

基于引力方法的复杂网络节点重要度评估方法*

阮逸润[†] 老松杨 汤俊 白亮 郭延明

(国防科技大学系统工程学院, 长沙 410073)

(2022 年 3 月 28 日收到; 2022 年 5 月 1 日收到修改稿)

如何用定量分析的方法识别复杂网络中哪些节点最重要, 或评价某个节点相对于其他一个或多个节点的重要程度, 是复杂网络研究的热点问题. 目前已有多种有效模型被提出用于识别网络重要节点. 其中, 引力模型将节点的核数 (网络进行 k -核分解时的 ks 值) 看作物体的质量, 将节点间的最短距离看作物体间距离, 综合考虑了节点局部信息和路径信息用于识别网络重要节点. 然而, 仅将节点核数表示为物体的质量考虑的因素较为单一, 同时已有研究表明网络在进行 k -核分解时容易将具有局部高聚簇特征的一类核团节点识别为核心节点, 导致算法不够精确. 基于引力方法, 综合考虑节点 H 指数、节点核数以及节点的结构洞位置, 本文提出了基于结构洞引力模型的改进算法 (improved gravity method based on structure hole method, ISM) 及其扩展算法 ISM₊. 在多个经典的实际网络和人工网络上利用 SIR (susceptible-infected-recovered) 模型对传播过程进行仿真, 结果表明所提算法与其他中心性指标相比能够更好地识别复杂网络中的重要节点.

关键词: 复杂网络, 传播影响力, 引力模型, H 指数, k -核分解**PACS:** 64.60.aq, 89.75.Hc, 89.75.Fb**DOI:** 10.7498/aps.71.20220565

1 引言

网络节点重要性排序是网络科学领域研究的重点和热点, 是为了挖掘能在更大程度上影响网络结构和功能的关键节点^[1]. 设计能够快速、准确地识别网络关键节点的算法在理论研究和生活实践上都具有重要意义. 例如对病毒传播网络, 有选择性地控制网络中的一些重要节点或改变其结构属性, 如接种疫苗、断边重连或漏洞修复等^[2,3], 就可以有效降低病毒的传播速度并减小扩散范围; 在军事供应链网络中, 寻找关键节点并进行重点保护, 可以提高物资保障的可靠性和效率, 有效完成后勤保障任务; 在社交网络中, 通过一定策略选择有影响力的用户 (如明星、网络红人等) 做新产品的推广和营销, 使产品信息在网络中得到大范围传播从而增加营收效益^[4].

关于如何挖掘网络关键节点, 已经有了许多研究成果, 典型的指标有度中心性 (degree)^[5]、半局部度 (semi-local)^[6]、接近中心性 (closeness)^[7]、介数中心性 (betweenness)^[8]、 k -核分解方法 (k -shell decomposition)^[9] 和 H 指数^[10] 等, 度中心性指标考虑了节点的直接邻居数量, 虽然简单直观, 但却把每一个邻居节点看作是同等重要的, 而实际上邻居节点间存在差异, 不同的邻居对于目标节点的重要性可能大不相同, 因而在很多场景下不够精确. 半局部度指标考虑了节点 4 层邻居的信息, 在提高算法精度的同时还兼顾了算法的效率. 接近中心性和介数中心性都假设网络中的信息是基于最短路径进行传播, 实际上多数真实场景下信息传播具有随机性. k -核分解方法认为网络节点的重要性由节点在网络中的位置所决定, 节点越接近核心层重要性越高, 边缘节点重要性最低. k -核分解方法计算复杂度低, 适用于大型复杂网络, 可以很好地应用于

* 国家自然科学基金 (批准号: 72101265) 资助的课题.

[†] 通信作者. E-mail: ruanyirun@163.com

寻找疾病传播网络中最有影响力的节点,但由于无法区分处于同一壳层节点的重要性,因此通常被认为是一种粗粒化的排序方法,随后提出了许多改进的策略,如领域核数算法^[11]及混合度分解 (mixed degree decomposition, MDD)^[12]等. H 指数表示一个节点的 H 指数如果是 h , 就说明这个节点至少有 h 个邻居, 且它们的度都不小于 h , H 指数在一些场景中的综合表现要好于度和核数.

最近有学者指出, 通过对不同的排序指标或策略进行融合可以获得更好的排序结果^[13]. 目前大多数指标都是从某一特定角度衡量节点重要性, 有一定适用性的同时也有一定的不足. 如果可以将一些从不同角度对节点重要性进行评价的指标进行融合, 则排序结果将更加全面和可信^[14]. 韩忠民等^[15]基于 ListNet 的排序学习方法融合结构洞、介数等 7 个度量指标, 能够较为全面地评估网络中节点的重要性. Wang 等^[16]设计了一种基于节点位置和邻域信息的多属性排序方法, 该方法利用 k -核分解中的迭代信息来进一步区分节点位置, 并充分考虑邻域对节点影响能力的作用, 具有较低的计算复杂度. 闫光辉等^[17]以网络模体^[18,19]为基本单元研究网络高阶结构, 并进一步引入证据理论^[20,21]设计了一种融合节点高阶信息和低阶结构信息的重要节点挖掘算法. 根据渗流理论^[22], 去除一个网络节点后, 剩余网络与原始网络之间存在传播阈值上的差异, Zhong 等^[23]认为这种传播阈值差异可以用于表征节点的全局影响力, 通过考虑传播阈值差异和度中心性, 提出了一种融合局部与全局结构的重要节点识别算法.

受到万有引力公式启发, Ma 等^[24]提出了一种综合考虑节点邻居信息和路径信息的引力方法, 其中节点核数被看作节点的质量, 节点间的最短距离看作物体间距离. 然而, 仅将核数表示为物体的质量, 考虑的因素较为单一. 此外, 算法利用节点与邻域节点间的相互作用力来量化节点的影响力, 容易将局部呈高聚簇特征的节点误判为重要度高的节点, 实际上传播从这类节点发起, 容易局限在小团体内部, 不利于传播快速向外部蔓延. 由此, 本文将节点核数作为度量节点全局重要性的指标, 融合节点 H 指数重新定义节点的质量, 并结合节点的结构洞特征, 设计了引力模型的改进算法 ISM 及 ISM₊. 在多个真实世界网络和人工网络中的实

验表明, 所提算法在识别节点影响力方面相比介数中心性、接近中心性、度中心性, 引力模型, MDD, 局部引力模型^[25]以及基于 k -核分解方法的引力模型 (KSGC) 指标^[26]等算法更有优势.

2 相关概念

对于给定的复杂网络 $G = (N, E)$, 其中 N 表示节点集, E 表示边集, 网络的拓扑结构通常用邻接矩阵 $A = (a_{ij})_{N \times N}$ 表示. 邻接矩阵中的元素 a_{ij} 可以描述节点之间的连接关系, $a_{ij} = 1$ 表示节点 i 和节点 j 之间存在连接边, 否则 $a_{ij} = 0$.

2.1 度中心性、接近中心性和介数中心性

度排序方法^[5]最为简单直观, 表示节点的邻居数量, 表示为

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

度指标反映了节点的直接影响力, 节点上的链接数越多, 节点度 k_i 越大, 因为只考虑了节点局部信息, 因而是一种局部中心性指标.

接近中心性^[7]认为一个节点与网络中其他节点的平均距离越小, 节点重要性越高, 表示为

$$CC_i = \frac{N}{\sum_{j=1}^N d_{ij}}, \quad (2)$$

其中, d_{ij} 代表节点 i 和 j 之间的距离, N 表示网络节点数.

介数中心性^[8]描述了节点对网络中沿最短路径传播的信息流的控制力, 定义为

$$BC_i = \sum_{s \neq i \neq t \in V} \frac{n_{st}^i}{g_{st}}, \quad (3)$$

其中, g_{st} 表示网络中除了节点 i 以外任意节点对 (如节点 s 和节点 t) 之间的最短路径数, n_{st}^i 表示当中经过节点 i 的最短路径数.

2.2 H 指数

H 指数^[10]最初用于度量一个科学家最多有多少篇论文且每篇被引用的次数都不少于这个篇数, Lü 等^[10]将其引用到网络中, 认为一个节点的 H 指数如果是 h , 就说明这个节点有 h 个邻居, 它们的

度都不小于 h , 表示为

$$H_i = \xi(k_{j_1}, k_{j_1}, \dots, k_{j_s}, \dots, k_{j_{k_i}}), \quad (4)$$

其中, k_{j_s} 表示节点 i 的第 s 个邻居的度数. 在 (4) 式中, 算子 H 返回最大整数 h , 使得节点 i 至少有 h 个邻居的度数不低于 h .

2.3 结构洞理论

结构洞^[27]指网络结构中不存在冗余联系的两个人之间的缺口, 网络中占据结构洞位置的个体相比其邻居节点可以获得更多的竞争优势, 包括信息优势和控制优势, 从而影响甚至控制社会关系与信息的传播. 为了量化结构洞节点对这些关系的控制, Burt^[27]提出网络约束系数这一定量化指标来衡量节点形成结构洞所受到的约束, 表示为

$$c_i = \sum_j \left(\mu_{ij} + \sum_{q(\neq i,j)} \mu_{iq} \mu_{qj} \right)^2, \quad (5)$$

其中, 节点 q 表示 i 和 j 之间的共同邻居, μ_{ij} 表示节点 i 为维持与节点 j 的关系而投入的精力占总精力的比例.

$$\mu_{ij} = z_{ij} / \sum_{j \in \Gamma(i)} z_{ij}, \quad (6)$$

式中, $\Gamma(i)$ 表示节点 i 的邻居集合, 当 i 和 j 之间存在连边时, $z_{ij} = 1$, 反之 $z_{ij} = 0$.

2.4 引力模型、局部引力模型以及 KSGC 指标模型

Ma 等^[22]认为如果节点的邻域节点具有更高的 ks 值, 则节点更有可能是网络中的核心节点; 另一方面, 两个节点之间的相互作用效应会随距离的增加而减小. 通过将节点的 ks 值看作节点的质量, 节点间的最短距离看作物体间距离, 提出了一种综合考虑节点邻居信息和路径信息的节点重要性排序指标, 表示为

$$G(i) = \sum_{j \in \varphi_i} \left(\frac{ks_i ks_j}{d_{ij}^2} \right), \quad (7)$$

其中, φ_i 表示距离节点 i 小于或等于给定值 r 的邻域节点集, ks_i 和 ks_j 分别表示节点 i 和 j 的 k -核分解值, d_{ij} 表示节点 i 到节点 j 的距离. 根据 (7) 式进一步扩展得到扩展引力中心性指标指数标记为 (Gravity₊), 其定义为

$$G_+(i) = \sum_{j \in A_i} G(j), \quad (8)$$

A_i 表示节点 i 的直接邻居.

类似于引力中心性指标, Li 等^[25]认为度大的节点往往有更大的影响力, 同时节点对其邻近节点的影响更大, 将节点的度看作物体的质量, 由此也提出了一种综合考虑节点邻居信息和路径信息的局部引力模型来评估网络节点的重要性, 定义为

$$\text{LGM}(i) = \sum_{d_{ij} \leq R} \frac{k_i k_j}{d_{ij}^2}, \quad (9)$$

其中, k_i 和 k_j 分别表示节点 i 和 j 的度, R 表示网络截断半径, 是网络最短路径平均值的一半.

Yang 等^[26]指出节点的位置是节点在网络中的一个重要属性, 而多数节点重要性评估算法却很少考虑节点的位置. 由此他们设计了一种基于 k -核分解方法的引力模型的改进方法 KSGC, 用于识别复杂网络中节点的传播影响力, 表示为

$$\text{KSGC}(i) = \sum_{d_{ij} \leq R} c_{ij} \frac{k_i k_j}{d_{ij}^2}, \quad (10)$$

其中, $c_{ij} = e^{\frac{ks_i - k_i}{ks_{\max} - ks_{\min}}}$, ks_{\max} 表示网络中最大的 ks 值, ks_{\min} 表示最小的 ks 值.

3 算法设计与评价标准

3.1 基于引力方法的节点重要性排序方法

引力模型仅将核数表示为物体的质量, 考虑的因素较为单一, 节点在网络中的位置, 是节点的重要属性, 这里的位置不仅指节点基于全局信息的 k -核中心性, 还包括基于局部信息的结构洞位置. 此外, H 指数也是一个很好的度量节点重要性的指标, 当一个节点核数和 H 指数较高, 同时还占据较多的结构洞时, 该节点往往具有更大的影响力. 基于以上分析, 本文构造了基于引力方法的节点重要度排序方法 ISM 及其扩展算法 ISM₊, 基本思想是: 综合考虑节点局部拓扑信息 (H 指数) 和全局位置信息 (k -核中心性) 并将其看作物体质量的同时, 融合节点的结构洞特征以此消减网络伪核心节点重要度排序虚高对算法排序准确性的影响, 利用节点与领域节点间的相互作用力来描述节点的传播影响力.

由于节点核数和 H 指数不是同一个量纲, 二者不能直接融合, 为了融合节点这两方面的结构特

征, 引入一个均衡因子 γ , 定义为网络平均核数值与网络平均 H 指数之比, 表达式为

$$\gamma = \frac{\langle ks \rangle}{\langle h \rangle}, \quad (11)$$

其中, $\langle ks \rangle$ 表示网络平均核数值, $\langle h \rangle$ 表示网络平均 H 指数. 由此, 将节点局部信息和节点全局位置信息进行融合, 得到节点 i 的质量 $m(i)$, 定义为

$$m(i) = ks_i + \gamma h_i. \quad (12)$$

Liu 等 [28] 指出 k -核分解方法分解网络时容易将类核团节点错误识别为网络核心, 类核团内节点彼此紧密相连, 与网络的其他部分几乎没有联系. 实际上 H 指数在衡量节点的传播影响力时也存在类似问题, 对于类核团节点, H 指数同样会赋予这个节点高 h 值. 而那些不仅彼此之间连接十分紧密, 且与核心之外的节点还存在大量连接的节点, 则是网络的真核心. 综上, 对于一个高 ks 值或高 h 值节点, 如果该节点同时占据着较多结构洞, 那么该节点很可能是网络的重要节点. 因此, 我们进一步引入网络约束系数 [27] 来度量节点的结构洞特征, 根据邻域节点间的连接情况对节点重要度排序值进行校正, 从而消减 k -核分解方法和 H 指数识别出的类核团节点重要度排序虚高对算法精度的影响, 节点 i 的重要度校正函数 $\omega(i)$ 定义为

$$\omega(i) = \frac{e^{-C_i}}{2}, \quad (13)$$

e 是自然常数, $0 < \omega(i) \leq 1$, C_i 表示节点形成结构洞所受到的约束 (见 (5) 式), 当节点 i 的度越大且占据的结构洞越多, 节点的网络约束系数 C_i 值越小, $\omega(i)$ 的值越大. 反之, 节点 i 的度越小且邻居之间的闭合程度越高, 节点网络约束系数 C_i 值越大, $\omega(i)$ 的值越小. 最后, 模拟万有引力公式的形式, 综合考虑节点 i 与邻域节点间的相互作用力, 定义节点 i 的重要度 $ISM(i)$,

$$\begin{aligned} ISM(i) &= \sum_{d_{ij} \leq \psi_i} \omega(i) \frac{m(i)m(j)}{d_{ij}^2} \\ &= \sum_{d_{ij} \in \psi_i} e^{-C_i} \frac{(ks_i + \gamma h_i)(ks_j + \gamma h_j)}{2d_{ij}^2}, \end{aligned} \quad (14)$$

其中, ψ_i 是到节点 i 的距离小于或等于给定值 r 的邻域节点集, 为了降低算法复杂度, 参照文献 [24] 将 r 值设为 3. 进一步, 本文设计了 ISM 的扩展算法 ISM_+ , 定义为

$$ISM_+(i) = \sum_{j \in \Gamma_i} ISM(j)^\theta, \quad (15)$$

其中, $0 \leq \theta \leq 1$, 对于较小的 θ , ISM_+ 方法会削弱具有较大 ISM 值的有影响力邻居的影响, 而较大的 θ 值则会增强具有较大 ISM 值的有影响力邻居的影响. 不失一般性, 后续实验中 θ 都取为 0.8.

相比引力模型只考虑节点核数及节点的路径信息, ISM 与 ISM_+ 算法在几乎不增加算法计算时间的情况下, 融合了节点的多种属性信息, 包括节点 H 指数、节点位置、节点结构洞特征和节点的路径信息, 从而可以更准确地对节点重要度进行排序.

3.2 评价标准

本文基于经典的 SIR (susceptible-infected-recovered) [2,29] 传播动力学模型模拟网络中信息传播过程. 在 SIR 模型中, 节点可能处于以下 3 种状态: 1) 易受感染 (susceptible, S) 状态; 2) 已被感染 (infected, I) 状态; 3) 恢复 (removed, R) 状态. 处于状态 I 的节点将以一定的传播率 β 将疾病传播给处于状态 S 的邻居节点, 节点被感染后以概率 λ 被治愈呈恢复状态 R, 此后不再被感染. 当网络中不再有状态 I 的节点出现时传播过程终止. 不失一般性, 本文所有实验均考虑恢复率 $\lambda = 1$ 的情况. 节点经过 M 次 SIR 信息传播实验后的传播能力定义为 $\Phi(i) = \frac{1}{M} \sum_{m=1}^M \Phi'(i)$, 其中 $\Phi'(i)$ 表示其中一次传播实验中, 节点 i 作为起始传播源传播过程终止时处于状态 R 的节点总数.

为了验证所提算法相比其他指标对于节点重要性排序结果的准确性, 本文采用 Kendall tau 相关系数 [30,31] 来度量不同重要性度量指标得到的节点重要性排序列表与基于 SIR 模型得到的节点传播影响力排序列表之间的相关性, 其表达式为

$$\tau(R_1, R_2) = \frac{n_c - n_d}{\sqrt{(n_t - n_u)(n_t - n_v)}}, \quad (16)$$

其中, R_1 与 R_2 代表 n 个节点的两种不同重要性排序序列, n_c 和 n_d 分别是这两种排列中同序对和异序对的数量, $n_t = n(n-1)/2$, $n_u = \sum_{i=1}^s u_i(u_i-1)$, $n_v = \sum_{i=1}^t v_i(v_i-1)$, n_u 与 n_v 分别是针对 R_1 与 R_2 计算得出, 以 n_u 作计算说明 (n_v 计算过程可类推), 将 R_1 中相同的元素组成小集合, 小集合数用 s , u_i 表示第 i 个集合中元素个数. 利用 Kendall tau 相关系数进行结果计算, τ 值越高表示节点重要性评

价指标的排序结果与 SIR 仿真结果越近似, 评估结果越准确.

4 实验数据集与结果分析

实验选取了 6 个来自不同领域的真实数据集, 分别是安然邮件网络 Enron^[32], Slavo Zitnik 的朋友圈关系网络 Facebook^[33], 科学家合作网络 Netscience^[34], 美国航空网络 USAir^[35], 人群感染网络 Infectious^[36] 以及网页网络 EPA^[34]. 表 1 列出这些网络的统计特征, 包括网络节点总数 N , 网络连边数 E , 节点间平均最短距离 $\langle d \rangle$, 节点平均度 $\langle k \rangle$, 网络集聚系数 C , 网络直径 D , 网络最大 ks 值 ks_{\max} , 信息传播阈值 $\beta_{\text{th}} = \langle k \rangle / \langle k^2 \rangle$ 以及信息传播率 β , 其中 $\langle k^2 \rangle$ 表示节点二阶平均度.

4.1 真实网络

首先使用第 3 节中介绍的 SIR 模型分析不同算法排序结果与节点真实传播能力之间的相关性, 按表 1 中的 β 值设置 6 个网络的感染概率, 独立运行 1000 次取平均结果, 相关程度越高, 表明相应算法得到的节点重要性排序结果越准确.

从图 1 可以观察到, 本文所提的 ISM 与 ISM₊ 方法与 SIR 传播过程中感染数量 ϕ 的大小高度相关, 尤其是 ISM₊ 方法在大多数情况下都优于其他算法, 说明所提算法相比其他指标能够较为准确地识别节点的传播影响力. 传统的度量方法如接近中心性和介数中心性指标与实际影响力之间相关性较弱, 结果较为发散, 尤其是介数中心性与 SIR 影响节点数的相关性最弱, 其原因与网络的社区化有关, 因为社区化的情况下节点间聚集程度高, 节点介数普遍很小, 导致利用介数进行传播影响力排序时节点间区分度不大. 造成这一结果的还可能是因为排名靠前的节点集中在同一个社区, 导致了信息

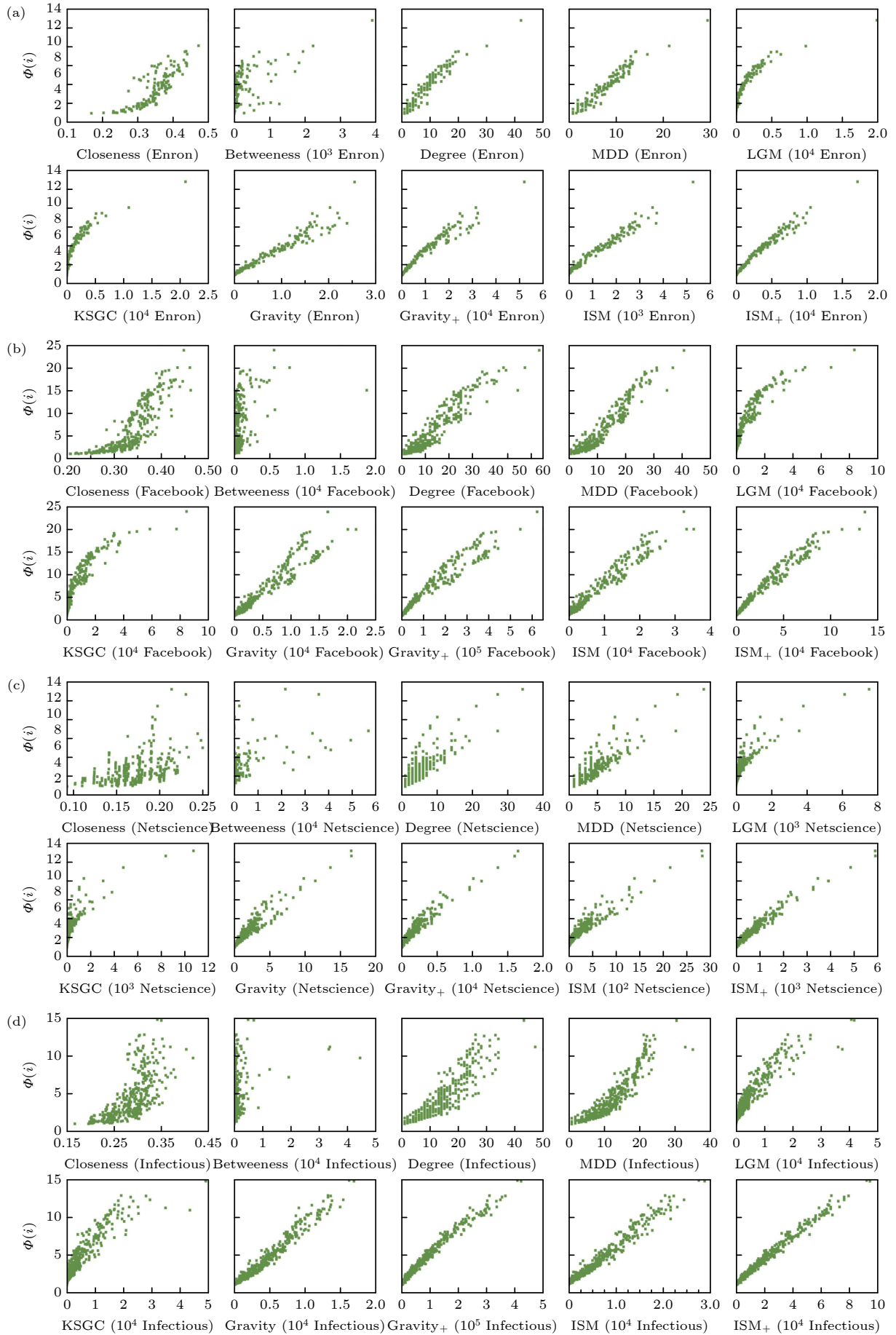
传播的局部性. KSGC 方法是针对 LGM 做的改进, 但在相关性实验中, 两种算法的结果较为接近.

在相关性实验中, 实验设置的传播率是固定的, 实验结果只反映了特定传播率下的静态状态. 为了更全面评价各个算法的节点重要性排序精度, 我们将 τ 值作为准确性度量值, 设置传播率区间为 $[|\beta_{\text{th}}| - 7\%, |\beta_{\text{th}}| + 7\%]$ (若 $\beta_{\text{th}} \leq 0.07$, 传播率区间设置为 $[0.01, 0.15]$). 结果如图 2 所示, 纵轴表示节点实际传播能力排序结果与不同中心性算法得到的节点重要性排序结果间的相关系数值, 该值越大表示对应排序算法越准确. 可以看出, 当传播率超过传播阈值 β_{th} (虚线表示不同网络的 β_{th} 值) 时, ISM 与 ISM₊ 方法表现一般都要优于多数算法, 尤其是 ISM₊ 方法表现更加突出, 同 SIR 模型模拟传播过程得到的节点传播能力有显著的相关性. 然而, 从图 2 可以清楚地看到, 尽管介数中心性和接近中心性方法是基于网络全局信息计算得到的, 但在识别这些网络中重要节点方面并不具有优势. 同时, 度中心性, MDD, LGM 和 KSGC 这类基于度的方法在传播率较小的情况下表现较好, 是因为当传播率较小时, 信息从节点发起容易局限于局部, 此时影响传播结果的主要因素是邻居节点数量, 即节点度越大感染到的节点也越多, 度中心性, MDD, LGM 和 KSGC 方法正好适合这一情况.

调整考察的节点范围进一步对 Kendall 相关系数的结果进行观察, 设置节点比例 L 的变化范围为 0.05—1.00, 图 3 给出了不同算法得到的不同比例排名靠前的节点与节点实际传播影响力排序之间的相关性结果. 不难看出当 L 较小时, 除了在 Enron 网络中 MDD, LGM 和 KSGM 表现要好于 ISM 与 ISM₊ 以外, 其他 5 个网络中, 本文提出的 ISM₊ 算法在不同比例节点时都可以获得较好的节点重要性排序结果, 并且能够在更大范围的 L 值下取得更好的评价结果.

表 1 6 个真实网络的拓扑统计参数
Table 1. Topological parameters of six real networks.

网络名	N	E	$\langle d \rangle$	β_{th}	β	$\langle k \rangle$	C	D	ks_{\max}
Enron	143	623	2.9670	0.0774	0.08	8.7133	0.4339	8	9
Facebook	324	2218	3.0537	0.0466	0.05	13.6914	0.4658	7	18
Netscience	379	914	6.0419	0.1250	0.13	4.8232	0.7410	17	8
USAir	453	2025	2.7381	0.0231	0.03	12.8072	0.6252	6	26
Infectious	410	2765	3.6309	0.0534	0.05	13.4878	0.4558	9	17
Web_EPA	4253	6258	4.5003	0.0366	0.08	4.1839	0.0714	10	6



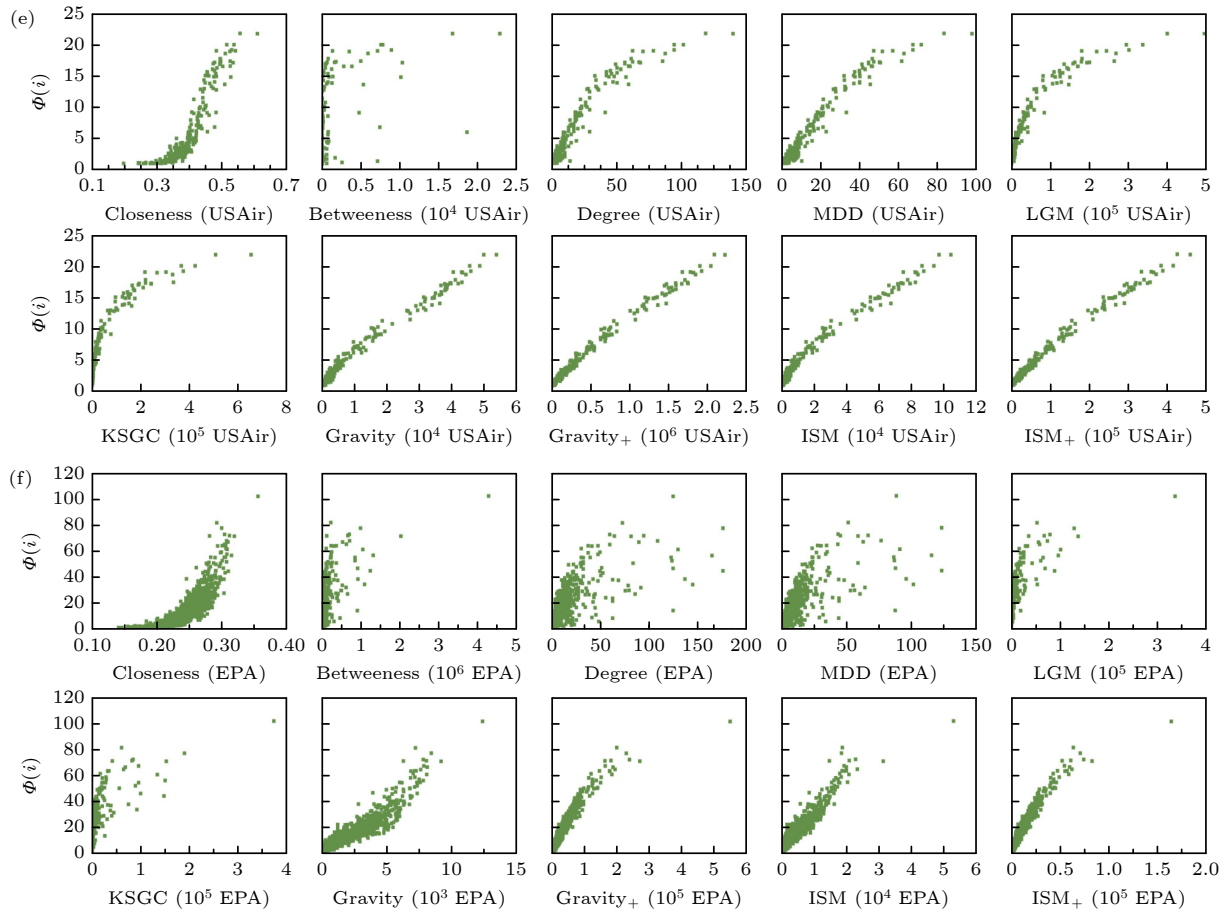


图 1 十种不同排序方法得到的排序结果与 SIR 传播过程感染节点数的相关性 (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA

Fig. 1. The correlation between the ranking results obtained by ten different ranking methods and the number of infected nodes in the SIR propagation process: (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA.

4.2 模拟数据集

除了 6 个真实网络数据外, 还在 Lancichinetti-Fortunato-Radicchi (LFR) [35] 模型生成的人工网络数据集上比较了不同传播率下 SIR 和不同评估算法间的 Kendall 相关系数. 通过设置不同的 LFR 参数, 生成拓扑特征不同的网络结构, 设置 LFR 模型参数为: 节点数 $N = 2000$, 社区的最小规模 $c_{\min} = 20$, 社区的最大规模 $c_{\max} = 50$, 网络的最大度 $k_{\max} = 30$, 混合参数 $\mu = 0.1$. 调整网络平均度 $\langle k \rangle$ 来调节网络的连接紧密程度, 分别生成 $\langle k \rangle = 5, 10, 15$ 的三个网络数据集. 设置传播率区间为 $[0.01, 0.15]$, 实验结果如图 4 所示, 当传播率超过传播阈值时, ISM₊ 实验结果明显优于其他 9 种算法, 尤其在集聚程度高的网络中, 如图 4(b), (c), 相比其他 9 种指标, ISM₊ 指标在更大范围的传播率下具有优势. 当传播率较小时, 度中心性, MDD, LGM 与 KSGC 算法表现相对较好, 这与真实数据

集上的结果类似, 其原因也是因为传播率偏小时, 节点的真实影响力主要由节点度大小决定.

4.3 ISM₊ 算法的最优 θ 值

不同的实际网络可能要求不同的 θ 值, 从而保证 ISM₊ 方法可以获得最佳性能, 实验取间隔为 0.02, 区间范围为 0.02—1.00 的多个 θ 值, 采用平均 Kendall tau 指标 $\langle \tau \rangle$ [37], 系统分析参数 θ 对 ISM₊ 算法性能的影响:

$$\langle \tau \rangle = \frac{1}{M} \sum_{\beta=\beta_{\min}}^{\beta=\beta_{\max}} \tau(\beta), \quad (17)$$

其中 β 表示传播率, β_{\min} 和 β_{\max} 分别表示最小和最大传播率, M 表示考察的传播率数量, $\tau(\beta)$ 表示当传播率为 β 时, ISM₊ 方法生成的节点重要性排序序列与 SIR 过程生成节点传播影响力排序序列之间的 Kendall 相关性 τ 值. 这里同样设置传播率区间为 $[\beta_{\text{th}} - 7\%, \beta_{\text{th}} + 7\%]$ (即除了 Netscience

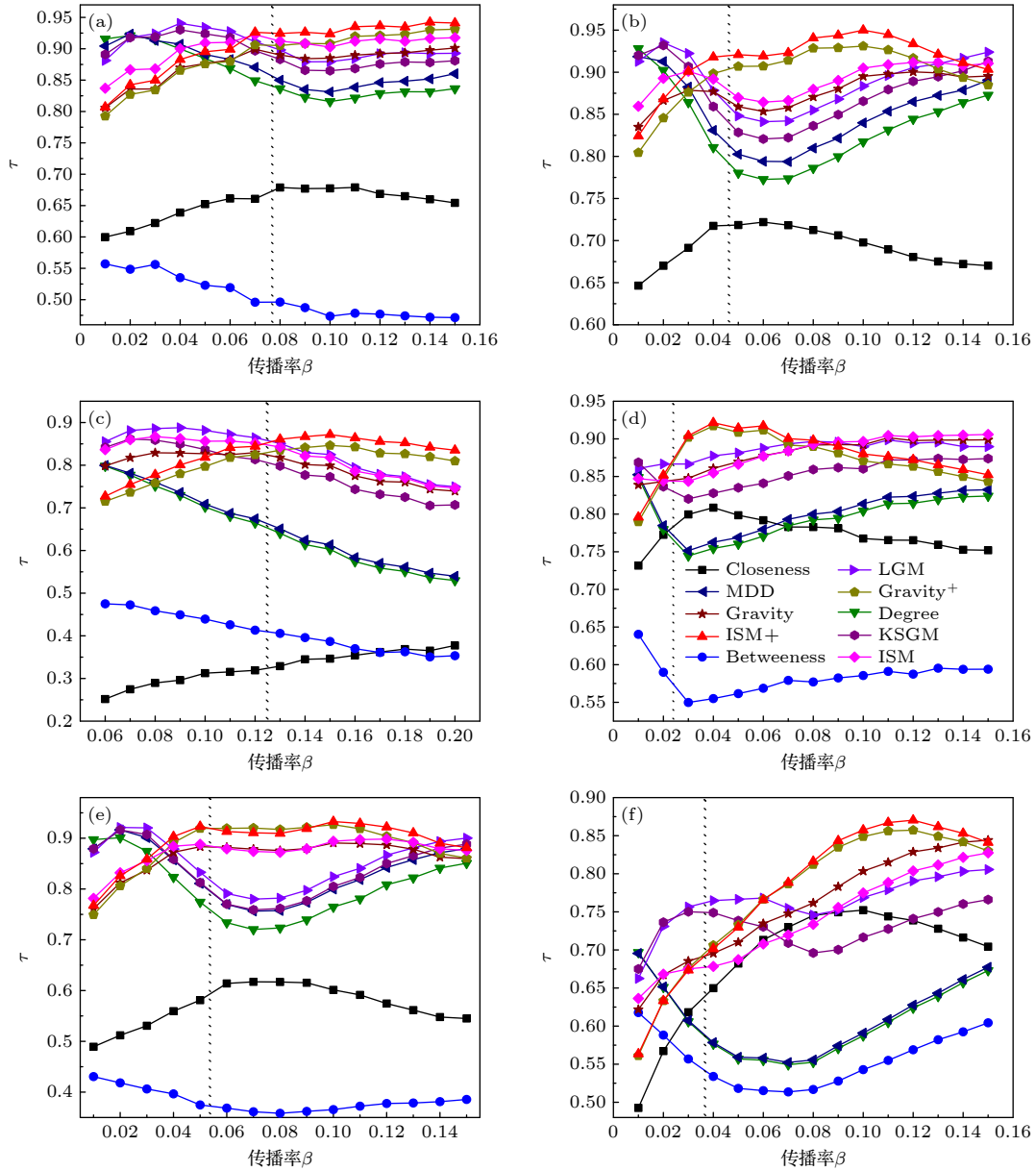


图2 6个真实网络数据集上十种不同排序方法排序准确性对比 (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA

Fig. 2. Comparison of sorting accuracy of ten different sorting methods on six real network datasets: (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA.

网络传播率区间设置为 $[0.06, 0.20]$ 以外, 其他网络的传播率区间均设置为 $[0.01, 0.15]$). $\langle \tau \rangle$ 值介于 -1 — 1 之间, 值越大意味着对应 θ 值的 ISM_+ 方法可以更准确地识别网络中具有传播影响力的重要节点. 实验结果如图5红色曲线所示, 对于每个网络, 都有一个最佳的 θ 值, 该值对应的 ISM_+ 方法可获得最大的 $\langle \tau \rangle$ 值. Enron, Facebook, Netscience, USAir, Infectious, EPA 以及平均 $\langle k \rangle$ 分别为 5, 10, 15 的 LFR 网络, 对应的最佳 θ 值分别为 0.60, 0.60, 0.56, 0.38, 0.60, 0.64, 0.46, 0.68 及 0.72, 多

数网络中最优 θ 值都超过 0.5. 由于 ISM_+ 算法的设计原理决定了其在信息传播率超过传播阈值时更具有优势, 因此我们进一步分析传播率超过 β_{th} 时, θ 的取值对 ISM_+ 算法性能的影响, 实验结果如图5中黑色曲线所示, Enron, Facebook, Netscience, USAir, Infectious, EPA 这 6 个真实网络传播率区间分别取 $[0.08, 0.15]$, $[0.05, 0.15]$, $[0.13, 0.20]$, $[0.03, 0.15]$, $[0.05, 0.15]$ 及 $[0.05, 0.15]$, 对应的最佳 θ 值分别为 0.70, 0.68, 0.76, 0.38, 0.76 及 0.64, 平均 $\langle k \rangle$ 为 5, 10, 15 的 LFR 网络的传播

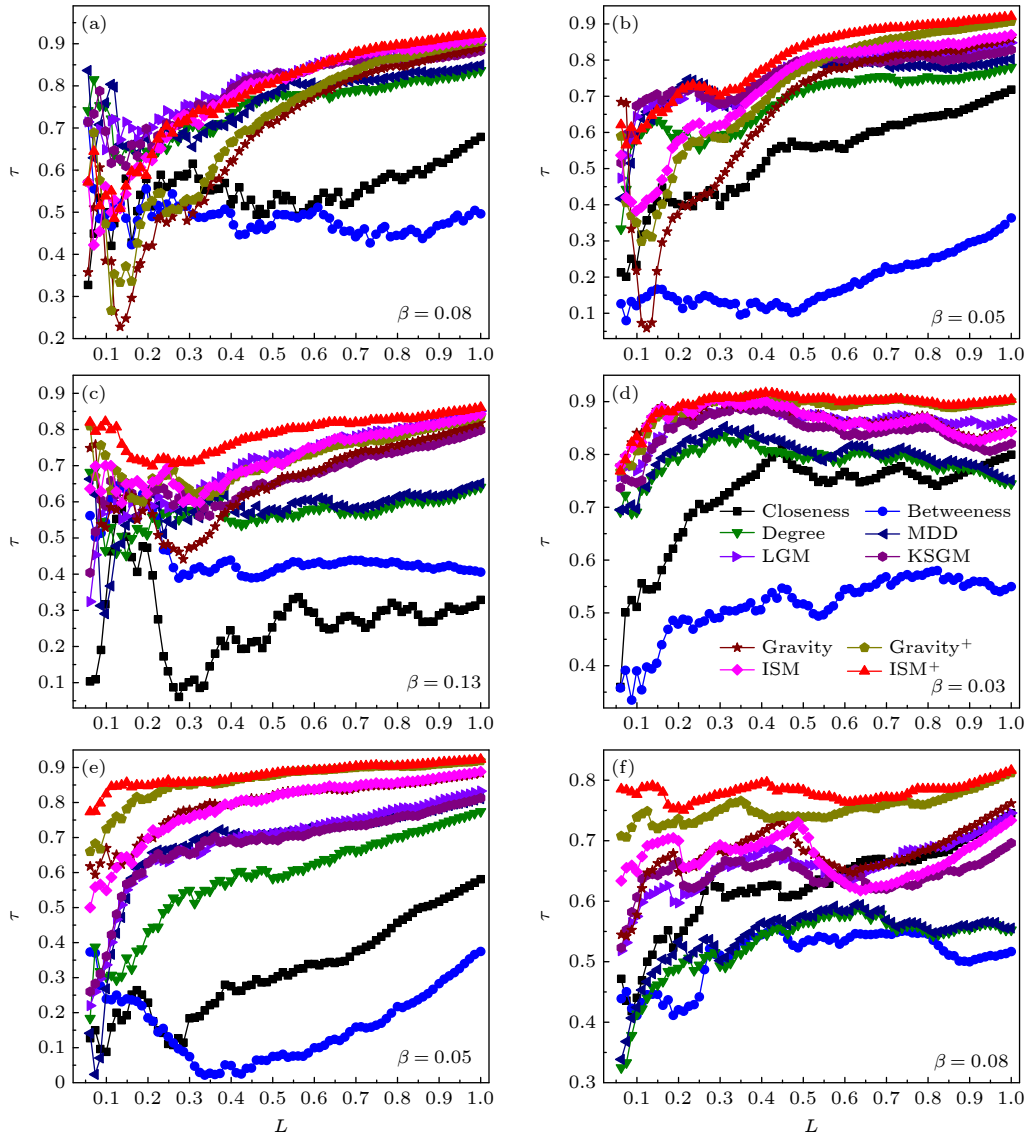


图 3 不同比例节点下十种评估算法的 Kendall 相关系数对比 (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA

Fig. 3. Comparison of Kendall correlation coefficients of ten node influence evaluation algorithms under different scale nodes: (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA.

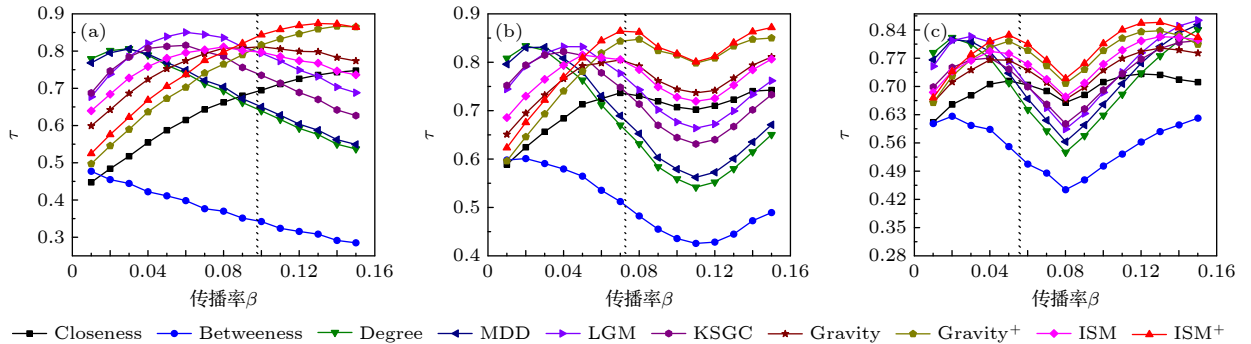


图 4 LFR 模拟数据集上十种评估算法的 Kendall 相关系数对比, 黑色虚线为三个网络的传播阈值 β_{th} (a) $\langle k \rangle = 5, \beta_{th} = 0.0984$; (b) $\langle k \rangle = 10, \beta_{th} = 0.0723$; (c) $\langle k \rangle = 15, \beta_{th} = 0.0577$

Fig. 4. Comparison of Kendall correlation coefficients of ten evaluation algorithms on the LFR simulation dataset, the black dashed line is the propagation threshold β_{th} of three different network: (a) $\langle k \rangle = 5, \beta_{th} = 0.0984$; (b) $\langle k \rangle = 10, \beta_{th} = 0.0723$; (c) $\langle k \rangle = 15, \beta_{th} = 0.0577$.

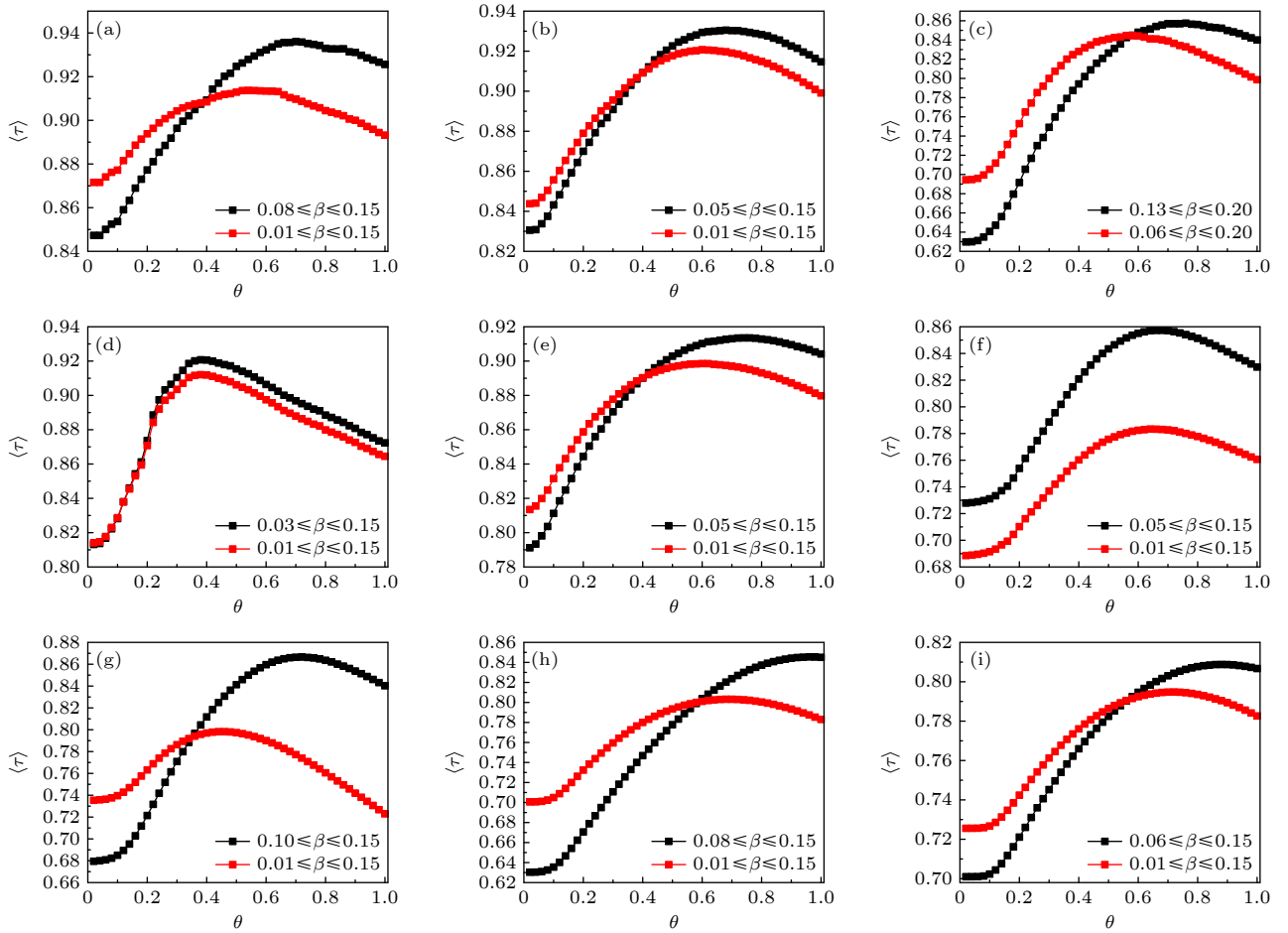


图 5 当 β 变化时, 不同 θ 值所对应的 ISM₊ 方法生成的节点重要性排序序列与 SIR 传播扩散过程生成的节点传播影响力排序序列之间的平均 Kendall's $\langle \tau \rangle$ 值 (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA; (g) LFR_k5; (h) LFR_k10; (i) LFR_k15

Fig. 5. The average Kendall's $\langle \tau \rangle$ obtained by comparing the ranking list generated by SIR spreading process and the ranking list generated by the ISM₊ methods with different θ when the β changes: (a) Enron; (b) Facebook; (c) Netscience; (d) Infectious; (e) USAir; (f) EPA; (g) LFR_k5; (h) LFR_k10; (i) LFR_k15.

区间分别取 $[0.10, 0.15]$, $[0.08, 0.15]$, $[0.06, 0.15]$, 对应的最佳 θ 值分别为 0.72, 0.96, 0.88, 可见当传播率超过 β_{th} 时, 强化具有较大 ISM 值的有影响力邻居的影响对于提高 ISM₊ 性能具有积极作用。

5 结 论

如何准确识别网络中具有传播影响力的重要节点, 是近年来网络科学研究的热点问题. 本文基于引力模型设计了 ISM 方法及其扩展算法 ISM₊, 可以有效地对复杂网络中的节点重要性进行评价和排序. 所提算法兼顾局部拓扑信息和全局位置信息, 基于牛顿力学中的引力公式, 融合了节点的多种属性信息包括节点 H 指数、 k 核中心性以及节点的结构洞特征, 弥补了现存方法评估角度片面的不

足, 可以更有效地对节点重要性进行评价. 在 6 个真实网络和 3 个 LFR 模拟数据集上的实验结果表明, 与其他评估方法 (如度中心性, 介数中心性, 接近中心性, MDD, LGM, KSGC 与引力模型等) 相比, 所提方法在识别网络节点重要性方面具有一定优势, 当传播率大于传播阈值时, 多数网络中算法在不同比例节点下都能更准确地评估节点的重要性. 本文所提算法参照引力模型, 仅将最短路径表示为节点间的路径信息, 实际上节点间除最短路径以外的其他可达路径对于衡量节点间的相互作用效应也有效, 未来的工作中我们将从这一角度出发进一步提升算法精度。

参考文献

- [1] Lü L Y, Chen D B, Ren X L, Zhang Q M, Zhang Y C, Zhou T 2016 *Phys. Rep.* **650** 1

- [2] Pastor-Satorras R, Vespignani A 2001 *Phys. Rev. Lett.* **86** 3200
- [3] Albert R, Barabási A L 2002 *Rev. Modern Phys.* **74** 47
- [4] Alshahrani M, Fuxi Z, Sameh A, Mekouar S, Huang S 2020 *Inform. Sciences* **527** 88
- [5] Albert R, Jeong H, Barabási A L 1999 *Nature* **401** 130
- [6] Chen D B, Lu L Y, Shang M S, Zhang Y C, Zhou T 2012 *Physica A* **391** 1777
- [7] Sabidussi G 1966 *Psychometrika* **31** 581
- [8] Freeman L C 1977 *Sociometry* **40** 35
- [9] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nat. Phys.* **6** 888
- [10] Lü L Y, Zhou T, Zhang Q M, Stanley H E 2016 *Nat. Commun.* **7** 10168
- [11] Bae J, Kim S 2014 *Physica A* **395** 549
- [12] Zeng A, Zhang C J 2013 *Phys. Lett. A* **377** 1031
- [13] Zareie A, Sheikhhahmadi A, Khamforoosh K 2018 *Expert Syst. Appl.* **108** 96
- [14] Fei L, Lu J, Feng Y 2020 *Comput. Ind. Eng.* **142** 106355
- [15] Hang 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]
- [16] Wang Z X, Du C J, Fan J P, X Y 2017 *Neurocomputing* **260** 466
- [17] Yan G H, Zhang M, Luo H, Li S K, Liu T 2019 *J. Communications* **40** 109 (in Chinese) [闫光辉, 张萌, 罗浩, 李世魁, 刘婷 2019 *通信学报* **40** 109]
- [18] Alon U 2007 *Nat. Rev. Genet.* **8** 450
- [19] Benson A R, Gleich D F, Leskovec J 2016 *Science* **353** 163
- [20] Li Y, Deng Y 2018 *Int. J. Comput. Commun. Control* **13** 792
- [21] Wang J, Qiao K Y, Zhang Z Y 2019 *Future Gener. Comp. Sy.* **91** 1
- [22] Morone F, Makse H A 2015 *Nature* **527** 544
- [23] Zhong L F, Liu Q H, Wang W, Cai S M 2018 *Physica A* **511** 78
- [24] Ma L L, Ma C, Zhang H F, Wang B H 2016 *Physica A* **451** 205
- [25] Li Z, Ren T, Ma X Q, Liu S M, Zhang Y X, Zhou T 2019 *Sci. Rep.* **9** 1
- [26] Yang X, Xiao F Y 2021 *Knowl-Based Syst.* **227** 107198
- [27] Burt R S 2004 *American J. Sociology* **110** 349
- [28] Liu Y, Tang M, Zhou T, Do Y 2015 *Sci. Rep.* **5** 9602
- [29] Newman M E J 2002 *Phys. Rev. E* **66** 016128
- [30] Kendall M G 1945 *Biometrika* **33** 239
- [31] Knight W R 1966 *J. Amer. Statist. Assoc.* **61** 436
- [32] Rossi R, Ahmed N 2015 *Twenty-ninth AAAI Conference on Artificial Intelligence* Austin, Texas, USA, January 4 2015, pp4292–4293
- [33] Blagus N, Šubelj L, Bajec M 2012 *Physica A* **391** 2794
- [34] Newman M E J 2006 *Phys. Rev. E* **74** 036104
- [35] Batagelj V, Mrvar A 1998 *Connections* **21** 47
- [36] Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J F 2011 *J. Theor. Biol.* **271** 166
- [37] Lin J H, Guo Q, Dong W Z, Tang L Y, Liu J G 2014 *Phys. Lett. A* **378** 3279

Node importance ranking method in complex network based on gravity method*

Ruan Yi-Run[†] Lao Song-Yang Tang Jun Bai Liang Guo Yan-Ming

(College of Systems Engineering, National University of Defense Technology, Changsha 410073, China)

(Received 28 March 2022; revised manuscript received 1 May 2022)

Abstract

How to use quantitative analysis methods to identify which nodes are the most important in complex network, or to evaluate the importance of a node relative to one or more other nodes, is one of the hot issues in network science research. Now, a variety of effective models have been proposed to identify important nodes in complex network. Among them, the gravity model regards the coreness of nodes as the mass of object, the shortest distance between nodes as the distance between objects, and comprehensively considers the local information of nodes and path information to identify influential nodes. However, only the coreness is used to represent the quality of the object, and the factors considered are relatively simple. At the same time, some studies have shown that the network can easily identify the core-like group nodes with locally and highly clustering characteristics as core nodes when performing k -core decomposition, which leads to the inaccuracy of the gravity algorithm. Based on the universal gravitation method, considering the node H index, the number of node cores and the location of node structural holes, this paper proposes an improved algorithm ISM and its extended algorithm ISM₊. The SIR model is used to simulate the propagation process in several classical real networks and artificial networks, and the results show that the proposed algorithm can better identify important nodes in the network than other centrality indicators.

Keywords: complex networks, spreading influence, gravity model, H index, k -shell method

PACS: 64.60.aq, 89.75.Hc, 89.75.Fb

DOI: [10.7498/aps.71.20220565](https://doi.org/10.7498/aps.71.20220565)

* Project supported by the National Natural Science Foundation of China (Grant No. 72101265).

[†] Corresponding author. E-mail: ruanyirun@163.com



基于引力方法的复杂网络节点重要度评估方法

阮逸润 老松杨 汤俊 白亮 郭延明

Node importance ranking method in complex network based on gravity method

Ruan Yi-Run Lao Song-Yang Tang Jun Bai Liang Guo Yan-Ming

引用信息 Citation: *Acta Physica Sinica*, 71, 176401 (2022) DOI: 10.7498/aps.71.20220565

在线阅读 View online: <https://doi.org/10.7498/aps.71.20220565>

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

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 <https://doi.org/10.7498/aps.67.20172295>

基于Tsallis熵的复杂网络节点重要性评估方法

A method of evaluating importance of nodes in complex network based on Tsallis entropy

物理学报. 2021, 70(21): 216401 <https://doi.org/10.7498/aps.70.20210979>

基于区域密度曲线识别网络上的多影响力节点

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

物理学报. 2018, 67(19): 198901 <https://doi.org/10.7498/aps.67.20181000>

基于加权 K -阶传播数的节点重要性

Node importance based on the weighted K -order propagation number algorithm

物理学报. 2019, 68(12): 128901 <https://doi.org/10.7498/aps.68.20190087>

基于多阶邻居壳数的向量中心性度量方法

Complex network centrality method based on multi-order K -shell vector

物理学报. 2019, 68(19): 196402 <https://doi.org/10.7498/aps.68.20190662>

动态复杂网络中节点影响力的研究进展

Node influence of the dynamic networks

物理学报. 2020, 69(4): 048901 <https://doi.org/10.7498/aps.69.20190830>