基于加权方法的节点重要性度量*

王 露[1]，郭 强[1]，刘建国[1,2†]

(1. 上海理工大学 复杂系统科学研究中心，上海 200093；2. 上海财经大学 科研实验中心，上海 200433)

摘 要：网络中节点重要性度量对于信息的扩散、产品的曝光、传染性疾病的检测等都具有重大的理论意义。为了度量节点重要性，基于网络拓扑结构考虑全局信息和局部信息提出了加权的节点重要性度量方法。对于一个无权网络，先考虑网络全局信息，计算出每个节点的特征中心向量值，将边两端节点值的和作为边的权重，从而构成一个加权网络；然后根据加权网络的局部信息求出加权网络的度。基于 SIR 模型的四个实证网络，实验结果表明加权方法比特征向量中心性、度中心性、紧密度中心性和介数中心性方法的效果更显著。

关键词：社交网络；节点重要性；加权方法

中图分类号：TP393 **文献标志码**：A **文章编号**：1001-3695(2018)05-1426-03

doi:10.3969/j.issn.1001-3695.2018.05.031

Measuring node importance based on weighted nonlinear method

Wang Lu[1], Guo Qiang[1], Liu Jianguo[1,2†]

(1. *Research Center for Complex Systems Science, University of Shanghai for Science & Technology, Shanghai* 200093, *China*; 2. *Laboratory Centre, Shanghai University of Finance & Economics, Shanghai* 200433, *China*)

Abstract: Identifying the node importance in the network is of significance for information diffusion, product exposure, contagious disease detection, and so on. In order to measure the node importance based on the network topology, this paper presented a weighted method by taking into account the global and local information. Specifically, for an unweighted network, the link weight was the sum of the eigenvector of a pair of nodes connected by this link. Based on the local information, it calculated the node's importance by the node's link weight. According to the SIR simulation on four real networks, the weighted method can rank the node influence more accurately than eigenvector centrality, degree, closeness centrality and betweenness centrality.

Key words: social network; node importance; weighted method(WM)

0 引言

在过去十几年中，复杂网络的研究已经引起许多领域的关注，比如社会科学[1]、计算机科学[2,3]和生物科学[4,5]。识别有影响力节点对于信息的扩散[6,7]、产品的曝光、传染性疾病的检测等[8]都具有重大的理论意义。在实际应用中能预防流行病的蔓延，控制电力网络或互联网中的级联效应等。

到目前为止，已有很多依赖于网络拓扑结构来评估节点重要性的方法，如度中心性、介数中心性、紧密度中心性、特征向量中心性和 k-核中心性等指标。不同指标在不同网络中的效果是不一样的，有的指标基于网络的局部属性，有的是全局属性，还有的是基于网络位置和随机游走，等等。因此，如何根据拓扑信息提高节点重要性排名的准确性是非常重要的一项工作。最近，有很多的研究提出了用多个中心性指标来识别有影响力的节点。Huang 等人[9]认为中心性指标可以分为成键特性、桥链接特性和拓扑特性三大类，通过分析中心性指标之间的联系，Huang 等人提出了一种新的方法识别有影响力的节点，三个有着强相关的中心性通过线性加和的方式来识别。类似地，文献[10,11]也提出了这种多个中心性指标相结合的方法。

这些方法只考虑了多个中心性指标的线性关系，并不适用于实证网络。在本文中讨论了有影响力的节点不仅仅取决于与它直接相连节点和网络局部信息，还取决于网络结构的拓扑性质。受此想法的启发，本文提出了加权方法(weighted method，WM)来识别有影响力的节点。经典的加权方法主要考虑了节点的局部属性对节点的重要性进行度量，如 2012 年 Chen 等人[12]认为节点的影响力不仅考虑节点最近邻居，还考虑次近邻居。文献[13,14]认为节点的影响力取决于其邻居节点 k-核值，这些方法主要通过对邻居节点的重要性进行线性叠加，度量目标节点的重要性。本文的方法首先根据网络的全局信息赋予每条边一个权值，从而构建成一个加权网络；接着，用每个节点的局部性信息作为识别有影响力节点的方法。通过比较 SIR 模型在四个实证网络中的传播过程，实验结果表明，本文方法相比于度中心性、介数中心性、紧密度中心性和特征向量中心性，更能准确地识别出有影响力的节点。

1 相关理论

1.1 网络的图表示

对于一个含有 $|V| = N$ 个节点和 $|E| = M$ 条边的无向网络

收稿日期：2017-01-10；**修回日期**：2017-03-09 **基金项目**：国家自然科学基金资助项目(71271126,71374177,61361125)；上海市曙光学者资助项目(14SG42)

作者简介：王露(1992-)，男，重庆大足人，硕士，主要研究方向为复杂网络；郭强(1975-)，女，辽宁大连人，教授，博士，主要研究方向为复杂网络、数据挖掘、科学知识图谱分析；刘建国(1979-)，男(通信作者)，山西临汾人，教授，博士(后)，主要研究方向为网络科学(liujg004@ustc.edu.cn).

$G=(V,E)$,可以用一个邻接矩阵 $A=\{a_{ij}\}$ 来表示。其中,$a_{ij}=1$ 表示节点 i 与 j 相互连接,$a_{ij}=0$ 表示两节点之间不存在连边。

节点的度 k_i 定义为与节点 i 直接相连接边的数目。节点 i 的接近中心性定义为该节点到网络中所有节点的距离平均值的倒数[15];节点 i 的介数中心性刻画为经过节点 i 的最短路径的数目。在一个复杂网络中,介数中心性测量了经过该节点最短路径所占比例大小[16]。特征向量中心性既考虑了邻居节点的数量,也考虑了邻居节点的重要性[17]。用 SIR 模型在一个网络中仿真,能够发现生成的中心性排序之间有很大的不同,这种现象表明没有一个中心性指标可以很全面地度量出一个节点传播影响力大小。为此,通过考虑多个中心性指标之间强的相关性和网络的拓扑性质,本文提出了一种加权的方法。之前的一些研究工作已经分析了多个指标之间的相关性,结果表明包括度中心性、特征向量中心性、介数中心性指标彼此之间呈正相关。所以在加权方法中,本文选择了特征中心向量和度中心性分别作为全局和局部测量指标。

1.2　加权方法

如图 1 所示,其中(a)表示每个节点通过邻接矩阵计算得到的特征向量值,(b)表示通过式(1)构建的一个加权网络。对于一个无权网络,本文首先根据两个节点的特征向量值的和赋予节点连边一个权值,有以下定义:

$$w_{ij}=EV_i+EV_j \tag{1}$$

其中:w_{ij}是连边 E_{ij}的权值;EV_i 和 EV_j 分别是节点 i 和 j 的特征向量值。根据式(1)算出边的权重,如图 1(b)所示。构建一个加权网络之后,本文考虑加权网络的局部信息,用度指标识别节点影响力大小,定义如式(2)所示。

$$s_i=\sum_j w_{ij} \tag{2}$$

其中:j 是节点 i 的邻居节点;w_{ij}是连边 E_{ij}的权值。

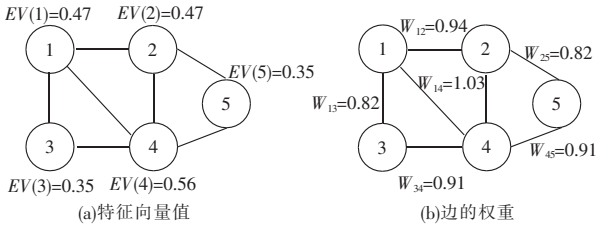

图 1　一个包含五个节点和七条边的网络实例

2　理论分析

2.1　数据分析

为了检验加权方法的效果,本文介绍了四个实证网络,即 Email[18]、Message、C. elegans、US air line 网络。Email 网络起源于 Rovira i Virgili 大学里成员之间邮件的往来,成员代表节点,成员之间的联系就表示两者之间有连边。Message 网络来源于美国尔湾市 California 大学学生的一个在线社区,数据集包括用户至少接收或发送一个信息。C. elegans 网络代表了线虫的一个神经网络,N 代表网络节点总数,E 代表网络的总边数,平均度$\langle k\rangle=\frac{1}{N}\sum_i k_i$,二阶平均度$\langle k^2\rangle=\frac{1}{N}\sum_i k_i^2$,传播阈值$\beta_{\text{thd}}\approx\langle k\rangle/\langle k^2\rangle$。这四个实证网络的统计特征如表 1 所示。

表 1　网络统计特征

网络	N	E	$\langle k\rangle$	$\langle k^2\rangle$	β_{thd}
Email	1 133	5 451	9.62	180	0.06
US air line	332	2 126	12.81	568	0.03
Message	1 266	6 451	10.19	279	0.04
C. elegans	297	2 148	14.46	377	0.04

2.2　SIR 模型

在本文中采用 SIR[19] 传染病模型仿真节点传播过程,SIR 模型被广泛地运用于模拟网络的传播过程。在这个模型中,节点状态可以分为三类[20]:a)易染状态 S,一个个体在感染之前是处于易感染状态的,即该个体有可能被邻居个体感染;b)感染状态 I,一个感染上某种疾病的个体就处于感染状态,该个体还会以一定概率感染其邻居个体;c)移除状态 R,也称为免疫状态或恢复状态,当一个个体经历过一个完整的感染周期后,该个体就不再被感染,因此就可以不再考虑该个体。在每个时间步长里,本文最初将除了初始节点外的所有节点都看成是易感染状态;接着感染的初始节点开始以传播率 β 感染其邻居节点,同时,感染状态下的节点以概率 μ 恢复到易感染状态;最后,如果感染的节点数量不再随着时间的变化而增加,传播过程达到一个稳定的状态,初始节点以及感染的节点数量表示该节点的传播影响力大小。本文定义了有效的传播效率 $\lambda=\beta/\mu$,恢复速率固定为 $\mu=1$[21],所有的结果都是取超过 10 000 次仿真实验的平均值来实现的。

在四个实证网络中对加权方法作了一个排序之后,本文也同时对度、紧密度中心性、介数中心性和特征向量中心性作了降序排列。为了评估加权方法与中心性指标性能,本文使用 Kendall's tau(肯德尔系数)τ 测量由 SIR 模型中产生的传播影响力大小和通过加权方法生成的结果之间的相关性,Kendall's tau τ 用来测量这两个排序列表之间的相关性。Kendall's tau τ 值为[-1,1],该值越大,说明这种方法识别节点传播影响力大小更精确。Kendall's tau τ 被定义为

$$\tau=\frac{2}{n(n-1)}\sum_{i<j}\operatorname{sgn}[(x_i-x_j)(y_i-y_j)] \tag{3}$$

其中:n 表示节点的总数;x_i 表示 SIR 模型中产生的传播影响力大小;y_i 表示通过加权方法、度中心性、紧密度中心性、介数中心性和特征向量中心性产生的排序列表。sgn(x)是一个分段函数,当 $x>0$ 时,sgn(x) = +1;当 $x<0$ 时,sgn(x) = -1;当 $x=0$ 时,sgn(x) =0。这个 τ 值越大,说明该中心性方法更加精确,性能更好。τ 的理想值是 $\tau=1$,此时通过中心性指标生成排序列表完全符合实际传播过程产生的排序列表;当 $\tau=-1$ 时,通过中心性指标生成排序列表与实际传播过程产生的排序列表完全相反。

3　数值结果

对于 Email、Message、C. elegans 和 US air line 网络,用中心性指标和加权方法的 Kendall's tau τ 值得到的结果如图 2 所示。在 Email、US air line、Message 和 C. elegans 四个网络中,Kendall's tau τ 值大小包括通过 SIR 传播过程产生的排序列表和通过加权方法、度 K、紧密度 C、介数 B、特征向量 eig 产生的排序列表。实验结果是 10 000 次独立实验结果的平均值和不

同传播率下的 20 个时间步长。结果表明加权方法总是比特征向量、介数、紧密度中心性指标效果更好。尽管在 US air line 和 Message 网络中,特征向量中心性指标实验效果与加权方法接近,这是由于这两个网络拓扑结构决定的,但是加权方法效果还是比特征向量 eig 效果要好。当传播速率 λ 小于传染病阈值 β_{thd} 时,SIR 传播过程将会停止,因此节点度越大将会感染更多的未被感染的节点,这就是为什么当传播速率 λ 小于传染病阈值 β_{thd} 时度的 Kendall's tau τ 值大的原因。然而当传播速率大于传染病阈值时,加权方法相比度中心性指标性能更好。

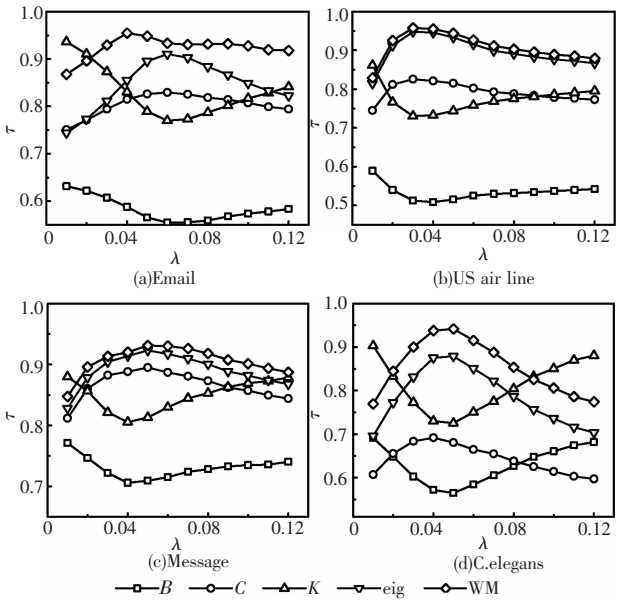

图 2 多个指标实验结果

4 结束语

在网络中识别出重要节点是一项非常重要的工作。最近提出了许多关于使用多个中心性指标方法来识别出有影响力节点的方法,然而,这些方法只考虑了多个中心性指标之间的线性关系且不适用于实证网络中。本文讨论了网络中有影响力的节点不仅取决于与它直接相连的邻居节点(局部信息),也取决于网络的拓扑结构。基于此想法,本文提出了加权的方法来识别有影响力节点。WM 是一种同时考虑了网络结构和节点的局部及全局信息的非线性耦合的方法。

为了评估加权方法的性能,本文使用 Kendall's tau 的度、介数、紧密度、特征向量四个指标在 SIR 模型中生成的排序列表进行比较。四个实证网络结果表明,加权方法总比特征向量、介数和紧密度指标的性能好;当传播速率 λ 大于传染病阈值 β_{thd} 时,加权方法相比度中心性指标性能也更好。

加权方法取决于一个节点的局部及全局信息,对于识别出有影响力节点是非常有帮助的。然而,一个节点传播影响力不仅取决于网络结构、节点的局部及全局信息,还受一个网络动力学的影响。如何结合网络结构和网络动力学来识别出影响力大的节点是值得深入研究的问题。

参考文献:

[1] Zhou Tao, Liu Jianguo, Bai Wenjie, *et al*. Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity[J]. Physical Review E, 2006, 74(5): 056109.

[2] Siganos G, Faloutsos M, Faloutsos P, *et al*. Power laws and the AS-level Internet topology[J]. IEEE/ACM Trans on Networking, 2003, 11(4): 514-524.

[3] 宋文君,郭强,刘建国,等. 一种改进的混合推荐算法[J]. 上海理工大学学报,2015,37(4):327-331.

[4] Albert R, Barabási A L. Statistical mechanics of complex networks [J]. Reviews of Modern Physics, 2002, 74(1): 47-97.

[5] Wang Xiaofan, Chen Guanrong. Synchronization in small-world dynamical networks[J]. International Journal of Bifurcation and Chaos, 2002, 12(1): 187-192.

[6] 郭强. 基于个体局部交互作用的舆情传播模型研究[J]. 计算机应用研究,2012,29(11):4085-4086,4112.

[7] 郭强,刘新惠,胡兆龙. 真实信息发布在谣言传播中的作用研究[J]. 计算机应用研究,2014,31(4):1031-1034,1050.

[8] 倪小军,王美娟. 无标度网络的三种病毒控制策略研究[J]. 上海理工大学学报,2006,28(3):249-252.

[9] Huang Shaobin, Lyu Tianyang, Zhang Xizhe, *et al*. Identifying node role in social network based on multiple indicators[J]. PLoS One, 2014, 9(8): e103733.

[10] Fu Y H, Huang C Y, Sun C T. Using global diversity and local topology features to identify influential network spreaders[J]. Physica A: Statistical Mechanics & Its Applications, 2015, 433(9): 344-355.

[11] Liu Zhonghua, Jiang Cheng, Wang Juyun, *et al*. The node importance in actual complex networks based on a multi-attribute ranking method [J]. Knowledge-Based Systems, 2015, 84(3): 56-66.

[12] Chen Duanbing, Lyu Linyuan, Shang Mingsheng, *et al*. Identifying influential nodes in complex networks[J]. Physica A: Statistical Mechanics & Its Applications, 2012, 391(4): 1777-1787.

[13] Bae J, Kim S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness[J]. Physica A: Statistical Mechanics & Its Applications, 2014, 395(4): 549-559.

[14] Lin Jianhong, Guo Qiang, Dong Wenzhao, *et al*. Identifying the node spreading influence with largest *k*-core values[J]. Physics Letters A, 2014, 378(45): 3279-3284.

[15] Freeman L C, Roeder D, Mulholland R R. Centrality in social networks: ii. experimental results[J]. Social Networks, 1980, 2(2): 119-141.

[16] Freeman L C. A set of measures of centrality based on betweenness [J]. Sociometry, 1977, 40(1): 35-41.

[17] Estrada E, Rodríguez-Velázquez J A. Spectral measures of bipartivity in complex networks [J]. Physical Review E, 2005, 72(4): 046105.

[18] Guimera R, Danon L, Diaz-Guilera A, *et al*. Self-similar community structure in a network of human interactions[J]. Physical Review E, 2003, 68(6): 065103.

[19] Shulgin B, Stone L, Agur Z. Pulse vaccination strategy in the SIR epidemic model [J]. Bulletin of Mathematical Biology, 1998, 60(6): 1123-1148.

[20] Anderson R M, May R M, Anderson B. Infectious diseases of humans: dynamics and control[M]. Oxford: Oxford University Press, 1992.

[21] Zhao Xiangyu, Huang Bin, Tang Ming, *et al*. Identifying effective multiple spreaders by coloring complex networks [J]. Europhysics letters, 2014, 108(6): 68005.