

基于区域密度曲线识别网络上的多影响力节点

康玲 项冰冰 翟素兰 鲍中奎 张海峰

Identifying multiple influential nodes based on region density curve in complex networks

Kang Ling Xiang Bing-Bing Zhai Su-Lan Bao Zhong-Kui Zhang Hai-Feng

引用信息 Citation: *Acta Physica Sinica*, 67, 198901 (2018) DOI: 10.7498/aps.67.20181000

在线阅读 View online: <http://dx.doi.org/10.7498/aps.67.20181000>

当期内容 View table of contents: <http://wulixb.iphy.ac.cn/CN/Y2018/V67/I19>

您可能感兴趣的其他文章

Articles you may be interested in

[基于复杂网络动力学模型的无向加权网络节点重要性评估](#)

Evaluation methods of node importance in undirected weighted networks based on complex network dynamics models

物理学报.2018, 67(9): 098901 <http://dx.doi.org/10.7498/aps.67.20172295>

[一种改进的基于信息传播率的复杂网络影响力评估算法](#)

An improved evaluating method of node spreading influence in complex network based on information spreading probability

物理学报.2017, 66(20): 208901 <http://dx.doi.org/10.7498/aps.66.208901>

[多层单向耦合星形网络的特征值谱及同步能力分析](#)

Synchronizability and eigenvalues of multilayer star networks through unidirectionally coupling

物理学报.2017, 66(18): 188901 <http://dx.doi.org/10.7498/aps.66.188901>

[基于领域相似度的复杂网络节点重要度评估算法](#)

Node importance measurement based on neighborhood similarity in complex network

物理学报.2017, 66(3): 038902 <http://dx.doi.org/10.7498/aps.66.038902>

[网络集聚性对节点中心性指标的准确性影响](#)

Effect of variable network clustering on the accuracy of node centrality

物理学报.2016, 65(2): 028901 <http://dx.doi.org/10.7498/aps.65.028901>

基于区域密度曲线识别网络上的多影响力节点*

康玲 项冰冰 翟素兰 鲍中奎† 张海峰

(安徽大学数学科学学院, 合肥 230601)

(2018年5月23日收到; 2018年6月27日收到修改稿)

复杂网络多影响力节点的识别可以帮助理解网络的结构和功能, 具有重要的理论意义和应用价值. 本文提出一种基于网络区域密度曲线的多影响力节点的识别方法. 应用两种不同的传播模型, 在不同网络上与其他中心性指标进行了比较. 结果表明, 基于区域密度曲线的识别方法能够更好地识别网络中的多影响力节点, 选中的影响力节点之间的分布较为分散, 自身也比较重要. 本文所提方法是基于网络的局部信息, 计算的时间复杂度较低.

关键词: 复杂网络, 多影响力节点, 区域密度曲线

PACS: 89.75.Fb, 89.75.Hc

DOI: 10.7498/aps.67.20181000

1 引言

随着信息技术的发展以及经济的全球化, 人类的社会活动日趋网络化, 像在线社交网络、科研合作网络、交通网络、电力网络以及与人自身相关的新陈代谢网络等不断进入人们的视野. 另外, 这些网络的数据规模与日剧增, 内在结构也变得日益复杂. 面对这些大规模的复杂网络, 能否有效识别其中具有影响力的节点, 具有重要的理论意义和实际应用价值, 已被广泛应用于疾病传播、谣言扩散、新产品的推广以及交通拥堵的治理等方面^[1-6].

一些中心性指标如度中心性^[7]、介数中心性^[8]、接近中心性^[9]、特征向量中心性^[10], k -shell分解等^[11]相继被提出来识别网络中的影响力节点. 近年来, Chen等^[12]考虑了节点的高阶邻居信息, 提出一种基于多级邻居信息的半局部指标来进行节点的排序; 考虑到网络的结构洞节点对网络传播的作用, 韩忠明等^[13]以及苏晓萍和宋玉蓉^[14]结合结构洞理论, 利用邻域的结构洞来探测网络中的

最具影响力节点; Radicchi和Castellano^[15]提出非回溯性指标并结合渗流理论来进行关键节点的识别. 由于信息在网络中的传播, 不仅与节点间的最短路径有关, 还与节点间最短路径的条数以及传播率有关. 阮逸润等^[16]在文献^[17]的基础上提出一种改进的SP(传播概率)指标来评估节点的影响力. 然而, 以上对影响力节点的研究主要基于单个影响力节点来开展, 但事实往往并非如此, 像一些疾病、谣言或广告的传播可能来自多个不同的传播源, 所以网络中往往存在多个影响力节点. Hu等^[18]在带有社团结构的网络中探讨了多影响力节点的识别问题, 并发现每个社团的hub点(大度节点)往往具有很强的传播能力. 然而, 当传播源的个数超过社团数目时, 该方法将无能为力. Zhao等^[19]将图论中的着色方法引入复杂网络中, 提出一种多影响力节点的识别方法, 但该方法有时不能保证传播源间的分布足够分散; Guo等^[20]提出了改进的着色方法. 可以看出, 对复杂网络影响力节点的识别研究还处于初始阶段, 仍有许多问题有待进一步改进.

* 安徽省自然科学基金(批准号: 1808085MF201)、安徽省高校自然科学基金(批准号: KJ2017A025)、浙江省海洋大数据挖掘与应用重点实验室(批准号: OBDMA201502)、安徽大学信息保障技术协同创新中心开放课题(批准号: ADXXBZ201608)和安徽大学研究基金(批准号: 01001951, 01005102)资助的课题.

† 通信作者. E-mail: zkbao@ahu.edu.cn

多影响力节点识别的指导性思想是: 选取的节点间的分布要较为分散, 且自身要足够重要, 这样能在保证传播非冗余的同时, 使得传播范围尽可能的广. 但两者之间要想同时满足几乎不可能, 只能在两者之间寻求一个平衡. 在网络的核心-边缘 (core-periphery) 和社团结构探测中, 我们提出一种统一的方法, 即通过绘制网络的区域密度曲线 (region density curve), 然后利用曲线的峰值点来确定网络的核 (core) 节点或者社团内部的节点, 进而用局部扩张的方法来获得网络的核心边缘 (core-periphery) 结构和社团结构 [21]. 通过对网络区域密度曲线的进一步分析, 发现区域密度曲线的波谷点正是连接核心与边缘 (core-periphery)、或者社团和社团之间的桥梁节点. 与其他节点相比, 桥梁节点在网络的传播过程中具有很重要的影响力 [22], 并且分布较为分散. 鉴于此, 本文提出基于区域密度曲线的多影响力节点识别方法 (RDC), 并在不同的网络上利用疾病和谣言两种不同的传播模型进行了实验分析, 结果表明, RDC 能够很好地识别网络中的多影响力节点, 而且能够保证选取的这些影响力节点之间分布较为分散, 自身也足够重要.

2 方法

假设网络是一个无权无向图 $G(V, E)$, 其中 V 表示网络中的节点, E 表示网络中的连边. 首先, 给出几种常用的中心性指标, 本文将以此作为所提方法的比较指标.

2.1 中心性指标

度中心性 (DC), 被定义为节点的邻居数目, 即

$$DC(i) = \sum_{j=1}^N a_{ij}. \quad (1)$$

介数中心性 (BC), 是以经过某个节点的最短路径的数目来刻画节点的重要性指标, 即

$$BC(i) = \sum_{j \neq i \neq k}^N n_{jk}^i / n_{jk}, \quad (2)$$

其中, n_{jk} 为从节点 j 到节点 k 的最短路径的数目, n_{jk}^i 为从节点 j 到节点 k 的条最短路径中经过节点 i 的最短路径的数目.

k -shell 方法 (KS) 的具体步骤如下: 首先把网络中度为 1 的节点及其所连接的边去掉, 剩下的网

络中会出现一些度为 1 的节点, 再将这度为 1 的节点去掉, 直到所剩的网络中没有度为 1 的节点, 则所有被去掉的节点称为 1-shell; 然后继续上面的方法, 去掉剩下的网络中度为 2 的节点及其相连的边, 直至不再有度值为 2 的节点为止, 则这一轮所有被去除的节点及其连边称为网络的 2-shell. 依次类推, 直到所有的节点都被分到相应的 k -shell.

度折扣方法 (DDC) 是由 Chen 等 [23] 提出的, 其思想是: 设 v 是节点 u 的邻居集, 如果 u 已被选作传播源, 当基于度中心性考虑 v 作为下一个传播源时, 为避免传播冗余, 我们不应该再考虑 uv 的边, 应对 v 的度进行折扣减 1.

将图论中的图着色方法引入复杂网络, Zhao 等 [19] 使用 Welsh-Powell 算法将网络分成若干个独立的子集, 以保证每一个独立子集中的任意两点都不相连. 然后基于某个中心性指标, 在最大的独立子集中选取按指定的中心性指标排序靠前的 m 个节点作为多传播源. 在文中, 我们比较了基于度中心性、介数中心性以及 k -shell 中心性等着色方法, 分别标记为 DCC, BCC 和 KSC.

2.2 基于区域密度曲线的多影响力节点探测

首先对网络中的节点进行重新排序, 使连接紧密的节点的次序也相近; 然后绘制网络的区域密度曲线, 并找出曲线上的波谷点; 最后在波谷点两侧选取一定比例的节点作为影响力节点, 具体步骤如下 [21].

1) 对节点进行排序, 使连接紧密的节点次序也相近

定义一个初始集合 $U = \phi$, 初始化 $V' = V$. U 作为存储有序序列, $V' = V/U$ 为待选集合. 首先, 从 V 中选择一个中心性指标比较好的节点 N_2 加入到序列 U 中; 然后, 计算 V' 中各节点的 $C(i, U)$ 值, $C(i, U)$ 值最大的节点作为节点序列的第二个节点 N_2 , 即选择与 U 中节点的数量最多的节点加入到 U 中. 若同时找到两个或两个以上的节点, 选择度大的节点添加到 U 中; 更新 V' 和 U , 按照上面的方法依次进行, 最后可得 $U = \{u_1, u_2, \dots, u_N\}$, $V' = \phi$.

$$C(i, U) = \sum_{j \in U} a_{ij} + \frac{d(i)}{d_{\max}}, \quad (3)$$

其中, a_{ij} 表示节点 i 与 j 是否相连, 如果相连 $a_{ij} = 1$, 否则 $a_{ij} = 0$; $d(i)$ 是节点 i 的度, d_{\max} 是网络中最大度节点的度.

2) 绘制网络的区域密度曲线

为了刻画网络中某个区域 S 内部点的连接密度, 定义 S 的区域密度 $CD(S)$ 为

$$CD(S) = \frac{2m'}{n'(n' - 1)}, \quad (4)$$

其中 n' 为区域 S 内节点的数量, m' 为区域 S 内连接的边的数量.

给定参数值 α , 即核的最小尺寸, 在文中取作网络的平均度. 在节点序列 U 中, 将排序为 r 的节点 U_r 的区域密度定义为序号为 $r - \alpha - 1$ 至 r 的节点组成的子图的区域密度, 即

$$RD(r) = CD(\cup_{i=r-\alpha+1}^r \{U_i\}). \quad (5)$$

然后, 将各节点的区域密度绘制在二维直角坐标系中, 即获得网络的区域密度曲线 (密度曲线如图 1(a) 所示). 在文献 [21] 中, 通过对该区域密度曲线的分析可以探测出核心-边缘 (core-periphery) 结构、社团结构等, 而且可以找出连接不同社团之间的桥梁节点. 据此, 我们利用此方法来探测网络中的多影响力节点.

3) 多影响力节点的选取

首先, 通过上面绘制的区域密度曲线, 找出处在波谷位置的节点. 注意到区域密度曲线的第 1 个节点的 RD 值为 0, 第 2 个会达到峰值 (因为两个节点有连接就是全连接), 故第一个节点的 RD 值不是有效值. 另外, 区域密度曲线的末端代表的都是些比较稀疏的节点, 与社团以及核心边缘结构等没有多大的联系, 因此选择多传播源时, 不考虑区域密度曲线这两个波谷位置的节点.

然后, 利用区域密度曲线, 计算出各谷点之间的节点数, 并确定各谷点处要选取的传播源的个数. 假设要求探测 m 个有影响力的节点, 通过区域密度曲线观测到波谷位置的节点序号为 $[N_0, N_1, \dots, N_k, N_{k+1}]$ (其中 N_0 和 N_{k+1} 为上面讨论的两类节点), 同时记录每个节点之间的节点数 $[n_1, n_2, \dots, n_{k+1}]$ (其中 n_i ($i = 1, 2, \dots, k$) 代表 N_i 和 N_{i+1} 之间的节点数, 如下例中图 1(a) 标注的 n_4), 则在各个谷点处选取传播源个数为 $[m_1, m_2, \dots, m_k]$, 其中第 i ($i = 1, 2, \dots, k$) 个谷

点 N_i 处的传播源个数 m_i 定义为

$$m_i = \frac{m(n_i + n_{i+1})}{n_1 + 2n_2 + 2n_3 + \dots + 2n_{k-1} + n_k}. \quad (6)$$

因为我们是从小谷点 N_i 所在位置的左边选取一定比例的传播源, 所以分母会出现系数 2, 相当于 $(n_1 + n_2) + (n_2 + n_3) + \dots + (n_{k-2} + n_{k-1}) + (n_{k-1} + n_k)$.

最后, 在各个谷点位置的左右两边各选取一半比例的传播源, 利用各谷点确定的传播源放在一起, 即是要探测的多传播. 具体算法如下.

算法 1 节点进行排序

输入: 网络 $G(V, E)$.

输出: 节点排序 U (有序集合).

1) 初始化: $U = \emptyset, V' = V$.

2) 选择最大接近中心性节点 $N_2, U \leftarrow N_2, V' = V'/N_2$.

3) WHILE $V' \neq \emptyset$ DO.

4) FOR $i \in \Gamma(U) \cup V'$ DO.

5) 通过 (1) 式计算 $C(i, U)$.

6) $N_2 = \arg \max_i (C(i, U))$ (若存在多个最大值, 取度最大的节点)

7) ENDFOR.

8) $U \leftarrow N_2, V' = V'/N_2$.

9) ENDWHILE.

算法 2 寻找多个影响力节点

输入: 节点排序 U 多影响力节点的个数 K .

输出: 多影响力节点的集合 C .

1) 初始化 $C = \emptyset$.

2) FOR $U_i \in U$ DO.

3) 通过 (3) 式计算 $RD(i)$.

4) ENDFOR.

5) 计算 RD 曲线的有效谷点序号为: $[N_0, N_1, \dots, N_k, N_{k+1}]$ (其中 N_0 和 N_{k+1} 为文中讨论的两类节点).

6) 统计谷点之间的节点数: $[n_1, n_2, \dots, n_{k+1}]$.

7) 通过 (4) 式计算每个谷点处要寻找的影响力节点的个数 $[m_1, m_2, \dots, m_k]$.

8) FOR $i \in [1, \dots, K]$ DO.

9) 在 N_i 谷点处前后共取 M_i 个节点记为 C_i .

10) $C = C \cup C_i$.

11) ENDFOR.

其中 $\Gamma(U)$ 表示 U 中所有节点的邻居. 下文用 RDC 来表示本文提出的方法.

下面以Football网^[24]为例来说明具体的探测方法, 假设要探测的传播源的个数 $m = 12$. 首先利用(1)式, 对网络节点进行重新编号, 并利用(2)式和(3)式, 计算每个节点的区域密度. 然后, 以横坐标为节点序列, 纵坐标为区域密度, 绘制出节点区域密度曲线图, 如图1(a)所示. 接下来, 寻找区域密度曲线的谷点, 得到谷点的节点序列[1, 77, 89, 115, 4, 36, 65, 22, 13, 43], 记录各谷点之间的节点数目, 记为[15, 15, 15, 17, 14, 9, 9, 13, 8]; 然后, 根据(4)式计算各谷点处应该选取的传播源个数, 记为[2, 2, 2, 2, 1, 1, 1, 1]; 最后, 在每个谷点的左右两边各取一半的传播源, 比如在第一个谷点77号

的前面选取1个, 那么在77号的后面选0个传播源, 因为谷点本身就已经作为传播源了, 以此类推, 最终选取12个传播源[96, 77, 37, 89, 84, 115, 85, 4, 36, 65, 22, 13]. 图1(c)所示为选取的影响力节点及其位置情况, 其中红色圆圈代表的节点就是我们找到的12个传播源, 可以看出, 这些节点并不是所谓的大度节点, 恰恰相反的是, 有些节点的度却很小, 比如图中的37号, 13号以及96号节点, 但这些节点往往都位于不同社团的交界处. 这也正好满足我们所寻找传播源的条件, 传播源之间能尽可能的分散, 避免传播的冗余.

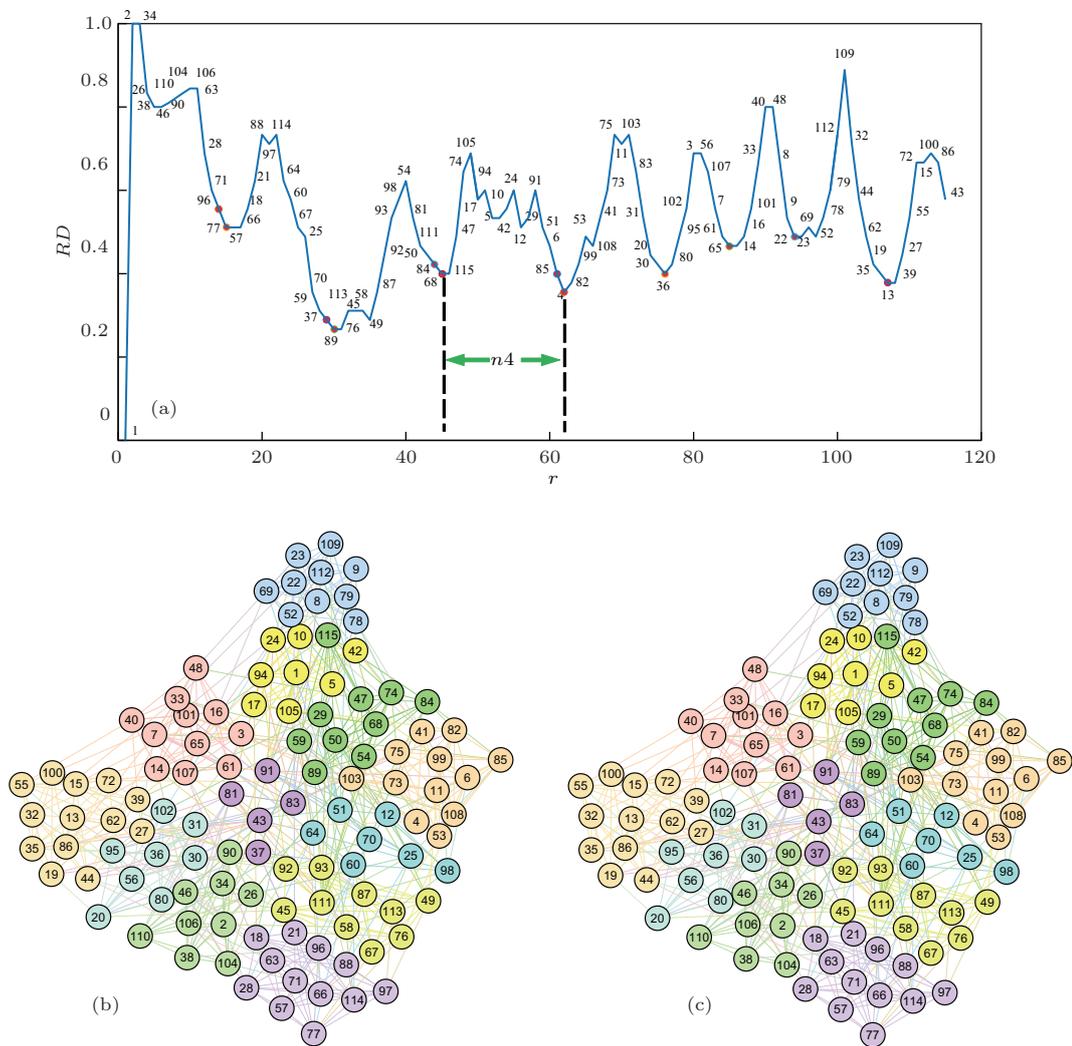


图1 Football网的多影响力节点识别 (a) 网络的区域密度曲线; (b) 网络的划分情况, 不同颜色代表不同的社团; (c) 红色的节点为RDC方法选取的12个影响力节点

Fig. 1. Identification of multiple influential nodes on the Football network: (a) Region density curve of the network; (b) original structure, where nodes with the same color are in the same community; (c) the twelve red nodes are influential nodes which are identified by the RDC method.

3 实验分析

3.1 传播模型

很多文献中已经指出, 节点的影响力是依赖于传播的动力学 [25,26]. 因此, 在本文分别考虑两种疾病传播动力学, 疾病传播的 SIR 模型和谣言传播模型, 进而更全面地来比较不同传播源探测方法.

在疾病传播模型中, 网络中的节点分为三类, 易感者 (S), 感染者 (I), 恢复者 (R). 一个感染者 (I) 以一定的传播率将疾病传染给易感的邻居 (S), 传播完自己的所有邻居后, 感染者以一定的恢复率变为恢复者 [27]. 在这里, 为了更清晰地区分不同的传播源探测方法, 仅考虑每个感染者每次仅与一个邻居接触, 即单传的 SIR 模型.

在谣言传播模型中, 将网络节点分为三类, 无知者 (S), 传播者 (I), 已知者 (R). 一个传播者 (I) 以一定的概率将谣言传播给相邻的无知者 (S), 当传播者接触到另一个传播者或已知者时, 最初的传播者就会失去传播谣言的兴趣, 以一定的概率变为已知者 [28]. 本文考虑恢复率 $\mu = 0.1$ 的情况.

3.2 数据描述

将不同的多影响力节点识别方法在六个真实网络中进行比较, 包括 Email, Yeast, SciMet, Kohonen, HEP, PGP. 表 1 是网络的一些基本信息, 其中 N , M 为网络的节点数和边数, $\langle k \rangle$ 和 C 为网络的平均度和聚类系数, L 和 D 为网络的平均短路长度和直径 [29].

表 1 网络的基本信息表
Fig. 1. Basic structural properties for networks.

	N	M	$\langle k \rangle$	C	L	D
Email	1133	5451	9.62	1.942	3.606	8
Yeast	2375	11693	9.8467	3.4756	5.0960	15
SciMet	2678	10368	7.7431	2.4265	4.1796	12
Kohonen	3704	12673	6.8429	9.3170	3.6703	12
HEP	5835	13815	4.7352	1.9262	7.0264	19
PGP	10680	24316	4.5536	4.1465	7.4855	24

3.3 结 果

为便于比较不同的多影响力节点识别方法, 首先选取 m 个传播源, 然后分别依据 SIR 疾病传播、谣言传播模型, 去模拟网络的动力学过程, 直到网络中没有处于感染态的节点, 最后以网络中恢复节

点的比例来衡量识别方法的性能. 为了保证实验结果的可靠性, 对传播过程中进行了 100 次平均.

为便于比较不同方法的结果, 定义一个相对比率指标 Δ 如下 [19]:

$$\Delta = (R_i - R_{DC})/R_{DC}, \quad (7)$$

其中, R_i 表示在某种传播源识别方法 (如 DCC, KS, KSC, BC, BCC, DDC, RDC) 下最终恢复节点的比例, R_{DC} 是采用度中心性方法时最终恢复节点的比例. $\Delta > 0$ 意味着当前使用的方法要优于度中心性的方法, $\Delta > 0$ 的值越大表示当前方法的优势越明显.

首先, 利用 SIR 疾病传播模型, 在六个实际网络上对不同的识别方法进行比较. 实验结果表明, 面对 SIR 疾病传播时, 本文提出的基于区域密度曲线的识别方法 (RDC) 都要优于其他方法, 尤其是当传播率较小时, 优势非常明显. 即使传播率增大到一定程度, RDC 方法还是要优于其他方法. 另外, 随着选取的传播源个数的增多, 这种优势越发明显 (见图 2 和图 3). 图 2 是在 SIR 疾病传播模型下, 不同识别方法相对于度中心性指标所得的结果, 其中传播率 β 从 0.05 到 0.5, 每次增加 0.05, 恢复率 $\mu = 0.1$. 图 2(a)—(f), (g)—(l), (m)—(r) 表示传播过程中选取的传播源个数分别为 30, 50, 90.

紧接着, 在上面的六个实际网络上, 考虑了不同识别方法在谣言传播模型中的效果. 实验结果表明, 在谣言传播机理下, RDC 方法仍然要优于其他的多影响力节点识别方法. 随着选取的传播源个数的增加, RDC 方法的优势越发明显. 图 3 是在谣言传播模型下的实验结果, 参数的设定与 SIR 疾病传播模型下的相同. 图 3(a)—(f), (g)—(l), (m)—(r) 表示传播过程中选取的传播源个数分别为 30, 50, 90.

接下来, 研究不同识别方法所得的影响力节点之间的平均距离和平均度的情况. 图 4 是考虑不同识别方法所获取的影响力节点之间的平均距离随传播源个数的变化情况, 其中传播源的个数 m 是从 20 开始, 每次增加 20, 一直到 200. 可以看出, 相比于其他识别方法, 利用 RDC 方法选取的传播源之间的平均距离要大很多, 并随选取的传播源个数的增加, 平均距离呈现增大的趋势. 由此可以看出, RDC 方法获取的传播源彼此之间较为分散, 能够很好地避免传播的冗余.

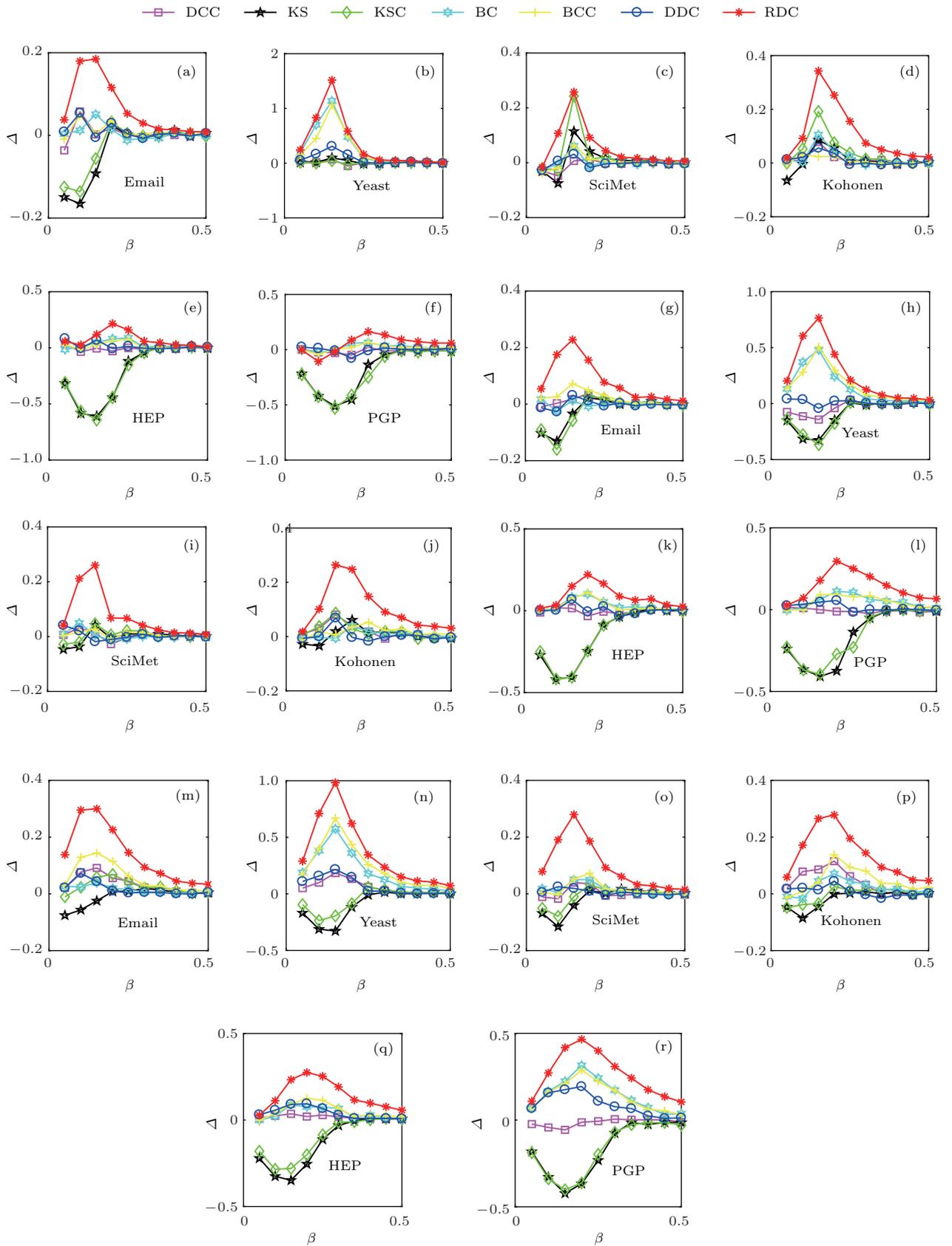


图2 对于 SIR 疾病传播模型, 不同识别方法与疾病传播率之间的关系

Fig. 2. For SIR model, the relative ratios Δ for different indices as functions of transmission rate β are compared in six real networks.

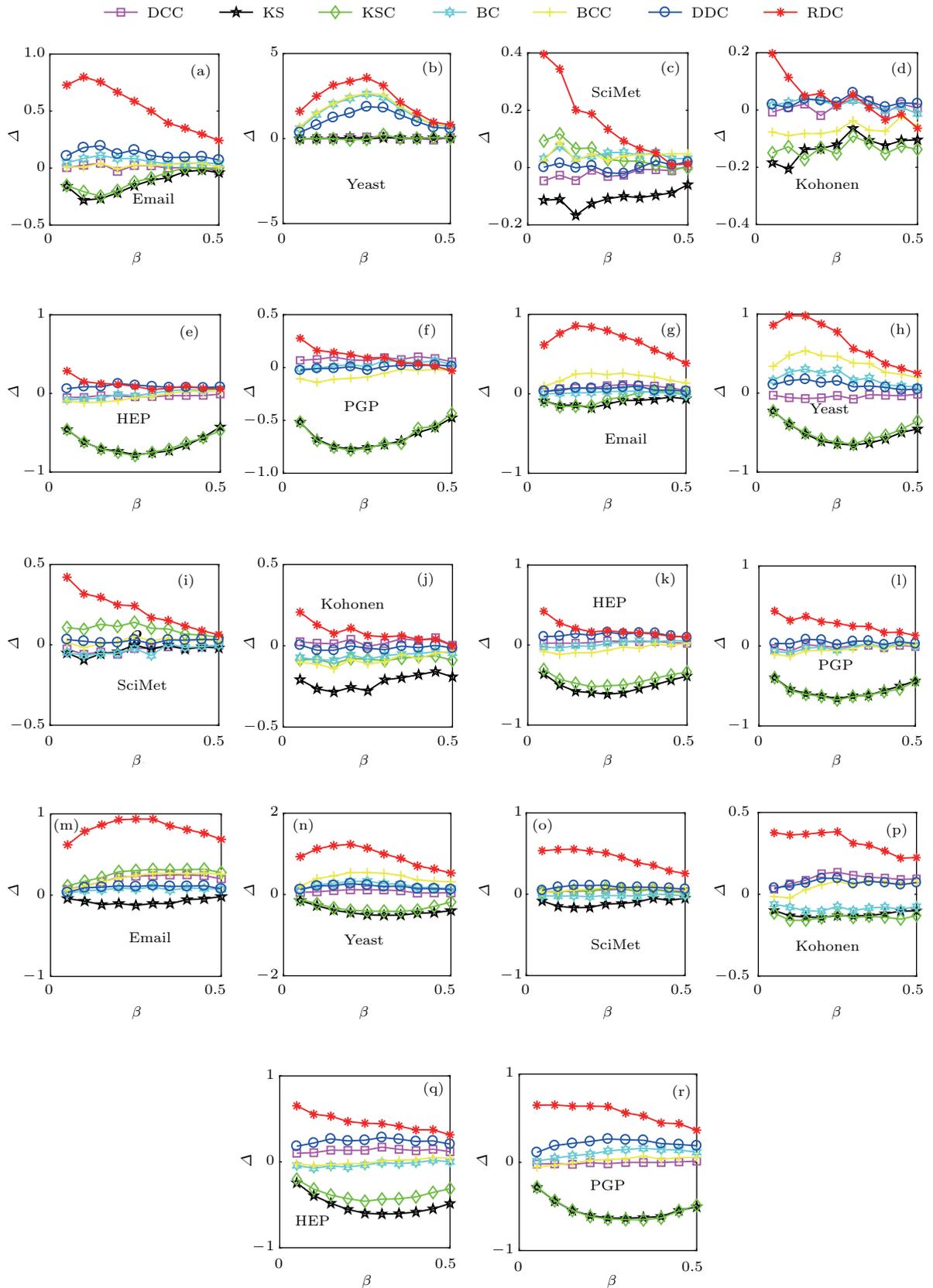


图3 对于谣言传播模型, 不同识别方法与疾病传播率之间的关系

Fig. 3. For rumor propagation model, the relative ratios Δ for different indices as functions of transmission rate β are compared in six real networks.

图5是不同识别方法所获取的影响力节点的平均度随传播源个数的变化情况,其中传播源个数 m 的设定与图4相同.可以看出,RDC方法选取的传播源的平均度要小于其他的识别方法.由选取传播源的指导思想可知,传播源间的分布分散和自身重要两者不可兼得.虽然RDC方法选取的传播源的平均度比其他方法都要小,但从图5中的

蓝色虚直线(即网络的平均度)可以看出,选择的这些传播源不是“不重要”的节点,这些节点的平均度比各自网络的平均度都要大.而且随着选取的传播源个数的增加,RDC方法的平均度基本保持不变.由此可以看出,RDC方法选取的传播源彼此间分布较分散,而且各个传播源自身也比较重要.

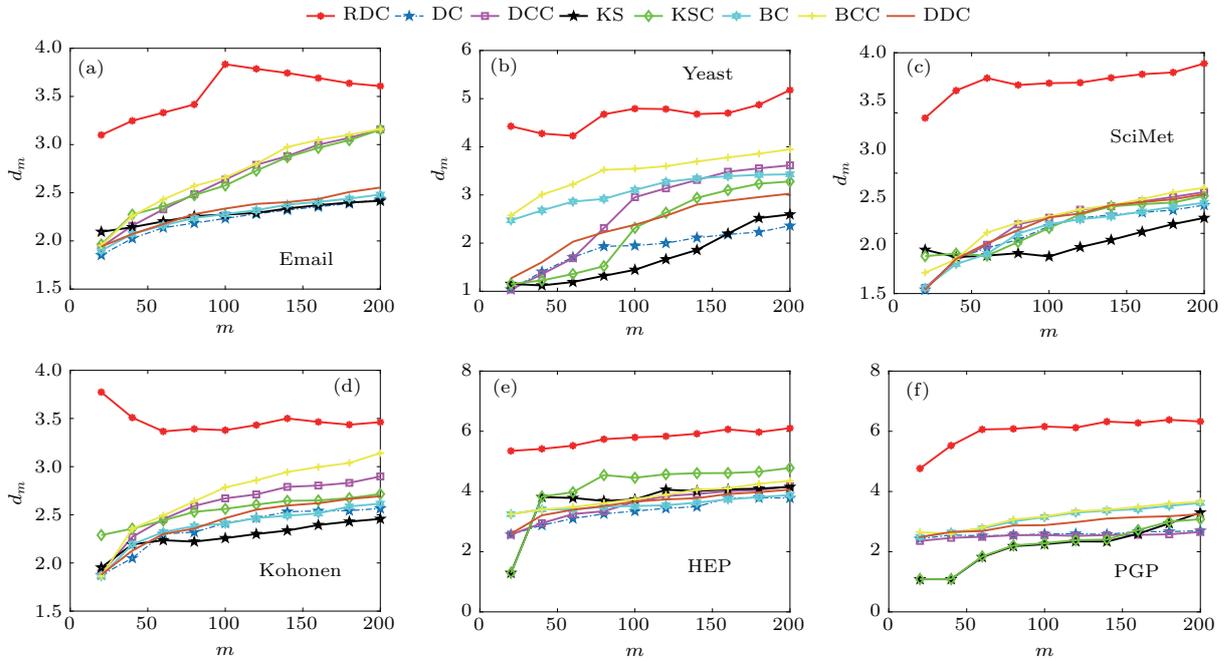


图4 不同方法选取的传播源间的平均距离与传播源数量间的关系

Fig. 4. The effects of the number of multiple spreaders m on the average distance d_m are compared in six real networks.

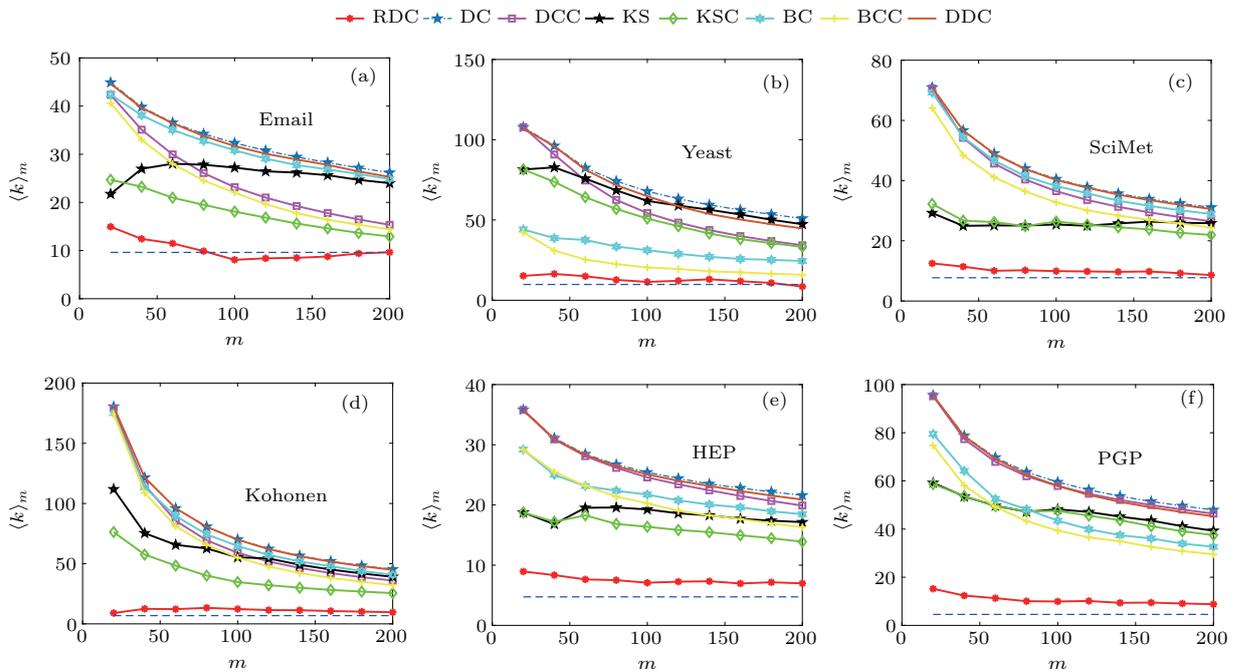


图5 不同方法选取的传播源的平均度与传播源数量间的关系,其中蓝色虚线代表每个网络的平均度

Fig. 5. The effects of the number of multiple spreaders m on the average degree $\langle k \rangle_m$ are compared in six real networks. Dotted line in each subfigure denotes the average degree of the network.

图6和图7讨论了不同的参数 α 对传播模型的影响,红色方框中的部分是本文采用的 α 值,可以发现随着 α 的变化,传播范围会呈现一定的无规则

的波动.因此本文采用了文献[21]提供的默认参数平均度,虽然在所有的网络都是最优的,但是总体表现较好.

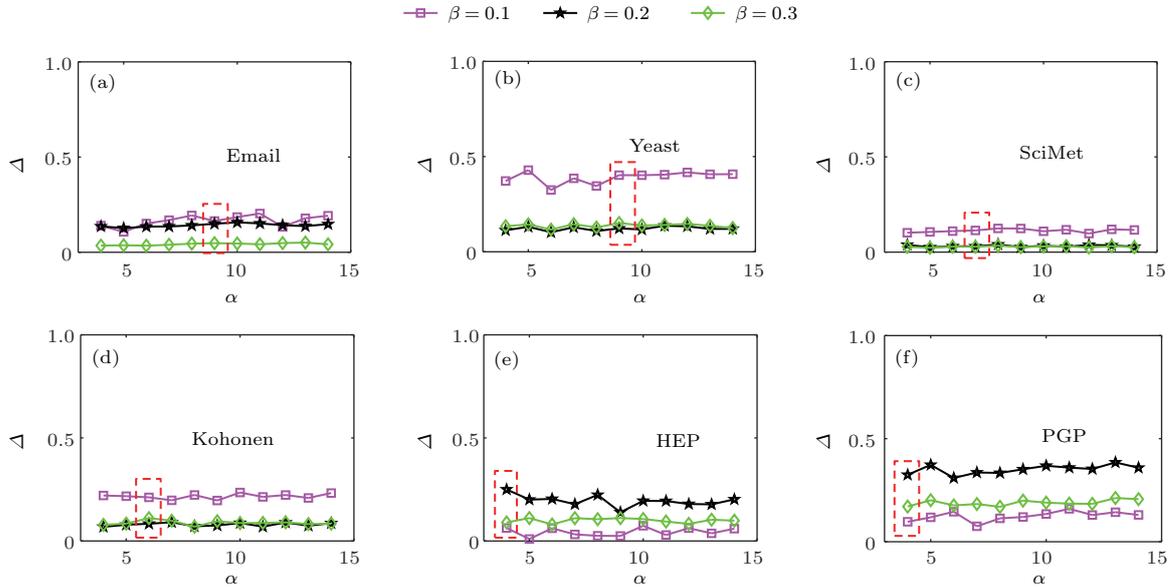


图6 对于SIR传播模型,不同传播率下参数 α 与相对比率 Δ 的关系

Fig. 6. For SIR model, the effects of parameter α on the relative ratios Δ for different transmission rates β are compared in six real networks.

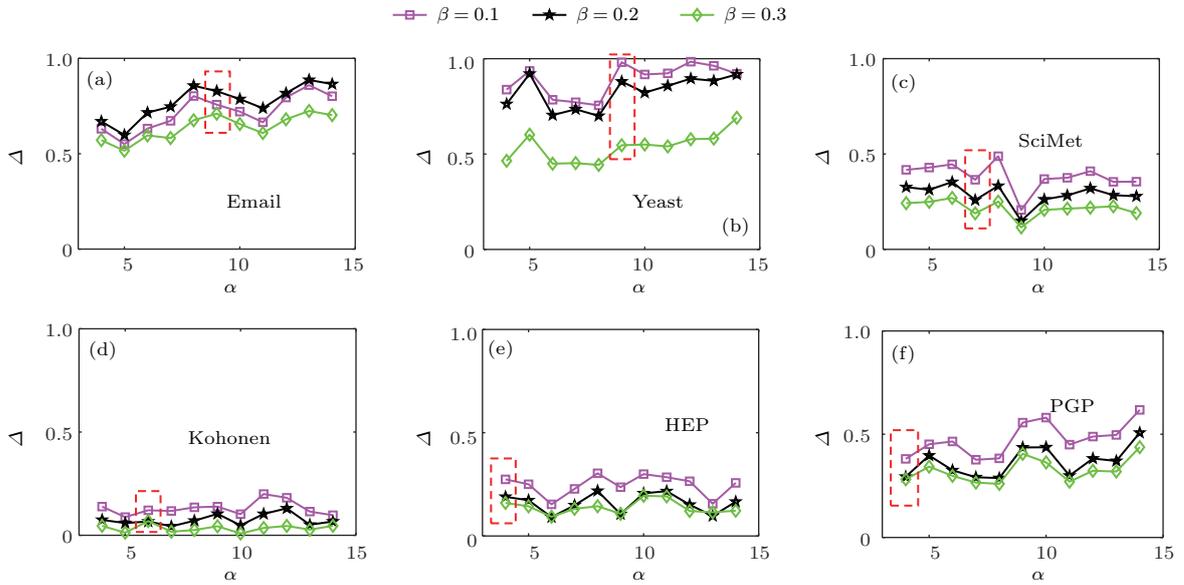


图7 对于谣言传播模型,不同传播率下参数 α 与相对比率 Δ 的关系

Fig. 7. For rumor propagation model, the effects of parameter α on the relative ratios Δ for different transmission rates β are compared in six real networks.

4 结 论

多影响力节点识别的指导性思想是:选取的节点彼此间的分布要较为分散,而且自身要足够重要,但两者不可兼得.本文提出一种基于网络区域

密度曲线的多影响力节点识别方法,该方法基于网络的局部信息,计算的时间复杂度较低.应用两种不同的传播动力学模型,在六个真实网络上进行了数据实验,结果表明本文提出的识别方法要优于其他的识别方法,而且选中的影响力节点之间的分布较为分散,自身也较为重要.本文仅考虑无向无权

网络, 下一步可考虑将本文的方法推广至有向有权网络中.

参考文献

- [1] 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]
- [2] Lü L Y, Chen D B, Ren X L, Zhang Q M, Zhang Y C, Zhou T 2016 *Phys. Rep.* **650** 1
- [3] Huang B, Zhao X Y, Qi K, Tang M, Do Y H 2013 *Acta Phys. Sin.* **62** 218905 (in Chinese) [黄斌, 赵翔宇, 齐凯, 唐明, 都永海 2013 物理学报 **62** 218905]
- [4] 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]
- [5] Shu P P, Wang W, Tang T, Shang M S 2015 *Acta Phys. Sin.* **64** 208901 (in Chinese) [舒盼盼, 王伟, 唐明, 尚明生 2015 物理学报 **64** 208901]
- [6] Liu Y, Tang M, Do Y H, Hui P M 2017 *Phys. Rev. E* **96** 022323
- [7] Freeman L C 1979 *Social Networks* **1** 215
- [8] Freeman L C 1977 *Sociometry* **40** 35
- [9] Sabidussi G 1966 *Psychometrika* **31** 581
- [10] Bonacich P 1972 *J. Math. Sociol.* **2** 113
- [11] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nature Phys.* **6** 888
- [12] Chen D B, Lü L Y, Shang M S, Zhang Y C, Zhou T 2012 *Physica A* **391** 1777
- [13] Han Z M, Wu Y, Tan X S, Duan D G, Yang W J 2015 *Acta Phys. Sin.* **64** 058902 (in Chinese) [韩忠明, 吴杨, 谭旭升, 段大高, 杨伟杰 2015 物理学报 **64** 058902]
- [14] Su X P, Song Y R 2015 *Acta Phys. Sin.* **64** 020101 (in Chinese) [苏晓萍, 宋玉蓉 2015 物理学报 **64** 020101]
- [15] Radicchi F, Castellano C 2016 *Phys. Rev. E* **93** 062314
- [16] Ruan Y R, Lao S Y, Wang J D, Bai L, Hou L L 2017 *Acta Phys. Sin.* **66** 208901 (in Chinese) [阮逸润, 老松杨, 王竣德, 白亮, 侯绿林 2017 物理学报 **66** 208901]
- [17] Bao Z K, Ma C, Xiang B B, Zhang H F 2017 *Physica A* **468** 391
- [18] Hu Z L, Ren Z M, Yang G Y, Liu J G 2014 *Int. J. Mod. Phys. C* **25** 1440013
- [19] Zhao X Y, Huang B, Tang M, Zhang H F, Chen D B 2015 *Eur. Phys. Lett.* **108** 68005
- [20] Guo L, Lin J H, Guo Q, Liu J G 2016 *Phys. Lett. A* **380** 837
- [21] Xiang B B, Bao Z K, Ma C, Zhang X Y, Chen H S, Zhang H F 2018 *Chaos* **28** 013122
- [22] Zhao Z Y, Yu H, Zhu Z L, Wang X F 2014 *Chin. J. Comp.* **37** 753 (in Chinese) [赵之滢, 于海, 朱志良, 汪小帆 2014 计算机学报 **37** 753]
- [23] Chen W, Wang Y, Yang S 2009 *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining Pairs*, France, June 28–July 01, 2009 p199
- [24] Newman M E J 2004 *Phys. Rev. E* **69** 066133
- [25] Liu Y, Tang M, Zhou T, Do Y H 2015 *Sci. Rep.* **5** 13172
- [26] Liu J G, Lin J H, Guo G, Zhou T 2016 *Sci. Rep.* **6** 21380
- [27] Li R Q, Wang W, Shu P P, Yang H, Pan L M, Cui A X, Tang M 2016 *Compl. Syst. Compl. Sci.* **3** 1 (in Chinese) [李睿琪, 王伟, 舒盼盼, 杨慧, 潘黎明, 崔爱香, 唐明 2016 复杂系统与复杂性科学 **3** 1]
- [28] Borge-Holthoefer J, Moreno Y 2012 *Phys. Rev. E* **85** 026116
- [29] Lü L Y, Zhou T 2013 *Link Prediction* (Beijing: Higher Education Press) p286 (in Chinese) [吕琳媛, 周涛 2013 链路预测 (北京: 高等教育出版社) 第286页]

Identifying multiple influential nodes based on region density curve in complex networks*

Kang Ling Xiang Bing-Bing Zhai Su-Lan Bao Zhong-Kui[†] Zhang Hai-Feng

(School of Mathematical Science, Anhui University, Hefei 230601, China)

(Received 23 May 2018; revised manuscript received 27 June 2018)

Abstract

Complex networks are ubiquitous in natural science and social science, ranging from social and information networks to technological and biological networks. The roles of nodes in networks are often distinct, the most influential nodes often play an important role in understanding the spreading process and developing strategies to control epidemic spreading or accelerating the information diffusion. Therefore, identifying the influential nodes in complex networks has great theoretical and practical significance. Some centrality indices have been proposed to identify the influential nodes in recent years, but most of the existing algorithms are only appropriate to the identifying of single influential node. Many times, spreading process is initiated by simultaneously choosing multiple nodes as the spreading sources, such as rumors, opinions, advertisements, etc. Therefore, it is necessary to develop efficient methods of identifying the multiple influential nodes in complex networks. In this paper, a method based on region density curve of networks (RDC) is proposed to identify the multiple influential nodes in complex networks. Firstly, we rearrange all nodes of network in a new sequence, and then plot the region density curve for network. Finally, we identify the multiple influential nodes based on the valley points of region density curve. Using two kinds of spreading models, we compare RDC index with other indices in different real networks, such as degree, degree discount, k -shell, betweenness and their corresponding coloring methods. The results show that the influential nodes chosen according to our method are not only dispersively distributed, but also are relatively important nodes in networks. In addition, the time complexity of our method is low because it only depends on the local information of networks.

Keywords: complex networks, multiple influential nodes, region density curve

PACS: 89.75.Fb, 89.75.Hc

DOI: [10.7498/aps.67.20181000](https://doi.org/10.7498/aps.67.20181000)

* Project supported by the Natural Science Foundation of Anhui Province, China (Grant No. 1808085MF201), the Natural Science Foundation of the Higher Education Institutions of Anhui Province, China (Grant No. KJ2017A025), the State Key Laboratory for Ocean Big Data Mining and Application of Zhejiang Province, China (Grant No. OBDMA201502), the Information Security Technology Collaborative Innovation Center of Anhui University, China (Grant No. ADXXBZ201608), and the Anhui University Foundation, China (Grant Nos. 01001951, 01005102).

[†] Corresponding author. E-mail: zkbao@ahu.edu.cn