



基于多阶邻居传播度量和拓扑特征的高影响力节点识别

罗余¹, 王建波^{1,2*}, 李平¹, 杜占玮², 许小可^{3,4*}

1. 西南石油大学计算机与软件学院, 成都 610500

2. 香港大学公共卫生学院, 香港 999077

3. 北京师范大学计算传播学研究中心, 珠海 519087

4. 北京师范大学新闻传播学院, 北京 100875

* 通信作者. E-mail: jwang@swpu.edu.cn, xuxiaoke@bnu.edu.cn

收稿日期: 2023-07-03; 修回日期: 2023-08-14; 接受日期: 2023-08-30; 网络出版日期: 2024-04-11

国家自然科学基金 (批准号: 62173065)、四川省自然科学基金 (批准号: 2023NSFSC0501) 和北京市自然科学基金 (批准号: 4242040) 资助项目

摘要 如何定量评估复杂网络中节点的影响力是一个重要的研究课题, 因为它有助于深入理解网络的结构和功能. 现有的多数方法主要基于网络固有拓扑的分析建立, 缺少对多阶邻居节点的传播性质和拓扑信息的综合利用, 然而它们对影响力节点识别有重要影响. 为此, 本文提出了一种综合多阶邻居传播度量和拓扑特征 (multi-order neighbor propagation metrics and topological features, MNPMTF) 的算法来有效识别复杂网络中的影响力节点. 首先, 该算法结合传播模型和最短路径来刻画邻居节点的传播概率, 从而量化节点之间信息传播的可能性. 其次, 考虑多阶邻居中的邻居重叠比形成邻居重叠度, 进而量化信息在邻居网络中的传播路径. 再次, 利用节点的 k 壳、 h 指数和聚类系数构成新指标 KHC 系数, 以此来描述节点的拓扑特征. 最后, 算法综合 3 阶邻居范围内的传播概率、邻居重叠度和拓扑特征以评估节点的影响力. 在 9 个真实网络上的大量实验表明, 所提算法在排序准确性、有效性和区分能力等多方面均优于 7 种具有代表性的方法, 为复杂网络中节点影响力评估提供了一种新的思路.

关键词 传播概率, 邻居重叠度, KHC 系数, 影响力节点, 复杂网络

1 引言

自然界的大多数真实系统都能使用复杂网络进行描述, 涉及各个领域^[1], 涵盖生物系统^[2]、社会系统^[3]和电力系统^[4]等. 复杂网络的异构性使得少数节点在网络中起主导作用. 在这些网络中, 这

引用格式: 罗余, 王建波, 李平, 等. 基于多阶邻居传播度量和拓扑特征的高影响力节点识别. 中国科学: 信息科学, 2024, 54: 944–959, doi: 10.1360/SSI-2023-0201
Luo Y, Wang J B, Li P, et al. Identification of high-influential nodes based on multi-order neighbor propagation metrics and topological features (in Chinese). Sci Sin Inform, 2024, 54: 944–959, doi: 10.1360/SSI-2023-0201

些节点被认为是高影响力节点,因为它们对网络的结构和功能以及信息传播有着重要影响.影响力节点的识别可以应用于许多真实场景,例如社交网络中谣言控制^[5]、生物网络中必需蛋白质^[6]、交通网络中关键站点^[7]和社区检测^[8]等.由于网络结构的复杂性和多样性,如何准确且有效地识别各类复杂网络中的影响力节点,是一个重要且富有挑战性的研究课题,备受学者们的关注^[9].

关于如何提出准确衡量网络中的影响力节点的统一标准,学者们还在探索中.各种具有代表性的方法被不断提出以识别网络中的高影响力节点.基于网络拓扑的方法是其中一类常见的方法,通过分析网络的结构特征来评估节点的影响力.基于网络拓扑特征的经典度量也是这方面研究的基础,如局部 h 指数^[10]、核心中心性^[11]、度中心性^[12]和 h 指数^[13]等.然而,仅仅依靠单一的拓扑特征评估节点影响力存在局限性,无法适用于多种复杂场景,且随着网络规模的增大,单一指标无法有效评估节点的影响.为了克服这些局限性,研究人员开始倾向于采用综合多种特征的中心性方法.尤其在考虑高阶邻居节点的影响时,这种方法变得尤为重要.在复杂网络中,节点的影响力可能远超其直接邻居的影响范围.这种情况在信息传播、疾病传播等动态过程中更为突出.因此,有必要综合考虑高阶邻居节点的影响,以更好地捕捉节点在网络中的地位和作用,从而更准确地评估节点的影响力.Zhang等^[14]采用模糊逻辑来测量节点的高阶邻居贡献,并设计了一种算法,根据局部模糊信息熵来识别有影响的节点.Namtirtha等^[15]分析了节点的高阶邻居信息,并定义了半全局三角形中心性算法.通过考虑高阶邻居节点,研究者们可以更准确地评估节点的传播能力和重要性.高阶邻居节点可能会通过间接连接网络产生影响,甚至在局部区域中缺乏显著度的节点也可能在全局范围内发挥重要作用.为此,学者们还设计了基于高阶邻居拓扑的新算法来识别复杂网络中的影响力节点,如引力模型.引力模型,作为一种融合多属性特征并考虑高阶邻居影响的研究框架,已经广泛应用于复杂网络的影响力节点识别.Curado等^[16]考虑反向随机游走和有效距离提出了一种改进的重力模型.Yang等^[17]基于自适应截断半径和全路径推断提出重力模型算法.此外,一些学者考虑了无权网络中边权重的差异,设计了新的基于边拓扑的影响力节点识别算法.例如Zheng等^[18]利用度中心性定义广义度,并利用其来衡量边权重.Meng等^[19]根据节点位置信息和节点领域信息定义了新的潜在边缘权重来识别影响力节点.多属性决策方法^[20,21]也已进一步扩展到复杂网络,并从多个维度评估节点的影响力.

尽管这些方法在评估节点影响力方面取得了较为丰硕的成果,但多数算法只考虑了网络中节点的拓扑结构特征.仅利用拓扑信息的算法虽然有效,但不能完整地描述节点在信息、病毒传播过程中的作用.在现实生活中,研究动态传播机制并识别传播过程中的传播者,可以有效地遏制疫情传播和指导疫情防控.因此,结合动态传播机制研究节点识别问题变得尤为重要,它有助于更好地理解和干预现实世界中的传播过程^[22].最近,基于传播动力学模型的方法已经在识别有影响的节点方面取得了一些进展.例如,结合节点的局部传播概率来识别有影响的传播者^[23],发现这种结合可以有效地提高识别算法的性能.通过融合节点的动态传播性质,可以在一定程度上提升复杂网络中影响力节点的识别准确度^[24].Liu等^[25]根据传播动力学模型识别网络中的冗余边,并在删除冗余边后执行 k 壳分解,以识别有影响的节点.Qu等^[26]将相邻节点的感染风险纳入结构中心性,提出了基于易感-感染-易感(susceptible-infectious-susceptible, SIS)模型的中心性方法.Ai等^[27]根据SI模型,考虑三阶邻居范围内的节点来识别影响力节点.Yu等^[28]提出了一种结合节点间直接和间接感染的非对称连接强度,对影响力节点进行识别和排序.Chen等^[29]分析了影响力节点传播的因素,利用节点的局部拓扑特征和传播速率,提出了一种基于网络抗毁性的影响力节点搜索策略.此外,节点与邻居之间的信息传播模式也是评估节点传播能力的关键.Chen等^[30]利用传播模型度量信息在节点与邻居节点传播的概率,从而评估节点的影响.上述方法都结合了传播动力学模型和节点间的传播关系,对于评估节点的影响具有重要意义.

然而, 上述方法只考虑从低阶邻居到高阶邻居的动态传播, 忽视了同阶邻居之间的关系对传播的影响或节点拓扑特征的作用. 研究表明^[31] 同阶邻居之间关系和节点拓扑会对信息传播造成影响, 因此有必要考虑邻居节点之间的影响和拓扑来评估节点的影响力. 在现实世界中, 信息的传递主要依赖于节点及其邻居之间、邻居与邻居之间的交互行为. 基于此, 本文提出一种基于多阶邻居传播度量和拓扑特征 (multi-order neighbor propagation metrics and topological features, MNPMTF) 的影响力节点识别算法, 其基本思想是节点的影响力取决于三阶邻居范围内的邻居节点的传播概率、同阶邻居节点之间的影响和邻居节点的拓扑结构信息. 在信息传递过程中, 算法将信息传播分为纵向和横向传播, 前者涉及不同阶邻居之间的传播 (传播概率), 后者涉及同阶邻居之间的传播 (邻居重叠度). 纵向传播揭示了信息从源节点向外扩散到不同阶次邻居节点的潜力, 而横向传播衡量了节点在同级别邻居中的传播影响. 综合考虑这两种传播方式, 能够更全面、准确地评估节点的影响力, 进而在实际网络中表现出更好的效果. 同时, 节点的位置属性 (k 壳)、领域属性 (h 指数) 和局部拓扑结构 (聚类系数) 也被考虑为节点影响力的重要组成部分.

2 算法设计与分析

本文从传播动力学的角度^[27] 出发, 设计了一种新的有效识别影响力节点算法, 即基于多阶邻居传播度量和拓扑特征的影响力节点识别算法 (MNPMTF 算法). 该算法综合利用了 3 阶邻居范围内的邻居节点的传播概率、邻居节点的局部拓扑以及同阶邻居节点之间的影响, 来评估节点的影响力. 具体来说, 当节点的传播概率较高、节点处于网络的核心位置、局部的拓扑结构较稀疏且邻居节点之间的影响较大时, 可以认为节点的影响力较高.

2.1 MNPMTF 算法

令 $G = (V, E)$ 表示无向无权网络, 其中 $|V| = N$ 为节点总数, $|E| = M$ 为边总数. $\Gamma_n(u)$ 表示 n 阶邻居网络, 即网络中节点与源节点 u 的最短路径长度为 n 的节点的集合. 进一步写成, $\Gamma_n(u) = \{x_n | (u, x_n), \text{dis}(u, x_n) = n\}$, 其中 x_n 代表源节点 u 的 n 阶邻居节点, $\text{dis}(u, x_n)$ 代表节点 u 和 x_n 之间的最短路径长度. 为了定量刻画节点的传播概率、同阶邻居之间的影响和节点的拓扑结构信息, 本文首先给出以下 3 个定义.

定义1 传播概率 (propagation probability) 表示的是局部范围内, 邻居节点 x 受到源节点 u 感染的概率, 从而量化节点之间传播的可能性. 因为随着节点间距离的增加, 节点被感染的概率呈指数级减少, 所以利用节点之间的最短路径长度来衡量节点之间的距离以及考虑节点之间最短路径的数量来刻画传播概率. 由此, 节点 x 的传播概率 $\text{PP}(x, l)$ 表示为

$$\text{PP}(x, l) = \begin{cases} \beta, & l = 1, \\ n_{ux}\beta^l, & l \geq 2, \end{cases} \quad (1)$$

其中 $\beta = \alpha \times \beta_{\text{th}}$ 表示感染率, n_{ux} 表示节点 u 到节点 x 的存在的最短路径的数量, l 表示节点 u 和 x 之间的最短路径长度, β_{th} 表示的是网络中的传播阈值, 其表达式为 $\frac{\langle k \rangle}{\langle k^2 \rangle}$, α 为一个可调参数. 在复杂网络中, 传播概率起着关键作用, 它对节点间信息传播过程进行建模, 并由定义的感染率 β 和传播阈值 β_{th} 来量化节点在传播过程中的激活情况, 从而加深对信息在网络中传播的理解. 传播概率在以下几个方面发挥重要作用: 首先, 传播概率有助于识别重要节点. 在信息传播过程中, 传播概率较高的节点在传播路径上可能扮演着关键角色, 它们更有可能成为信息的传播源头或传播的关键节点. 因此,

对这些节点的重视有助于优化网络的稳定性和效率. 其次, 传播概率的定义使得能够研究不同节点之间的传播特性. 通过计算不同节点间的传播概率, 可以了解网络中不同部分之间信息传播的异质性并有助于揭示网络的结构和功能, 从而更好地理解和分析复杂网络中的信息传播过程.

定义2 (邻居重叠度 (neighbor overlap)) 在现实世界中, 信息不仅在不同年龄阶段传播, 也可能在同龄人之间进行传播. 因此, 给定源节点, 信息可能在源节点中的邻居网络中进行传播. 这种传播流向不仅在一定程度上加速了信息的传播, 而且在评估源节点的重要性时, 起到不可替代的作用. 对于邻居网络中的一个节点 x , 如果 x 节点在邻居网络拥有更多关系近的邻居节点, 则表示该节点越有可能向其邻居传播信息. 为此, 为了测量网络中节点之间的联系, 引入邻居重叠比 Q_{xy} ^[32] 对节点之间的关系远近进行衡量. 邻居重叠比 Q_{xy} 表示为

$$Q_{xy} = \frac{C_{xy}}{k_x - 1 + k_y - 1 - C_{xy}}, \quad (2)$$

其中 k_x 表示节点 x 的度, C_{xy} 表示节点 x 和 y 之间的公共邻居的数量. 在邻居网络中, 如果两个节点的公共邻居越多, 则表示两个节点越相似, 可以认为两个节点的关系越近. 为了衡量节点 x 在邻居网络中对信息传播的影响, 并量化信息在邻居网络的传播路径, 因此定义 $\text{NO}(x)$ 为邻居节点 x 的邻居重叠度, 其表达式定义如下:

$$\text{NO}(x, l) = \sum_{x, y \in \Gamma_l(u)} Q_{xy} + 1, \quad (3)$$

其中 x, y 是相同邻居网络中的节点. 通过这一度量指标, 能够更全面地理解节点在复杂网络信息传播中的位置和作用, 尤其是在涉及不同层次邻居网络的情境中. 在现实世界中, 这些不同邻居网络之间的联系和信息传播路径对于信息传播的速度和效果产生着重要影响. 因此, 引入邻居重叠度来评估节点在邻居网络中的联系紧密度和信息传播路径, 对于深入理解节点的重要性以及其在信息传播中所扮演的角色具有重要意义. 这种综合考虑节点在邻居网络中的位置和联系的方法, 旨在更加精准地评估节点的重要性、影响力以及其在信息传播中的贡献. 总体而言, 这一定义的目的是提供一种更全面、更综合的节点评估方法, 以更好地解析和分析复杂网络中的信息传播过程.

定义3 (KHC 系数) k 壳方法^[33] 是一种可以区分网络层次信息的经典方法, 其基本思想是将节点的度作为指标, 不断地删除节点并更新度值来测量节点位于网络中的位置. 如果节点位于较高的核心层, 则表征在信息传播的过程中处于有利的地位. h 指数是一种考虑节点邻域特征的中心性方法, 节点的 h 指数是指该节点周围邻居节点中, 至少有 h 个节点的度大于 h . 聚类系数^[34] 表示在节点的局部范围内形成三元组的趋势. 通常来说, 节点的聚类系数越高, 代表节点之间的拓扑连接越紧密, 其信息容易被重复传播, 节点的传播能力可能会受到限制. 相反, 如果节点之间拓扑连接越稀疏, 则表示信息会以更快的速度传播到整个网络. 因此, 节点的聚类系数对信息传播有负面效应. 考虑上诉因素, 给出一个新的衡量指标 (KHC 系数) 来刻画某个源节点的邻居节点 x 的拓扑特征对信息传播的影响, 其表达式如下

$$\text{KHC}(x) = \frac{\text{ks}(x) \times \text{hindex}(x)}{1 + C(x)}, \quad (4)$$

其中 $\text{ks}(x)$ 表示节点 x 的 k 壳值, $\text{hindex}(x)$ 表示节点 x 的 h 指数值, $C(x)$ 表示节点 x 的聚类系数. KHC 系数从拓扑结构的角度衡量节点在信息传播中的重要性和影响力. 通过综合考虑节点的度分布、位置、领域特征以及信息传播的效率, KHC 系数有助于更好地理解节点在网络中的信息传播过程中扮演的关键角色, 并能够更全面地评估节点在网络中的重要性, 尤其在信息传播中的影响力. KHC 系数还提供了一种综合性的指标来帮助深入理解网络的层次结构、信息传播效率以及节点在信息传播

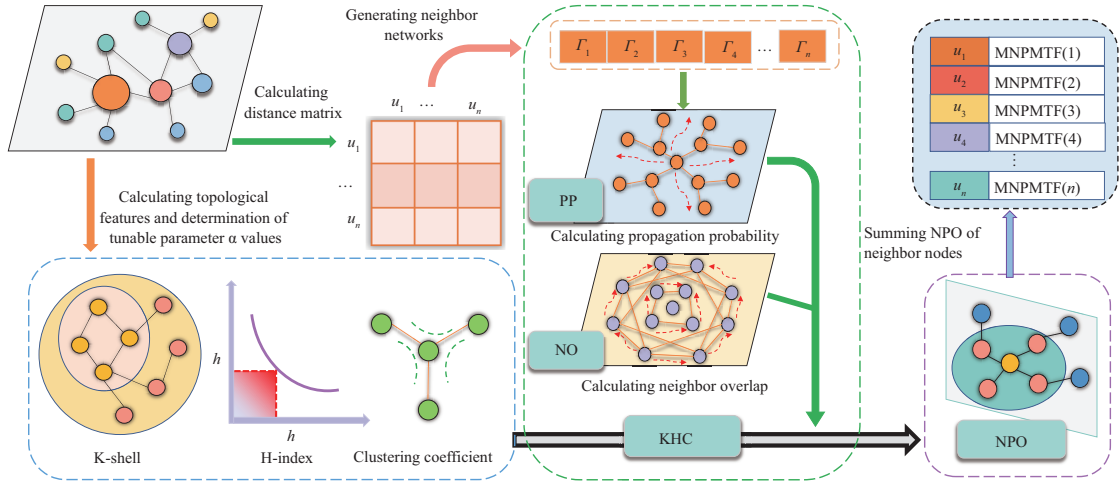


图 1 (网络版彩图) MNPMTF 算法框架图

Figure 1 (Color online) Framework diagram of MNPMTF algorithm

中的位置和作用. 较高的 KHC 系数意味着该节点在信息传播中处于更有利的位置, 其 k 壳值高、 h 指数值高、聚类系数低, 这意味着节点的核心层级较高, 其邻域节点有较多的高度连接节点, 且节点之间的拓扑连接相对稀疏, 有利于信息在网络中快速传播.

结合节点的传播特性和拓扑特征, 可以进一步计算源节点 u 的邻居节点 x 综合得分,

$$\text{NPO}(u, x, l) = \text{PP}(x, l) \times \text{NO}(x, l) \times \text{KHC}(x). \quad (5)$$

算法将节点的影响考虑为其多阶邻居的综合得分之和. 随着节点之间距离的增加, 节点的影响力会逐渐减弱, 更高阶的邻居节点会导致算法的计算复杂度增加, 特别是在大规模网络中, 会显著增加计算时间和资源消耗, 但基本不会再影响算法的精确度. 为了既考虑节点更广泛的局部影响力, 又保证算法具有优良效果, 同时还要避免更高阶邻居节点带来的巨大计算复杂度, 因此主要考虑 1 阶、2 阶和 3 阶邻居节点的影响, 从而提出如下 MNPMTF 算法:

$$\text{MNPMTF}(u) = \sum_{l=1}^{r=3} \sum_{x \in \Gamma_l(u)} \text{NPO}(u, x, l). \quad (6)$$

图 1 描述了 MNPMTF 算法框架. 首先, 该算法计算节点的拓扑结构特征 (k 壳、 h 指数和聚类系数 C), 并融合成节点的 KHC 系数, 这是因为在网络中, 节点的拓扑结构对其影响力有着重要影响. 综合多种拓扑特征, 可以更全面地描述节点在网络中的位置和连接情况, 从而更准确地评估其重要性. 同时, 算法考虑了高阶邻居节点的影响, 通过计算距离矩阵来生成节点的邻居网络. 在生成邻居网络后, 算法会计算邻居网络中的节点的传播概率 (propagation probability, PP) 和邻居重叠度 (neighbor overlap, NO). 这样做的理由是, 高阶邻居节点在信息传播和病毒传播等过程中可能发挥着重要的作用, 但传统算法往往忽视了这一点. 通过考虑高阶邻居节点的传播概率和重叠度, MNPMTF 算法更全面地评估了节点的传播能力, 提高了节点影响力的识别准确度.

其次, 算法采用综合评估的方法, 将 KHC 系数、传播概率和邻居重叠度相结合, 以得出邻居综合得分 NPO, 这种做法源于仅仅依靠单一的拓扑特征或传播概率, 评估节点影响力存在局限性, 综合多种特征可以更全面地评估节点在网络中的影响力. 最后算法由聚合邻居节点的综合得分得到源节点的

算法 1 MNPMTF 算法**输入:** 网络图 $G = (V, E)$;

```

1: 初始化网络结构信息, 确定可调参数  $\alpha$  取值;
2: for all node  $u$  in  $V$  do
3:   计算节点  $u$  的  $k$  壳值,  $ks(u)$ ;
4:   计算节点  $u$  的  $h$  指数值,  $hindex(u)$ ;
5:   计算节点  $u$  的聚类系数,  $C(u)$ ;
6: end for
7: for all node  $u$  in  $V$  do
8:    $MNPMTF(u) \leftarrow 0$ ;
9:   for  $l = 1$  to  $r$  do
10:    生成源节点  $u$  的邻居网络  $\Gamma_l(u)$ ;
11:    for all node  $x \in \Gamma_l(u)$  do
12:     计算邻居网络中的节点  $x$  的传播概率,  $PP(x, l)$ ;
13:     计算邻居网络中的节点  $x$  的邻居重叠度,  $NO(x, l)$ ;
14:     计算邻居网络中的节点  $x$  的 KHC,  $KHC(x)$ ;
15:     计算邻居网络中的节点  $x$  的综合得分,  $NPO(u, x, l)$ ;
16:      $MNPMTF(u) \leftarrow MNPMTF(u) + NPO(u, x, l)$ ;
17:    end for
18:   end for
19: end for
20: 对最终的结果  $MNPMTF(u)$  进行降序排序;

```

输出: 节点影响排序列表: $\{MNPMTF(u)|u \in V\}$.

节点影响, 这是因为在复杂网络中, 节点的影响力受其邻居节点的影响较大. 通过聚合邻居节点的综合得分, MNPMTF 算法更好地考虑了节点在网络中的传播能力和重要性, 从而提高了节点影响力的识别准确度. MNPMTF 算法的伪代码如算法 1 所示.

2.2 复杂度分析

对于给出的无权无向网络, 节点总数为 N , $\langle k \rangle$ 为平均度. 根据算法 1 的 1~5 行, 计算网络中所有节点的 k 壳值, h 指数值和聚类系数的复杂度 $O(N)$. 为了降低计算节点的 1 阶邻居、2 阶邻居和 3 阶邻居的时间复杂度, 首先计算网络中每个节点的距离矩阵, 复杂度为 $O(N \log N)$, 然后以复杂度 $O(N)$ 寻找节点距离小于等于 3 的元素. 由此, 生成节点的 1 阶邻居网络、2 阶邻居网络和 3 阶邻居网络的复杂度约为 $O(N \log N + N)$. 接下来, 计算 PP, NO 和 NPO 的复杂度为 $O(\langle k \rangle)$, $O(\langle k \rangle^2)$ 和 $O(\langle k \rangle)$, 所以算法中的 11~15 行的时间复杂度为 $O(\langle k \rangle^2)$. 综上, MNPMTF 算法的时间复杂度为 $O(N + N \log N + N + N \times \langle k \rangle^2) \approx O(N(\log N + \langle k \rangle^2))$. 对于大型网络来说, $\langle k \rangle$ 远小于 N , 因此 MNPMTF 算法可以应用于大规模稀疏网络.

3 实验设计与结果分析**3.1 数据集**

实验选取了 9 个不同的真实复杂网络数据集 (Jazz, USAir, NS, EEC, Email, PB, Yeast, GrQc, Router). 其中, Jazz 是爵士乐手的协作网络; USAir 是美国航空网络; NS 是从事网络科学的科学家的合作网络; EEC 描述了一个大型欧洲研究机构的机构成员之间的电子邮件交流; Email 是 Universitat

表 1 9 个真实网络的数据特征
Table 1 Nine data features of the real network^{a)}

Network	$ N $	$ E $	$\langle d \rangle$	$\langle k \rangle$	$\langle C \rangle$	β_{th}	α
Jazz	198	2742	2.2350	27.6970	0.6174	0.0258	2
USAir	332	2126	2.7381	12.8072	0.6252	0.0225	0.5
NS	379	914	6.0419	4.8232	0.7412	0.1247	2.5
EEC	986	16687	2.5869	33.8479	0.4070	0.0132	2.5
Email	1133	10903	3.606	9.6247	0.2202	0.0535	1.5
PB	1222	16714	2.7375	27.3552	0.3203	0.0123	0.5
Yeast	2361	7182	4.3763	6.0839	0.1301	0.0548	1.5
GrQc	4158	13422	6.0494	6.456	0.5569	0.0556	2.5
Router	5022	6528	2.4922	6.4488	0.0116	0.0729	2.5

a) $|N|$ denotes the number of nodes, $|E|$ denotes the number of edges, $\langle d \rangle$ denotes the average path strength length, $\langle k \rangle$ denotes the average degree, $\langle C \rangle$ denotes the average clustering coefficient, β_{th} denotes the propagation threshold, and α denotes the optimal parameters of MNPMTF and SPC.

Rovira i Virgili 的教职工和学生等用户之间的电子邮件交流网络; PB 是美国政治博客网络; Yeast 是芽殖酵母蛋白质相互作用网络; GrQc 是 arXiv 类别中广义相对论和量子宇宙学的电子版文章的合作网络; Router 是自治系统级别互联网结构的对称快照. 这些数据集可以从 <http://snap.stanford.edu/data> 和 <https://github.com/MLIF/Network-Data> 获得. 表 1 展示了这 9 个真实网络数据集的网络参数特征.

3.2 基准方法

为了验证所提出 MNPMTF 算法的性能, 将其与 7 种具有代表性的算法进行比较. 其中, 有 6 种算法是新文献提出的算法, 有 1 种是经典的网络参数算法. 这 7 种算法分别是 (1) 基于扩散概率的算法 SPC, 该算法^[27] 利用节点之间的扩散概率计算节点的重要性; (2) 基于潜在边缘权重的算法 HIC, 该算法^[19] 利用节点的位置信息、领域信息和局部拓扑结构信息来定义潜在边缘权重并评估节点的重要性; (3) 改进重力中心性的算法 KSGC, 该算法^[35] 在重力中心性的基础之上, 考虑节点的位置来评估网络中节点的重要性; (4) 基于局部和全局属性信息的算法 LGC, 该算法^[36] 考虑节点的局部属性信息和全局属性信息来衡量节点的重要性; (5) 基于模糊信息熵的算法 LFIC, 该算法^[14] 利用节点之间的最短路径形成模糊信息熵来评估节点的重要性; (6) 基于局部结构相似性的算法 ALSI, 该算法^[37] 考虑与邻居节点的相似度以及节点的位置等信息来衡量节点的重要性; (7) 度中心性算法 DC, 该算法^[12] 利用邻居节点的数量来衡量节点的重要性.

3.3 实验设置与评价指标

3.3.1 传播模型

易感 - 感染 - 恢复 (susceptible-infectious-recovered, SIR) 模型, 作为一个经典的传播模型, 既可以描述节点的病毒传播也可以描述信息传播. 在该模型中, 网络中的每个节点被划分为 3 种状态, 即易感 (S)、感染 (I) 和恢复 (R) 状态. 值得注意的是, 从 I 状态恢复为 R 状态的节点不能再被感染. 具体步骤如下: 在初始网络中, 只有一个源节点被感染, 它处于 I 状态, 而其余的节点是易感的 (健康的).

初始节点将感染其处于 S 状态的邻居, 感染概率为 β . 这时, 节点从 S 状态转变为 I 状态. 接下来, I 状态下的节点将恢复为 R 状态. 为简化问题, 设置恢复的概率 $\gamma = 1$. 这个过程一直持续到没有节点处于 I 状态的节点为止. 令 $F(t)$ 表示在 t 时间上处于 I 和 R 状态的累计节点数量. 随着时间的演化, 网络中的各状态的节点比例趋向于稳定状态, 此时, 用 F 表示稳定状态下的 $F(t)$, 它体现了初始节点的影响力. 在每个网络中进行 100 次独立的 SIR 仿真模拟实验并求其平均值, 以获得客观的初始节点影响力.

选择 SIR 模型的理由如下: 首先, 作为传播领域的经典模型, SIR 模型广泛应用于病毒传播和信息传播等现象, 因其建立在传统疾病传播理论之上, 具备实际意义. 其次, SIR 模型作为常见的评估工具, 帮助研究者深入了解传播过程的关键参数, 如感染率和恢复率. 此外, SIR 模型的适用范围广泛, 包括疾病传播、信息传播以及社交网络传播等场景. 因此, 在研究节点影响力时, 采用 SIR 模型能够涵盖多种传播情境.

3.3.2 评价指标

在此介绍 4 种标准评价指标用于后文的实验结果评价.

(1) Kendall 系数 τ 是用于衡量两个随机序列之间的相关性的统计值, 其定义为一致对和不一致对之间的差与总对数的比率. Kendall 系数 τ 的表达式如下

$$\tau = \frac{2(N_C - N_D)}{N(N-1)}, \quad (7)$$

其中 N_C 表示一致对的数量, N_D 表示不一致对的数量, N 表示序列的长度. 具体而言, 假设有通过特定算法产生的排名列表 A 和通过 SIR 模型进行模拟并达到稳态 F 时产生的影响力排名列表 B . 这两个序列中的第 i 个值分别用 a_i 和 b_i 来表示. 任意两个对应的值 (a_i, b_i) , (a_j, b_j) 组成一对序列, 如果满足 $(a_i - a_j)(b_i - b_j) > 0$, 则称这两个序列对是一致的; 相反, 如果满足 $(a_i - a_j)(b_i - b_j) < 0$, 则表示不一致. 在该指标中, 如果 τ 值越高, 表示两个序列的排名一致性越高, 相关性也越高.

(2) 综合累计分布函数 (comprehensive cumulative distribution function, CCDF)^[38] 是衡量算法的节点区分能力和分辨率的有效工具, 可以用以下形式表示:

$$\text{CCDF}(r) = \frac{N - \sum_{i=1}^r n_i}{N}, \quad (8)$$

其中 n_i 表示等级为 i 的节点数量. 当多个节点有相同的排名时, 很难区分这些节点的影响力差异. 因此, 从节点区分的角度来看, 如果一种算法能使处于同一等级的节点数相对较少, 那么该算法的 CCDF 曲线下降更慢, 斜率更平. 换言之, 如果一种算法 CCDF 曲线下降越缓慢, 则代表该算法的节点区分能力越强.

(3) 单调性也是衡量算法的节点区分能力和分辨率的有效措施, 其表达式为

$$\text{Mon}(R) = \left(1 - \frac{\sum_{r \in R} N_r(N_r - 1)}{N(N-1)}\right)^2, \quad (9)$$

其中, R 是由算法计算出的每个节点的影响力排名向量, N_r 是具有相同影响力的节点的数量. 具体来说, $\text{Mon} \in [0, 1]$. 当 $\text{Mon} = 1$ 时, 表明算法计算的节点等级都是不相同的. 当 $\text{Mon} = 0$ 时, 意味着所有节点的节点等级都是相同的值. 如果一种算法拥有更高的单调性, 则表示这种算法具有更好的节点区分能力.

(4) 不精确函数^[33]是衡量算法有效性的一个定量指标, 它被用来衡量由特定算法计算出的排名靠前的节点的传播能力和由 SIR 模型模拟并达到稳态 F 后获得的排名靠前的节点的传播能力之间的差异, 其描述为

$$\varepsilon(p) = 1 - \frac{S(p)}{S_{\text{eff}}(p)}, \quad (10)$$

其中 p 代表的是顶级节点的比例. q 表示排序后的前 $p \times N$ 个节点, $S(p)$ 表示的是由特定算法得到的前 q 个节点的平均影响, $S_{\text{eff}}(p)$ 表示的是由 SIR 模型模拟的前 q 个节点的平均影响. 从式 (10) 中可以看出 $\varepsilon(p)$ 的值越小, 表示该方法的效果越好.

3.4 实验结果与分析

本节首先进行一组实验, 以确定在不同复杂网络上所提出的 MNPMTF 算法最优参数 α 的值. 然后, 将 MNPMTF 算法与 7 种算法进行多个评价指标的比较, 以评估其综合性能.

3.4.1 最优参数 α 的确定

如前所述, 在 MNPMTF 算法中, 式 (1) 中的可调参数 α 用来控制网络中节点之间的感染概率. 由于不同网络的结构存在差异, 需要用一组实验来确定不同真实网络的最优参数 α . 参照传播阈值 β_{th} , 遍历 β_{th} 周围的感染概率来确定不同网络中的最优参数 α . 如果 α 过小, 会导致信息在节点之间的传播受阻; 如果 α 过大, 导致信息很容易传播到整个网络, 无法轻易区分节点的传播能力. 因此, 将 α 设置为 0.5~2.5 之间, 这是为了覆盖传播阈值附近一个相对广泛的范围, 以探索在不同感染率下算法的性能表现. 这样设置旨在寻找在实际网络中合适的参数取值, 使得信息传播过程的模拟更加准确和有效. 通过在这个范围内选择最优参数 α , 可以更好地适应不同网络的传播特征, 提高算法的预测能力. 进一步, 为了平衡复杂性和有效性, 以 0.5 的步长设置可调参数 α , 即 $\alpha \in [0.5, 1, 1.5, 2, 2.5]$. 接下来, 使用肯德尔系数 τ 来评估不同可调参数 α 下所提算法的准确性. 图 2 展示了 MNPMTF 算法在 9 个真实网络上通过调整可调参数 α 获得的肯德尔系数 τ . 为了显示普遍性, 表 2 列出了不同真实网络中通过不同可调参数 α 获得的平均 τ 值, 而具有最高平均 τ 值对应的可调参数 α 为该网络的最优参数 α .

从图 2 和表 2 中可以看出, 提出的算法在 USAir, EEC 和 PB 这些网络上的性能近似, 换言之, 提出的算法在这些网络上对参数 α 不太敏感. 而对于其他网络, 当 α 值改变, 对应的 τ 值也会发生显著改变. 表 2 展示了 9 个真实世界网络上的平均 τ 值, 并且在每个网络上加粗了最优平均 τ 值. 从表 2 中可以看出, 在不同的真实世界网络中, 对应的最优参数 α 是不一样的. 为了更准确地识别网络中的影响力节点, 选取每个网络中的最优参数 α 作为 MNPMTF 算法和 SPC 算法的最优参数.

3.4.2 肯德尔系数

在这个实验中, SIR 模型中感染概率 β 是一个重要参数. 也就是说, 当感染概率 β 过小时, 节点之间的感染将局限在局部范围内, 无法传播到整个网络. 当感染概率 β 过大时, 节点会很容易感染网络中的大量节点. 也就是, 过大或过小的感染概率 β 都无法正确衡量节点的传播能力. 因此, 这组实验将节点之间的感染概率设置在网络的传播阈值附近. 具体来说, 对于不同的网络, 将感染概率 β 设置为网络传播阈值的倍数 $p \times \beta_{\text{th}}$, 其中倍数 p 的范围设置为 0.6~2. 为了验证 MNPMTF 算法的性能, 将其与 7 种代表性算法进行肯德尔系数的比较. 如图 3 所示, 显然, 提出的算法的性能优于其他方法. 其中, MNPMTF 算法在 Jazz, NS, EEC 和 Email 这些网络中表现出最佳的性能; 在 USAir, PB, Yeast, GrQc 和 Router 中, 当感染概率 $\beta > \beta_{\text{th}}$ 时, 提出的算法取得最佳的性能. 这是由于此刻, 节点之间的传播不再局限于局部范围, 能够更好地体现网络中节点的传播能力. 为了展示算法的普遍性,

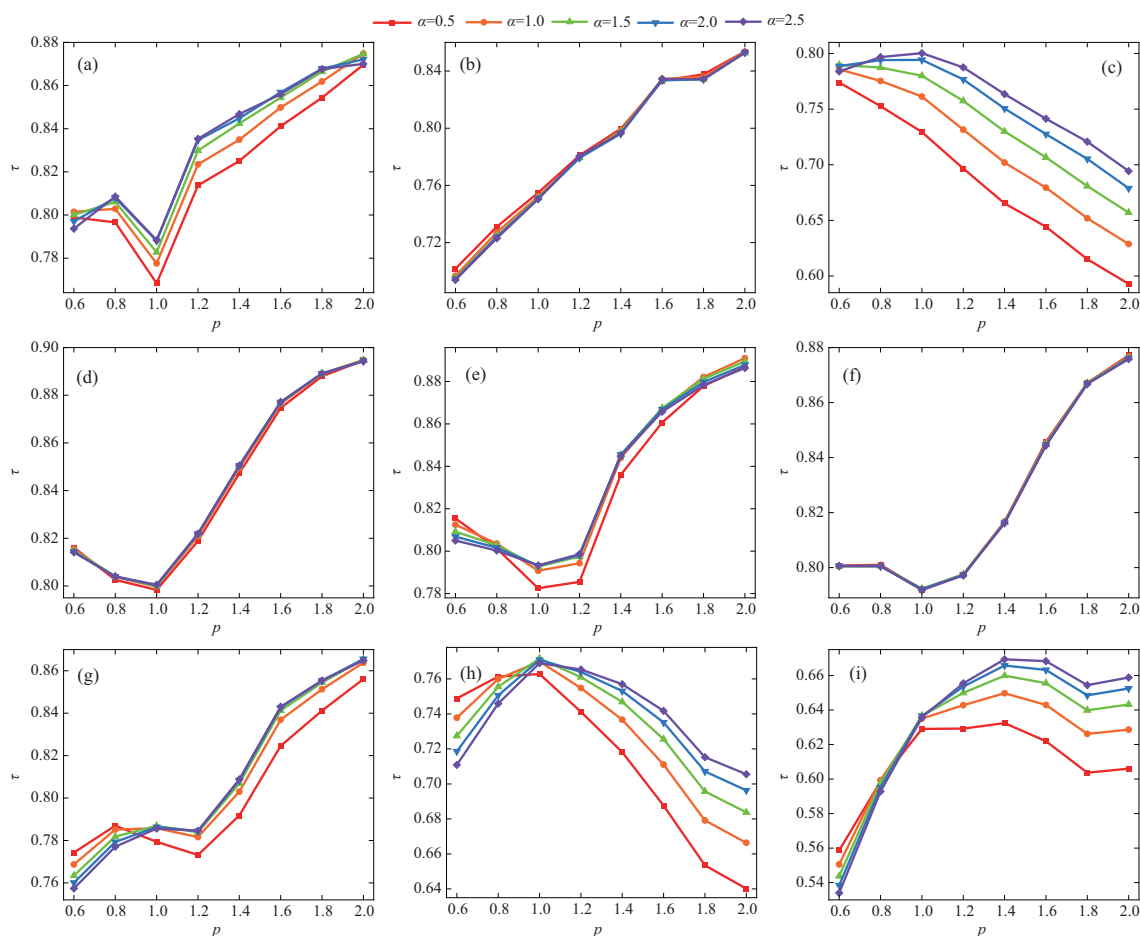


图 2 (网络版彩图) 通过 MNPMTF 算法获得的排名列表与通过 SIR 模拟产生的实际排名列表之间的肯德尔系数 τ , p 表示网络传播阈值 β_{th} 的倍数, α 表示可调参数

Figure 2 (Color online) Kendall coefficient τ between the ranked list obtained by the MNPMTF algorithm and the actual ranked list generated by the SIR simulation. p denotes the multiple of the network propagation threshold β_{th} . α denotes the adjustable parameter. (a) Jazz; (b) USAir; (c) NS; (d) EEC; (e) Email; (f) PB; (g) Yeast; (h) GrQc; (i) Router

表 3 中列出了不同算法在 9 个真实网络上的平均 τ 值. 表 3 中对最优 τ 值进行了加粗. 可以看出提出的 MNPMTF 算法明显优于对比的其他方法. 以上结果说明, MNPMTF 算法相比于其他比较算法具有更好的准确性和普遍性.

3.4.3 综合累计分布函数与单调性

在该实验中, 如果一种算法的 CCDF 曲线下降越缓慢, 斜率越平缓, 则表示该算法能够更好地区分节点. 不同的算法在 9 个网络的 CCDF 如图 4 所示, 可以看出提出的 MNPMTF 算法在 9 个网络中下降缓慢, 斜率相对平缓. 具体来说, 在 USAir, NS, Yeast, GrQc 和 Router 网络中, 可以明显看出 MNPMTF 算法的曲线下降最慢. 在 Jazz, EEC, Email 和 PB 网络中, 从局部图看, MNPMTF 算法的 CCDF 曲线下降最慢. 尽管在一些网络中的曲线早期阶段, MNPMTF 算法并不是最优, 但是随着节点等级的增加, MNPMTF 最终也会到达最优的效果.

除了综合累计分布函数以外, 单调性也是一种衡量节点区分能力的一种指标. 单调性越高, 表示

表 2 9 个真实网络上通过不同参数 α 计算的 MNPMTF 算法的平均 τ 值

Table 2 Average τ values of the MNPMTF algorithm calculated by different parameters α on nine real networks

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$
Jazz	0.820978	0.828352	0.832088	0.833669	0.833283
USAir	0.786599	0.784697	0.783593	0.783128	0.783255
NS	0.683774	0.714496	0.736122	0.751986	0.761054
EEC	0.842535	0.843675	0.84392	0.84393	0.843935
Email	0.830804	0.835624	0.835736	0.834956	0.834068
PB	0.824736	0.824574	0.824473	0.824286	0.824164
Yeast	0.803446	0.809537	0.810515	0.810265	0.80961
GrQc	0.714194	0.726966	0.733444	0.736967	0.738834
Router	0.609924	0.621936	0.628275	0.631735	0.633668

表 3 8 种算法在 9 个网络中的平均 τ 值

Table 3 Average τ values of eight algorithms in nine networks

Network	SPC	HIC	KSGC	LGC	LFIC	DC	ALSI	MNPMTF
Jazz	0.78310	0.82463	0.80098	0.78761	0.75590	0.78295	0.80593	0.83366
USAir	0.73575	0.78285	0.76065	0.73293	0.64703	0.72261	0.74628	0.78659
NS	0.66617	0.68638	0.70829	0.62488	0.65166	0.57556	0.62865	0.76105
EEC	0.82505	0.84157	0.83320	0.81769	0.79183	0.81688	0.82396	0.84393
Email	0.80807	0.82963	0.80135	0.78155	0.79327	0.77844	0.79329	0.83573
PB	0.80420	0.82382	0.81754	0.80410	0.79263	0.80925	0.81208	0.82473
Yeast	0.78836	0.80361	0.78650	0.73621	0.69503	0.70732	0.69556	0.81051
GrQc	0.71073	0.71771	0.71467	0.6483	0.65356	0.62711	0.64508	0.73883
Router	0.61457	0.59803	0.60736	0.52876	0.61681	0.39077	0.53506	0.63362

表 4 8 种算法在 9 个真实网络上的单调性

Table 4 Monotonicity of eight algorithms on nine real networks

Network	SPC	HIC	KSGC	LGC	LFIC	DC	ALSI	MNPMTF
Jazz	0.99897	0.99928	0.99928	0.99928	0.99887	0.96594	0.99928	0.99958
USAir	0.99413	0.99402	0.99500	0.99494	0.99410	0.76420	0.99403	0.99731
NS	0.99508	0.99508	0.99424	0.99508	0.99308	0.85858	0.99447	0.99807
EEC	0.99989	0.99970	0.99950	0.99989	0.99989	0.88736	0.99944	0.99996
Email	0.99986	0.99986	0.99985	0.99987	0.99988	0.95710	0.99984	0.99989
PB	0.99930	0.99926	0.99926	0.99956	0.99930	0.93284	0.99925	0.99970
Yeast	0.99682	0.99163	0.99682	0.99683	0.99671	0.73412	0.99671	0.99719
GrQc	0.99936	0.99811	0.99937	0.99939	0.99936	0.79160	0.99749	0.99951
Router	0.99594	0.86859	0.99676	0.99704	0.99639	0.28861	0.93866	0.99739

节点的区分能力越好. 表 4 展示了 8 种算法在 9 个网络上的单调性结果, 并对每个网络上的最高单调性值进行加粗. 从表 4 中可以看出, MNPMTF 算法在所有网络中都具有最高的单调性值. 这表明

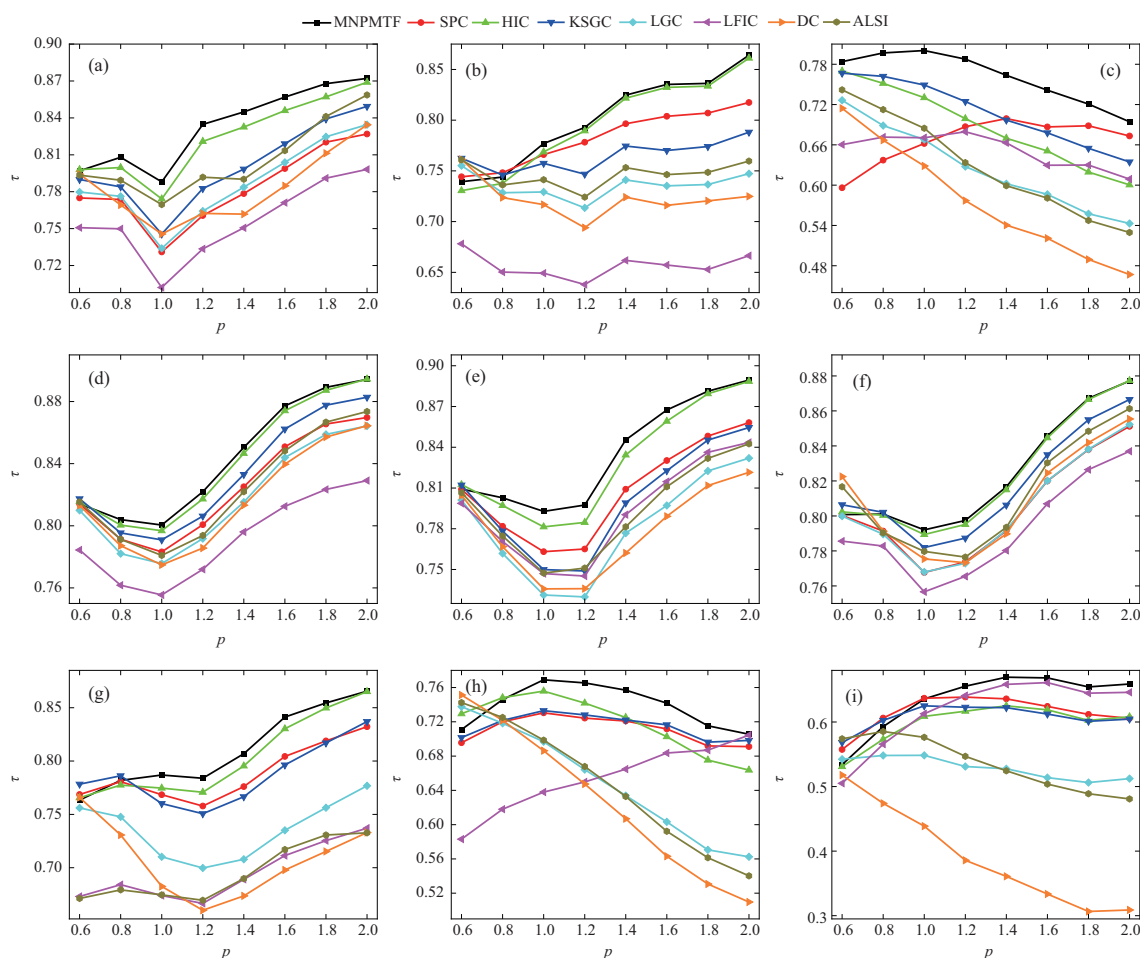


图 3 (网络版彩图) 8 种算法获得的排名列表与通过 SIR 模拟产生的实际排名列表之间的肯德尔系数 τ . p 表示网络传播阈值 β_{th} 的倍数

Figure 3 (Color online) Kendall coefficient τ between the ranked list obtained by the eight algorithms and the actual ranked list generated by SIR simulation. p denotes the multiple of the network propagation threshold β_{th} . (a) Jazz; (b) USAir; (c) NS; (d) EEC; (e) Email; (f) PB; (g) Yeast; (h) GrQc; (i) Router

MNPMTF 算法在识别影响力节点方面具有更高的节点区分能力. 以上两个实验证明了 MNPMTF 算法具有优越的节点区分能力.

3.4.4 不精确函数

为进一步证明 MNPMTF 算法能够有效地识别网络中的影响力节点, 比较了不同方法的不精确函数值. 图 5 展示了 8 种算法在 9 个不同网络上的不精确函数值. 如图 5 所示, 在 NS, Email, Yeast, GrQc 和 Router 网络中, 所提出 MNPMTF 算法的不精确函数值 $\varepsilon(p)$ 都是最低的. 这显示了 MNPMTF 算法的最佳性能, 也表明了该算法能够更加有效地识别网络中排名靠前的影响力节点. 此外, 就 Jazz 网络而言, 相对于其他网络, 曲线的波动较大. 排除了实验次数的原因后 (实验次数增加到 5000 结果仍然类似), 认为是 Jazz 网络规模较小所致. 当网络规模较小且选取的节点数较少时, 网络中的节点之间的连接和信息传播行为可能更加敏感. 因此, 在节点数较少的情况下, 即使出现微小的变化, 也可能导致曲线出现较大的波动. 对于 Jazz, USAir, EEC 和 PB 网络来说, MNPMTF 算法的 $\varepsilon(p)$ 不是一直处

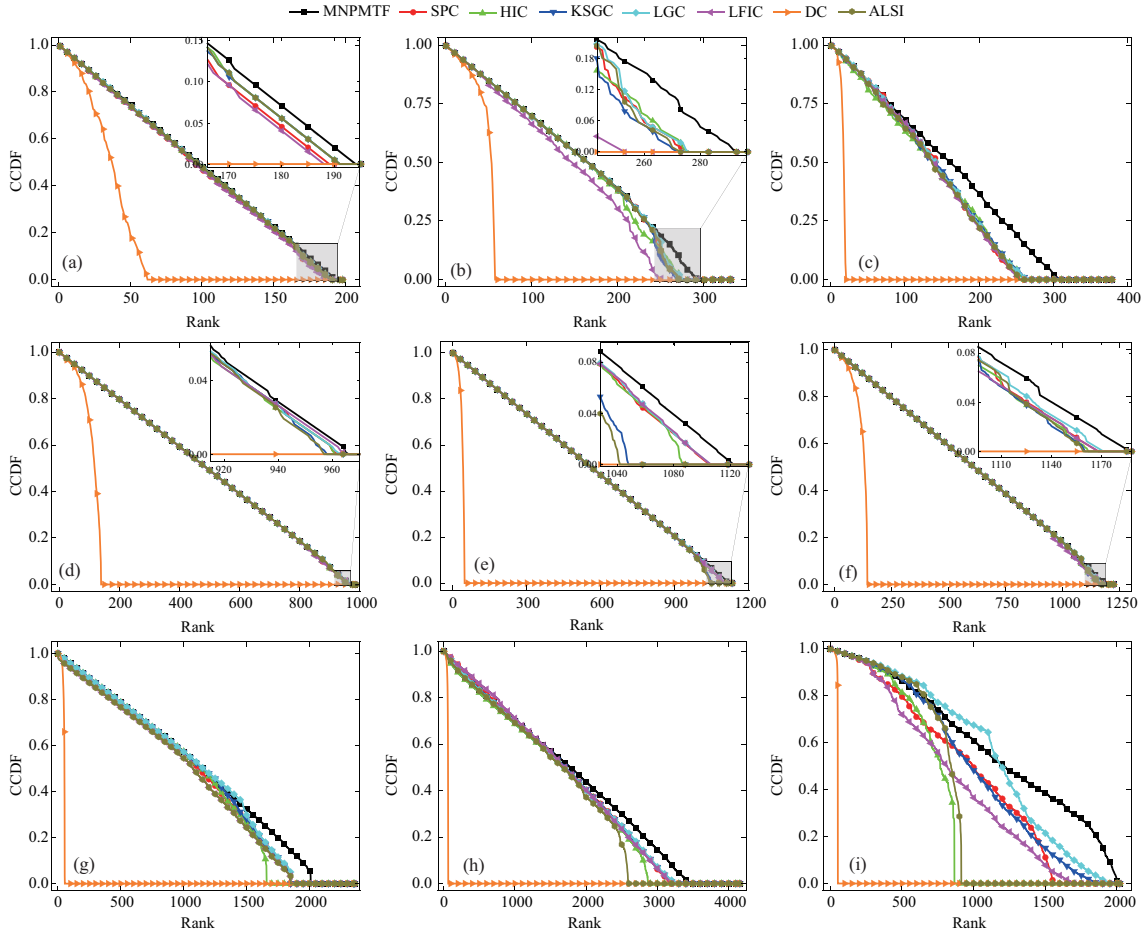


图 4 (网络版彩图) 9 个网络中 8 种算法生成的 CCDF 曲线

Figure 4 (Color online) CCDF curves generated by eight algorithms in nine networks. (a) Jazz; (b) USAir; (c) NS; (d) EEC; (e) Email; (f) PB; (g) Yeast; (h) GrQc; (i) Router

于最低状态,但是在大多数尺度上是最底的. 因为 MNPMTF 算法考虑的信息传播过程不仅仅是从低阶邻居到高阶邻居,而且还考虑了节点邻居对信息传播过程的影响,以及节点的局部拓扑结构,更符合现实世界信息传播过程,所以 MNPMTF 算法可以更准确地识别大多数网络中排名靠前的节点.

4 结论

本文提出了一种新的识别影响力节点的算法 (MNPMTF), 用于有效识别影响力节点, 并对其进行了实验验证和分析. MNPMTF 算法通过综合考虑节点的 3 阶邻居范围内的邻居节点的多种信息, 包括邻居节点的传播概率、邻居重叠度和拓扑结构信息, 来准确评估节点的影响力. 在 9 个真实网络上进行了多组实验, 并将 MNPMTF 算法与其他 7 种代表性算法进行了综合比较. 实验结果表明, MNPMTF 算法在排序精度、区分能力和有效性等方面表现出优于其他算法的性能, 验证了该算法的优势. 此外, MNPMTF 算法还具有较低的计算复杂度, 适用于处理大规模稀疏网络. 这使得该算法在实际应用中具备更好的可扩展性和效率.

MNPMTF 算法在节点影响力识别方面展现出了良好的潜力, 为未来研究提供了新的思路. 在应

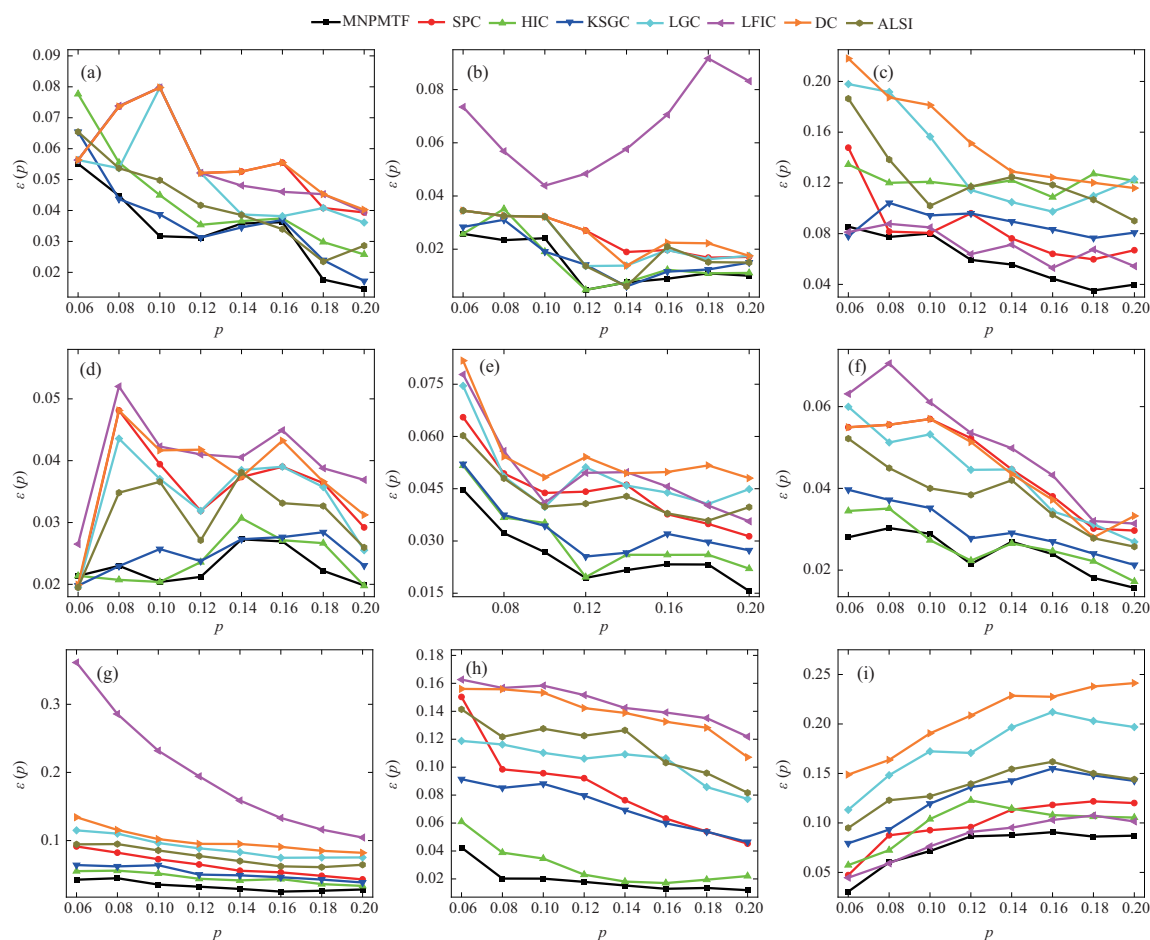
图 5 (网络版彩图) 9 个网络中通过 8 种算法生成的 $\varepsilon(p)$ 曲线

Figure 5 (Color online) $\varepsilon(p)$ curves generated by eight algorithms in nine networks. (a) Jazz; (b) USAir; (c) NS; (d) EEC; (e) Email; (f) PB; (g) Yeast; (h) GrQc; (i) Router

用方面, MNPMTF 算法既适用于无向网络也适用于有向网络, 可以通过考虑链路方向和引入权重值来衡量节点传播概率. 在节点路径方面, 虽然本文考虑了节点之间的最短路径来衡量传播概率, 但其他非最短路径也可能对节点的传播概率产生影响, 因此可以进一步探索其他非最短路径的影响. 以上这些方向可能会对 MNPMTF 算法的改进和拓展提供有价值的参考.

参考文献

- Li M, Liu R R, Lü L, et al. Percolation on complex networks: theory and application. *Phys Rep*, 2021, 907: 1–68
- Yang H C, Liu Y J, Wan Q, et al. A bio-inspired optimal network division method. *Phys A-Stat Mech Its Appl*, 2019, 527: 121259
- Wang G, Wang Y, Li J, et al. A multidimensional network link prediction algorithm and its application for predicting social relationships. *J Comput Sci*, 2021, 53: 101358
- Boers N, Goswami B, Rheinwalt A, et al. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 2019, 566: 373–377
- Choi D, Oh H, Chun S, et al. Preventing rumor spread with deep learning. *Expert Syst Appl*, 2022, 197: 116688
- Ahmed H, Howton T C, Sun Y, et al. Network biology discovers pathogen contact points in host protein-protein

- interactomes. *Nat Commun*, 2018, 9: 2312
- 7 Gao C, Jiang S H, Wang Z, et al. A novel method to identify influential stations based on dynamic passenger flows. *Sci Sin Inform*, 2021, 51: 1490–1506 [高超, 蒋世洪, 王震, 等. 基于动态客流的城市轨道交通关键站点识别. *中国科学: 信息科学*, 2021, 51: 1490–1506]
- 8 Xu G, Guo J, Yang P. TNS-LPA: an improved label propagation algorithm for community detection based on two-level neighbourhood similarity. *IEEE Access*, 2020, 9: 23526–23536
- 9 Guerreiro L, Silva F N, Amancio D R. A comparative analysis of knowledge acquisition performance in complex networks. *Inf Sci*, 2021, 555: 46–57
- 10 Liu Q, Zhu Y X, Jia Y, et al. Leveraging local h-index to identify and rank influential spreaders in networks. *Phys A-Stat Mech Its Appl*, 2018, 512: 379–391
- 11 Bae J, Kim S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Phys A-Stat Mech Its Appl*, 2014, 395: 549–559
- 12 Freeman L C. Centrality in social networks conceptual clarification. *Soc Network*, 1978-1979, 1: 215–239
- 13 Lü L, Zhou T, Zhang Q M, et al. The H-index of a network node and its relation to degree and coreness. *Nat Commun*, 2016, 7: 10168
- 14 Zhang H, Zhong S, Deng Y, et al. LFIC: identifying influential nodes in complex networks by local fuzzy information centrality. *IEEE Trans Fuzzy Syst*, 2021, 30: 3284–3296
- 15 Namtirtha A, Dutta B, Dutta A. Semi-global triangular centrality measure for identifying the influential spreaders from undirected complex networks. *Expert Syst Appl*, 2022, 206: 117791
- 16 Curado M, Tortosa L, Vicent J F. A novel measure to identify influential nodes: return random walk gravity centrality. *Inf Sci*, 2023, 628: 177–195
- 17 Yang P, Meng F, Zhao L, et al. AOGC: an improved gravity centrality based on an adaptive truncation radius and Omni-channel paths for identifying key nodes in complex networks. *Chaos Solitons Fractals*, 2023, 166: 112974
- 18 Zheng J, Liu J. A new scheme for identifying important nodes in complex networks based on generalized degree. *J Comput Sci*, 2023, 67: 101964
- 19 Meng L, Xu G, Yang P, et al. A novel potential edge weight method for identifying influential nodes in complex networks based on neighborhood and position. *J Comput Sci*, 2022, 60: 101591
- 20 Zhou M Y, Wu X Y, Cao Y, et al. A novel method to identify multiple influential nodes in complex networks. *Sci Sin Inform*, 2019, 49: 1333–1342 [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]
- 21 Dong C, Xu G, Meng L, et al. CPR-TOPSIS: a novel algorithm for finding influential nodes in complex networks based on communication probability and relative entropy. *Phys A-Stat Mech Its Appl*, 2022, 603: 127797
- 22 Li X, Li C, Wang J B. *Theory of Spreading on Complex Networks: Hidden Rules of Epidemics*. Beijing: Higher Education Press, 2020. 10 [李翔, 李聪, 王建波. 复杂网络传播理论 —— 流行的隐秩序. 北京: 高等教育出版社, 2020. 10]
- 23 Xu G, Meng L. A novel algorithm for identifying influential nodes in complex networks based on local propagation probability model. *Chaos Solitons Fractals*, 2023, 168: 113155
- 24 Qu J Y, Tang M, Liu Y, et al. The relative importance of structure and dynamics on node influence in reversible spreading processes. *Front Phys*, 2021, 16: 51503
- 25 Liu Y, Tang M, Zhou T, et al. Improving the accuracy of the k-shell method by removing redundant links: from a perspective of spreading dynamics. *Sci Rep*, 2015, 5: 13172
- 26 Qu J, Tang M, Liu Y, et al. Identifying influential spreaders in reversible process. *Chaos Solitons Fractals*, 2020, 140: 110197
- 27 Ai J, He T, Su Z, et al. Identifying influential nodes in complex networks based on spreading probability. *Chaos Solitons Fractals*, 2022, 164: 112627
- 28 Yu S, Gao L, Xu L, et al. Identifying influential spreaders based on indirect spreading in neighborhood. *Phys A-Stat Mech Its Appl*, 2019, 523: 418–425
- 29 Chen G, Zhou S, Liu J, et al. Influential node detection of social networks based on network invulnerability. *Phys Lett A*, 2020, 384: 126879
- 30 Chen D B, Sun H L, Tang Q, et al. Identifying influential spreaders in complex networks by propagation probability

- dynamics. *Chaos*, 2019, 29: 033120
- 31 Tong T C, Dong Q, Sun J S, et al. Vital spreaders identification synthesizing cross entropy and information entropy with Kshell method. *Expert Syst Appl*, 2023, 224: 119928
 - 32 Meo P D, Ferrara E, Fiumara G, et al. On Facebook, most ties are weak. *Commun ACM*, 2014, 57: 78–84
 - 33 Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. *Nat Phys*, 2010, 6: 888–893
 - 34 Zhang D, Jia L, Ning J, et al. Power grid structure performance evaluation based on complex network cascade failure analysis. *Energies*, 2023, 16: 990
 - 35 Yang X, Xiao F. An improved gravity model to identify influential nodes in complex networks based on k-shell method. *Knowledge-Based Syst*, 2021, 227: 107198
 - 36 Ullah A, Wang B, Sheng J F, et al. Identifying vital nodes from local and global perspectives in complex networks. *Expert Syst Appl*, 2021, 186: 115778
 - 37 Wang F, Sun Z, Gan Q, et al. Influential node identification by aggregating local structure information. *Phys A-Stat Mech Its Appl*, 2022, 593: 126885
 - 38 Helton J C. Probability, conditional probability and complementary cumulative distribution functions in performance assessment for radioactive waste disposal. *Reliab Eng Syst Saf*, 1996, 54: 145–163

Identification of high-influential nodes based on multi-order neighbor propagation metrics and topological features

Yu LUO¹, Jianbo WANG^{1,2*}, Ping LI¹, Zhanwei DU² & Xiaoke XU^{3,4*}

1. *School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China;*

2. *School of Public Health, The University of Hong Kong, Hong Kong 999077, China;*

3. *Computational Communication Research Center, Beijing Normal University, Zhuhai 519087, China;*

4. *School of Journalism and Communication, Beijing Normal University, Beijing 100875, China*

* Corresponding author. E-mail: jwang@swpu.edu.cn, xuxiaoke@bnu.edu.cn

Abstract How to quantitatively evaluate the influence of nodes in complex networks is an important research topic because it helps to understand the structure and functionality of networks. Most existing methods are primarily based on analyzing the network's inherent topology, needing more comprehensive utilization of propagation properties of multi-order neighboring nodes and topological information, significantly impacting identifying influential nodes. To this end, this paper proposes an algorithm called multi-order neighbor propagation metrics and topological features (MNPMTF) to effectively identify influential nodes in complex networks. Firstly, the algorithm combines a propagation model and shortest paths to characterize the propagation probability of neighboring nodes, thereby quantifying the likelihood of information spreading between nodes. Secondly, the algorithm considers the neighbor overlap ratio among multi-order neighbors to form the neighbor overlap degree, thereby quantifying the propagation paths of information in the neighbor network. Thirdly, the algorithm utilizes the k-shell index, h-index, and clustering coefficient to form a new index called the KHC coefficient, which describes the topological features of nodes. Finally, the algorithm integrates the propagation probability, neighbor overlap degree, and topological features within a 3-order neighbor range to evaluate the influence of nodes. Extensive experiments on nine entire networks demonstrate that the proposed algorithm outperforms seven representative methods regarding ranking accuracy, effectiveness, and discriminability, providing a new approach for assessing node influence in complex networks.

Keywords propagation probability, neighbor overlap, KHC coefficient, influential nodes, complex networks