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

利用邻域``结构洞"寻找社会网络中最具影响力节点

苏晓萍　宋玉蓉

Leveraging neighborhood ``structural holes" to identifying key spreaders in social networks

Su Xiao-Ping　Song Yu-Rong

引用信息 Citation: Acta Physica Sinica, 64, 020101 (2015)　DOI: 10.7498/aps.64.020101
在线阅读 View online: http://dx.doi.org/10.7498/aps.64.020101
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2015/V64/I2

利用邻域“结构洞”寻找社会网络中最具影响力节点*

苏晓萍[1)] 宋玉蓉[2)†]

1) (南京工业职业技术学院计算机与软件学院, 南京 210046)

2) (南京邮电大学自动化学院, 南京 210003)

(2014 年 5 月 19 日收到; 2014 年 9 月 9 日收到修改稿)

识别复杂网络中的关键节点对网络结构优化和鲁棒性增强具有十分重要的意义. 经典的关键节点测量方法在一定程度上能够辨识网络中影响力节点, 但存在一定局限性: 局部中心性测量方法仅考虑节点邻居的数目, 忽略了邻居间的拓扑关系, 不能在计算中反映邻居节点间的相互作用; 全局测量方法则由于算法本身的复杂性而不能应用于大规模社会网络的分析, 另外, 经典的关键节点测量方法也没有考虑社会网络特有的社区特征. 为高效、准确地辨识具有社区结构的社会网络中最具影响力节点, 提出了一种基于节点及其邻域结构洞的局部中心性测量方法, 该方法综合考虑了节点的邻居数量及其与邻居间的拓扑结构, 在节点约束系数的计算中同时体现了节点的度属性和“桥接”属性. 利用 SIR(易感- 感染-免疫) 模型在真实社会网络数据上对节点传播能力进行评价后发现, 所提方法可以准确地评价节点的传播能力且具有强的鲁棒性.

关键词: 复杂网络, 结构洞, 社团结构, 节点中心性测量

PACS: 01.75.+m, 89.75.Hc, 89.75.Fb **DOI:** 10.7498/aps.64.020101

1 引 言

复杂的交互系统通常可以用复杂网络来建模与刻画: 顶点代表系统组件而边则表示组件间的关系与相互作用. 对网络中节点重要性的评价与度量对提高系统鲁棒性[1]、抗毁性[2]意义重大, 也是加快信息传播、谣言与病毒抑制[3], Web 页面排序[4]以及生物医药学中的药物标靶发现[5,6]等不同学科的重要研究内容. 经典的基于节点属性和网络位置的关键节点测量方法[7,8]包括: 度中心性[9]、介数中心性 (Betweenness centrality)[10]、紧密度指标 (closeness centrality)[11], K-Shell (KS 指标)[12]. 基于节点属性和网络位置的关键节点测量方法可分为网络局部属性、网络全局属性两类, 基于网络局部属性的度量方法是度中心性, 它仅考虑节点自身信息和其邻居信息, 具有计算简单、时间复杂度低的特点, Albert 等[9]利用度中心性寻找社会网络中最具影响力的节点, 他们指出: 在异质的无标度网络中度大的 Hubs 节点传播影响力也大. 然而, 度中心性度量方法仅考虑了节点邻居的数目却忽略了邻居间的拓扑关系, 也没有考虑节点在网络中的位置, 因此不能在计算中反映邻居节点间的相互作用, 计算结果不够准确[12]. Chen 等[13,14]提出半局部中心性方法辨识网络中最具影响力节点: 定义节点的度和其邻居的度之和为节点的中心性值[13], 该方法不但考虑节点的度还考虑了节点的邻域信息, 但是没有考虑邻居之间的拓扑结构. 为了在评价指标中考虑邻居节点间的拓扑关系, 他们又提出利用聚类系数和度两个参数联合评价节点的传播能力[14], 发现度相同的节点其传播能力随聚类系数的增加而下降, 即节点的聚类系数与节点重要性

* 国家自然科学基金 (批准号: 61373136, 61103051)、教育部人文社会科学研究项目 (批准号: 12YJAZH120) 和南京工业职业技术学院重大项目 (批准号: Yk13-02-03) 资助的课题.

† 通信作者. E-mail: songyr@njupt.edu.cn

之间具有负相关性. 这一研究也间接证明节点的中心性不但与节点在网络中的位置有关, 还与网络的群聚特性有着密切联系. 胡庆成等[15]认为节点的影响力不只是由节点内部属性决定, 节点所在社区的大小以及社区内节点连接紧密程度也是决定节点中心性的关键因素. 他们通过社区发现算法将节点划分到不同社区, 利用节点在社区中与其余节点连接的紧密程度和它的KS值联合判断节点的重要程度, 该算法需要调整两个参数. 而Cheng等[16]则从具有社区结构的网络中边的重要性出发, 提出采用确定连接不同社区间的"桥接"边的方法, 寻找传播中的关键路径, 认为与这些关键路径连接的节点就是传播中的关键节点. 该算法仅使用网络的局部拓扑信息, 计算量小.

Betweenness[10]利用节点到达网络其余节点的最短路径数目衡量节点的重要性, Betweenness刻画了某一节点在"最短路径传输"原则下对信息的控制能力, 通常介数大的节点位于多个社区的"桥接"处, 对网络的鲁棒性作用显著; 而closeness[10]是通过网络平均路径长来衡量节点的重要性, 它与Betweenness均属于网络全局属性的中心性测量方法. 该类方法需要获得整个网络的拓扑特征, 计算复杂度高且仅对全连通网络有效, 因此不适用于大型社会化网络.

Kitsak等[12]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, 指出度和Betweenness有时不能精确描述一个节点的影响力, 经过K-Shell分解得到的节点的核数更好地刻画了节点的传播能力. 该方法的提出引发了节点中心性测量方法研究的热情, 提出了一系列扩展和改进KS指标的方法[17−19]. Bae等[17]提出利用节点及其邻居的KS值之和度量节点的重要性, 实质上是在K-Shell的基础上考虑节点的度信息使处于网络边缘的度大节点获得正确评价. Liu等[18]则综合考虑了节点自身的KS值及其与网络中最大KS值节点的距离, 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. Zeng和Zhang[19]提出了混合度分解方法(MDD), 在KS的计算中考虑了移除节点的度信息, 获得了较高的影响力排名精度. 任卓明等[20]则对处于网络边缘KS值最小节点的传播能力进行区分与评价. 然而不能回避的一点就是K-Shell方法能够精确评价在单一传播源情形下节点的传播能力, 但是在多传播源情形和网络谣言传播模型下则不适用[21], 该方法同样不能在计算中反映邻居节点间的相互作用和邻居的拓扑结构.

综上所述, 要准确度量节点传播能力不但要考虑节点自身的属性, 还需考虑其与邻域节点间的拓扑关系. 另外, 许多网络存在着明显的群聚特性[1,22], 以往在研究中通常会把网络节点重要性评价与网络社区发现作为两个不同的问题分别加以考虑. 事实上, 最具影响力节点在强社团结构的网络中应具有以下特征: 社区中心和在各社区中起到"桥接"作用的节点. 如何在计算中同时体现社区中心性和社区间"桥接"性是发现具有社团结构的社会网络中最具影响力节点的关键. 已有最新研究[23]利用与节点直接相连的社团的数目(V_c)衡量该节点的传播能力, 经易感- 感染-免疫(SIR)模型验证以V_c值大的节点作为感染源, 感染速度更快且范围更广. 该方法的准确性依赖于社区划分算法的精确程度. 本文提出一种新的局部中心性方法, 用于评价网络节点的重要性. 该方法利用节点及其邻居拓扑结构共同计算结构洞约束值, 以该值作为评价节点重要性的指标, 节点形成结构洞受到的约束越大, 节点在信息传播中越不利. 该方法在计算复杂性与排名精度上取得了很好的平衡.

2 理论与方法

2.1 结构洞理论

结构洞是Burt[24,25]研究社会网络中竞争关系时提出的经典社会学理论, 从社会学角度看, 结构洞是非冗余联系人之间存在的缺口(如图1 (a)中的A, B间没有冗余联系), 由于结构洞的存在, 洞两边的联系人可以带来累加而非重叠的网络收益, 从图1 (a)明显看到, 节点"Ego"充当了中间人角色, 因此获得了较其3个邻居更多的网络收益, 即节点"Ego"在网络中的重要性要大于其他节点. 若B, C间发生联系则将减少"Ego"获得的网络收益. 从复杂网络角度看, 拥有较多结构洞的网络节点更有利于信息的传播.

Burt提出用网络约束系数(network constraint)来衡量网络节点形成结构洞时所受到的约束:

$$C_i = \sum_{j\in\Gamma(i)}\Big(p_{ij}+\sum_{q}p_{iq}p_{qj}\Big)^2 \quad q\neq i,j. \tag{1}$$

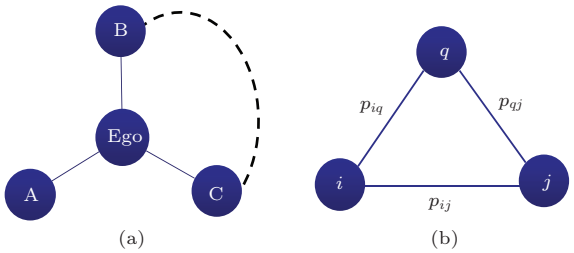

图 1 结构洞概念 (a) 节点 “Ego” 的结构洞; (b) 评价节点 i 对节点 j 投入精力

如图 1 (b) 所示, p_{ij} 表示节点 i 为维持与节点 j 的邻居关系所投入的精力占总精力的比例, p_{iq} 和 p_{qj} 分别是节点 i, j 与共同邻居 q 维持关系投入的精力占其总精力的比例.

$$p_{ij} = z_{ij} \Big/ \sum_{j\in\Gamma(i)} z_{ij},$$

其中

$$z_{ij} = \begin{cases} 1, & i \text{ 到 } j \text{ 有链接}, \\ 0, & i \text{ 到 } j \text{ 没有链接}, \end{cases}$$

p_{iq} 和 p_{qj} 计算方法与 p_{ij} 相似. 根据 (1) 式可知: C_i 的值越小, 形成结构洞所受的约束越小. 根据图 2 计算节点 i 与节点 A 间的约束系数, 已知节点 i 邻居集合 $\Gamma(i) = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$, 因此对于任意邻居, $p_{i\mathrm{A}} = 1/6$ (i 有 6 个邻居, 维持每个邻居所需精力为总精力的 1/6), i 与 A 的共同邻居有 B, E, F. 于是可得:

$$\sum_{q\in\{\mathrm{B},\mathrm{E},\mathrm{F}\}} p_{iq}p_{q\mathrm{A}} = \frac{1}{6}*\frac{1}{3} + \frac{1}{6}*\frac{1}{3} + \frac{1}{6}*\frac{1}{3} = \frac{1}{6},$$

$C_{i\mathrm{A}} = (1/6 + 1/6)^2$. 同理可求得 $C_{i\mathrm{B}}$, $C_{i\mathrm{C}}$, $\cdots$, 求和后可得 C_i. 从 C_i 的计算过程可以看出, C_i 的值能够综合评价节点的邻居数目以及它们之间连接的紧密程度, 节点 i 的度越大, p_{ij} 值越小, 说明度大的 Hubs 节点容易形成结构洞. $\sum_q p_{iq}p_{qj}$ 的值由节点 i, j 的共同邻居 q 的数量决定, i, j, q 连接越紧密, 它们之间形成的闭合三角形越多, $\sum_q p_{iq}p_{qj}$ 值越大, 形成结构洞的机会就越小. 可见 C_i 值的计算综合考虑了节点度和节点邻居拓扑关系信息, 网络约束系数值越大说明该节点邻居数量少且与其邻居间的闭合程度高. 这样的节点由于不易获得新的关系资源使得它在竞争中处于不利地位. 反之, 网络约束系数值越小, 结构洞形成机会就越大, 越有利于获得新的关系资源. 从复杂网络的观点看, 网络约束系数利用了网络局部属性评价节点重要性, 在计算量上有优势, 约束系数小的节点在信息传播中具有较大影响力. 这一观点与最新研究 [26,14] 得到的结论一致: 节点与其邻居间连接紧密不利于信息的传播.

2.2 基于邻域 “结构洞” 的节点中心性评价方法 (N-Burt)

约束系数只衡量了节点与其最近邻节点间的关系, 没有进一步考虑邻居节点与其余节点相连的拓扑结构对该节点的影响, 该指标不能发现一些重要的 “桥接” 节点. 如图 2 所示.

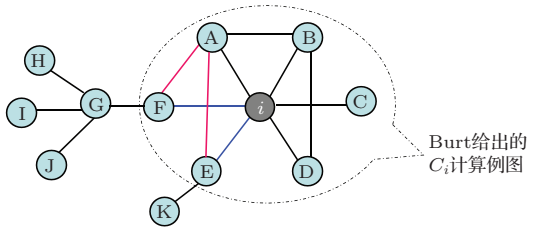

图 2 约束系数计算示意

图 2 虚线中的部分为 Burt 在文献 [24] 中用于说明节点 i 与邻居间关系的例图, 节点 E 和 F 与节点 i 有共同邻居 A, A 的度为 4, 于是根据约束系数的定义可得: $C_{i\mathrm{E}} = C_{i\mathrm{F}} = (1/6 + 1/6 * 1/4)^2 = 0.0434$, 从节点 i 的角度看, 节点 E 和 F 具有同等重要的地位, 若 E, F 分别有两个不与节点 i 相连的邻居 K 和 G 时, 情况显然与约束系数计算不同: 由于 F 有较 E 更好的关系, 对于 i 来说, 保持与 F 的关系应比保持与 E 的关系要付出更多精力. 这一事实在 Burt 的约束系数的计算中无法体现, 因为 E 和 F 与节点 i 有共同邻居仍然只有 A. 因此, 需要改进网络约束系数的计算方法, 更精确地衡量节点在网络中的地位.

设有无向网络 $G = (V, E)$, $V = \{v_1, v_2, v_3, \cdots, v_n\}$ 是网络中节点的集合, $|V| = n$, $E = \{e_1, e_2, e_3, \cdots, e_m\} \subseteq V \times V$ 是边的集合, $|E| = m$. 节点 i 的度可以表示为: $k(i) = \sum_{j\in G} a_{ij}$, 其中

$$a_{ij} = \begin{cases} 1, & \text{若 } i \text{ 与 } j \text{ 有边相连}, \\ 0, & \text{若 } i \text{ 与} j\text{没有边相连}, \end{cases}$$

节点 i 的邻接度被定义为 $Q(i) = \sum_{w\in\Gamma(i)} k(w)$, 其中 $\Gamma(i)$ 为节点 i 的邻居的集合. 定义:

$$p_{ij} = \frac{Q(j)}{\sum\limits_{v\in\Gamma(i)} Q(v)}. \tag{2}$$

节点i的约束系数仍表示为(1)式. (2)式改进了为维持与其邻居的关系而投入的精力占比的计算方法,使约束系数能够更真实地反映节点对其邻居投入的精力. 仍以图2中节点i为例, 在新定义下: $C_{i\mathrm{E}} = (p_{i\mathrm{E}} + p_{i\mathrm{A}} * p_{\mathrm{AE}})^2 = 0.0444$, $C_{i\mathrm{F}} = (p_{i\mathrm{F}} + p_{i\mathrm{A}} * p_{\mathrm{AF}})^2 = 0.0719$, 发现节点$i$在总精力不变的情况下, 向节点F投入了比节点E更多的精力, 从图中看到节点F具有比节点E更"好"的关系, 节点i有理由花更多精力维持与F的联系将获得更多网络回报: 以节点i为信息源, 通过节点F能够传播更远. (2)式的改进使对节点向其邻居投入精力的计算由原来的简单平均变为引入二次邻居的拓扑信息, 从而能更精确地评价节点在网络中的重要程度. 在改进后的节点约束系数计算中同时体现了节点的度信息和节点与其邻居拓扑结构的信息, 是解决具有社区结构的社会网络的有效方法. 该方法不但能够正确评价节点在网络中的地位, 还可以为链路预测以及在社会关系中找到重要联系人提供帮助.

3 仿真与结果分析

3.1 节点传播能力评价

进行节点影响力分析时, 很多学者采用SIR模型来模拟信息、病毒的传播过程, 这是由于SIR模型可以映射为边渗流过程, 运用这一数学工具能够很好地理解传播过程并对传播过程进行理论分析得到精确解. 近年来, 学者将SIR模型进行了一系列的改进、推广, 使之能够更接近真实传播规律, 并在带权图、有向图上使用. 这些分析与研究得到了许多关于传播特征的最新结论[27−29].

基于以上分析, 采用SIR[30]模型能够正确评价节点的传播能力. 一般地, SIR模型将网络节点划分为3类:

1) 易感(susceptible)节点, 健康但缺乏免疫能力的节点;

2) 感染(infected)节点, 已经感染病毒病具有病毒传播能力的节点;

3) 免疫(recovered)节点, 已治愈并获得免疫能力的节点或者已经死亡、不再对相应动力学行为产生影响的节点. 在SIR模型中, 感染节点i作为传染源, 在单位时间内以概率β向相邻的易感节点进行病毒传播, 每个感染节点则以概率γ治愈或死亡.

采用Kendall相关系数τ[31]评价排序结果的正确性. τ的定义如下:

$$\tau(R_1, R_2) = \frac{N_{\mathrm{c}} - N_{\mathrm{d}}}{\sqrt{(N_t - N_{t1})(N_t - N_{t2})}}, \quad (3)$$

R_1和R_2为拥有N个元素两种不同排名序列; N_{c}为具有一致性排序顺序的元素对数; N_{d}为具有不一致性排序顺序的元素对数;

$$N_t = n(n-1)/2, \quad N_{t1} = \sum_i t_i(t_i - 1)/2,$$

$$N_{t2} = \sum_j t_j(t_j - 1)/2,$$

i, j为R_1和R_2中并列排名位次的数目. 用该指标可以量化两种排名次序的相似性, τ的取值在$[-1, 1]$之间$\tau = 1$为完全正相关, $\tau = -1$为完全负相关, $\tau = 0$为不相关.

3.2 数据集及相关参数

首先使用图2所示的小型网络验证所提算法在小规模数据集上的可行性, 然后选取4个真实社会网络作为算法验证的数据集基于SIR模型进一步与相关算法进行比较. 4个网络分别是: 空手道俱乐部社会关系网karate、小学生面对面接触网PriCon、洛维拉·依维尔基里大学成员邮件通信关系网E-mail、科学家合著关系网(选择其中最大连通子图) Netscience[32], 表1给出了各网络的统计特征.

表1 实验网络统计特性

网络名	节点数n	边数m	平均度$\langle k\rangle$	聚类系数C	平均路径长$\langle d\rangle$
karate	34	78	4.59	0.588	2.4
PriCon	236	5899	49.99	0.502	1.86
Netscience	379	914	4.82	0.798	6.04
E-mail	1133	5451	4.81	0.570	3.606

从以上统计特征看到, 4个网络均具有明显的社区群聚特性和较短的平均路径长度, 在该类社会网络中最具影响力节点应同时具有社区中心属性和在各社区中起到“桥接”作用的特征.

3.3 实验结果与分析

对所提N-Burt方法与Burt方法、Betweenness、度中心性(k)和半局部中心性方法(N-k)进行比较. 之所以选择以上4种方法作为比较, 基准是因为Burt方法、度中心性(k)和半局部中心性方法(N-k)与本文所提N-Burt方法均为基于网络局部属性的度量方法, 该类方法最大特点是不需要获得整个网络的拓扑信息, 计算简单适用于大规模社会网络. 而之所以要将所提N-Burt方法算法与Betweenness方法进行比较是因为研究发现: 节点的介数刻画了某一节点在“最短路径传输”原则下对信息的控制能力, 通常介数大的节点位于多个社区的“桥接”处, 对网络的鲁棒性作用显著. 但是Betweenness方法的最大问题是它属于网络全局属性的中心性测量方法, 计算复杂无法在大规模社会网络或不连通网络中使用. 若能找到与Betweenness方法结果近似、但计算量较小的局部性计算方法将有利于大规模社会网络中重要节点的发现.

4 小规模数据集实验结果分析

分别使用本文所提N-Burt方法、Burt方法、Betweenness、度中心性(k)、半局部中心性方法(N-k)[13]对图2中节点的中心性排序, 排序Top-5结果见表2, 括号中的值是各算法计算结果. N-Burt和Burt关于节点中心性的计算具体见2.1节和2.2节, 度中心性以节点邻居数目$k(i)$作为评价节点中心性的指标, N-k通过计算节点的二阶邻居节点数目作为评价节点中心性的依据, 节点v的半局部中心性值$C_L(v)$被定义为

$$Q(u)=\sum_{w\in\Gamma(u)}N(w),\quad C_L(v)=\sum_{u\in\Gamma(v)}Q(u),$$

其中$\Gamma(u)$为节点u以及它的邻居节点组成的集合, $N(w)$为节点w的邻居以及邻居的邻居数目.

表2 图2中Top-5排序结果

排序方法	Rank1	Rank2	Rank3	Rank4	Rank5
N-Burt	G(0.36)	i(0.44)	F(0.59)	A(0.62)	E(0.67)
Burt	G(0.26)	i(0.37)	E(0.48)	F(0.50)	A(0.54)
k	i(6)	A, G(4)		B, F, E(3)	
N-k	i(116)	A (106)	F(96)	B(88)	E(81)
Betweenness	F(56)	G(54)	i(52)	E(20)	A(17)

在小规模的图2中, 我们容易发现其具有两个社区, 社区的中心点分别是i和G, 度中心性(k)和半局部中心性方法(N-k)考虑度大节点重要性强, 于是认为节点i和A相对重要, N-k方法下排名前5的节点均在同一个社区中, 没有找到另一个社区的中心点, 度中心性方法无法区分度相同节点的传播差异, 因此A和G, B和E, F排名相同. N-Burt, Burt, Betweenness方法均成功地发现社区的“桥接”节点; Burt, N-Burt找出两个社区的中心节点; Betweenness方法对节点的测量仅关注“桥接”特性, 忽视节点在社区中的“中心”地位, 于是节点F排名更靠前; 而Burt, N-Burt方法则由于综合考虑“桥接”特性和“中心性”, 使既有桥接特性也具有社区中心性的节点G, i排名靠前, 从图2可见, 这一排名显然更加合理; 同时, 在N-Burt方法中, 由于更详细地考虑了节点与其邻居拓扑结构间的关系, 使排序更合理: 节点F由于具有更“好”的关系而排名先于节点E, N-Burt方法还充分利用了节点邻域信息, 相较于度中心性方法具有更好的区分度.

5 大规模真实社会网络结果分析

1) 算法排序结果与节点真实传播能力的相关性分析

使用3.1节介绍的SIR模型验证所提方法在大规模社会网上的有效性, 取网络中任意一个节点作为初始传播源, 定义在规定传播时间(t=10)后感染与免疫的节点总数作为该节点的实际传播影响力:

$\overline{S}_i = \dfrac{1}{M}\sum_{m=1}^{M} S_i$, 其中 M 为对节点 i 进行重复实验的次数, 文中取 $M = 100$, 为保证传播能够进行, 取传播率 $\beta \sim \langle k\rangle/\langle k^2\rangle$, 治愈概率为固定值 $\gamma = 0.01$. 根据 $\overline{S}_i$ 值得到节点实际影响力排名, 这一排名体现节点真实的传播能力. 在 4 个真实社会网络数据集上对所提方法与其他各方法得到的排序结果进行相关性分析, 可以获得节点中心性计算方法与真实传播能力间的相关程度, 相关程度越高, 算法对节点中心性测量越准确.

图 3 给出了不同网络中心性评价指标与实际影响力 $\overline{S}_i$ 之间的相关性结果 (限于篇幅仅给出网络规模较大的 Netscience, E-mail 相关性结果图). 发现 N-Burt 和结构洞 Burt 方法与 $\overline{S}_i$ 呈负相关: 随着约束系数的增加节点影响力下降, 这一现象在节点数目较多的 E-mail 网中表现更为明显. 度中心性、半局部中心性、Betweenness 方法则与 $\overline{S}_i$ 呈正相关. Betweenness 方法发现的介数大的 "桥接" 节点数目较少, 绝大多数节点的介数很小, 这一结果符合社会化网络的社区特点. 但是仅靠介数衡量节点的重要性并不准确, 因此介数相近的节点传播能力有较大差异, 所以相关性曲线发散. "度" 越大的节点越重要这一原则依然有效, 充分利用二阶邻域信息有利于提高排名精度, 半局部中心性方法效果好于度中心性方法, 然而另一种可能的问题是当社团结构不平衡时, 排名靠前的节点将集中于同一个社区, 有可能导致信息传播的局部性.

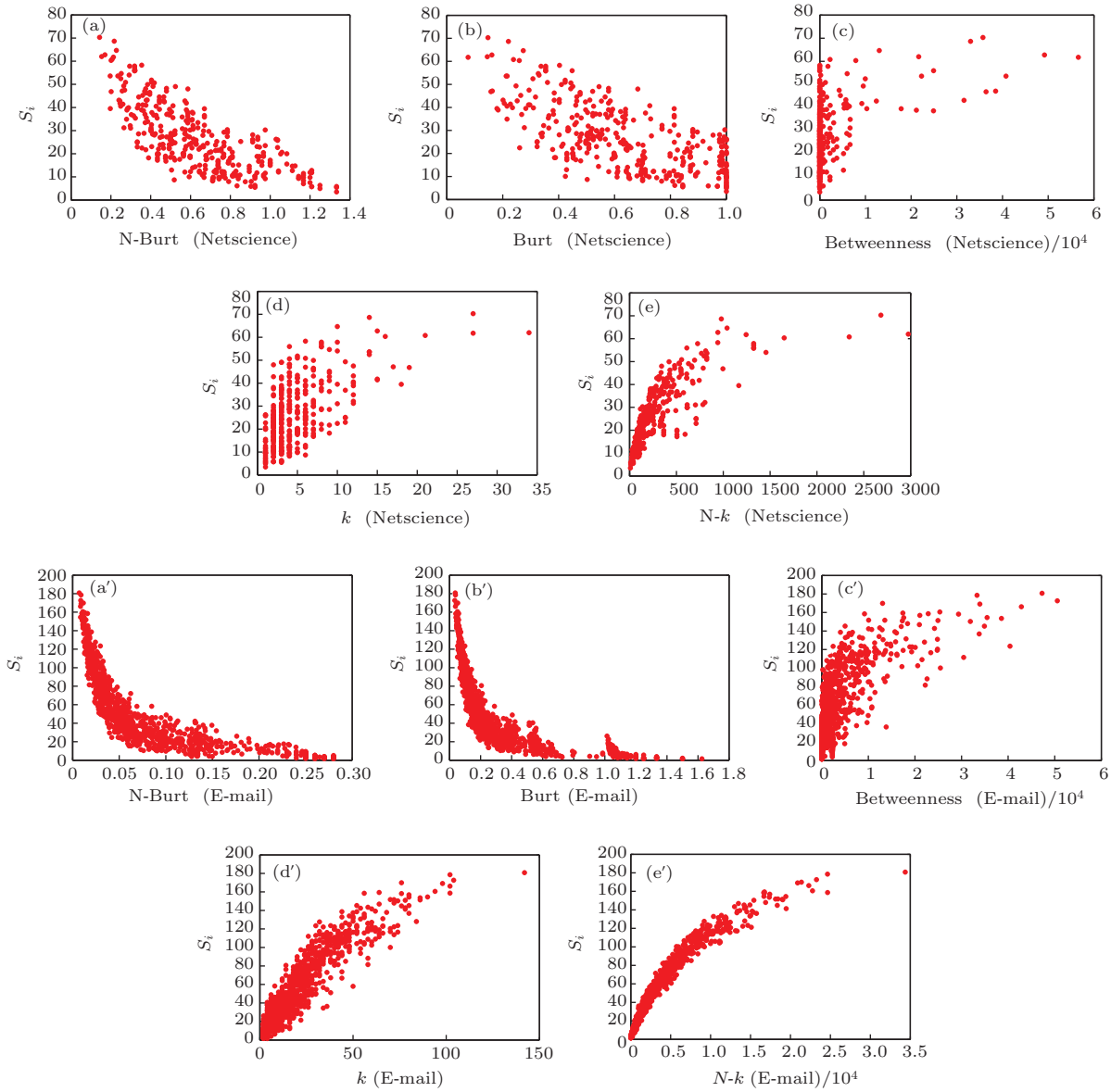

图 3 各方法与实际影响力的相关性分析 (a), (a′) N-Burt; (b), (b′) Burt; (c), (c′) Betweenness; (d), (d′) 度中心性; (e), (e′) N-k ((a)—(e) 为 Netscience 网络的实验结果, (a′)—(e′) 为 E-mail 网络的实验结果)

N-Burt 方法对节点重要性的评价综合考虑了节点度信息和邻居拓扑关系且新的约束系数能够更真实地反映节点对其邻居投入的精力, 因此获得了较其余方法更理想的结果. 表 3 详细说明了算法与实际传播能力的关系. 表 3 的第 1 列给出了传播阈值 $\beta_{\rm th} = \langle k\rangle/\langle k^2\rangle$; 第 2 列 β 值为传播实验中实际选取的传播概率, $\beta > \beta_{\rm th}$ 从而保证传播能够在网络中正常进行, 表 3 的 3—5 列给出节点实际传播能力排名结果与不同中心性计算方法下排名结果间的相关系数值, τ 值越大说明实际传播能力与算法排名越相近, 则算法排名准确度越高. 根据表 3, N-Burt 方法在 4 个网络上的结果均好于 Betweenness、度中心性 (k)、半局部中心性方法 (N-k).

图 4 分析了 N-Burt 方法与 Betweenness 间的关系, 图 4 的横轴为 N-Burt 计算得到的网络各节点约束系数, 纵轴为各节点的介数, 颜色坐标代表 SIR 模型仿真得到的实际影响力. 结果显示, N-Burt 值与介数间存在强的相关性, 说明在结构洞计算方法中不但考虑节点的度信息, 还考虑节点的邻域信息, 所以有能力找到具有“桥接”作用的节点. 同时, 在具有社区结构的社会网络中, 尽管“桥接”节点在社区间的信息传递起到重要作用, 但是社区内的信息传播关键应是社区的 Hubs 节点, 仅靠介数衡量节点的重要性并不准确, 因此从图 4 的结果中可以发现有一些节点尽管其介数大, 但其实际影响力并不很大, 所以尽管 N-Burt 值与介数间存在强的相关性在寻找具有社区结构特点的社会网络中最具影响力节点上, N-Burt 方法较 Betweenness 方法更有优势.

表 3 各算法与实际影响力的 Kendall 相关系数 $|\tau|$

网络	$\beta_{\rm th}$	β	$\tau(\overline{S}_i, k)$	$\tau(\overline{S}_i$, N-$k)$	$\tau(\overline{S}_i$, Betweenness)	$\tau(\overline{S}_i$, Burt)	$\tau(\overline{S}_i$, N-Burt)
karate	0.129	0.15	0.9162	0.9198	0.8378	0.8798	0.9551
PriCon	0.018	0.02	0.9507	0.9761	0.9349	0.9408	0.9913
Netscience	0.125	0.15	0.9130	0.9443	0.8911	0.9704	0.9503
E-mail	0.026	0.05	0.6556	0.6899	0.6919	0.6513	0.6956

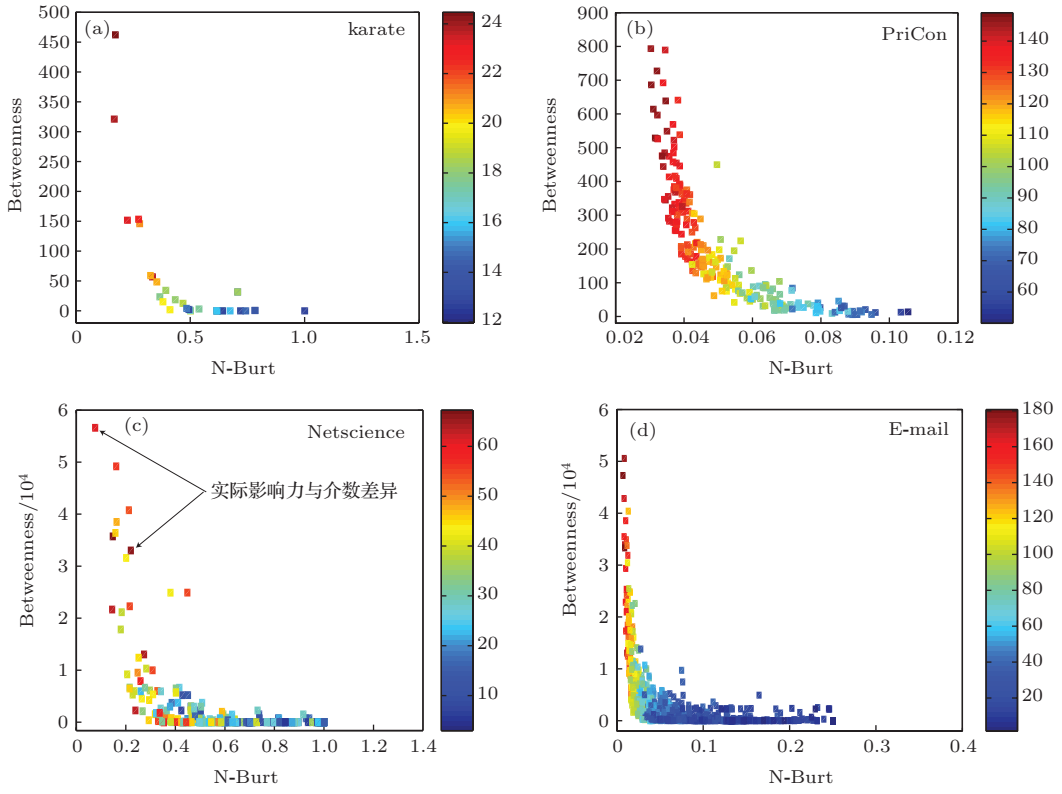

图 4 (网刊彩色) N-Burt 与 Betweenness 传播影响力比较 (a) karate; (b) PriCon; (c) E-mail; (d) Netscience

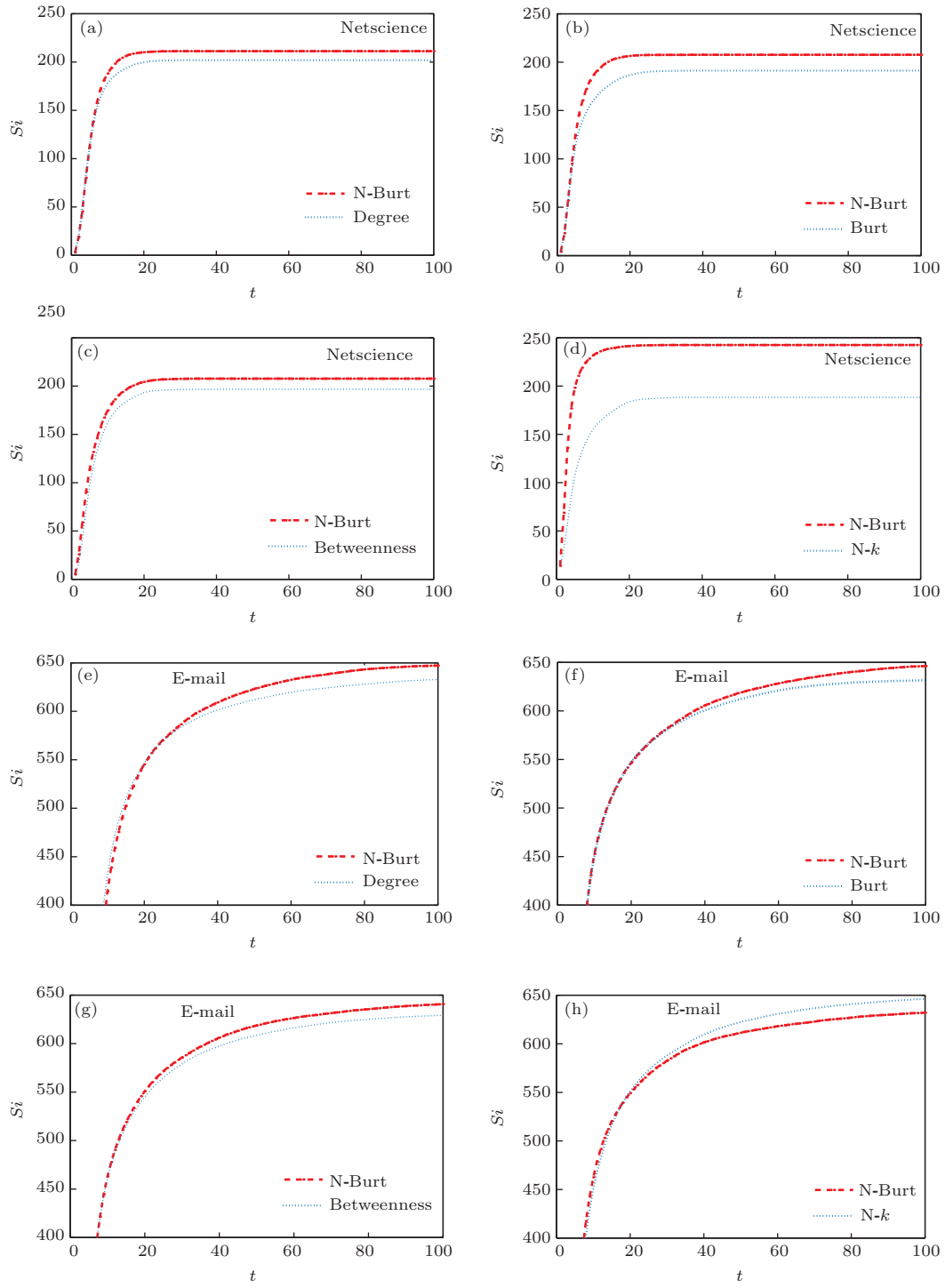

图 5 不同时出现在各算法排名 TOP-20 中的节点传播影响力差异 (a)—(d) Netscience 网络的实验结果; (e)—(h) E-mail 网络的实验结果

2) 各算法排名差异性分析

为进一步清楚地显示各算法对节点中心性测量结果之间的差异, 分别取不同时间出现在各算法排名 TOP-20 中的节点, 基于 SIR 模型进行传播分析. 这些仅出现在一种算法的 TOP-20 中的节点恰好反映了不同算法评价节点中心性侧重点的不同.

取传播时间 $t = 100$ 传播达到稳定状态, 以感染与免疫的节点总数 $\overline{S}_i$ 评价节点的传播影响力. 图 5 给出了这些差异节点传播能力曲线 (限于篇幅仅给出网络规模较大的 Netscience, E-mail 实验结果, 其余两个网络得到相似的结果).

图 5 结果显示: N-Burt 算法排名前 20 的节点与其他算法排名前 20 的节点均有不同, 这些仅被 N-Burt 算法发现的节点传播能力比 Burt 方法、Be-

tweenness、度中心性 (k) 方法发现的差异节点有着更强的传播能力, 仅在 E-mail 网络中差于 N-k 方法, 而 N-k 方法却在 Netscience 网络中表现逊于其余算法, 这说明 N-k 方法对网络拓扑结构敏感.

3) 鲁棒性分析

社会网络上通常存在恶意 Sybil 攻击行为: 用户会通过以虚假身份添加好友获得偏高的社会评价或网络水军利用该方法进行谣言的快速扩散, 这些 Sybil 攻击有可能使节点中心性评价算法得到与事实不符的排序结果. 为评价算法在受到攻击时的鲁棒性 [33], 采用在真实原网络中增加 20% 的虚假节点的方法模拟网络攻击, 对受到攻击的网络进行节点中心性评价并与原网络排名进行比较, $y = x$ 为比较基准, 即受到网络攻击后排名与原排名结果相同, 排名变化越小说明算法的鲁棒性越好, 反之则算法鲁棒性越差. 图 6 给出了在规模较大的 Netscience 网络和 E-mail 网络上的原排名与受攻击后新排名的波动曲线. 实验结果显示: 本文所提算法在受到网络攻击的情况下相较其余算法排名的波动较小, 在 Netscience 网络中, 仅在排名 95 附近有明显波动, 其余波动均很小; 在 E-mail 网络中

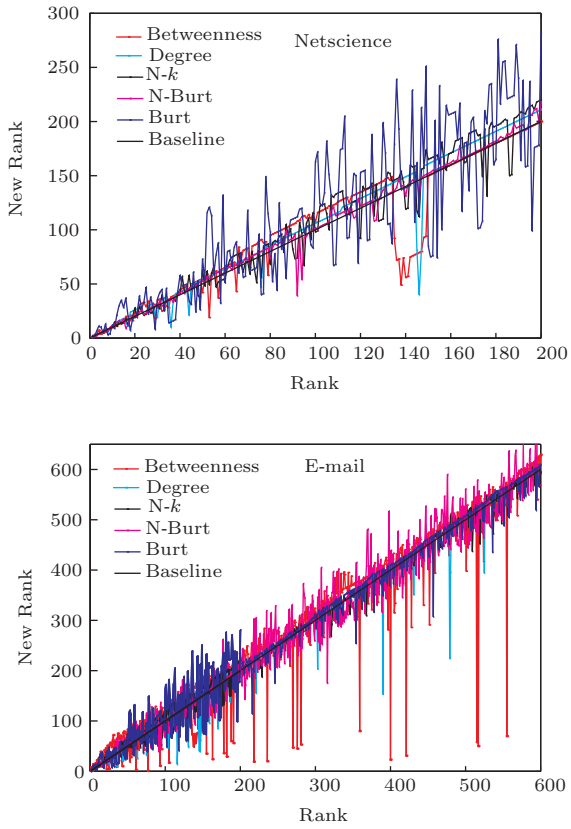

图 6 (网刊彩色) 鲁棒性分析

N-Burt 算法略逊于度中心性和 N-k 方法, 但是比 Betweenness 和 Burt 方法好, 尤其是 Burt 方法在排名靠前的节点附近波动较大, 这使得算法不能对具有较强影响力节点正确评价, 而这些排名靠前节点正是整个网络传播的关键. 综上所述, N-Burt 算法能够抵抗来自虚假用户的攻击, 具有较强的鲁棒性.

6 总结与展望

识别复杂网络中的关键节点是网络相继故障检测和网络攻击与信息传播控制等领域的核心问题. 实践证明, 网络中少数节点对网络有着很强的控制力和影响力, 关键节点挖掘的研究受到来自信息科学、物理以及管理科学等众多学者的关注, 并提出许多解决该问题的方法, 这些方法从不同应用领域、不同角度出发探索不同背景下节点的重要性, 因此各有千秋. 其中基于网络局部属性的关键节点测量方法由于计算简单、时间复杂度低而适用于大规模社会网络的挖掘工作. 以往在研究中通常会把网络节点重要性评价与网络社区发现作为两个不同的问题分别加以考虑. 事实上, 最具影响力节点在具有强社团结构下应具有以下特征: 社区中的 Hubs 和在各社区中起到“桥接”作用的节点. 如何在计算中同时体现中心性和“桥接”性是找到社会网络中最具影响力节点的关键. 基于该思想, 提出了一种新的局部中心性方法评价网络节点的重要性, 该方法在“结构洞”基础上改进了评价节点邻居重要程度的方法, 利用节点及其邻居结构洞约束值作为评价节点重要性的指标. 在该评价方法中同时考虑了节点 Hubs 属性和“桥接”属性, 因此, 在计算复杂性与排名精度上取得了很好的平衡, 其排序效果与 Betweenness 有很强的相关性, 同时相较于 Betweenness 方法, N-Burt 方法能够更准确地评价节点的重要程度且计算仅基于网络局部信息. 实验结果还显示, N-Burt 方法具有较强的鲁棒性, 在面对虚假用户的 Sybil 攻击时依然能够对节点的中心性做出正确的评价.

本文的研究表明, 在社会网络中, 节点的重要性评价同时依赖于节点在社区内的中心性和社区间的“桥接”性, 但是这两种因素是谁在什么社区结构特征下起关键作用? 社区结构又与挖掘结果间有怎样的联系? 故社区结构与关键节点挖掘间的关系值得进一步研究.

参考文献

[1] Wang L, Wang J, Shen H W, Cheng X Q 2013 *Chin. Phys. B* **22** 108903

[2] Iyer S, Killingback T, Sundaram B, Wang Z 2013 *PloS one* **8** e59613

[3] Konstantin K, Ángeles S M, San M M 2012 *Scientific Reports* **2** 292

[4] Page L, Brin S, Motwani R, Winograd T 1999 *Stanford InfoLab*

[5] Overington J P, Al-Lazikani B, Hopkins A L 2006 *Nature Reviews Drug Discovery* **5** 993

[6] Yıldırım M A, Goh K I, Cusick M E, Barabási A L, Vidal M 2007 *Nature Biotechnol.* **25** 1119

[7] Liu J G, Ren Z M, Guo Q, Wang B H 2013 *Acta Phys. Sin.* **62** 178901 (in Chinese) [刘建国, 任卓明, 郭强, 汪秉宏 2013 物理学报 **62** 178901]

[8] Ren X L, Lü L Y 2014 *Chin. Sci. Bull.* **59** 1175 (in Chinese) [任晓龙, 吕琳媛 2014 科学通报 **59** 1175]

[9] Albert R, Jeong H, Barabási A L 2000 *Nature* **406** 378

[10] Freeman L C 1977 *Sociometry* **40** 35

[11] Krackhardt D 1990 *Administr. Sci. Quart.* **35** 342

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

[13] Chen D B, Lü L Y, Shang M S, Zhang Y C, Zhou T 2012 *Physica A: Statist. Mech. Appl.* **391** 1777

[14] Chen D B, Gao H, Lü L Y, Zhou T 2013 *PloS one* **8** e77455

[15] 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]

[16] Cheng X Q, Ren F X, Shen H W, Zhang Z K, Zhou T 2010 *J. Statist. Mech.: Theory and Experiment* **2010** P10011

[17] Bae J, Kim S 2014 *Physica A: Statist. Mech. Appl.* **395** 549

[18] Liu J G, Ren Z M, Guo Q 2013 *Physica A: Statist. Mech. Appl.* **392** 4154

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

[20] Ren Z M, Liu J G, Shao F, Hu Z L, Guo Q 2013 *Acta Phys. Sin.* **62** 108902

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

[22] Palla G, Barabási A L, Vicsek T 2007 *Nature* **446** 664

[23] Zhao Z Y, Yu H, Zhu Z L, Wang X F 2014 *Chin. J. Comput.* **37** 753 (in Chinese) [赵之滢, 于海, 朱志良, 汪小帆 2014 计算机学报 **37** 753]

[24] Burt R S 2009 *Structural Holes: The Social Structure of Competition* (London: Harvard University Press) pp53–58

[25] Burt R S, Kilduff M, Tasselli S 2013 *Ann. Rev. Psychol.* **64** 527

[26] Ugander J, Backstrom L, Marlow C, Kleinberg J 2012 *PNAS* **109** 5962

[27] Sun Y, Liu C, Zhang C, Zhang Z 2014 *Phys. Lett. A* **378** 635

[28] Liu C, Zhang Z 2014 *Commun. Nonlinear Sci. Numer. Simulat.* **19** 896

[29] Zhang Z K, Zhang C X, Han X P, Liu C 2014 *PloS one* **9** e95785

[30] Pastor-Satorras R, Vespignani A 2001 *Phys. Rev. Lett.* **86** 3200

[31] Knight W R 1966 *J. Amer. Statist. Associat.* **61** 436

[32] Wikipedia editor 2014 http://wiki.gephi.org/index.php/Datasets[2014.8.1]

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

Leveraging neighborhood "structural holes" to identifying key spreaders in social networks*

Su Xiao-Ping[1)] Song Yu-Rong[2)†]

1) (*School of Computer and Software Engineering, Nanjing Institute of Industry Technology, Nanjing 210046, China*)

2) (*College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*)

(Received 19 May 2014; revised manuscript received 9 September 2014)

Abstract

The identifying of influential nodes in large-scale complex networks is an important issue in optimizing network structure and enhancing robustness of a system. To measure the role of nodes, classic methods can help identify influential nodes, but they have some limitations to social networks. Local metric is simple but it can only take into account the neighbor size, and the topological connections among the neighbors are neglected, so it can not reflect the interaction between the nodes. The global metrics is difficult to use in large social networks because of the high computational complexity. Meanwhile, in the classic methods, the unique community characteristics of the social networks are not considered. To make a trade off between affections and efficiency, a local structural centrality measure is proposed which is based on nodes' and their 'neighbors' structural holes. Both the node degree and "bridge" property are reflected in computing node constraint index. SIR (Susceptible-Infected-Recovered) model is used to evaluate the ability to spread nodes. Simulations of four real networks show that our method can rank the capability of spreading nodes more accurately than other metrics. This algorithm has strong robustness when the network is subjected to sybil attacks.

Keywords: complex networks, structural holes, community structure, influential node centrality measure

PACS: 01.75.+m, 89.75.Hc, 89.75.Fb **DOI:** 10.7498/aps.64.020101

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61373136, 61103051), the Ministry of Education Research in the Humanities and Social Sciences Planning Fund Project, China (Grant No. 12YJAZH120) and the Nanjing Institute of Industry Technology Major Programs, China (Grant No. Yk13-02-03).

† Corresponding author. E-mail: songyr@njupt.edu.cn