更加依赖于已知重要节点集，具有对已知重要节点集的偏好依赖。该方法已被广泛采用[42-45]。

3) PHITS方法[17, 42]

文献[17]将HITS方法进行扩展，得到PHITS方法。高权威性网页(authority)和高中心性(hub)网页的已知重要节点集可以不同，但一般选择两个节点集相同，都为R。节点权威性值au_i和中心性值h_i的迭代公式如下：

$$\begin{cases}\mathrm{au}_i(k+1)=s\sum_{j=1}^{N}a_{ji}\frac{h_j(k)}{H(k)}+(1-s)\frac{v}{|R|}\\ h_i(k+1)=\sum_{j=1}^{N}a_{ij}\frac{\mathrm{au}_j(k)}{\mathrm{AU}(k)}+(1-s)\frac{v}{|R|}\\ H(k)=\sum_{i=1}^{N}\sum_{j=1}^{N}a_{ji}h_j(k),\ \ \mathrm{AU}(k)=\sum_{i=1}^{N}\sum_{j=1}^{N}a_{ij}\mathrm{au}_j(k)\end{cases} \tag{8}$$

式中，a_{ji}与v的含义与式(7)相同；s为一参数。该方法中，如果节点为已知重要节点，计算该节点的权威性值和中心性值时，以概率$(1-s)v/|R|$跳转回已知重要节点，从而使结果更依赖于已知重要节点。

4) CHITS方法[46]

该方法基于文献统计学提出，目的是提高作者认为重要的文献的权威性。给定每个节点j的初始权威性值au_j和初始中心性值h_j。首先按照HITS算法进行如下迭代：

$$\mathrm{au}_i=\sum_j a_{ji}h_j \qquad h_i=\sum_j a_{ij}\mathrm{au}_j \tag{9}$$

迭代收敛后有$\mathrm{au}^+=\mathrm{au}$，满足$\mathrm{au}_j^+=\sum_i(\sum_k a_{ki}a_{kj})\mathrm{au}_j$。将HITS迭代收敛后节点$j$的权威值$\mathrm{au}_j^+$对任意一个$a_{ki}$求导，得到：

$$\forall_{k,i}\Delta a_{ki}=\frac{\partial \mathrm{au}_j^+}{\partial a_{ki}}=a_{kj}\mathrm{au}_j \tag{10}$$

为了提高节点j的权威性，执行梯度上升法，做如下更新并归一化：

$$a_{ki}^+=a_{ki}+\gamma\Delta a_{ki}/\sum_i\Delta a_{ki} \tag{11}$$

式中，归一化后要求$a_{ki}\geqslant 0$。最后用更新并归一化的A计算新的权威值。该方法中HITS迭代与梯度更新可以循环多次，或者直到矩阵a^+趋于稳定，根据最终的权威值计算节点相对重要性。

5) DK方法[47](diffusion kernel)

文献[47]利用扩散核方法查找致病基因。扩散核定义为：

$$K \equiv \exp(-\beta \boldsymbol{L}) = \lim_{n\to\infty}\left(\boldsymbol{I} - \frac{\beta}{n}\boldsymbol{L}\right)^n \tag{12}$$

式中，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$ 为拉普拉斯矩阵，$\boldsymbol{D}$ 为对角矩阵，对角元素为节点的度 k_i，即节点的邻居数，$\boldsymbol{A}$ 为邻接矩阵；β 为控制扩散程度的正常数，当 β 足够小时，可以看作是一个懒惰随机游走的过程[47-48]；节点以概率 $\frac{\beta}{n}(\beta \leqslant 1/\max\{k_i\})$ 转移到某个邻居节点，每一步的转移概率矩阵为 $\boldsymbol{I}-\frac{\beta}{n}\boldsymbol{L}$，$n$ 为随机游走步数，n 步后转移概率矩阵收敛到 $\lim_{n\to\infty}\left(\boldsymbol{I}-\frac{\beta}{n}\boldsymbol{L}\right)^n$。因此，$K_{ij}$ 代表从节点 i 随机游走到 j 的概率，也称为这两个节点间的扩散核距离。最终利用扩散核距离计算出相对重要性：$I(t\,|\,R)=\sum_{r\in R}K_{rt}$。

6) KSMar方法[17,42,49](*K*-step Markov)

K 步马尔科夫方法(KSMar)利用 *K* 步随机游走后暂态概率分布的累积和来衡量节点相对重要性，当 *K* 足够大时，暂态分布接近于稳态分布。具体地，节点 t 对已知重要节点集 R 的相对重要性 $I(t\,|\,R)$ 由下面的公式计算：

$$I(t\,|\,R)=[\boldsymbol{TP}+\boldsymbol{T}^2\boldsymbol{P}+\cdots+\boldsymbol{T}^K\boldsymbol{P}]_t \tag{13}$$

式中，$\boldsymbol{T}$ 为 $n\times n$ 阶转移概率矩阵，每一个元素表示节点 j 到 i 的转移概率，通常取为 $1/k_j$；$\boldsymbol{P}$ 为 $n\times 1$ 的矢量，代表初始概率分布矢量。$I(t\,|\,R)$ 是和矢量中第 t 个元素。

7) SigPS方法[50](sum of significant paths)

文献[50]提出了路径概率求和方法，即SigPS方法。该方法首先将节点 s' 相对于最近邻邻居节点 s 重要性定义为随机游走过程中节点 s 跳到节点 s' 的概率：

$$P(s,s')=\begin{cases}(1-f)\dfrac{w(s,s')}{\sum_{v\in N(s)}w(s,v)} & s\neq s'\\ 1 & s=s'\end{cases} \tag{14}$$

式中，$w(s,s')$ 是 s 与 s' 的连边权重；$w(s,v)$ 与之类似；f 表示随机游走中信息的损失，称为损失因子，取值为0~1。

在跳转概率基础上，从已知重要节点 r 到节点 t 的一条简单路径(即不含闭环路径)的概率定义为：

$$P(r,t)=P(r,v_1)\left(\prod_{i=1}^{m-1}P(v_i,v_{i+1})\right)P(v_m,t) \tag{15}$$

该路径概率的取值范围是[0,1]。路径概率越大，r 与 t 连接越紧密。

最终，节点 t 相对于已知重要节点 r 的相对重要性由 r 到 t 的所有大于某一阈值 c 的路径概率之和来衡量：

$$I(t\,|\,r)=\sum_g P_g(r,t) \qquad P_g(r,t)\geqslant c \tag{16}$$

式中，$P_g(r,t)$ 为 r 到 t 的一条路径 g 的路径概率。如果 r 到 t 的路径中，有概率大于 c 的路径，t 有一定的相对重要性；如果 r 到 t 的所有路径概率都小于 c，则节点 t 相对于 r 是不重要的。

8) PSP方法[51](particle-swarm propagation)

为了找到合适的同行评审专家，文献[51]提出集群粒子传播法来评价节点的相对重要性，称其为PSP方法。该方法的主要思想为：对已知重要节点集中的每一个节点分配 p 个粒子。每个粒子具有初始能量 $\varepsilon_i=1$，这些粒子通过随机过程在网络中传播。传播过程中每一时步，节点上的每一个粒子按照归一化边权概率随机跳到一个邻居节点上。邻居节点的能量按照如下公式更新：

$$e_{c_i(t')}(t'+1)=e_{c_i(t')}(t')+\varepsilon_i(t') \tag{17}$$

式中，$c_i(t')\in N$ 为该粒子在 t' 时步所在位置，即传播到的节点。

节点能量更新后，该粒子的能量衰减为：

$$\varepsilon_i(t'+1)=(1-\delta_i)\varepsilon_i(t') \tag{18}$$

式中，$\delta_i\in[0,1]$ 为粒子 i 的能量随时间的衰减因子。粒子能量更新后，进行下一时步的随机传播和更新，直到指定的时步 k。然后计算每个节点拥有的能量值 $e_{c_i(t')}$，按照该能量值衡量节点相对重要性。即节点 t 对根节点集的相对重要性为：$I(t\,|\,R)=e_{c_i(t')=t}$。

2 准确性评价指标

衡量复杂网络中节点相对重要性的指标和方法已经应用于一些真实数据集。但是，对这些指标和方法本身，却只有少数文献对它们进行了客观的比较和分析。文献[17,50]对相对重要性指标或方法的top-*k*排序清单计算了*K*-min指标[52]，得到了方法之间的相关性。而precision[53]、recall[19,54]、AUC[55]等指标则可以对相对重要节点挖掘结果进行定量的准确性评价。为了比较各种指标和方法的优劣，本文对这3种定量的准确性评价指标进行简要描述。

2.1 precision指标

该指标只考察排在前 L 位的节点是否预测准确。定义为前 L 个节点中预测准确的比例，则：

$$\text{precision}=N_r/L \tag{19}$$

式中，N_r 为前 L 个节点在未知重要节点集中出现的

个数。

2.2 recall指标

该指标用相对重要性指标或方法找到的top-k个节点中未知重要节点个数n_r与未知重要节点集U中节点个数的比值作为衡量指标。具体可表示为：

$$\text{recall} = n_r / |U| \tag{20}$$

2.3 AUC指标

该方法从整体上衡量算法的准确度。具体计算过程为：每次从未知重要节点集和非重要节点集中各选择一个节点，比较两个节点的相对重要性，如果两节点有同样的重要性，则记0.5分，如果未知相对重要节点集中选择的节点相对重要性大于非重要节点集中选择的节点，则记1分。独立比较n次(n为遍历比较的次数)，其中有n_1次得0.5分，n_2次得1分。

$$\text{AUC} = (0.5n_1 + n_2)/n \tag{21}$$

3 实证比较与分析

本文针对不同的应用场景在多个复杂网络上对上述节点相对重要性指标和方法的准确性进行比较分析。

3.1 网络数据集及重要节点集的选取

表1　网络结构特征量及重要节点数N_1

网络	N	M	N_1	$\langle k\rangle$	C_1	L	C_2	B
yeast	5 093	24 743	1 167	9.72	0.1	3.78	0.21	0.08
Genepath	6 306	57 340	51	18.19	0.32	3.42	0.29	0.06
PPI	9 642	40 513	284	8.40	0.12	3.97	0.24	0.05
SARS	224	2 247	18	20.06	0.65	2.34	0.42	0.16

实验在4个无权无向网络上展开，可分为两种场景：1) 场景一：给出网络结构和已知重要节点集，查找相对重要节点。这一场景有3个网络，包括Yeast网络[56]、人类基因信号通路网①Genepath和人类蛋白质相互作用网络②PPI。分别以Yeast网络中的重要蛋白质[56]及Genepath网络中的阿尔兹海默疾病基因[57]，以及PPI网络中的心脏病基因③为重要节点集；2) 场景二：给出网络结构和已知重要节点，查找传染病可能到达的风险节点。这种场景中，本文以国际航空网络④为数据集，SARS⑤爆发早期传播到的节点(国家)为重要节点集。4个网络的网络结构参数和重要节点数如表1所示。结构特征量包括节点数N，边数M，平均度$\langle k\rangle$[58]，簇系数C_1[58]，平均路径长度L[58]，接近度中心性C_2[4]，介数中心性B[4]。

3.2 准确性评价与比较

在4个网络上，随机选取20组已知重要节点集，每个已知重要节点集包含两个节点。利用节点相对重要性指标和方法，计算节点的相对重要性，根据相对重要性所得排序序列计算precison、recall、AUC对20组已知重要节点集的平均值。在4个网络上将各种方法的参数调至接近最优，具体取值如下：WSP方法中取$\lambda = 6$；Katz指标中取$\varphi = 0.000\,1$；PPR、PHITS方法分别取$s = 0.75$、$s = 0.1$；DK方法中取$\beta = 0.1$；KSMar方法中取$K = 3$；SigPS方法中取$f = 0.3$，$c = 0.000\,1$；PSP中取$\delta = 0.009$，传播时步为100；CHITS中迭代次数为2，$\gamma = 0.8$。

对4个网络取前$N/8$ (四舍五入取整)的节点作为预测的相对重要节点。计算结果如图1～图3所示。总体来看，PPR、PHITS、KSMar方法和Katz指标具有较高的准确性。表2中列出了按照precision和recall的平均值排名处于前6的方法。显然，PPR、PHITS、KSMar方法和Katz指标至少在3个网络上都排在前5，它们的AUC平均值也大都排列在前5。因此，这4种方法在网络数据集上展现了很好的准确度。准确度次之的是MarC、SigPS方法和BTW指标，该方法在两个网络上排名进入前6。总之，在基于网络结构的指标中，大部分指标和方法都考虑了节点与已知重要节点之间距离的影响，由于Katz指标不仅考虑了节点间的所有路径，而且对较短路径赋予更大权重，不同长度路径的重要程度可以通过调节参数决定，因此优于其他基于结构的指标。在基于随机游走的指标和方法中，PPR、PHITS方法分别是PageRank方法和HITS方法对已知重要节点的偏离，KSMar方法为从已知节点出发的PageRank值的累计量。这3种方法能够较大限度地保持原始网络上偏好于已知重要节点的随机游走概率分布，即最接近于主特征矢量对于已知重要节点集的偏离，因此有较好的准确度。此外，根据各种方法和指标得到的平均precision和平均recall值，按照式$F1 = 2\cdot\text{precision}\cdot\text{recall}/(\text{precision}+\text{recall})$计算$F1$ [59-60]，并得到它在4种网络上的平均值，对这些指标和方法进行排序，发现PPR、PHITS、KSMar方法和Katz指标仍然比其他方法和指标具有更大的优势。他们的F1平均值都排在前5名。

① http://www.cancer-systemsbiology.org/dataandsoftware.htm
② http://www.hprd.org/sentDataRequest
③ http://www.disgenet.org/web/DisGeNET/menu/downloads;jsessionid=b9mnw0dwia7keelp7k55ec9e
④ https://openflights.org/data.html
⑤ http://www.who.int/influenza/en/

表2 precision和recall的平均值排名前6的方法和指标

网络	1	2	3	4	5	6
yeast	MarC	**PPR**	**PHITS**	**Katz**	**KSMar**	BTW
Genepath	**KSMar**	**PPR**	BTW	**Katz**	SigPS	KD
PPI	**PPR**	MarC	**PHITS**	CHITS	NN	PSP
SARS	**Katz**	**PHITS**	**KSMar**	**PPR**	SigPS	RD

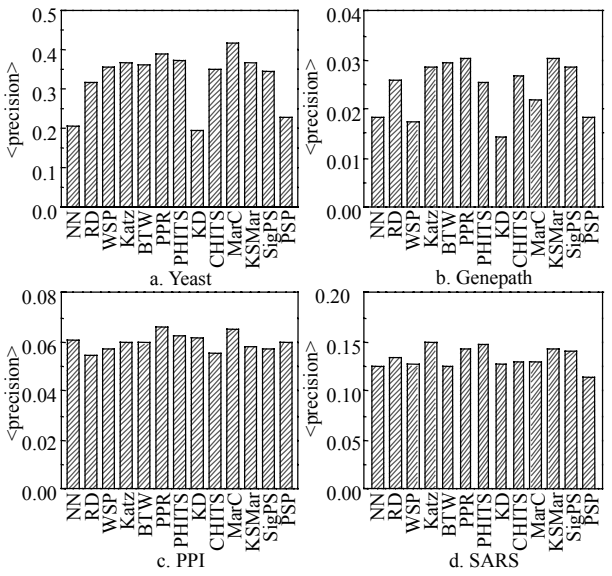

图1 节点相对重要性指标和方法的平均precision

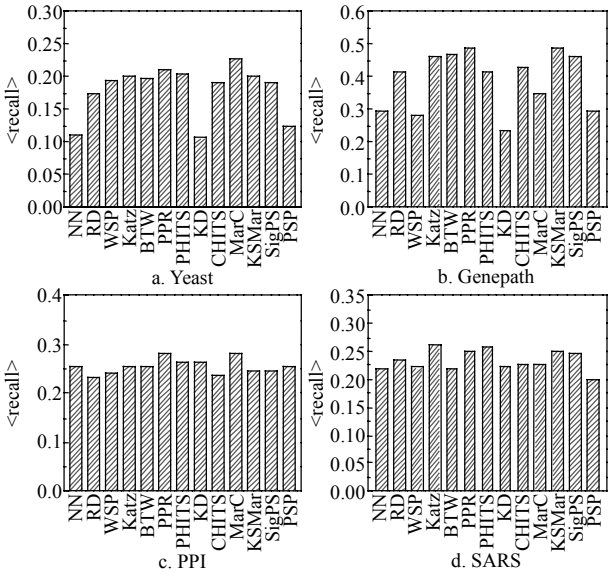

图2 节点相对重要性指标和方法的平均recall

最后，本文考察了PPR、PHITS、KSMar方法和Katz指标在各种网络上准确度的表现，发现这4种方法在4个网络上的平均precision准确度由高到低排名都为：Yeast、SARS、PPI、Genpath网络。其中一个可能的原因是重要节点数比例，4个网络重要节点数的比例依次为：0.231 6、0.081 5、0.029 5、0.008 0。可以看到，随着重要节点数比例依次变小，3种方法的平均precision依次变小。如果按照平均recall排名，则PPR、PHITS方法的准确度由高到低排名为：Genpath、PPI、SARS、Yeast网络。KSMar方法和Katz指标的准确度由高到低排名为：Genpath、SARS、PPI、Yeast网络，其中，PPI网络和SARS网络的平均recall值大小相差不多。由于计算recall值时未知重要节点集在网络中的相对大小不同(4个网络中未知重要节点数的比例依次为0.007 8、0.029 2、0.071 4、0.228 7)，而未知重要节点数对应recall值的分母，对recall值的影响很大，因此，recall值会有上述排名。方法的平均AUC值排序为：Genpath、PPI、Yeast、SARS(除KSMar方法外，KSMar方法在PPI、Yeast、SARS网络上的平均AUC值相差不多)，这也跟未知重要节点数比例大致成正比。当然从更深层次讲，或许这些方法的优劣很大程度上决定于网络结构本身，这也是今后值得研究的一个问题。

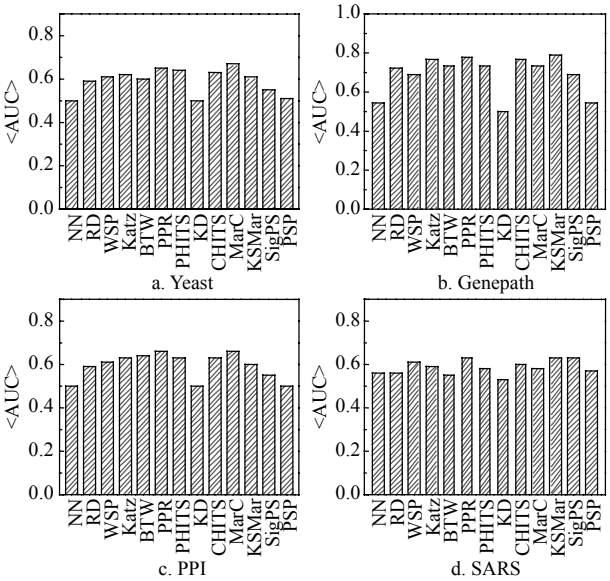

图3 节点相对重要性指标和方法在4个网络上的平均AUC

4 结 束 语

目前，复杂网络中节点相对重要性识别方法已经取得了比较显著的成果，查找相对重要节点的框架已形成，相对重要性识别指标和方法也已经发展了很多。这些方法可以帮助我们查找复杂网络中的相对重要节点，有助于在实际问题中削减经济开支，减少时间花费。本文利用precision、recall和AUC评价指标，对节点相对重要性指标和方法在不同网络上进行了对比和分析。综合来讲，在我们的网络数据集上，PPR、PHITS、KSMar方法和Katz指标具有较高的准确性。

尽管相对重要节点挖掘指标或方法已有较多成

果，但仍然存在以下需要进一步探讨的问题。1) 方法的效率和准确度有待进一步提高。随着大数据的发展[61]，海量数据中查找兴趣节点必然受到关注，怎样设计高效高准确度的相对重要节点挖掘指标和方法至关重要。首先，可以通过粗粒化或社团化将网络进行层次划分[62]，从局域化的角度来设计相对重要节点挖掘方法，也可以先对网络进行合理的抽样，得到相对重要的子网络，在子网络中进行相对重要节点挖掘。其次，可以通过平衡矢量维度与计算效率之间的关系[63]或者进行矩阵分解[64]等方法，达到更高的计算效率。最后，在不同的网络上，方法和指标的参数如何设置[65-66]，才能使相对重要性方法有更高的准确度，从而发挥更大的效力。此外，方法的选取与网络结构的关系[67]又是如何？这些指标或方法的内在作用机制也需要进一步研究。2) 虽然这些方法已经应用到恐怖分子查找[68]、生物领域合作机构关系的研究[17]、重要文献的发现[46]、重要科学家的挖掘[50]、致病基因查找[22,26,32,42,56]、案件中罪犯的揭示[19]等实际应用中，但还有很多领域可以推广，如人才挖掘、计算机病毒传播网络中相对风险节点识别[69]、网络控制中驱动节点的查找[70]、主流新闻引用量的预测[71]，以及网络抽样[72]、关键词提取[73]等。在这些场景中，人工查找相对重要节点耗时费力，应用相对重要节点挖掘指标和方法可以事半功倍。然而，有些情况下应用过程并非直接将方法应用于这样的场景这么简单，在一些场景中，如计算机病毒，传染病传播中风险节点的识别，需要实时的计算，而考虑时序因素的相对重要节点挖掘方法还没有展开。3) 评价指标的问题。在很多场景中，未知重要节点集很难获得，而当前衡量指标大都需要未知重要节点集作为衡量基准。是否可以通过其他方法找到更好或更合适的衡量基准？此外，评价指标本身该怎样选取，哪些指标更为客观公正？上面提出的问题都是需要进一步研究并澄清的开放性问题。

参 考 文 献

[1] ALBERT R, JEONG H, BARABÁSI A L. Error and attack tolerance of complex networks[J]. Nature, 2000, 406: 378-382.

[2] COHEN R, EREZ K, BEN-AVRAHAM D, et al. Breakdown of the Internet under intentional attack[J]. Phys Rev Lett, 2001, 86: 3682.

[3] LÜ L Y, CHEN D B, REN X L, et al. Vital nodes identification in complex networks[J]. Phys Rep, 2016, 650: 1-63 .

[4] FREEMAN L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3): 215-239.

[5] FREEMAN L C. A set of measures of centrality based upon betweenness[J]. Sociometry, 1977, 40: 35-41.

[6] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Comput Netw ISDN Syst, 1998, 30(1): 107-117.

[7] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[8] MORONE F, MAKSE H A. Influence maximization in complex networks through optimal percolation[J]. Nature, 2015, 524: 65-68.

[9] CHEN D, LÜ L, SHANG M S, et al. Identifying influential nodes in complex networks[J]. Physica A, 2012, 391(4), 1777-1787.

[10] CHEN D B, GAO H, LÜ L, et al. Identifying influential nodes in large-scale directed networks: The role of clustering[J]. PLoS One, 2013, 8(10): 77455.

[11] LIU Y, TANG M, ZHOU T, et al. Identify influential spreaders in complex networks, the role of neighborhood[J]. Physica A, 2016, 452: 289-298.

[12] LEE Y L, ZHOU T. Fast asynchronous updating algorithms for k-shell indices[J]. Physica A, 2017, 482: 524-531.

[13] LIU J G, LIN J H, GUO Q, et al. Locating influential nodes via dynamics-sensitive centrality[J]. Scientific Reports, 2016, 6: 21380.

[14] ZHANG C, LIU C, YU L, et al. Identifying the academic rising stars via pairwise citation increment ranking[C]//The 1st Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data. Beijing: Springer international Publishing, 2017: 475-483.

[15] 王伟, 杨慧, 龚凯, 等. 复杂网络上的局域免疫研究[J]. 电子科技大学学报, 2013, 42(6): 817-830.
WANG Wei, YANG Hui, GONG Kai, et al. Local immunization algorithm on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(6): 817-830.

[16] 朱为华, 刘凯, 闫小勇, 等. 识别流网络关键节点的虚拟外界投入产出分析法[J]. 电子科技大学学报，2018, 47(2): 292-297.
ZHU Wei-hua, LIU Kai, YAN Xiao-yong, et al. Identification of critical nodes in flow network by a virtual external input-output analysis[J]. Journal of University of Electronic Science and Technology of China, 2018, 47(2): 292-297.

[17] WHITE S, SMYTH P. Algorithms for estimating relative importance in networks[C]//The 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D.C, USA: ACM, 2003: 266-275.

[18] 赫南, 李德毅, 淦文燕, 等. 复杂网络中重要性节点发掘综述[J]. 计算机科学, 2007, 34(12): 1-5.
HE Nan, LI De-yi, GAN Wen-yan, et al. Mining vital nodes in complex networks[J]. Computer Science, 2007, 34(12): 1-5.

[19] ALZAABI M. CISRI: A crime investigation system using

the relative importance of information spreaders in networks depicting criminals communications[J]. IEEE T INF FOREN SEC, 2015, 10(2): 2196-2211.

[20] MAGALINGAM P, DAVID S, RAO A. Ranking the importance level of intermediaries to a criminal using a reliance measure[DB/OL]. [2015-07-07]. https://arxiv.org/abs/1506.06221v3.

[21] MAGALINGAM P. Complex network tools to enable identification of a criminal community[J]. Bull Aust Math Soc, 2016, 94: 350-352.

[22] 赵静, 林丽梅. 基于分子网络的疾病基因预测方法综述[J]. 电子科技大学学报, 2017, 46(5): 755-765.
ZHAO Jing, LIN Li-mei. A survey of disease gene prediction methods based on molecular networks[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(5): 755-765.

[23] 周涛, 汪秉宏, 韩晓璞, 等. 社会网络分析及其在舆情和疫情防控中的应用[J]. 系统工程学报, 2010, 25(6): 742-754.
ZHOU Tao, WANG Bing-hong, HAN Xiao-pu, et al. Social network analysis and its application in the prevention and control of propagation for public opinion and the epidemic[J]. Journal of Systems Engineering, 2010, 25(6): 742-754.

[24] TIV M, SNEL B, HUYNEN M A, et al. Predicting disease genes using protein-protein interactions[J]. Journal of Medical Genetics, 2006, 43(8): 691-698.

[25] KRAUTHAMMER M, KAUFMANN C A, GILLIAM T C, et al. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease[J]. PNAS, 2004, 101(42): 15148-15153.

[26] GUNEY E, OLIVA B. Exploiting protein-protein interaction network for genome-wide disease-gene prioritization[J]. PloS One, 2012, 7(9): e43557.

[27] BIAGIONI R, VANDENBUSSCHE P Y, NOVACEK V. Finding explanations of entity relatedness in graphs: A survey[DB/OL]. [2018-08-09]. https://arxiv.org/abs/1809.07685v1.

[28] MAGALINGAM P, DAVIS S, RAO A. Using shortest path to discover criminal community[J]. Digital Investigate, 2015, 15: 1-17.

[29] LANGOHAR L. Methods for finding interesting nodes in weighted graphs[D]. Finland: University of Helsinki, 2014.

[30] WU W, JIANG R, ZHANG M Q, et al. Network-based global inference of human disease genes[J]. Molecular Sysemst Biology, 2008, 189(4): 1-11.

[31] KATZ L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.

[32] ZHAO J, YANG T H, HUANG Y, et al. Ranking candidate disease genes from gene expression and protein interaction: A Katz-centrality based approach[J]. PloS One, 2011, 6(9): e24306.

[33] SINGH-BLOM U M, NATARAJAN N, TEWARI A, et al. Prediction and validation of gene-disease associations using methods inspired by social network analyses[J]. PloS One, 2013, 8(5): e58977.

[34] BRANDES U. A faster algorithm for betweenness centrality[J]. Journal of Mathematical Sociology, 2001, 25(2): 163-177.

[35] GEISBERGER R, SANDERS P, SCHULTES D. Better approximation of betweenness centrality[C]//Proceedings on Algorithm Engineering & Experiments. Philadelphia: [s.n.], 2008: 90-100.

[36] LEE Z Q, HSU W J, LIN M. Efficient algorithm for ranking of nodes' importance in information dissemination[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). Beijing: IEEE, 2014: 89-92.

[37] LIN Y, ZHANG Z. Mean first-passage time for maximal-entropy random walks in complex networks[J]. Scientific Reports, 2014(5365): 1-7.

[38] LIN Y，ZHANG Z. Non-Backtracking centrality based random walk on networks[J]. The Computer Journal, 2019, 62(1): 63-80.

[39] CRNOVRSANIN T, CORREA C D, MA K L. Social network discovery based on sensitivity analysis[C]//International Conference on Advances in Social Network Analysis and Mining. Athens: IEEE, 2009: 107-112.

[40] HAVELIWALA T H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search[J]. IEEE Trans Knowl Data Eng, 2003, 15(4): 784-796.

[41] JEN G, WISDOM J. Scaling personalized Web search [C]//The 12th International Conference on World Wide Web. Budapest: ACM, 2003: 271-279.

[42] CHEN J, ARONOW B J, JEGGA G. Disease candidate gene identification and prioritization using protein interaction networks[J]. BMC Bioinformatics, 2009, 10(73): 1-14.

[43] GORI M, PUCCI A. Research paper recommender systems: A random-walk based approach[C]//IEEE/WIC/ACM International Conference on Web Intelligence. Hong Kong, China: IEEE, 2006: 778-781.

[44] LIGETI B, VERA R, LUKACS G. et al. Predicting effective drug combinations via network propagation[C]// Biomedical Circuits and Systems Conference (BioCAS). Rotterdam: IEEE, 2013: 10.1109/BioCAS. 2013. 6679718.

[45] CARDENTE J. Using centrality measures to identify key members of an innovation collaboration network[EB/OL]. [2013-09-16]. http://snap.stanford.edu/class/cs224w-2012/projects/cs224w-043-final.pdf.

[46] CHANG H, COHN D, MCCALLUM A. Creating customized authority lists[C]//The 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2000: 167-174.

[47] KÖHLER S, BAUER S, HORN D, et al. Walking the interactome for prioritization of candidate disease genes[J]. The American Journal of Human Genetics, 2008, 82(4): 949-958.

[48] ZHANG S, NING X M, ZHANG X S. Graph kernels, hierarchical clustering, and network community structure: Experiments and comparative analysis[J]. The European Physical Journal B-condensed Matter and Complex

Systems, 2007, 57(1): 67-74.
[49] SONG M, BLEIK S, YU H, et al. Extracting biomedical concepts from fulltext by relative importance in a graph model[C]//IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). Atlanta: IEEE, 2011: 586-593.
[50] WANG H, CHANG C K, YANG H I, et al. Estimating the relative importance of nodes in social networks[J]. J Inf Process, 2013, 21(3): 414-422.
[51] RODRIGUEZ M A, BOLLEN J. An algorithm to determine peer-reviewers[C]//The 17th ACM Conference on Information and Knowledge Management. Napa Valley: ACM, 2008: 319-328.
[52] FAGIN R, KUMAR R, SIVAKUMAR D. Comparing top *k* lists[J]. SIAM J Discrete Math, 2003, 17(1): 134-160.
[53] ANDREW T, FALK S. User performance versus precision measures for simple search tasks[C]//The 29th Annual International ACM SIGIR Conference on Research and Development in information Retrieval. Seattle: ACM, 2006: 11-18.
[54] PERRY J W, KENT A, BERRY M M. Machine literature searching X machine language, factors underlying its design and development[J]. American Documentation, 1955, 6(4): 242-254.
[55] HANELY J A, MCNEIL B J. The meaning and use of the area under a receiver oprating charicteristic (ROC) curve [J]. Radiology, 1982, 143: 29-36.
[56] LI M, ZHANG H H, WANG J X, et al. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data[J]. BMC Systems Biology, 2012, 6: 15.
[57] KRAUTHAMMER M, KAUFMANN C A, GILLIAN T C, et al. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimers disease[J]. PNAS, 2004, 101(42): 15148-15153.
[58] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012.
WANG Xiao-fan, Li Xiang, CHEN Guan-rong. Network science: An introduction[M]. Beijing: Higher Education Press, 2012.
[59] VAN RIJSBERGEN C J. Information retrieval[M]. Springer, 1998.
[60] SASAKI Y. The truth of the F-measure[EB/OL]. (2007-10-26). https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.
[61] 周涛, 张子可, 陈关荣, 等. 复杂网络研究的机遇与挑战[J]. 电子科技大学学报, 2014, 43(1): 1-5.
ZHOU Tao, ZHANG Zi-ke, CHEN Guan-rong, et al. The opportunities and challenges of complex network research[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(1): 1-5.
[62] ZHANG W, WANG Q. A hierarchical method for estimating relative importance in complex networks[C]// International Symposium on Computer Science and Computational Technology. Shanghai: IEEE, 2008: 63-65.
[63] VIAL D, SUBRAMANIAN V. Personalized pagerank dimensionality and algorithmic implications[DB/OL]. [2018-04-09]. http://export.arxiv.org/pdf/1804.02949.
[64] MAEHARA T, AKIBA T, IWATA Y, et al. Computing personalized PageRank quickly by exploiting graph structures[C]//Proceedings of the VLDB Endowment. Hangzhou: VLDB Endowment, 2014: 1023-1034.
[65] NATHAN E, SANDERS G, FAIRBANKS J. Graph ranking guarantees for numerical approximations to Katz centrality[J]. Procedia Computer Science, 2017, 108: 68-78.
[66] NATHAN E, SANDERS G, HENSON V E. Numerically approximating centrality for graph ranking guarantees[J]. Journal of Computational Science, 2018, 26: 205-216.
[67] DANIEL V, VIJAY S. Personalized PageRank estimation for many nodes: The impact of clustering on complexity[EB/OL]. [2017-10-24]. https://midas.umich.edu/wp-content/uploads/sites/3/2017/10/24-Vial-Daniel-Methodology.pdf.
[68] KREBS, V E. Mapping networks of terrorist cells[J]. Connections, 2001, 24(3): 43-52.
[69] COHEN F. Computer viruses: Theory and practice[J]. Computers & Security, 1987, 6: 22-35.
[70] ZHANG X, WANG H, LV T. Efficient target control of complex networks based on preferential matching[J]. PLoS One, 2017, 12(4): e0175375.
[71] TIMILSINA M, DAVIS B, TAYLOR, M. et al. Predicting citations from mainstream news, weblogs and discussion forum[C]//The International Conference on Web Intelligence. Leipzig: ACM, 2017: 37-244.
[72] SUH C, TAN V Y F, ZHAO R. Adversarial top-k ranking[J]. IEEE Transactions on Information Theory, 2017, 63(4): 2201-2225.
[73] ZHANG Z, PETRAK J, MAYNARD D. Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms[J]. Semantics, 2018, 37: 102-108.

编 辑 蒋 晓