物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于扩展度的复杂网络传播影响力评估算法

闵磊 刘智 唐向阳 陈矛 刘三妍

Evaluating influential spreaders in complex networks by extension of degree

Min Lei Liu Zhi Tang Xiang-Yang Chen Mao Liu San-Ya

引用信息 Citation: Acta Physica Sinica, 64, 088901 (2015) DOI: 10.7498/aps.64.088901
在线阅读 View online: http://dx.doi.org/10.7498/aps.64.088901
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2015/V64/I8

基于扩展度的复杂网络传播影响力评估算法*

闵磊 刘智 唐向阳 陈矛† 刘三妍

(华中师范大学, 国家数字化学习工程技术研究中心, 武汉 430079)

(2014 年 9 月 4 日收到; 2014 年 11 月 17 日收到修改稿)

对网络中节点的传播影响力进行评估具有十分重要的意义, 有助于促进有益或抑制有害信息的传播. 目前, 多种中心性指标可用于对节点的传播影响力进行评估, 然而它们一般只有当传播率处于特定范围时才能取得理想的结果. 例如, 度值中心性指标在传播率较小时较为合适, 而半局部中心性和接近中心性指标则适用于稍大一些的传播率. 为了解决各种评估指标对传播率敏感的问题, 提出了一种基于扩展度的传播影响力评估算法. 算法利用邻居节点度值叠加的方式对节点度的覆盖范围进行了扩展, 使不同的扩展层次对应于不同的传播率, 并通过抽样测试确定了适合于特定传播率的层次数. 真实和模拟数据集上的实验结果表明, 通过扩展度算法得到的扩展度指标能在不同传播率下对节点的传播影响力进行有效评估, 其准确性能够达到或优于利用其他中心性指标进行评估的结果.

关键词: 复杂网络, 传播影响力, 扩展度

PACS: 89.75.Hc, 89.75.Fb **DOI:** 10.7498/aps.64.088901

1 引 言

现实世界中许多事物都是相互影响、彼此关联的, 它们通常能以复杂网络的形式呈现, 比如社交网络、论文合著网络以及互联网等[1−3]. 这些网络上经常存在着信息 (如流言、知识或病毒等) 的传播[4−6], 因此为了促进有益或抑制有害信息的传播, 通常需要对网络中节点的传播影响力进行评估.

目前, 多种中心性指标或方法可被用来对节点的传播影响力进行评估[7−13], 例如度中心性 (Degree)、接近中心性 (Closeness)[7]、半局部中心性 (SemiLocal)[8]、K-核 (K-Shell)[9,10] 和核心 (Coreness) 中心性等[11]. 其中以度中心性最为简单, 但它只考虑了源节点在一步邻域范围内的传播能力, 在传播率较小时能取得较高的评估准确性, 但对于稍大一些的传播率则不一定适合. 接近中心性从全局范围考虑了网络的拓扑结构, 适合于在传播率较为适中的情况下对影响力进行评估, 但由于涉及节点对之间最短路径的计算, 因此时间复杂度较高[14,15]. 为了在准确性和时间复杂度之间做出平衡, Chen 等[8] 提出了半局部中心性指标, 该指标扩大了源节点邻域的覆盖范围, 在保证较低时间复杂度的前提下, 尽可能提高评估的准确性. 另外, 利用 K-核指标只能将节点的传播影响力区分到几个有限的等级, 粒度较粗[16], 并且其通常只是用来确定影响力最大的节点. 核心中心性指标综合考虑了节点邻域的影响范围及这一范围内节点的 K-核值, 认为影响力大的节点应与 K-核值较大的节点间拥有更多的连接[11]. 关于中心性指标和网络传播方面更详细的介绍, 可以参见文献 [15, 17—19].

以上这些中心性指标虽然原理各异, 但它们都只考虑了网络的拓扑结构. 然而事实上, 即便是在同一网络中, 对于不同的传播率, 相同节点所表现出的传播重要性也不一定完全相同. 因此, 为了能在不同传播率下对节点的传播影响力进行准确评估, 本文提出了一种基于扩展度的传播影响力

* 国家科技支撑计划 (批准号: 2013BAH72B01)、教育部新世纪优秀人才支持计划 (批准号: NCET-11-0654) 和教育部 -中国移动科研基金 (2012) 研发项目 (批准号: MCM20121061) 资助的课题.

† 通信作者. E-mail: eitecpaper@gmail.com

评估算法 (ExDegree 算法, 简称扩展度算法), 将通过该算法得到的影响力度量值称为扩展度中心性 (ExDegree 中心性). 实验表明, 扩展度中心性指标对传播率具有较强的鲁棒性, 能在不同传播率下对节点的传播影响力进行有效评估, 其准确性能够达到或优于相同传播率下的其他中心性指标.

2 评估结果准确性的度量

为了度量评估结果的准确性, 本文使用了易染-感染-免疫 (susceptible-infected-recovered, SIR) 传播仿真模型 [20,21]. 该模型能对真实传播现象进行模拟, 但时间复杂度较高, 因此并不适合直接对传播影响力进行评估, 而是通常被作为衡量其他指标准确性的一个基准.

对于需要评估的节点, 根据评估指标得到一组影响力评估值, 构成一个等级序列; 同时根据仿真传播过程稳定后的 SIR 影响的节点数, 按照相同的节点次序构成另一等级序列. 如果两序列的相关度越高, 则说明评估指标的结果与 SIR 仿真结果越接近, 即也可近似地认为该指标用于传播影响力评估的准确性越高.

等级序列的相关度采用 Kandall's tau 等级相关系数 [14,22] 来衡量, tau 值的取值范围在 -1—1 之间, 值越大说明两等级序列的相关度越高, 这里使用文献 [14] 中用到的公式计算 tau 值:

$$\begin{aligned} &tau(Seq_X, Seq_Y) \\ =&\frac{2}{T(T-1)}\sum_{i<j}\mathrm{Sgn}\big((x_i-x_j)(y_i-y_j)\big), \end{aligned} \tag{1}$$

其中, Seq_X 和 Seq_Y 为两等级序列; x_i 为序列 Seq_X 中元素 i 的等级值; T 为序列长度. $\mathrm{Sgn}(x)$ 为符号函数, 当 $x>0$ 时, 函数值为 $+1$, 表示元素 i 和 j 在两序列中的等级排序次序一致 (不包含等级次序相等); $x<0$ 时, 值为 -1, 表示次序相反; $x=0$ 时, 值为 0, 表示除前面两种场合之外, 即包含等级次序相等的情况.

3 基于扩展度的传播影响力评估算法

本文所述实验均在无权无向网络上进行. 约定 $G=(V,E)$ 表示网络, 其中, V 为节点集合, E 为边集合, $\langle k\rangle$ 为节点平均度, 节点数 $N=|V|$, 边数 $M=|E|$.

算法中, 源节点扩展度的覆盖范围随传播率动态确定, 对较小的传播率选择较小的范围, 略大的则匹配稍大的范围, 覆盖范围与扩展层次相对应, 每个节点在各层次中的扩展度值反映其在特定传播率下的影响力. 设层次数为 0, 1, 2, $\cdots$, 对应的扩展度值分别为 $ExDegree^0_{(i)}$, $ExDegree^1_{(i)}$, $ExDegree^2_{(i)}$, $\cdots$. 初始设置各节点在 0 层的扩展度值为 1, 之后各层次中的扩展度值根据 (3) 式进行计算.

$$ExDegree^0_{(i)}=1, \tag{2}$$

$$\begin{cases} ExDegree^1_{(i)}=\displaystyle\sum_{t\in(\Gamma(i)\cup i)} ExDegree^0_{(t)}, \\ ExDegree^2_{(i)}=\displaystyle\sum_{t\in(\Gamma(i)\cup i)} ExDegree^1_{(t)}, \\ \quad\cdots \\ ExDegree^L_{(i)}=\displaystyle\sum_{t\in(\Gamma(i)\cup i)} ExDegree^{L-1}_{(t)}, \end{cases} \tag{3}$$

其中, $\Gamma(i)$ 表示节点 i 的一步直接邻域; $ExDegree^L_{(i)}$ 为 i 在层次 L 时的扩展度值, 它是 $\Gamma(i)\cup i$ 范围内的节点在 $L-1$ 层扩展度值的叠加. 叠加原理如图 1 所示, 图中的网络层次即为扩展层次, 示例中选 0 号节点为源节点.

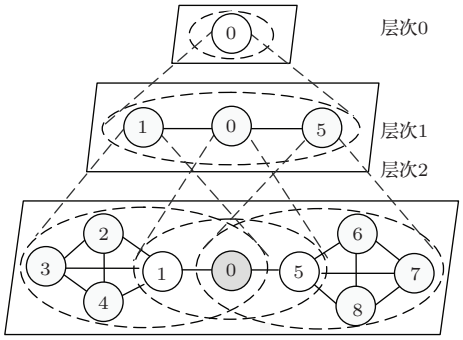

图 1 扩展度的层次图. 更大层次中节点的扩展度具有更大的覆盖范围

Fig. 1. Levels of extension. Nodes in larger levels have larger coverage area.

事实上, 扩展度在覆盖范围未到达网络边界时, 之所以能对影响力准确评估, 除了该范围与实际传播感染范围 (主要被感染节点所在的范围) 匹配外, 更主要的是因为总能找到这样一个扩展层次, 使各节点的权重与它们实际被感染概率之间存在较近似的拟合, 这种权重由节点在叠加过程中被重复计数的次数决定. 图 1 中, 在层次为 2 时, 0, 1, 5 这三个节点在一步邻域范围内包含的节点若分别被计数一次, 则 0 号源节点会被重复计数三次, 1, 5 号节点二次, 其余节点一次, 这样就会按照与源节点的远近 (本质上并非绝对距离, 因为链接密度对感染率也存在影响) 形成节点权重递减的趋势. 另

外, 实际传播中节点被感染的概率也是由近及远递减的. 因此, 若某扩展层次下这两种趋势越接近, 则用扩展度对影响力的评估越准确. 图2 (a) 和图2 (b) 分别描述了节点的被感染率和权重, 其中图2 (b) 中所示的权重被映射到了0—1的范围. 很明显, 在扩展范围未达到网络边界时, 对于某个传播率, 总存在一个适合的扩展层次能产生与实际被感染率最为接近的节点权重分布, 从而使得评估准确性相对较高.

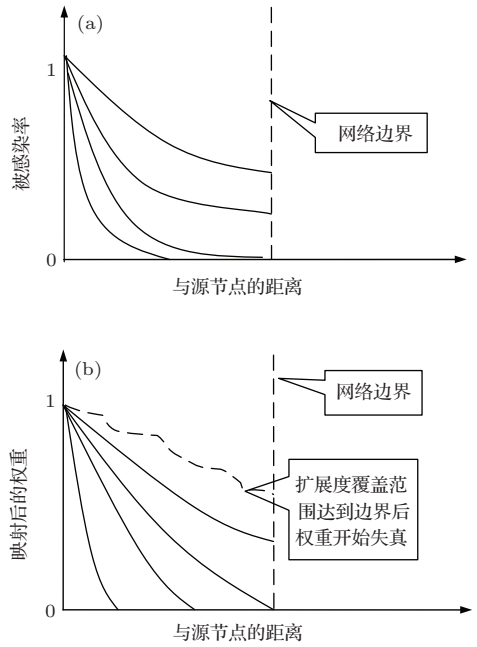

图2 与源节点不同距离的节点的被感染率和基于扩展度的权重 (a) 节点的被感染率, 越靠上的曲线对应着越大的传播率; (b) 不同扩展层次中节点被赋予的权重, 越靠上的曲线对应着越大的扩展层次数

Fig. 2. Infected probabilities and weights of nodes at different distances from source. (a) Infected probabilities of nodes. The curve closer to the top means the larger infection rate. (b) Weights of nodes. The curve closer to the top means the larger level.

而当扩展度覆盖范围到达网络边界后, 对于更大的传播率、更大的扩展层次则通常会导致更低的准确性. 因为此时由扩展过程所产生的越远离源节点权重越小的现象将逐渐淡化, 而网络中高密度区域内节点的权重将逐渐增强. 理由是高密度区域内的叠加更为密集, 该因素在节点权重的决定中将逐渐起主导作用. 所以对于过大的扩展层次, 会产生节点权重与被感染率间不匹配的失真现象.

若将各节点按照与源节点的远近进行划分, 则可近似认为源节点在传播影响力上的差异取决于不同距离节点的数目以及这些节点在平均被感染率上的差异性. 为方便描述, 这里将更靠近源节点的节点称为近源节点, 反之称为远源节点. 设与源节点距离为d的节点表示为$nodes_d$、平均被感染率为$\beta(d)$, 距离$d+1$节点的描述相似. 如果$nodes_{d+1}$的个数越多或平均度$\langle k\rangle$越大, 那么从$nodes_{\rm d}$到$nodes_{d+1}$的路径越多. 由于传播路径越多, 节点被感染的概率越大, 因此$\beta(d+1)$除了与$\beta(d)$有关外, 还与$nodes_d$的个数和$\langle k\rangle$有关. 因此可以认为, 近源节点在个数和$\langle k\rangle$上的差异性直接影响着远源节点的被感染率. 另外从图2 (a) 可以看出, 随着传播率的增大, 不同距离的节点在被感染率上的差异性逐渐减小, 因此在不考虑近源节点的个数和$\langle k\rangle$对远源节点被感染率影响的前提下, 不同距离节点在个数上的差异性对源节点影响力的区分能力将逐渐减小. 而前面提到, 源节点的影响力取决于不同距离节点的数目以及它们在被感染率上的差异性, 因而相同情况下, 不同距离节点在被感染率上的差异性对源节点的影响力区分所起的作用就会相对得到加强. 于是我们推测, 当传播率增大到一定程度后, 近源节点在数目和$\langle k\rangle$上的差异性对源节点影响力的区分将逐渐起主导作用. 若此推测成立, 那么当扩展度覆盖范围到达网络边界后, 很可能只需适当增强近源节点的权重就可提高对源节点传播影响力的区分能力, 从而提高评估准确性, 而在扩展度算法中增强近源节点的权重只需减小扩展层次数就能做到. 因此, 此时更大的传播率对应着更小的扩展层次以增强近源节点的权重, 用于弥补节点权重与实际被感染率不匹配所带来的准确性降低.

按照以上分析, 针对逐渐增大的传播率, 最优扩展层次数变化的趋势将是先增大然后减小, 后文的图11对此进行了验证. 然而, 最优扩展层次数变化的趋势只能辅助我们对其进行判断而不能用来确定其确切值. 因此, 本文另一个较为重要的工作就是根据传播率动态地确定最优扩展层次数.

尽管SIR模型由于时间复杂度较高, 当待评估节点较多时并不适合直接用于影响力的评估, 但可以利用其准确性较高的特点, 通过抽样测试的方式确定最优扩展层次. 首先随机选取少量样本节点, 对它们进行传播仿真, 得到SIR影响力等级序列. 同时根据扩展度算法计算这些节点在不同扩展层次时的扩展度值, 得到与之对应的扩展度值等级序列. 然后从这些序列中, 找到与SIR影响力序列相关度最大的一个, 其对应的层次即被作为最优扩展层次. 最后就可以利用最优扩展层次数确定适合于具体传播率的扩展度值, 即扩展度中心性.

$$tau_{\max} = \max\{tau(Seq^S_{level}, Seq^S_{SIR})|level = 1, 2, 3, \cdots, L\}, \tag{4}$$

$$level_{best} = \{level|tau(Seq^S_{level}, Seq^S_{SIR}) = tau_{\max}|level = 1, 2, 3, \cdots, L\}. \tag{5}$$

(4) 和 (5) 式中, Seq^S_{SIR} 和 Seq^S_{level} 分别为 S 个样本节点通过SIR模型和扩展度算法得到的等级序列; $level$ 为扩展层次; L 为扩展层次上限; $level_{\text{best}}$ 为计算出的最优扩展层次数. 序列的相关度用 (1) 式的 tau 值公式进行描述.

通过实验发现 (图 7), 样本节点数 S 通常无须很大就可较准确地找到最优扩展层次数, 若无特殊说明本文取 $S = 50$. 另外, 根据六度分割理论, 网络中某节点到其他任意节点的距离一般为较小的数, 因此扩展层次数无需很大就可使扩展度的影响力覆盖到很大范围, 文中取扩展层次上限 $L = 15$.

算法中, 当前层的扩展度值根据上一层的结果得到, 在每一层的计算中, 时间增量为 $O(N \times \langle k\rangle)$, 因此扩展过程总的时间为 $O(L \times N \times \langle k\rangle)$. 另外, S 个样本节点进行SIR仿真的时间为 $S \times O(\text{SIR})$, 其中 $O(\text{SIR})$ 为对单个节点仿真所用的时间. 所以, 除去一些额外开销, 扩展度算法总的时间复杂度大致为 $O(L \times N \times \langle k\rangle) + S \times O$ (SIR). 假设待评估节点数为 Z, 直接利用SIR仿真消耗的时间为 $Z \times O(\text{SIR})$. 当 S 远小于 Z 时, $S \times O$ (SIR) 也远小于 $Z \times O(\text{SIR})$, 又因为 $O(L \times N \times \langle k\rangle)$ 与网络规模 N 呈线性关系, 因此, 扩展度算法的时间复杂度远小于SIR模型, 基本在可以接受的范围内, 具有实际可行性.

4 实验与分析

本文在真实和模拟数据上将扩展度中心性与度中心性、接近中心性、半局部中心性、K-核和核心中心性进行了实验对比. 实验中, 如果传播率很大, 整个网络将很快被感染, 从而很难区分单个个体的传播影响力 [23], 因此文中设置传播率限定在 $0.01 \leqslant \beta \leqslant 0.15$ 范围之间, 恢复率设置为0.8, SIR仿真重复1000次结果取平均. 另外, 由于样本节点采取随机选择策略, 因此若无特殊说明, 文中扩展度中心性及相关计算结果均由30次独立实验取平均得到.

4.1 实验数据集

实验中用到的真实网络包括: Blogs [24], Netscience [25], PPI [26], E-mail [27] 和PGP [28]. 这些网络的详细参数如表1所列. 其中, N 为网络中的节点数, M 为边数, $\langle k\rangle$ 为节点的平均度, $\beta_{\text{th}} = \langle k\rangle/\langle k^2\rangle$ 为传播阈值 [29].

表 1 真实数据集的拓扑参数

Table 1. Topological parameters of real datasets.

网络	参数			
	N	M	$\langle k\rangle$	β_{th}
Blogs	3982	6803	3.417	0.072
Netscience	379	914	4.823	0.125
PPI	2445	6265	5.125	0.079
E-mail	1133	5451	9.622	0.054
PGP	10680	24316	4.554	0.053

另外, 实验还使用了 Lancichinetii-Fortunato-Radicchi (LFR) 基准数据模型 [30], 它能生成与现实世界中网络结构较为接近的模拟数据. 各实验设置的LFR参数中, 相同部分有: $N = 5000$, $\text{min}c = 30$, $\text{max}c = 100$, $t1 = 2$, $t2 = 1$, $mu = 0.1$. 其中, $\text{min}c$ 和 $\text{max}c$ 分别是社团的最小和最大规模, $t1$ 和 $t2$ 代表节点度值和社区大小的幂率指数, mu 为混合参数. 实验中, 通过调整平均度 $\langle k\rangle$ 来调节网络的紧密程度.

4.2 评估准确性实验

SIR模型能较为合理地对传播过程仿真, 这里将SIR影响节点数作为基准与通过各指标得到的结果进行对比. 图3在Blogs数据集上通过散布图的形式描述了各种评估值与SIR影响节点数的相关性, 实验中设置传播率为传播阈值 β_{th}. 可以看出, 度中心性、接近中心以及K-核的评估值与SIR影响节点数的相关性相对较弱, 并且利用K-核得到的评估值等级较少, 对节点影响力的区分度不强. 半局部中心性、核心中心性和扩展度中心性的评估值与SIR影响节点数则有相对较强的正相关性, 说明扩展度中心性和较优的半局部中心性和核心中心性一样, 能较为准确地对节点的传播影响力进行评估.

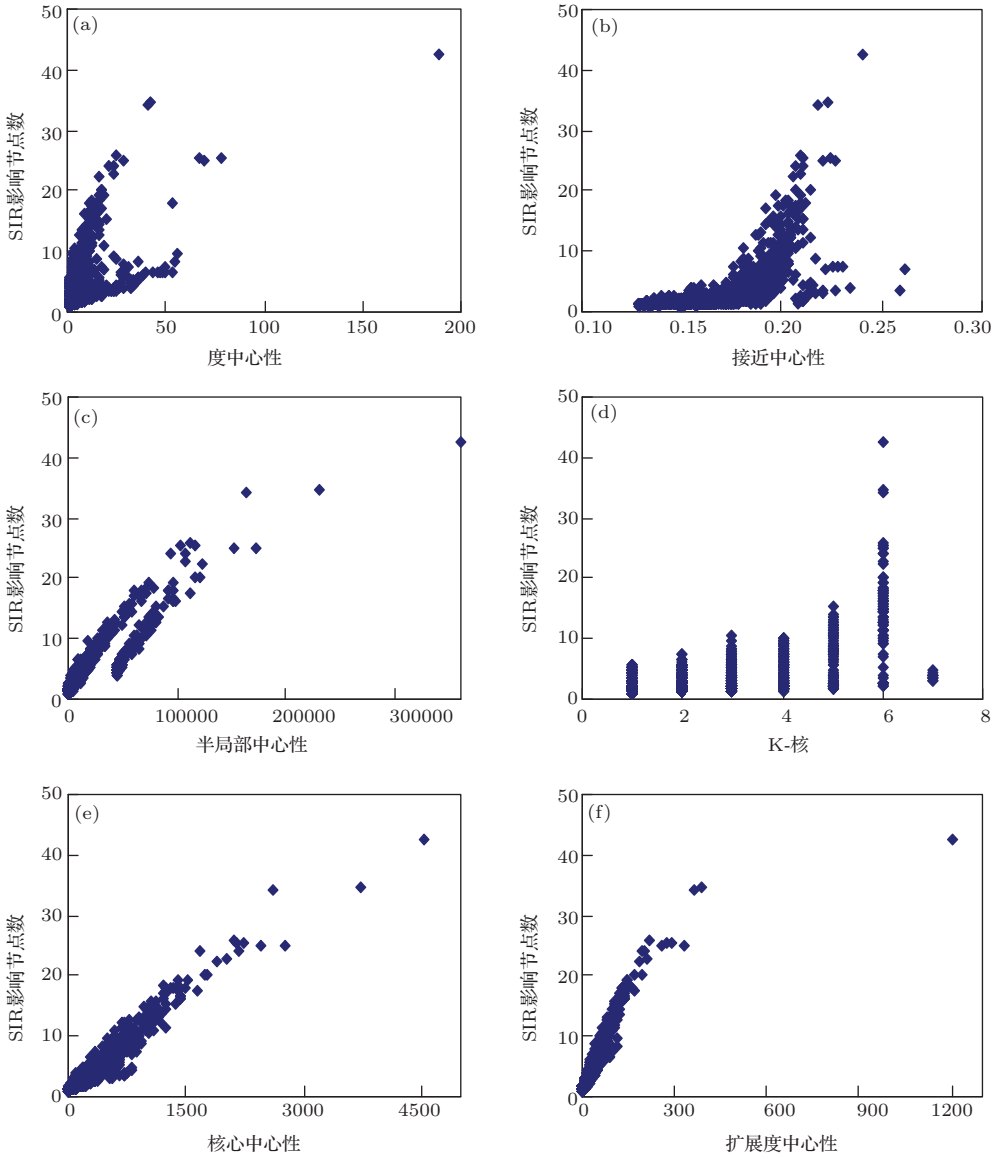

图 3 Blogs 数据集上不同指标评估值与 SIR 影响节点数的相关性 (a) 度中心性; (b) 接近中心性; (c) 半局部中心性; (d) K-核; (e) 核心中心性; (f) 扩展度中心性

Fig. 3. Correlation of evaluation values and number of SIR infected nodes for Blogs: (a) Degree; (b) Closeness; (c) SemiLocal; (d) K-Shell; (e) Coreness; (f) ExDegree.

上面实验设置的传播率是固定的, 只能反映一个静态的状态. 为了动态地分析各指标评估的准确性随传播率的变化, 我们顺序取 0.01 至 0.15 之间, 间隔为 0.01 的多个传播率, 并且将 tau 值作为准确性度量值进行实验. 实验在 5 个真实数据集和 LFR 模拟数据集上进行, 结果分别如图 4 和图 5 所示.

真实数据集上的实验结果如图 4 所示, 可见除扩展度中心性外的其他所有指标的准确性对传播率的变化都较为敏感. 其中, 度中心性在传播率较小时的准确率较高, 这主要是因为当传播率较小时, 源节点的影响力只能覆盖较小范围, 而度值正好适合这一情况. 另外, 由于考虑了更广范围的拓扑信息, 接近中心性、核心中心性和半局部中心性指标则在传播率适中或更大一些时效果较为理想. 而对于这些范围之外的传播率, 则会导致主要被影响的节点与各指标评估值的主要贡献节点在覆盖范围及权重上的差异性更大, 这可能会使评估准确性降低.

与对传播率敏感的评估指标不同, 由于考虑了传播过程中的具体传播率, 扩展度中心性在不同传播率下的评估准确性不会出现太大起伏, 并且准确性通常高于其他指标. 传播率较小时, 结果与度中心性较为接近, 当传播率逐渐增大时, 准确性逐渐过渡到与半局部中心性和核心中心性相当甚至更优.

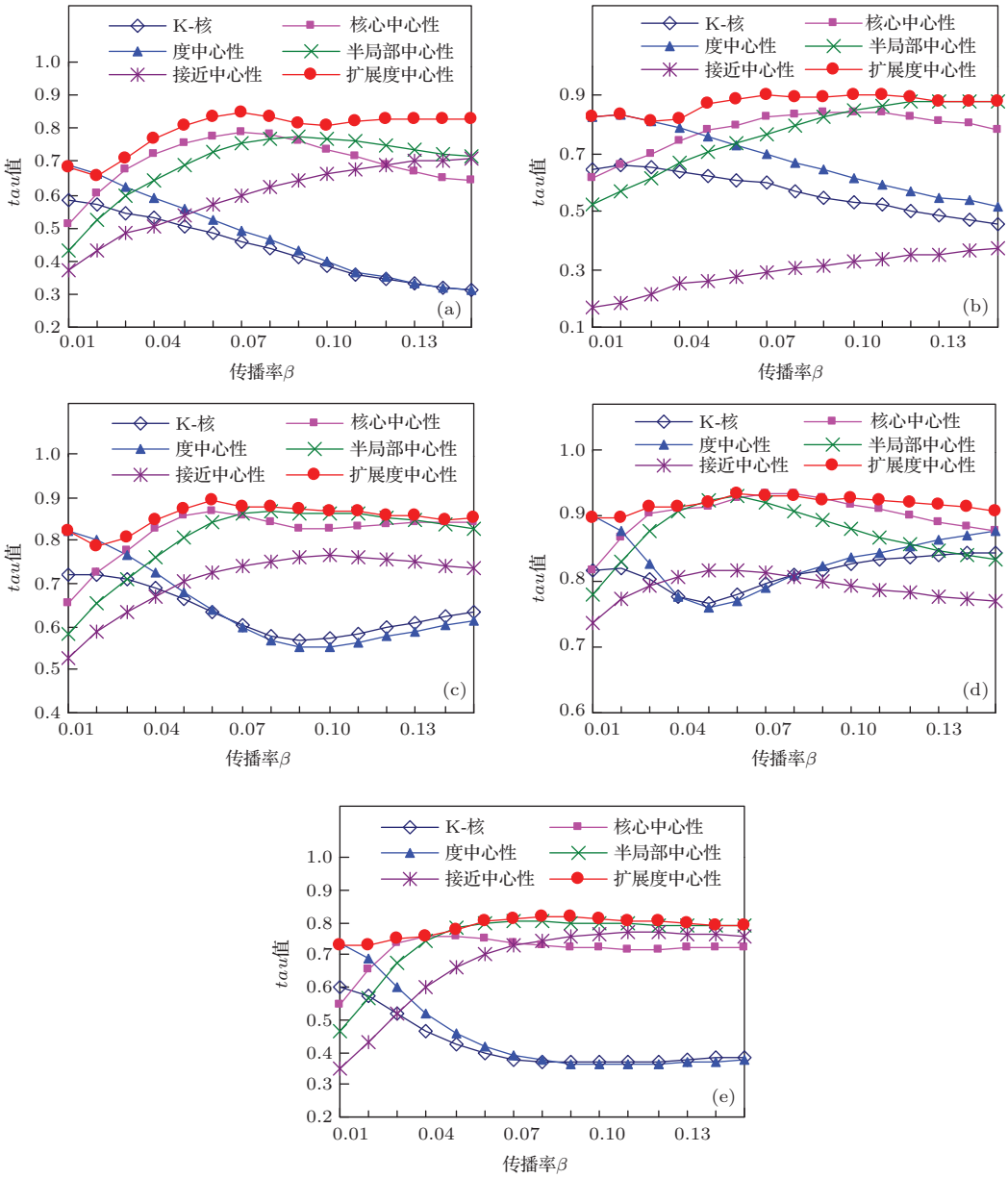

图 4 (网刊彩色) 五组真实数据集上各指标评估准确性的对比 (a) Bolgs 数据集; (b) Netscience 数据集; (c) PPI 数据集; (d) E-mail 数据集; (e) PGP 数据集

Fig. 4. (color online) Comparison of accuracy evaluation among various centralities for five real datasets: (a) Blogs; (b) Netscience; (c) PPI; (d) E-mail; (e) PGP.

LFR 模拟数据集上的实验结果如图 5 所示, 与真实数据集上的结果类似, 无论传播率如何, 扩展度中心性均能取得较优的结果. 按照传统方法, 若希望在某一特定传播率下得到较优的评估结果, 则必须找到针对这一传播率的合适指标. 虽然一般情况下可以将传播阈值 $\beta_{\rm th}$ 作为大致分界点, 比如传播率小于阈值时选择度中心性, 略大于阈值时选择核心中心性或半局部中心性指标. 但阈值计算只涉及节点度值和度值平方的平均值, 而各种实际传播中可能有其他参数未被考虑, 这会导致难以估计的误差, 例如实验中设置的恢复率就使得传播阈值与实际最优指标的分界点存在较大出入. 图 5 中, 三组数据的传播阈值分别为 0.087, 0.067, 0.051, 但实际分界点明显小于这些值. 此外, 传统指标评估的准确性也并非随传播率单调变化, 而且变化的趋势也各不相同, 因此分界点可能并不惟一, 例如 $\langle k\rangle = 15$ 的数据中, 当传播率在 $0.01 \leqslant \beta \leqslant 0.02$ 和 $0.11 \leqslant \beta \leqslant 0.15$ 两个区间时, 度中心性都优于其他指标, 存在两个分界点. 这些都说明, 通过传统方法很难在不同传播率下均取得较为准确的结果. 而从实验可以看出, 扩展度中心性则不存在这样的问题, 基本能够在不同传播率下都取得较为稳定和准

确的评估结果.

除了对所有节点的传播影响力分析外, 本文还对最具影响力的节点进行了实验. LFR 实验数据中, 设置 $\langle k\rangle = 10$, 选择各指标中评估值最大的 10 个节点分别作为源节点, 记录在 SIR 仿真中处于不同时刻的平均累积影响节点数作为测量值, 然后取平均. 针对不同传播率的对比结果如图 6 所示. 从图 6 可看出, 根据扩展度中心性和最优指标得到的这些节点在表现上较为接近, 这说明扩展度中心性同样适合于确定最具影响力的节点.

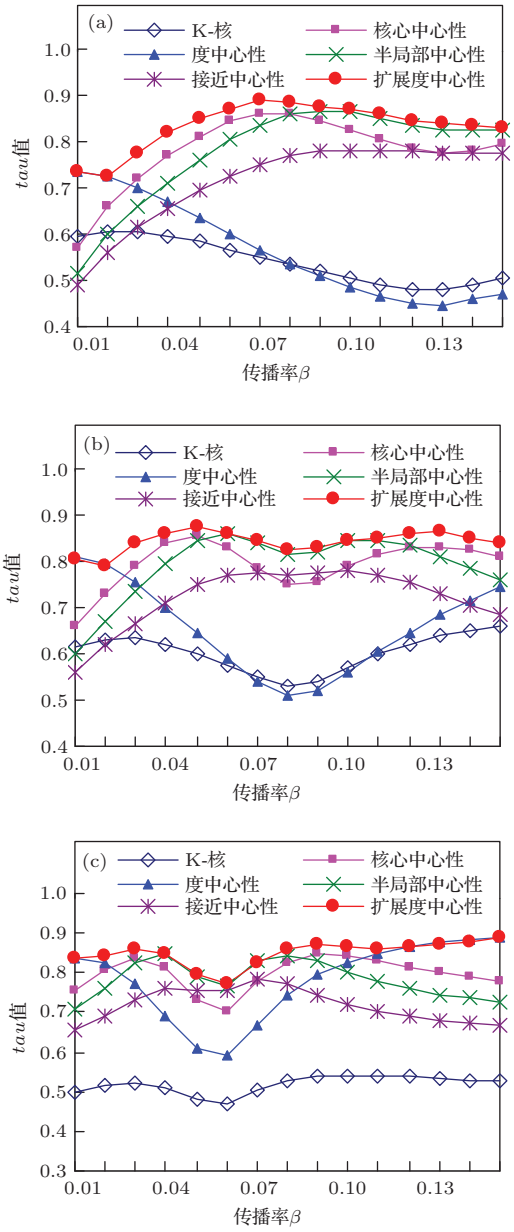

图 5 (网刊彩色) 三组 LFR 数据集上各指标评估准确性的对比 (a) $\langle k\rangle = 5$; (b) $\langle k\rangle = 10$; (c) $\langle k\rangle = 15$

Fig. 5. (color online) Comparison of accuracy evaluation among various centralities for three LFR datasets: (a) $\langle k\rangle = 5$; (b) $\langle k\rangle = 10$; (c) $\langle k\rangle = 15$.

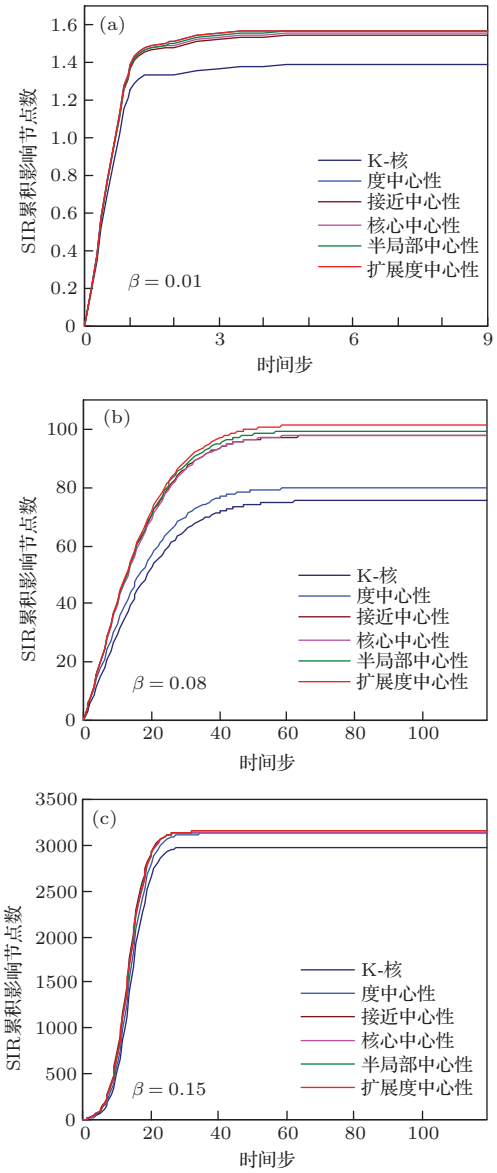

图 6 (网刊彩色) SIR 累积影响节点数随时间的变化, 对于每个评估指标均选择影响力最大的 10 个节点进行实验 (a) $\beta = 0.01$; (b) $\beta = 0.08$; (c) $\beta = 0.15$

Fig. 6. (color online) Variation of cumulative number of infected nodes with time. The most influential 10 nodes for each centralities are selected. (a) $\beta = 0.01$; (b) $\beta = 0.08$; (c) $\beta = 0.15$.

4.3 抽样准确性实验及参数分析

抽样方法的准确性直接决定了传播影响力评估的有效性, 因此这里对其随样本节点数及传播率变化的情况进行了实验. 另外, 也对扩展度算法的相关参数进行了实验分析. 实验所用的 LFR 数据分别取 $\langle k\rangle = 5$, 10 和 15.

抽样方法的准确性由最优扩展层次的误差决定, 最终反映为 tau 值的误差, 因此我们用如下两个指标衡量抽样的准确性: 1) 实际最优扩展层次数与

抽样得到的最优扩展层次数之间的差异性, 表示为 $Diff_{level} = |level_{\text{realbest}} - level_{\text{best}}|$; 2) 实际最优扩展层次对应的 tau 值与抽样方法得到的 tau 值之间的差异性, 表示为 $Diff_{tau} = tau_{\text{realbest}} - tau_{\text{best}}$. 其中, $level_{\text{realbest}}$ 和 tau_{realbest} 分别表示实际最优扩展层次数及其对应的 tau 值; $level_{\text{best}}$ 和 tau_{best} 分别表示通过抽样方法得到的最优扩展层次数及其 tau 值. 对于这两个指标, 值越小则说明抽样准确性越高.

抽样准确性随样本节点数的变化如图 7 所示, 这里取传播率 $\beta = 0.08$. 可以看出, 随着样本节点数的增加, 准确性呈现出提高的趋势, 这与直观逻辑相符. 当样本节点数达到 50 时, tau 值的误差就已经达到非常小的程度, 而这正是本文其他实验中所采用的样本个数. 图 8 描述了抽样准确性随传播率的变化, 从中我们几乎看不出什么规律性, 因此本文未对其做出任何规律性假设.

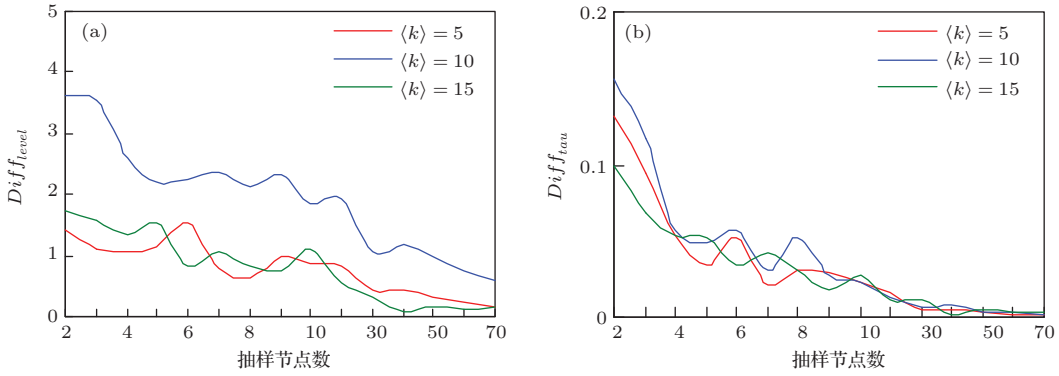

图 7 (网刊彩色) 抽样准确性随抽样节点数的变化 (a) $Diff_{level}$; (b) $Diff_{tau}$

Fig. 7. (color online) Variation of sampling accuracy with different number of sampling nodes: (a) $Diff_{level}$; (b) $Diff_{tau}$.

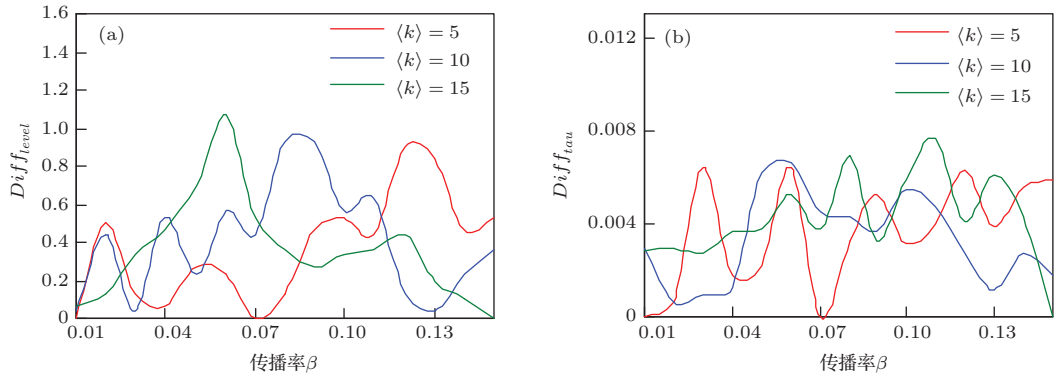

图 8 (网刊彩色) 抽样准确性随传播率的变化 (a) $Diff_{level}$; (b) $Diff_{tau}$

Fig. 8. (color online) Variation of sampling accuracy with different spreading possibilities: (a) $Diff_{level}$; (b) $Diff_{tau}$.

以上抽样准确性的结果为多次重复测试取平均得到, 只能反映总体趋势, 而不能对单次实验的稳定性给予说明. 下面取抽样节点数为 50, 并独立实验 30 次, 得到它们的最优扩展层次数及对应 tau 值随传播率的变化趋势, 同时也给出实际最优扩展层次数及对应 tau 值的曲线, 以方便做稳定性分析. 实验在 $\langle k\rangle = 10$ 的 LFR 数据上进行, 结果见图 9. 可以看到, 就最优扩展层次数而言, 通过抽样得到的结果与实际值的偏差, 基本浮动在一个层次左右, 而反映在 tau 值上的偏差则更小, 这说明在抽样节点数为 50 时, 抽样方法就已经可以达到较稳定的程度.

在扩展度算法的原理描述部分, 我们对最优扩展层次数随传播率增大所可能出现的变化趋势进行了理论分析, 但其中存在一些推测因素, 因此通过实验对这种分析进行验证. 同时, 也将扩展度值作为评估指标, 对评估准确性随扩展层次数变化的情况进行了实验. 不同扩展层次下的准确性实验结果如图 10 所示, 随着扩展层次的增大, 准确性先上升, 到达极值后开始缓慢下降. 这是因为, 初始时更大的扩展层次数会使得扩展度的覆盖范围与实际传播的影响范围更接近, 到达极值后这种接近程

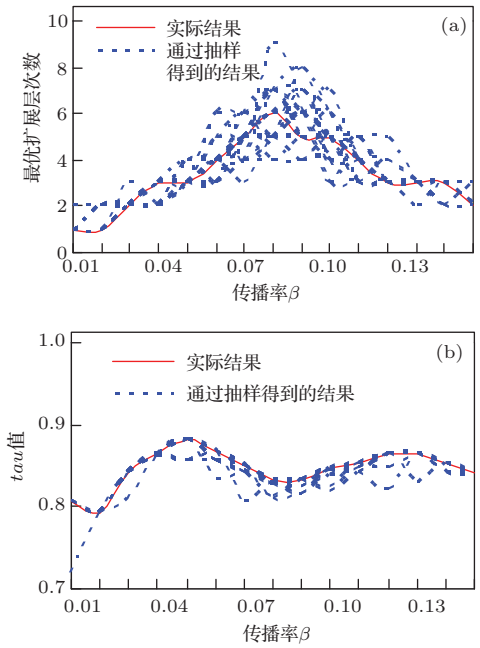

图 9 (网刊彩色) 独立抽样实验 30 次得到的结果 (a) 最优扩展层次数; (b) 通过最优扩展层次数得到的 *tau* 值

Fig. 9. (color online) Results from 30 independent sampling experiments: (a) optimal extension levels; (b) values of *tau* obtained from optimal extension levels.

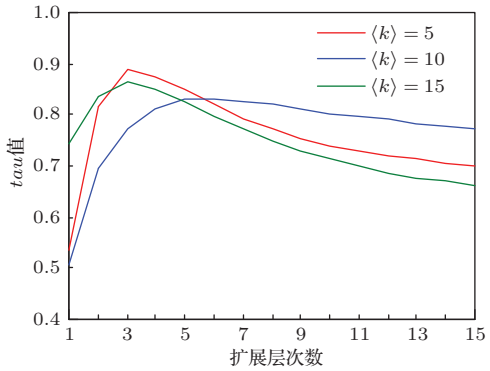

图 10 (网刊彩色) 扩展度值作为评估值时, 准确性随扩展层次数变化的趋势 ($\beta = 0.08$)

Fig. 10. (color online) Variation of accuracy with different extension levels when extension degree is taken as evaluation value ($\beta = 0.08$).

度会逐渐减小. 而下降比上升缓慢, 原因是初始时层次数比较小, 单位层次数的变化会带来评估值较为剧烈的变化, 层次数逐渐增大后则反之. 最优扩展层次数随传播率变化趋势的实验结果见图 11, 随着传播率的增大, 最优扩展层次数先增大然后减小 (在 $\langle k\rangle = 5$ 的数据上, 传播率为 0.15 时其还未减小), 这与前面给出的理论解释相符合. 另外, 不同传播率下, 最优扩展层次数取得最大值的情况, 应该近似地发生在扩展度的覆盖范围到达网络边界时, 而此时这个最大值应该与源节点到网络边界的距离近似, 从图 11 可以看出这个峰值在 5, 6 附近, 正好与六度分割理论相符合.

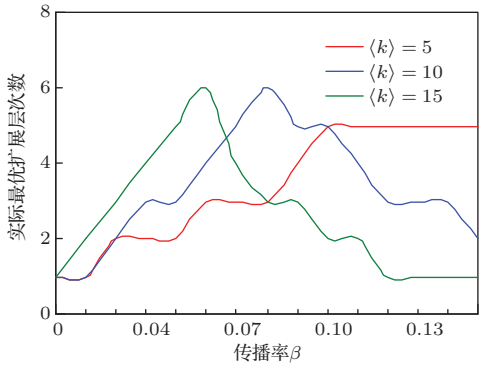

图 11 (网刊彩色) 实际最优扩展层次数随传播率变化的趋势

Fig. 11. (color online) Variation of real optimal extension level with different spreading possibilities.

5 结 论

本文提出了一种基于扩展度的节点传播影响力评估算法, 算法利用逐层叠加的方式对节点度的覆盖范围进行了扩展, 并通过抽样测试确定了适合于特定传播率的扩展层次数. 通过该算法得到的扩展度中心性指标解决了传统评估指标对传播率敏感的问题, 具有较高的稳定性, 且能够取得较为准确的评估结果.

算法中, 对节点度覆盖范围进行扩展的时间复杂度较低, 能取得与网络规模呈线性的关系, 但是通过抽样测试确定最优扩展层次数的过程依然比较耗时. 因此, 在今后的工作中我们将对扩展层次与传播率和网络结构的关系进行更加深入的研究, 用更为准确和省时的方式确定最优扩展层次数, 进一步降低算法的时间复杂度, 使其能更加适合于实际应用.

参考文献

[1] Zhang W, Bai S Y, Jin R 2014 *Int. J. Mod. Phys. B* **28** 1450136
[2] Newman M E J 2003 *SIAM Rev.* **45** 167
[3] Albert R, Barabasi A L 2002 *Rev. Mod. Phys.* **74** 47
[4] Wu Y, Hu Y, He X H, Deng K 2014 *Chin. Phys. B* **23** 060101
[5] Balthrop J, Forrest S, Newman M E J, Williamson M M 2004 *Science* **304** 527
[6] Li K Z, Xu Z P, Zhu G H, Ding Y 2014 *Chin. Phys. B* **23** 118904
[7] Freeman L C 1978–1979 *Soc. Networks* **1** 215
[8] Chen D B, Lu L Y, Shang M S, Zhang Y C, Zhou T 2012 *Physica A* **391** 1777

[9] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nat. Phys.* **6** 888
[10] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E 2007 *Proc. Natl. Acad. Sci. USA* **104** 11150
[11] Bae J, Kim S 2014 *Physica A* **395** 549
[12] Gao S, Ma J, Chen Z M, Wang G H, Xing C M 2014 *Physica A* **403** 130
[13] Du Y X, Gao C, Hu Y, Mahadevan S, Deng Y 2014 *Physica A* **399** 57
[14] Ren Z M, Liu J G, Shao F, Hu Z L, Guo Q 2013 *Acta Phys. Sin.* **62** 108902 (in Chinese) [任卓明, 刘建国, 邵凤, 胡兆龙, 郭强 2013 物理学报 **62** 108902]
[15] Ren X L, Lü L Y 2014 *Chin. Sci. Bul.* **59** 1175 (in Chinese) [任晓龙, 吕琳媛 2014 科学通报 **59** 1175]
[16] Zeng A, Zhang C J 2013 *Phys. Lett. A* **377** 1031
[17] Liu Y, Tang M, Zhou T, Do Y 2014 arXiv:1409.5187v1 [physics. soc-ph]
[18] Wang W, Tang M, Yang H, Do Y, Lai Y C, Lee G W 2014 *Sci. Rep.* **4** 5097
[19] Wang W, Tang M, Zhang H F, Gao H, Do Y, Liu Z H 2014 *Phys. Rev. E* **90** 042803
[20] Newman M E J 2002 *Phys. Rev. E* **66** 016128
[21] Pastor-Satorras R, Vespignani A 2001 *Phys. Rev. Lett.* **86** 3200
[22] Kendall M G 1938 *Biometrika* **30** 81
[23] Hu Q C, Yin Y S, Ma P F, Gao Y, Zhang Y, Xing C X 2013 *Acta Phys. Sin.* **62** 140101 (in Chinese) [胡庆成, 尹龑燊, 马鹏斐, 高旸, 张勇, 邢春晓 2013 物理学报 **62** 140101]
[24] Xie N 2006 *M. S. Dissertation* (Bristol: University of Bristol)
[25] Newman M E J 2006 *Phys. Rev. E* **74** 036104
[26] Palla G, Derenyi I, Farkas I, Vicsek T 2005 *Nature* **435** 814
[27] Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A 2003 *Phys. Rev. E* **68** 065103
[28] Boguna M, Pastor-Satorras R, Diaz-Guilera A, Arenas A 2004 *Phys. Rev. E* **70** 056122
[29] Castellano C, Pastor-Satorras R 2010 *Phys. Rev. Lett.* **105** 218701
[30] Lancichinetti A, Fortunato S, Radicchi F 2008 *Phys. Rev. E* **78** 046110

Evaluating influential spreaders in complex networks by extension of degree*

Min Lei Liu Zhi Tang Xiang-Yang Chen Mao† Liu San-Ya

(*National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China*)

(Received 4 September 2014; revised manuscript received 17 November 2014)

Abstract

Evaluating influential spreaders in networks is of great significance for promoting the dissemination of beneficial information or inhibiting the spreading of harmful information. Currently, there are some central indices that can be used to evaluate spreading influence of nodes. However, most of them ignore the spreading probability and take into consideration only the network topology or the location of source node, so the excellent results can be achieved only when the spreading probability is in a specified range. For example, the degree centrality is appropriate for a minor spreading probability, but to ensure the accuracy, semi-local and closeness centralities are more suitable for a slightly larger one. To solve the sensitivity problem of spreading probability, a novel algorithm is proposed based on the extension of degree. In this algorithm, the coverage area of degree is recursively extended by the overlapping of degree of neighbors, which makes different extension levels correspond to different spreading probabilities. For a certain spreading probability, the proper level index is calculated by finding the most correlate ranking sequences of sampling nodes, which is obtained by matching the results of different spreading levels and SIR simulation. In this paper, the relationship between extension level and spreading probability is explained by the theory of fitting the weight and infected possibility of nodes, and the feasibility of the sampling method is verified by the computational experiments. The experimental results on both real and computer-generated datasets show that the proposed algorithm can effectively evaluate the spreading influences of nodes under different spreading probabilities, and the performance is close or even superior to that evaluated by using other central indices.

Keywords: complex network, spread influence, extension degree

PACS: 89.75.Hc, 89.75.Fb **DOI:** 10.7498/aps.64.088901

* Project supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2013BAH72B01), the Program for New Century Excellent Talents in University of Ministry of Education of China (Grant No. NCET-11-0654), and the Scientific Research Foundation of Ministry of Education of China and China Mobile Limited (Grant No. MCM20121061).

† Corresponding author. E-mail: eitecpaper@gmail.com