中国科学: 信息科学 2019 年 第 49 卷 第 10 期: 1333–1342
SCIENTIA SINICA Informationis
论文

基于群体影响力的网络传播关键节点选择策略

周明洋[1,2], 吴向阳[2], 曹扬[1], 罗燎[1], 李晓宇[2], 廖好[1,2*], 汪秉宏[3]

1. 中电科大数据研究院有限公司, 贵阳 550022
2. 深圳大学计算机与软件学院, 深圳 518060
3. 中国科学技术大学近代物理系, 合肥 230027
* 通信作者. E-mail: jamesliao520@gmail.com

收稿日期: 2019–02–21; 接受日期: 2019–04–25; 网络出版日期: 2019–10–17

摘要 网络中的少量关键节点对政务舆情、病毒等信息扩散具有重要影响, 因此关键节点挖掘是网络科学的关键问题之一. 经典方法通过单个节点重要性指标选择关键节点, 而多个重要节点的综合影响力更值得讨论. 基于此, 本文从多节点的综合影响力角度出发, 基于 Rayleigh 熵机制, 首先分析了个体影响力和多节点的综合影响力之间的关系, 指出多节点的综合影响力小于单节点的影响力之和. 然后提出了一个指标刻画多节点的综合影响力, 并设计了一个高效的贪婪算法选择一组关键节点, 从而最大化多节点的综合影响力. 4 个真实网络上的信息传播实验验证了新算法的有效性.

关键词 传播, 传播源, 综合影响力, 复杂网络, 网络科学

1 引言

大量自然和人工系统的结构都可以抽象为由点和线组成的网络, 包含食物链网络、生态系统网络、交通网络、社交网络等 [1]. 一方面, 这些网络具有相似的拓扑结构, 比如无标度特征、群落结构、等级结构等 [2]; 另一方面这些网络也具有一些相似的动力学行为 [3]. 信息传播和扩散是网络中常见的一种动力学行为, 在不同类型的传播中 (SIS、SIRS、线性阈值模型 LT) [4~6], 信息以一定的概率从一个节点传递到另外一个节点, 传播概率一般存在一个临界阈值 [6], 如果传播概率高于该临界阈值, 初始少量信息会迅速扩散到整个网络; 反之, 信息在传递过程中会迅速消亡 (与初始状态无关). 在实际网络的拓扑结构分析中发现少量节点对整个网络的连通性具有重要影响, 并且网络中的信息传递主要通过这些关键节点的转发进行 [4,7]. 如果要对网络中的信息传播行为进行控制, 只需要对关键节点进行控制或者保护即可. 因此如何寻找高影响力节点是问题的关键.

仅根据网络拓扑结构选择若干个高影响力的节点是一个 NP 问题 [8,9]. 经典方法解决此类问题主要有两种方案: 启发式算法和基于目标函数的优化算法 [4]. 启发式算法一般根据节点的重要性进行

引用格式: 周明洋, 吴向阳, 曹扬, 等. 基于群体影响力的网络传播关键节点选择策略. 中国科学: 信息科学, 2019, 49: 1333–1342, doi: 10.1360/N112019-00041
Zhou M Y, Wu X Y, Cao Y, et al. A novel method to identify multiple influential nodes in complex networks (in Chinese). Sci Sin Inform, 2019, 49: 1333–1342, doi: 10.1360/N112019-00041

选择, 节点的重要性可通过节点度、介数、聚类系数、PageRank 指标进行确定 [7]. 根据使用信息的差异, 启发式算法可分为基于局部信息的重要性指标和基于全局拓扑信息的重要性指标 [4,10]. 基于局部信息的方法确定节点的重要性时仅根据节点的邻居节点或次近邻节点确定, 包含节点度、2 – 近邻、聚类系数等. 而基于全局拓扑信息的方法计算节点的重要性时需要整个网络的拓扑信息, 包括介数、PageRank、非回溯矩阵、随机游走等指标. 但基于启发式的算法结果和最优值差异较大, 并且启发式算法仅根据单个节点的影响力进行选择, 忽略了节点之间的耦合效应 [8], 例如在 K – 核算法 [11] 中, 单个高 K – 壳的节点影响力很大, 但多个高 K – 壳的节点综合影响力较小, 主要由于这些高影响力的节点之间具有相似的邻居, 影响范围有较大程度的重叠.

基于此, 一个合理的方法是考虑多节点的综合影响力, 通过设计一个综合影响力指标函数, 对其进行优化. 然而节点的综合影响力难以准确刻画. Morone 等 [8] 和 Szolnoki 等 [12] 通过分析稀疏网络并忽略网络中的环路, 理论上定义了节点的个体影响力, 但该方法在模型网络中可达到最优值, 在部分实际网络中效果较差, 因为实际网络中的环路较多, 忽略信息传播的环路效应导致理论分析与实际差异较大 [13]. Zhou 等 [14] 设计一个目标函数, 定义了节点之间的重叠影响力, 通过考虑重叠影响力最大化一组节点的综合影响力. 但该方法重叠影响力的定义理论不完善. 此外, 基于边渗流和博弈的方法选择重要节点均得到了广泛的研究 [15,16].

本文针对当前研究现状, 综合传播理论和重叠影响力分析, 将个体影响力的精确描述扩展到多节点的综合影响力, 在此基础上重新分析了多传播源的综合影响力. 然后通过设计一个贪婪算法选择关键节点用以最大化综合影响力. 算法结果在 SIS 传播模型中进行了验证, 在真实网络的实验表明本文提出的算法极大提高了多传播源的综合影响力. 并且指出多节点的综合影响力并不是单个节点的影响力线性累加, 需要考虑节点间的影响力重叠效应.

2 SIS 传播模型

经典的疾病传播模型主要指 SI, SIS, SIR 或 SIRS, 而信息传播模型主要指线性阈值模型等. 由于线性阈值模型的非线性和非连续特征导致解析困难, 本文主要分析连续时间 SIS 模型. 网络中的节点通常包含以下几个状态: S (susceptible) – 易感状态, I (infected) – 感染状态, R (recovered 或 removed) – 移除状态 [17]. 给定一个网络 $G=(V,E)$, 其中 V 表示顶点的集合, E 表示边集合. 网络 G 一般通过邻接矩阵 $A=(a_{ij})$ 表示, $a_{ij}=1$ 表示节点 i 和 j 之间有连接边, 否则 $a_{ij}=0$. 如果 G 是加权网络, a_{ij} 表示二个点之间的权重. 本文主要分析 SIS 传播模型, 在 SIS 传播模型中仅存在二个状态: 易感状态 S 和感染状态 I. 初始时刻只有非常小一部分节点处于感染状态 I, 其他节点处于易感状态 S. 在信息传播过程中个体只能将信息传递到它所能接触到的个体, 即通过邻居节点传播. 感染节点将信息传递到易感节点, 并以一定的概率恢复进入 S 状态. 假设节点 i 处于感染状态的概率为 $\rho_i(t)$, 节点将感染状态传递给邻居节点的概率为 ν, 同时感染节点的恢复速率为 μ, 理论研究表明整个网络是否处于感染状态只和 $\beta=\nu/\mu$ 相关 [18]. 实际的传播演化方程为

$$\frac{\mathrm{d}\rho_i(t)}{\mathrm{d}t}=\mu\rho_i(t)+\nu(1-\rho_i(t))\sum_{j\in N_i}\rho_j(t), \tag{1}$$

其中 N_i 表示节点 i 的邻居节点集合, β 表示信息传播率. 式 (1) 右边第 1 项表示节点 i 的恢复, 第 2 项表示节点 i 从邻居节点获得的传播信息. 理论研究表明如果 β 超过一临界阈值信息将扩散到整个网

络. 该临界阈值决定了信息是否会在网络中扩散或消亡, 即为 [19]

$$\beta_c = 1/\lambda_A, \tag{2}$$

其中 λ_A 为矩阵 A 的最大特征值. 注意到在网络的牵制控制问题中, 牵制控制能力与 Laplace 矩阵的非零最大特征根密切相关, 牵制控制采用网络的 Laplace 矩阵刻画, 而在本文传播问题中通过网络邻接矩阵刻画. Rong 等 [20, 21] 指出牵制节点选择需要考虑牵制集合与非牵制集合距离, Amani 等 [22] 进一步指出牵制控制中节点的重要性由 Laplace 矩阵最大特征值对应的特征向量决定, 进一步可得出单个节点的重要性可由最大特征值对应的特征向量 (eigenvector centrality, EC) 算法准确刻画, 但多个节点的综合影响力算法无法准确描述.

实证研究表明实际网络的 β_c 极小, 由于实际网络的无标度特征, 在网络规模增大时 $\beta_c \to 0$ [23], 表明实际网络的脆弱性, 这也解释了疾病的爆发行为、计算机病毒的快速扩散等. 为了提高网络的可靠性, 通常需要采取措施对网络进行保护, 一个合理的方式是仅保护网络中的关键节点, 问题的关键是如何定义关键节点以及如何选择这些节点.

通常对网络中的部分节点施加保护策略后, 这些受保护的节点不会再受到感染, 即永久性处于易感状态, $\rho_i(t) \to 0$. 实际上这些节点相当于不会再参与病毒的传播, 网络的传播行为受删除这些节点 (及其相连的边) 的剩余网络决定, 假设剩余网络为 A', 则网络的传播临界阈值变为 $\beta_c' = 1/\lambda_{A'}$. 寻找关键节点的行为变为如何最小化 $\lambda_{A'}$.

3 基于 Rayleigh 熵的重叠影响力算法

3.1 Rayleigh 熵与连接矩阵最大特征值

矩阵的 Rayleigh 熵 [24] 定义为

$$R(A, x) = \frac{x^{\mathrm{T}} A x}{x^{\mathrm{T}} x}, \tag{3}$$

如果 x 为矩阵 A 最大特征值对应的特征向量, 则 $R(A, x) = \lambda_A$, 否则 $R(A, x) < \lambda_A$, 这里 x 可以为任意向量. 如果将 A 替换为 A^k, 则可以推出

$$\lim_{k\to\infty} x^{\mathrm{T}} A^k x/(x^{\mathrm{T}} x) = \lambda_A^k \alpha, \tag{4}$$

其中 α 为向量 x 沿矩阵 A 最大特征向量的分量. 对关键节点进行免疫后, 即删除关键节点对应的连接关系, 矩阵 A 退化为 A', 剩余矩阵对应的 Rayleigh 熵为 $R(A', x)$. 寻找最优节点要求最小化 $R(A', x)$ 的最大特征值, 即 $\min_{A'}\{\max_x R(A', x)\}$. 也意味着对于给定的任意向量 x, 要求最小化 $R(A'^{,k}, x)$, 即最小化式 (4), $x^{\mathrm{T}} A'^{,k} x$, $|x| = 1$.

这里选定一组特殊值 $x = [d_1, d_2, \ldots, d_n]'$, 其中 d_i 为节点 i 的度, 实际上 $x = A\mathbf{1}$, $\mathbf{1}$为全 1 向量. 当仅有 x 为矩阵 A 的最大特征值对应的特征向量时 $x^{\mathrm{T}} A^k x$ 才能达到极值, 这里选择 x 的元素为节点的度主要由于节点度和矩阵最大特征向量有很大的相关性 [25~27], 这组特殊值也具有较好的效果. 这里的目标为最小化 $x^{\mathrm{T}} A'^{,k} x$, $k \to \infty$. 实际上 k 比较大时难以计算, 这里设置 $k = 2$, 这时

$$x^{\mathrm{T}} A^2 x = \sum_{i,j,k} a_{ij} a_{jk} x_i x_k, \tag{5}$$

实际上 a_{ij} 和 a_{jk} 表示网络中相邻的边 – 对数量 (i 和 k 可相等).

3.2 节点的个体影响力和节点之间的重叠影响力

现在考虑关键节点选择策略, 最小化矩阵 A' 的最大特征值实际上要求最小化式 (4), 在 $k=2$ 下, 可近似表达为最小化式 (5). 现在考虑单个节点的影响力, 首先定义边的影响力如下所示.

定义1 一条边的影响力定义为删除该边后, $x^{\mathrm{T}}A^2x$ (式 (5)) 下降的数量, $x=A\mathbf{1}$. 根据定义, 边 e_{ij} 的影响力为

$$S_{e_{ij}}=d_i\sum_{k\in N_j}d_k+d_j\sum_{k\in N_i}d_k. \tag{6}$$

删除网络中的一个节点实际上是删除连接该节点的所有边, 因此单个节点的影响力定义如下.

定义2 单个节点的影响力定义为删除该节点后, $x^{\mathrm{T}}A^2x$ (式 (5)) 下降的数量, $x=A\mathbf{1}$. 根据定义单个节点 m 的影响力实际上是连接该节点的边所有影响力之和,

$$S_m=d_m\sum_{j,k}a_{mj}a_{jk}d_k+\sum_{j,k}a_{jm}a_{mk}d_jd_k. \tag{7}$$

这里节点 k 指随机游走过程中经过 2 步能够跳转到的节点, 包含该节点本身, 因为 2 步随机游走可以从邻居节点返回该节点本身.

在完全稀疏的网络中忽略网络中边环路的情况下可推出节点的影响力为 [8]

$$S_m=(d_m-1)\sum_{j\in\text{Ball-2}}(d_j-1), \tag{8}$$

其中 Ball-2 是距离节点 m 为 2 的点集合, 式 (8) 是根据边渗流理论得出的节点影响力指标, 而式 (7) 是根据 Rayleigh 熵得出的针对任意网络下的单节点影响力指标. 实际上式 (8) 是式 (7) 在树状网络下的特殊情况.

以图 1 为例, 这里人工构造一个小型树状结构网络 (图 1(a) 所示), 在经典的渗流理论中忽略网络中的环路, 节点影响力由式 (8) 刻画, 而在存在环路的网络中 (图 1(b) 所示) 节点影响力由式 (7) 刻画, 实际信息传播过程中, 以节点 1 作为传播源, 信息沿着 $1\to4\to6$ 传播, 也可以反向沿着 $1\to6\to4$ 传播, 当环路的数量继续增加时, 信息传递方向是无序的, 本文提出的式 (7) 能够刻画这种无序行为, 而式 (8) 是在忽略网络环路, 即忽略无序传递信息情况下的简化形式.

接下来分析节点间的重叠影响力, 假设删除的两个节点间距离大于 2, 在计算式 (5) 过程中两个节点之间不会重现效应, 根据式 (7) 的分析, 两个节点的影响力等于单节点的影响力之和. 当两个节点间的距离小于 2 时, 比如删除节点 i 和 k, 通过中间节点 j, $a_{ij}\ a_{jk}\neq0$, 单独计算节点 i 和 k 的影响力时, $a_{ij}a_{jk}d_id_k$ 各计算一次, 而在计算两节点的综合影响力时, $a_{ij}a_{jk}d_id_k$ 只被计算一次, 因此重叠了一次. 更进一步, 当两个节点之间存在连边时 (距离为 1), 单个节点的影响力等于该节点依附的边影响力之和, 多节点的影响力等于该节点集合依附的边影响力之和, 而共同连边的影响力在计算时仅被计算一次, 因此也重叠了一次. 因而在选择关键节点时应该考虑整体的综合影响力, 而不是个体的单独影响力.

根据综合影响力的定义和重叠影响力分析, 这里设计一个贪婪算法选择关键节点. 假设节点数量 L, 初始时刻关键节点集合为空, 然后通过贪婪策略, 每次新增一个节点, 这个新增加的节点具有最大的个体影响力 (式 (7)), 具体如算法 1 所示.

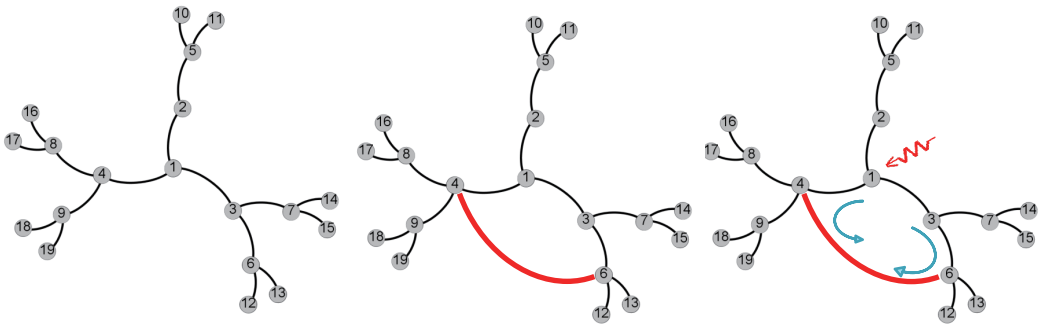


图 1 (网络版彩图) 稀疏网络和稠密网络的节点影响力对比图. (a) 人工树状网络结构图, 网络中不存在环路, 单节点的影响力为式 (8); (b) 增加一条边 (红色), 网络中存在环路, 网络的影响力由式 (7) 刻画; (c) 以节点 1 作为传播源, 信息沿着环路会出现反向传播, 而在树状网络中不会出现回环效应.

Figure 1 (Color online) Comparison of the information spread in networks with loops. (a) An artificial network; (b) the back tracking edge (red) that is neglected in Eq. (7); (c) the back path of the information spread in networks with loops.

Algorithm 1 The pseudo code to compute the set of influential nodes

Input: The adjacent matrix $A = (a_{ij})_{N\times N}$ and the size of influential nodes L.

1: Data preparation. Extract the giant component of the original network and remove the isolate nodes and other small nodes clusters because nodes located in distinct clusters could not exchange information.
2: Initialize the algorithm and calculate the importance of nodes based on Eq. (7).
3: Rank nodes based on the importance of nodes and choose the particular node with the largest importance score.
4: If the size of influential nodes is smaller than L, remove the edges attached to the influential nodes and go to step 2; otherwise go to step 5.
5: Return the index of the chosen influential nodes.

Output: The index of the selected influential nodes.

4 基准方法和评价指标

4.1 基准算法

这里选择常用的 5 种方法作为衡量算法性能的基准: 节点度、节点介数、K – 壳、稀疏矩阵综合影响力、非回溯矩阵分析 [4], 下面分别进行介绍.

节点度 (HD, high degree). 该算法根据节点的度进行排序, 依次选择度大的节点.

节点介数 (BW, betweenness) [4,7]. 节点的介数定义为网络中所有最短路径中经过该节点的路径的数量占最短路径总数的比例. 该算法依次选择介数大的节点作为关键节点.

K – 壳 (K-shell) [11]. K-shell 和 K-core 成对出现. 网络中反复删除度小于 K 的节点, 直至所有节点的度大于等于 K 为止, 剩余的网络即为该网络的 K-core. 如果一个节点存在于 K-core 中而不存在于 $(K+1)$-core 中, 则该节点位于 K-shell 中. 该算法根据节点的 K-shell 依次选择 K-shell 大的节点.

稀疏矩阵综合影响力 (CI, collective influence) [8]. 该算法通过边渗流理论, 在忽略网络中环路的情况下, 仅考虑网络为树状网络推导出单个节点的影响力为式 (8). 这里该算法选择关键节点时根据单个节点的影响力进行排序, 依次选择关键节点.

表 1　数据集特征基本描述 (节点数 N, 边数 E, 异质性指标 $H=\langle k^2\rangle/\langle k\rangle^2$, 同配异配指数 (度度相关性) r, 聚类系数 $\langle C\rangle$, 节点间平均距离 $\langle d\rangle$, 稀疏性 Sparsity $=2E/(N(N-1))$)

Table 1 Structural properties of the different real networks, including network size (N), link number (E), degree heterogeneity ($H=\langle k^2\rangle/\langle k\rangle^2$), degree assortativity (r), average clustering coefficient ($\langle C\rangle$), average shortest path length ($\langle d\rangle$) and sparsity

Network	N	E	H	r	$\langle C\rangle$	$\langle d\rangle$	Sparsity
Airtraffic	1225	2399	1.87	−0.0177	0.011	5.93	3.2×10^{-3}
Bitcoin	5880	21228	10.89	−0.163	0.022	3.59	1.2×10^{-3}
ca-HepTh	5039	11346	2.03	0.195	0.101	6.54	8.9×10^{-4}
Reactome	3851	140106	1.99	0.316	0.508	0.508	1.9×10^{-2}

非回溯矩阵分析 (NBM, non-backtracking matrix) [28]. 非回溯矩阵分析了基于边的传播特征, 定义了一个矩阵 $B_{2E\times2E}$, 其中 E 为网络中边的数量, 对于每一个元素 $B_{i\leftarrow j,k\leftarrow l}=\delta_{jk}(1-\delta_{il})$, $\delta_{jk}=1$ 如果 $j=k$, 否则 $\delta_{jk}=0$. 节点的影响力为其连接的边影响力之和, 矩阵 B 维度太高通常难以计算, 经过化简, 可得节点的影响力为以下 M 矩阵最大特征值对应特征向量的前 n 个元素:

$$M=\begin{pmatrix} A & I-D \\ I & 0 \end{pmatrix}, \tag{9}$$

NBM 方法根据 M 矩阵计算的节点影响力排序获取关键节点.

4.2　算法评价指标

评价关键节点影响力的核心标准是删除关键节点后剩余邻接矩阵的最大特征值, 即式 (2).

在实际网络的分析中发现基于 K – 壳的分解中具有高 K – 壳的节点影响力大. 但实证数据表明多个高 K – 壳节点影响力较小, 主要由于高 K – 壳节点之间的距离很近, 节点之间的重叠影响力较大, 从而高 K – 壳节点综合影响力并不大. 因此这里通过关键节点的平均距离进行分析, 描述不同方法选择的关键节点特征, 这里在计算平均距离时, 任意两节点的距离通过未删除关键节点的原始网络进行计算.

5　实验结果和分析

本文提出的算法 (算法 1) 分别在 4 个数据集上进行了验证. 这 4 个数据集来自 Stanford 和 Konect 的开放数据库 [29, 30], 分别为: Airtraffic, Bitcoin, ca-HepTh, Reactome. Airtraffic 是美国联邦航空管理局开放的航空信息数据, 网络中的每个节点代表一个机场或者服务中心, 每条边代表该中心推荐的服务关系. Bitcoin[1)]是用户和用户之间的相互依赖关系网络, 这是一个带权重的符号关系网络. ca-HepTh 是 Arxiv 上的高能物理领域的作者合作关系网络. Reactome 是智人的蛋白质关系网络. 不同类型的网络数据的边有可能存在有向边和加权边, 甚至存在自环 (指向自身的边), 实验中对数据进行了预处理, 忽略边的方向和权重, 同时删除网络中的孤立节点和孤立簇, 仅保留网络中的最大连通子图. 经过处理后网络结构特征如表 1 所示.

本文算法基于 Rayleigh 熵, 实验中用 EC 表示算法 1 的结果. 首先分析剩余网络的最大特征值与关键节点数量的关系 (图 2 所示), $\delta=L/N$, 随着免疫关键节点数量的增加, 剩余网络最大特征值逐渐

1) https://www.bitcoin-otc.com/.

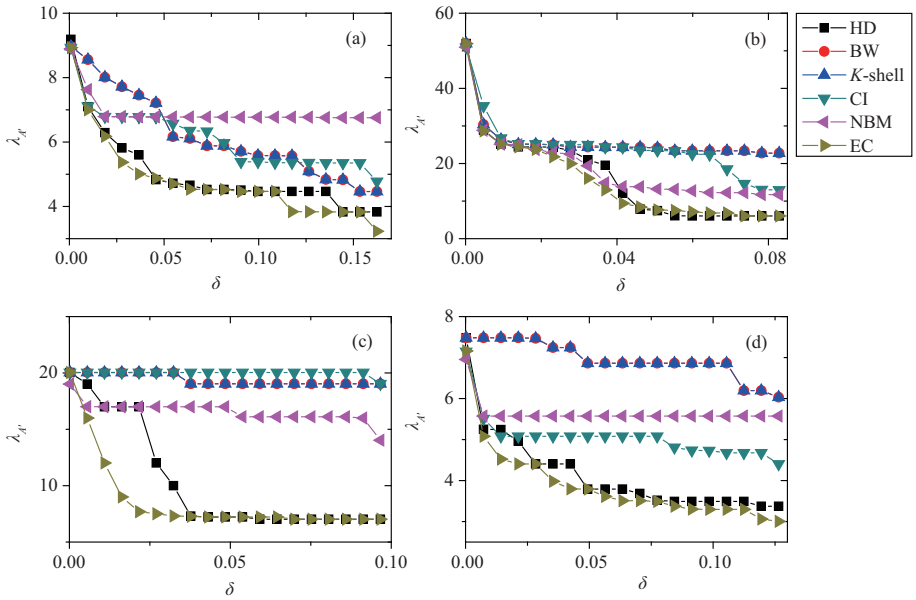


图 2 (网络版彩图) 剩余网络最大特征值 $\lambda_{A'}$ 与关键节点比例 δ 的关系

Figure 2 (Color online) The principle eigenvalues of the remaining networks that are obtained by removing the edges attached to the influential nodes. (a) Airtraffic network; (b) Bitcoin network; (c) ca-HepTh network; (d) Reactome network.

下降 ($\lambda_{A'}$ 越小越好). 注意到本实验中 CI 方法效果较差, 而在稀疏网络中效果很好, 主要由于实际网络具有复杂的内部结构, 比如聚类特征、度度相关性、富人俱乐部特征等, 导致实际网络中会形成局部稠密的子网络, 而在 CI 算法中 (式 (8)) 忽略了网络中信息传递的环路特征 (图 1(c)), 因此效果较差. 而本文提出的算法 (式 (7)) 考虑了环路信息的传递, 因此整体的性能相比 CI 算法得到了提高. 另外, 注意到基于节点度的算法 HD 也具有较好的实验结果 (图 2(c) 除外), 但基于节点度的方法仅在部分网络表现较好, 随着网络结构的变化, 该算法性能稳定性不高 [7]. 同时在 4 个网络中基于介数的方法和 K – 壳算法重叠, 这种重叠行为也会随着网络的变化而变化. 注意到在图 2(c) 中本文提出的算法在 $\delta < 0.025$ 时 $\lambda_{A'}$ 有很大降低, 表明在一些特定结构下算法能够挖掘出较好的关键节点, 而在一般情况下该算法也具有较高的稳定性和较好的性能.

进一步, 图 3 通过计算机仿真了实际网络的传播过程, 在免疫关键节点后的剩余网络中, 初始时刻假设网络中 0.1% 的节点处于感染状态, 在传播模型 (式 (1)) 中设置 $\beta = 0.6$, 仿真过程中时间间隔 $\delta t = 0.001$, 实验为 50 次实验平均值. 图 3 的结果和图 2 基本一致, 本文提出的 EC 算法选择的关键节点具有较好的性能, 通过免疫这些关键节点, 在临界点处剩余节点受感染的速度和比例比经典方法低, 表明了 EC 算法针对 SIS 传播模型具有较好的保护性能.

最后分析关键节点的特征, 这里主要分析关键节点之间的平均距离, 如图 4 所示. 综合图 2 和 4 发现基于非回溯矩阵的方法 NBM 选择的节点平均距离小, 性能较差, 在图 2(c) 中 CI 方法效果最差, 同时在图 4(c) 中平均距离最小. 表明关键节点之间的平均距离越小, 综合影响力越低. 因此为了提高多节点的综合影响力, 应该适当增加传播源之间的距离. 但是距离并不是越大越好, 新算法 EC 在图 2 中具有最优的 $\lambda_{A'}$, 在图 4 中平均距离并不是最大, 因此在节点的平均距离和综合影响力之间需要合适的平衡机制, 增加节点平均距离有助于降低重叠影响力, 在考虑重叠影响力的同时也需要分析综合影响力.

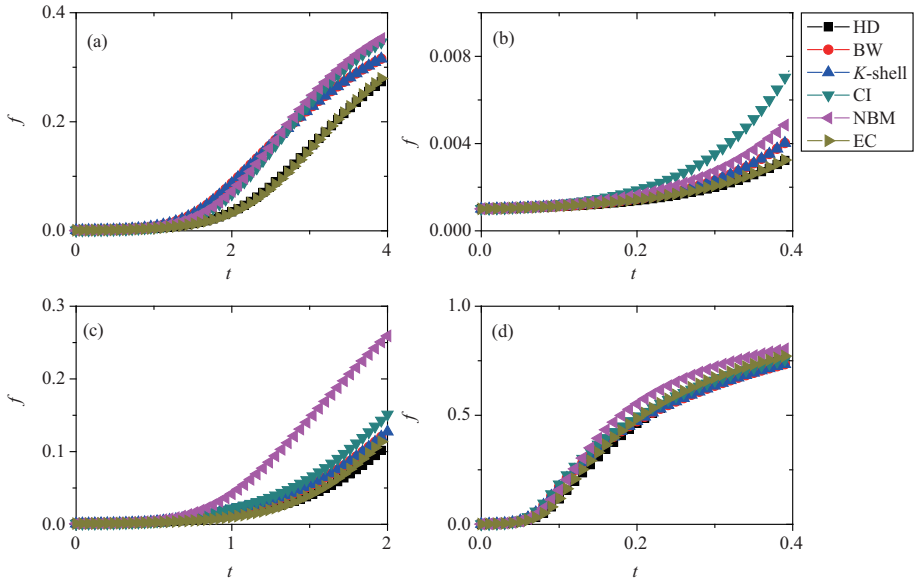


图 3 (网络版彩图) 稀实际网络的传播过程仿真, 纵坐标为整个网络受感染节点比例, 实验为 50 次实验平均值

Figure 3 (Color online) The evolving paths of the spread as a function of time. In the experiments, every node has a probability of 0.001 to be an infected node initially. The results are the average of 50 independent simulations. (a) Airtraffic network; (b) Bitcoin network; (c) ca-HepTh network; (d) Reactome network.

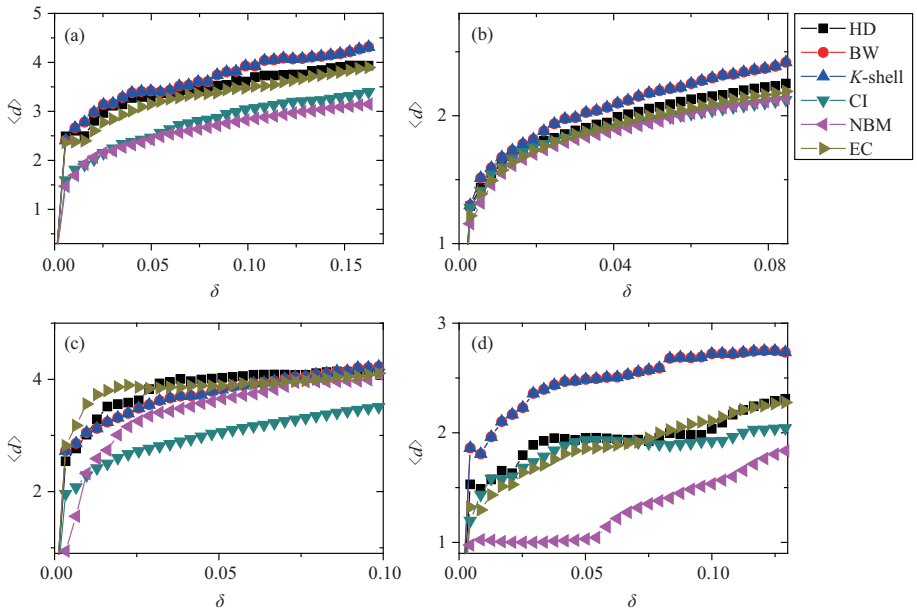


图 4 (网络版彩图) 关键节点之间的平均距离与关键节点比例 δ 的关系

Figure 4 (Color online) The average distance between influential nodes. (a) Airtraffic network; (b) Bitcoin network; (c) ca-HepTh network; (d) Reactome network.

6 总结

本文基于 SIS 传播模型分析了传播的关键节点选择策略, 在 Rayleigh 熵基础上考虑网络邻接矩

阵的高阶近似, 分析了网络最优节点选择策略和 Rayleigh 熵的关系. 在此基础上提出了关键节点的综合影响力和重叠影响力, 并分析了二者之间的关系. 进一步, 基于多节点的综合影响力, 本文提出了一个新算法选择传播过程中的关键节点, 通过与经典算法进行对比发现本方法优于目前经典的基于综合影响力和基于非回溯矩阵的方法. 同时论证了多节点的综合影响力小于单节点的影响力之和, 这是由于节点影响力之间的重叠效应. 本文从新的角度和思路分析传播中的关键节点选择问题: 从多节点的综合影响力和重叠影响力角度分析节点集合的传播问题, 为关键节点挖掘提供了一个新的视野, 本文的相关结果也适用于线性阈值模型等其他信息传播模型, 为政务舆情、口碑等消息的使用提供了更有效的方法.

参考文献

1 Wang X F, Li X, Chen G R. Network Science: An Introduction. Beijing: Higher Education Press, 2012 [汪小帆, 李翔, 陈关荣. 网络科学导论. 北京: 高等教育出版社, 2012]
2 Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics. Phys Rep, 2006, 424: 175–308
3 Liu Y Y, Barabási A L. Control principles of complex systems. Rev Mod Phys, 2016, 88: 035006
4 Lü L Y, Chen D, Ren X L, et al. Vital nodes identification in complex networks. Phys Rep, 2016, 650: 1–63
5 Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003. 137–146
6 Nowzari C, Preciado V M, Pappas G J. Analysis and control of epidemics: a survey of spreading processes on complex networks. IEEE Control Syst, 2016, 36: 26–46
7 Liao H, Mariani M S, Medo M, et al. Ranking in evolving complex networks. Phys Rep, 2017, 689: 1–54
8 Morone F, Makse H A. Influence maximization in complex networks through optimal percolation. Nature, 2015, 524: 65–68
9 Goyal A, Lu W, Lakshmanan L V. SIMPATH: an efficient algorithm for influence maximization under the linear threshold model. In: Proceedings of the 11th International Conference on Data Mining, 2011. 211–220
10 Zhang Z K, Liu C, Zhan X X, et al. Dynamics of information diffusion and its applications on complex networks. Phys Rep, 2016, 651: 1–34
11 Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. Nat Phys, 2010, 6: 888–893
12 Szolnoki A, Perc M. Collective influence in evolutionary social dilemmas. EPL-Europhys Lett, 2016, 113: 58004
13 Liu Y, Tang M, Zhou T, et al. Identify influential spreaders in complex networks, the role of neighborhood. Phys A-Stat Mech Appl, 2016, 452: 289–298
14 Zhou M Y, Xiong W M, Wu X Y, et al. Overlapping influence inspires the selection of multiple spreaders in complex networks. Phys A-Stat Mech Appl, 2018, 508: 76–83
15 Zdeborová L, Krzakala F. Statistical physics of inference: thresholds and algorithms. Adv Phys, 2016, 65: 453–552
16 Wang W, Tang M, Stanley H E, et al. Unification of theoretical approaches for epidemic spreading on complex networks. Rep Prog Phys, 2017, 80: 036603
17 Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks. Phys Rev E, 2001, 63: 066117
18 Zhou M Y, Xiong W M, Liao H, et al. Analytical connection between thresholds and immunization strategies of SIS model in random networks. Chaos, 2018, 28: 051101
19 Chakrabarti D, Wang Y, Wang C, et al. Epidemic thresholds in real networks. ACM Trans Inf Syst Secur, 2008, 10: 1–26
20 Lu W, Li X, Rong Z. Global stabilization of complex networks with digraph topologies via a local pinning algorithm. Automatica, 2010, 46: 116–121
21 Rong Z H, Li X, Lu W L. Pinning a complex network through the betweenness centrality strategy. In: Proceedings of IEEE International Symposium on Circuits and Systems, 2009. 1689–1692
22 Amani A M, Jalili M, Yu X H, et al. Finding the most influential nodes in pinning controllability of complex networks. IEEE Trans Circ Syst II, 2017, 64: 685–689
23 Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys Rev Lett, 2001, 86: 3200–3203
24 Parlett B N. The rayleigh quotient iteration and some generalizations for nonnormal matrices. Math Comput, 1974,

28: 679
25 Sarkar C, Jalan S. Spectral properties of complex networks. Chaos, 2018, 28: 102101
26 Bonacich P. Some unique properties of eigenvector centrality. Soc Netw, 2007, 29: 555–564
27 Restrepo J G, Ott E, Hunt B R. Approximating the largest eigenvalue of network adjacency matrices. Phys Rev E, 2007, 76: 056119
28 Krzakala F, Moore C, Mossel E, et al. Spectral redemption in clustering sparse networks. Proc Natl Acad Sci USA, 2013, 110: 20935–20940
29 Leskovec J, Krevl A. SNAP datasets: Stanford large network dataset collection. 2014. http://snap.stanford.edu/data/
30 The Koblenz network collection. 2016. http://konect.uni-koblenz.de/networks/

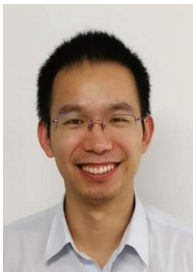
A novel method to identify multiple influential nodes in complex networks

Mingyang ZHOU[1,2], Xiangyang WU[2], Yang CAO[1], Liao LUO[1], Xiaoyu LI[2], Hao LIAO[1,2*] & Binghong WANG[3]

1. *CETC Big Data Research Institute Co., Ltd., Guiyang* 550022*, China*;
2. *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen* 518060*, China*;
3. *Department of Modern Physics, University of Science and Technology of China, Hefei* 230027*, China*
* Corresponding author. E-mail: jamesliao520@gmail.com

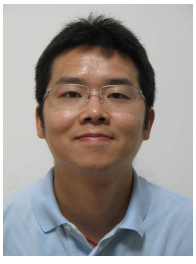
Abstract In the spread of information, a small fraction of nodes play an important role in the dynamics; therefore, detecting these influential nodes helps control the spread of information. However, classical methods can choose a single influential node yet fail in the case of multiple influential nodes. In this paper, we analyze the collective influence and overlapping influence of multiple spreaders. Further, based on the overlapping influence between spreaders, we clarify the problem of why multiple influential spreaders may have a low collective influence. Finally, a new algorithm is proposed to choose multiple spreaders, and the performance of the algorithm is validated on real networks.

Keywords spreading, spreaders, influence of grouping, complex networks, network science



Mingyang ZHOU was born in 1987. He got the bachelor's and Ph.D. degrees from the University of Science and Technology of China in 2010 and 2016, respectively. Since 2016, he has been working at Shenzhen University as an assistant professor. His research interests include complex network, network control, and data mining.

Liao LUO was born in 1991. He received the master's degree from Nankai University, Tianjin, China. He joined the Common Technology Department of CETC Big Data Research Institute Co., Ltd. as an algorithm engineer. His research interests mainly include natural language processing and data mining.



Hao LIAO was born in 1987. He got the Ph.D. degree from Fribourg University, Switzland in 2015. He is now a teacher at Shenzhen University. His main research interests include information networks, data mining, and complex networks.

Binghong WANG was born in 1944, and is now a professor at the University of Science and Technology of China. He has taken charge of about 30 research projects including projects supported by the National Natural Science Foundation of China or National 973 Plan . His research interests mainly include statistic physics, nonlinear dynamics, adaptive complex system, economic physics, traffic flow and particle flow, and chaotic system.