基于模体结构和度信息的关键节点组识别

杨云云，张辽，于海龙，王力
（太原理工大学电气与动力工程学院，山西 太原 030024）

摘 要：为了探究具有更小规模的高阶结构对关键节点组的影响，以优化网络传播为目标，提出了一种基于模体结构和度信息的关键节点组识别算法。基于模体结构对节点影响力进行评估，挖掘模体结构的核心节点，使用多准则妥协解排序（VIKOR）法将其与度信息进行融合，并利用种子排除算法对种子节点的邻居进行排除，有效减小影响力重叠问题。在 SIR 传播模型的基础上，选取 6 个不同的无向网络与 4 种基准算法进行比较，实验结果表明，所提算法在准确性和稳定性方面表现出更好的性能。

关键词：模体；关键节点组；影响力最大化
中图分类号：TP391
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024052

Identification of key node groups based on motif structure and degree information

YANG Yunyun, ZHANG Liao, YU Hailong, WANG Li
School of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China

Abstract: In order to explore the impact of higher-order structures with smaller scales on key node group mining problems and with the goal of optimizing network propagation, a key node group recognition algorithm was proposed based on motif structure and degree information. Firstly, the influence of nodes was evaluated based on the motif structure, and the core nodes of the motif structure were excavated. Then, the VIKOR method was used to fuse it with degree information. Finally, the seed exclusion algorithm was used to exclude the neighbors of the seed nodes, effectively reducing the problem of influence overlap. Based on the SIR propagation model, six different undirected networks were selected for comparison with four benchmark algorithms. The experimental results show that the proposed algorithm performs better in terms of accuracy and stability.

Keywords: motif, key node group, influence maximization

0 引言

影响力最大化问题作为复杂网络分析的关键问题之一，与多个领域紧密相关，引起许多研究人员的关注。面对大规模的复杂网络，有效识别网络中有影响力的节点，在市场营销[1]、抑制谣言传播[2]、疾病控制[3]等方面具有重要的理论意义和实际意义。如何设计一个有效的策略来确定一组有影响力的节点作为初始传播源使它能尽快地传播信息并最终覆盖到整个网络[4]，对理解和控制传播过程至关重要。

对于如何准确地挖掘网络中的一组关键节点，Kempe 等[5]用贪婪（Greddy）算法证明了这是一个 NP-hard 问题，该算法的准确率为 63%，且具有较长的运行时间。Leskovec 等[6]进一步提出了 CELF（cost-effective lazy forward）算法，基于传播模型的子模特性，减少对网络中一些节点的重复计算，使算法的运行时间提高了约 700 倍。Goyal 等[7]对 CELF

收稿日期：2023−10−12；**修回日期：**2024−01−17
基金项目：国家自然科学基金资助项目（No.62006169）
Foundation Item: The National Natural Science Foundation of China (No.62006169)

算法进一步优化，在一轮迭代中进行两次模拟，提高了算法的运行效率。Borgs 等[8]提出了反向影响力采样方法，使用反向采样生成大量的随机反向可达集，然后使用贪婪算法，选择节点加入关键节点组。考虑到贪婪算法的复杂度较高，许多研究人员提出了通过启发式算法来解决关键节点组挖掘问题。Chen 等[9]提出了度折扣（DegreeDiscount）算法，该算法结合传播动力学，考虑传播概率对节点的度值进行折扣，从而提高了算法的准确性。Wang 等[10]在 DegreeDiscount 算法的基础上，通过考虑邻居中种子的折扣差异和传播过程中的冗余削弱机制，提出了一种高效的启发式算法 RWTDD（redundancy weakening and two types of seeds into degree discount）。Zhao 等[11]引入图着色的方法来挖掘关键节点集合。Liu 等[12]提出 LIR（local index rank）算法，挖掘局部度值最大节点加入关键节点组，该算法具有极高的效率，但不够稳定。Zhang 等[13]利用投票的思想提出 VoteRank 算法，赋予每个节点投票能力和得分能力，对节点的得分能力进行重复迭代计算。高菊远等[14]将节点覆盖范围作为节点选取的中心性评价指标，提出了基于节点覆盖范围的影响力最大化算法 NCA（node coverage algorithm）。Liu 等[15]认为节点的初始投票能力应该根据重要性进行量化，同时节点对其邻居的投票数目会根据邻居对该节点的吸引力而变化，有效提高 VoteRank 算法的性能。Wang 等[16]通过节点的边缘权重来计算信息熵值，以选择有影响的节点，同时考虑了节点的第一级边缘和第二级边缘。此外，社区结构也经常被用来挖掘种子节点，Shang 等[17]对其进行改进，提出了 CoFIM 算法，将传播分为 2 个阶段，第一阶段是从种子到其邻居节点，第二阶段是从邻居节点到社区中的其他节点，其准确性得到了提升，且运行时间取决于种子节点的数量。Bao 等[18]提出 HC（heuristic clustering）算法，对网络进行聚类，选取每个节点簇的中心节点加入关键节点组。Bozorgi 等[19]引入了竞争传播模型，并在该模型下计算了节点在自己社区内的局部扩散。Beni 等[20]合并了具有相似信息扩散结构的群落，并将低扩展群落排除在种子选择之外。

相较于启发式方法，贪婪算法能更好地识别一组重要节点，但由于其时间和空间复杂度很高，不便于在大规模实际网络上应用。基于社区结构的关键节点组识别算法作为启发式算法的重要组成部分，能够有效挖掘网络中的影响力节点，但算法的准确性依赖于社区划分的质量，导致基于社区的关键节点组识别算法具有一定的局限性。因此从实用性出发，本文提出基于模体结构和度信息的启发式算法，首先利用多准则妥协解排序（VIKOR）方法将基于模体的评估指标与对传播具有重要影响的度信息进行融合，然后基于种子排除理论避免种子节点密集造成的影响力重叠问题，提出了一种基于模体结构和度信息的关键节点组挖掘算法。实验结果表明，本文算法与 4 种基准算法相比能更好地识别出网络中的关键节点组。

1　模体

网络模体是指相较于随机网络，实际网络中频繁出现的连通子结构。研究表明[21]，与一般的网络子结构相比，模体对网络的结构及功能具有重要意义。对于给定的连通子结构，Z 得分是识别其是否为模体的重要指标，当 Z 得分的值大于 0 时，将该连通子结构确定为模体。Z 得分的计算式为

$$Z=\frac{N_{\text{real}}-\langle N_{\text{rand}}\rangle}{\sigma_{\text{rand}}} \tag{1}$$

其中，N_{real} 表示该子结构在实际网络中出现的次数，$\langle N_{\text{rand}}\rangle$ 表示该子结构在随机网络出现的平均次数，σ_{rand} 表示该子结构在随机网络中出现的标准差。

根据模体结构中所包含的节点个数，可以将模体分为不同的阶数。如果网络中检测到的一类模体结构包含 3 个节点，则称其为三阶模体。图 1 为网络中可能存在的部分无向三阶和四阶模体结构。对于不同节点个数的模体，模体存在的类型会随着节点个数的增加而迅速增长，使模体挖掘的计算复杂度增加。研究表明，三阶模体和四阶模体对网络整体特性及组成结构有较大的影响[22]。相较于三阶模体和四阶模体，拥有更多节点的模体对网络信息挖掘并没有表现出更好的效果，这可能是由于模体节点个数越多，其拓扑结构中可能包含的噪声也越多，从而性能下降[23]。

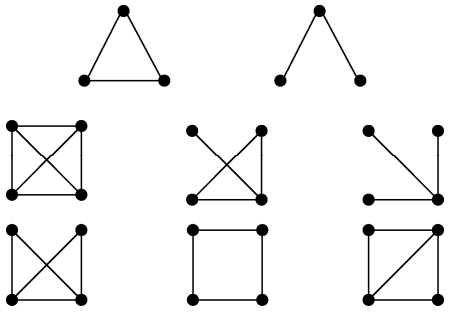


图 1　网络中可能存在的部分无向三阶和四阶模体结构

另外，在大部分网络中，三阶模体广泛存在，具有普遍性、代表性。因此，本文选用无向三阶模体为研究对象。

2 方法介绍

2.1 节点影响力计算

基于社团划分影响力最大化算法为基于网络模体结构识别关键节点组提供了思路，即可以利用一定的策略选择网络中局部更有影响力的节点加入关键节点组，同时避免了关键节点组内节点之间平均路径较小而导致的影响力重叠问题，相较于传统的启发式算法提高了效率和精度。

模体结构相较于只存在点对点交互的连边，还有多节点之间的交互模式，所以传播在模体内部更易进行。模体中一旦某个节点被激活，模体中的其他节点也更易被激活。模体顶点度指标[24]从节点参与模体的数目来对节点的传播影响力进行评价，可能导致最终选择的节点组中部分节点共同参与模体的次数较多，造成节点之间影响力重叠问题。本文对每个模体中的核心节点进行挖掘，对于模体中度值最大的节点，其不仅与模体内部节点有联系，还与模体之外其他节点之间存在连接，相较于模体内部节点，模体之外其他节点在网络信息的传播上具有更大优势，因此对于每个模体结构，选取其中度值最大的节点，称其为该模体的核心节点。由于每个节点可能在网络中参与不止一个模体结构，遍历网络中存在的模体结构，节点每作为一个模体中的核心节点，其基于模体的重要性值就加 1。节点基于模体的影响力评估计算式为

$$\delta(i,M_k^j)=\begin{cases}1, & i=\arg\max_{\text{node}}\left\{\text{DC(node)}\middle|\text{node}\in M_k^j\right\}\\ 0, & i\neq\arg\max_{\text{node}}\left\{\text{DC(node)}\middle|\text{node}\in M_k^j\right\}\end{cases} \tag{2}$$

$$\text{MC}(i)=\sum_{j=1}^{s}\delta(i,M_k^j) \tag{3}$$

其中，如果节点 i 是模体 M_k^j 中的核心节点，则 $\delta(i,M_k^j)=1$，反之 $\delta(i,M_k^j)=0$；s 表示网络中模体 M_k 的个数；MC 表示节点基于模体结构的重要性值；DC 表示节点的度重要性。

尽管模体对网络信息的传播具有重要的作用，但网络中仍然存在一些不属于模体结构且对网络信息十分重要的节点，对于这类节点，度中心性是一个重要的信息传播指标。通过 VIKOR 算法[25]对 2 个重要性指标进行结合，得到距离理想值最近的可行解 MD，具体过程如下。

1) 根据式(3)提出的重要性值构造初始决策矩阵 $\boldsymbol{D}$

$$\boldsymbol{D}=d(i,j)_{n\times 2}=\begin{bmatrix}\text{MC}_1 & \text{DC}_1\\ \text{MC}_2 & \text{DC}_2\\ \vdots & \vdots\\ \text{MC}_n & \text{DC}_n\end{bmatrix} \tag{4}$$

其中，MC_i 表示节点 i 基于模体结构的重要性值，DC_i 表示节点 i 的度重要性，$i=1,2,\cdots,n$。

2) 对决策矩阵 $\boldsymbol{D}$ 进行标准化，构建标准化决策矩阵 $\boldsymbol{R}$

$$\boldsymbol{R}=\begin{bmatrix}r_{11} & r_{12}\\ r_{21} & r_{22}\\ \vdots & \vdots\\ r_{n1} & r_{n2}\end{bmatrix},\ r_{ij}=\frac{d_{ij}}{\sqrt{\sum_{j=1}^{2}d_{ij}}} \tag{5}$$

3) 确定 2 个指标的权重，首先计算第 j 个属性信息熵

$$E_j=-\frac{1}{\ln n}\sum_{i=1}^{n}r_{ij}\ln r_{ij} \tag{6}$$

根据信息熵计算 2 个属性的权重

$$W_j=\frac{1-E_j}{\sum_{k=1}^{2}(1-E_k)} \tag{7}$$

网络中一些节点不参与模体结构，导致其基于模体的重要性值可能为 0，因此无法适用熵权法来对 2 个指标权重进行计算，这里依据主观因素设置对 2 个指标进行赋权，经过多次实验选取权重矩阵为 $\boldsymbol{W}=[0.5,0.5]$。

4) 确定正理想解 R^+ 和负理想解 R^-

$$R^+=\left\{r_1^+,r_2^+\right\}=\left\{(\max_i r_{ij}\,|\,j\in J_+)\text{or}(\min_i r_{ij}\,|\,j\in J_-)\right\} \tag{8}$$

$$R^-=\left\{r_1^-,r_2^-\right\}=\left\{(\min_i r_{ij}\,|\,j\in J_+)\text{or}(\max_i r_{ij}\,|\,j\in J_-)\right\} \tag{9}$$

其中，J_+ 代表效益型属性，属性值越高，节点越重要；J_- 代表成本型属性，属性值越低，节点越重要。在这里，正理想解是指在所有指标上取最优值的方案，负理想解是指在所有指标上取最劣值的方案。

5) 计算各节点的群体效用值 S_i 和个体遗憾值 R_i

$$S_i = \sum_{j=1}^{2} w_j \frac{(r^+ - r_{ij})}{(r^+ - r^-)} \tag{10}$$

$$R_i = \max_j \left[w_j \frac{(r^+ - r_{ij})}{(r^+ - r^-)} \right] \tag{11}$$

6) 计算节点最终的重要性 MD_i

$$\mathrm{MD}_i = b\frac{(S_i - S^+)}{(S^- - S^+)} + (1-b)\frac{(R_i - R^+)}{(R^- - R^+)} \tag{12}$$

$$S^+ = \min_i(S_i),\ S^- = \max_i(S_i)$$
$$R^+ = \min_i(R_i),\ R^- = \max_i(R_i) \tag{13}$$

其中，v 为决策机制系数。如果 $v>0.5$，则根据最大化群体效应决策机制决策；如果 $v<0.5$，则根据最小化个体遗憾值的决策机制决策；如果 $v=0.5$，则根据协商达成最大群体效应和最小遗憾值同等重要的决策机制进行决策。为了不失一般性，本文取 $v=0.5$。

7) 根据计算得到的节点重要性进行降序排列，得到节点影响力排序。

2.2　种子节点选择

对于识别网络中的关键节点组来使网络传播最大化的研究，传统的基于节点中心性指标的方法根据网络中的节点中心性结果从大到小选取前 k 个节点作为最终的关键节点组。将这些方法产生的结果在传播模型中进行模拟，传播效果往往并不理想。在现实网络中，经常出现富人俱乐部现象，即网络中的高度节点往往处于网络的核心层，同时彼此之间存在连接。这些现象导致传统的中心性方法挖掘的关键节点组中的节点距离过近，同时种子节点之间的共同邻居过多，从而影响传播的最终范围。基于此，文献[26]提出“种子排除”理论，通过对种子节点的连接关系进行排除，从而使节点组中的节点保持一定的距离，减少种子节点之间的影响力重叠问题，提高传播效果。种子节点指初始时作为传播者或感染者的节点。

种子排除算法的核心思想如下：如果选择节点 v 作为新种子，那么节点 v 对后续种子节点选择的边际影响应该最小化。一方面，节点 v 的邻居可以自己激活；另一方面，选择节点 v 的邻居作为种子将导致其影响范围重叠。因此，当节点 v 被选为种子节点时，其邻居的影响应尽可能地最小化。根据三度影响规则，节点对其不超过三阶邻居的节点产生影响，因此种子节点的排除范围不超过其三阶邻居。本文选择种子节点的排除范围为节点的一阶邻居。

根据 2.1 节利用 VIKOR 算法对网络中的节点影响力进行评估，最后利用种子节点排除算法挖掘符合条件的节点加入种子节点集合。

2.3　算法描述

基于模体结构的关键节点组挖掘算法具体流程主要包括以下 4 个阶段。

第一阶段：根据实际网络构建复杂网络模型，然后使用模体检测算法对网络中的模体进行检测。

第二阶段：遍历网络中的每个模体，根据式(12)对节点影响力进行计算，同时计算节点的度重要性。

第三阶段：基于第二阶段计算得到的 2 个节点重要性属性构建初始决策矩阵，再利用 VIKOR 方法计算每个节点的 MD 重要性值，并对得到的节点重要性进行降序排列。

第四阶段：根据节点的排序选取 MD 重要性值最高的节点加入种子节点集合 S，根据种子节点排除理论，将该节点及其邻居节点移出候选节点集合，再对候选节点集合中的节点进行重新排序，继续选取重要性值最高的节点加入种子节点集合 S，重复上述步骤，直到种子节点集合的节点数目满足 $|S|=k$。其中，种子节点集合指种子节点构成的集合，候选节点集合指除种子节点外的节点所组成的集合。随着种子节点集合中加入新的节点 i，候选节点集合就将节点 i 的邻居节点从候选节点集合中剔除。

3　数据集及评价标准

3.1　实验数据集

为了验证所提算法的有效性，选取了使用 6 个不同规模的无权无向网络，包括 USAir[27]、Netscience 网络[28]、C.Elegans 网络[29]、Email 网络[30]、GrQc 网络[31]和 PGP 网络[32]。经检测，6 个网络中的三阶

模体均存在闭合三角形，由于三角形模体在网络信息传播中的重要作用[33]以及在网络中存在的广泛性，这里选取三角形模体作为研究基础。6 个网络的拓扑特征如表 1 所示，其中，N 表示节点总数，M 表示连边总数，$<k>$ 表示网络的平均度，$<d>$ 表示平均最短路径长度，β_{th} 表示 SIR 模型的传播阈值[34]，其定义为

$$\beta_{\text{th}}=\frac{\langle k\rangle}{\langle k^2\rangle} \tag{14}$$

其中，$<k^2>$ 表示网络的二阶平均度。

表 1　6 个网络的拓扑特征

网络	N	M	$<k>$	$<d>$	β_{th}
USAir	332	2 126	12.807	2.74	0.022 5
Netscience	379	914	4.82	6.04	0.124 7
C.Elegans	453	2 025	8.94	2.46	0.024 9
Email	1 133	5 451	9.62	3.61	0.053 5
GrQc	4 158	13 425	4.46	6.05	0.059
PGP	10 680	24 316	4.54	7.46	0.053

3.2　传播模型

本文使用 SIR 传播模型[32]来验证 MDS 算法对识别重要节点的效果。在 SIR 传播模型中，节点一共有 3 种状态，分别是易感态（S）、感染态（I）以及恢复态（R）。在每个时间步 t，感染态节点会以传播概率 β 对其处于易感状态的邻居节点进行感染，同时感染态节点也以恢复概率 γ 转变为恢复态，转变为恢复态的节点将不再具有感染其他节点的能力，同时也不再具有被感染的风险。图 2 为 SIR 传播模型示意。将需要判断的节点 v_i 作为初始感染节点对传播进行迭代，直到网络中不存在感染态节点，此时将网络中处于恢复态的节点数目作为节点 v_i 的传播能力。为了不失一般性，本文设恢复概率 $\gamma=1$，传播概率 β 选取阈值附近的值[35]。

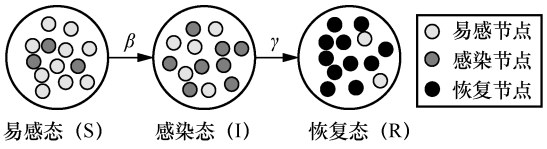


图 2　SIR 传播模型示意

3.3　评价标准

由于多个重要节点识别问题的研究目标与单个节点重要性并不一致，因此需要基于多个重要节点识别的角度来对实验方法进行评价。

1) 影响范围随种子节点数目的变化情况。基于 SIR 模型，选取间隔为 5 的 5~50 个种子节点进行传播，每种方法的影响范围为 1 000 次传播后最终为恢复态的节点数的平均值，观察其随种子节点数目的变化情况，以此判断各方法的准确性。

2) 影响范围随传播率的变化情况。对每个网络选取固定的种子节点数，然后改变 SIR 模型的传播概率，进行 1 000 次传播，将其平均值作为最终的传播范围，观察模型传播概率的变化后各个方法的稳定性。

3) 种子节点之间的平均最短路径长度。选取间隔为 5 的 5~50 个种子节点，计算任意 2 个节点之间的最短路径之和后取平均作为种子节点的平均最短路径长度，节点之间的最短路径长度越长，说明选取的种子节点越分散，对于削弱富人俱乐部现象导致的影响力重叠问题有积极作用。

3.4　基准算法

本节将本文所提算法与 4 种基准算法进行对比，首先对 4 种基准算法进行介绍。

3.4.1　VoteRank 算法

VoteRank 算法[13]引入投票的思想，通过邻居节点对网络中有影响力的节点进行选举，对于一个具有 N 个节点、M 条边的网络 $G(N,M)$，给定需要的关键节点个数 k。VoteRank 算法主要分为以下 5 个步骤。

Step 1　初始化。对网络中的每个节点 v 都赋予一组值 (S_v,V_v)，分别代表节点 v 的得分以及投票能力，初始阶段，每个节点的得分均为 0，同时投票能力均为 1，即 $(S_v,V_v)=(0,1)$。

Step 2　投票阶段。每个节点向邻居节点进行投票后，更新每个节点的投票得分。对于每个节点 v，其得分 S_v 的计算式为

$$S_v=\sum_{u\in\Gamma_v}V_u \tag{15}$$

其中，Γ_v 表示节点 v 的邻居节点集合，V_u 表示节点 u 的投票能力。

Step 3　选择阶段。对网络中所有节点的投票得分更新后，选择投票得分最高的节点加入关键节点集合中，同时将该节点的投票能力设为 0，并且不再参与后续节点选择的投票和得分计算。

Step 4　更新阶段。对于加入了关键节点集合的节点，对其邻居节点的投票能力进行一定的削弱，从而避免出现关键节点过于密集的问题。削弱后的节点 i 的投票能力为 $V_i - \frac{1}{\langle d \rangle}$。

Step 5　重复迭代。重复 Step 2 和 Step 3，直到选出 k 个关键节点。

3.4.2　DegreeDiscount 算法

DegreeDiscount 算法[10]有效解决了影响力重叠的问题，其基本思想是利用度中心性来对网络中的关键节点进行挖掘，当一个节点被选择加入关键节点组时，其邻居节点在下一轮的节点选择中需要对其重要性（基于度中心性）进行削弱。具体而言，给定一个网络 G，对于网络中的任意一个未加入关键节点组的节点 i，其度值为 $d(i)$，其邻居节点中已经加入关键节点组的节点个数为 n_i，设网络中的传播概率为 β，其被邻居节点激活的概率为 $1-(1-\beta)^{n_i}$，此时由于节点 i 被邻居节点激活，把节点 i 加入关键节点组并没有增加关键节点组的总体影响力。当节点未被其邻居节点激活时，将其加入节点组才能够增加传播范围，故节点的期望影响力计算式为

$$\begin{aligned}&(1+(d(i)-n_i)\beta)(1-\beta)^{n_i} \approx \\ &1+\left(d(i)-2n_i-(d(i)-n_i)n_i\beta\right)\beta\end{aligned} \tag{16}$$

其中，$(1-\beta)^{n_i}$ 表示节点未被邻居节点激活的概率，1 表示节点自身，$(d(i)-n_i)\beta$ 表示邻居节点中被节点激活的数量。

利用节点 i 没有邻居节点在关键节点组中的期望影响力与式(16)进行计算，最终得到节点 i 的 DegreeDiscount 值为

$$\mathrm{dd}(i) = d(i) - 2n_i - (d(i)-n_i)n_i\beta \tag{17}$$

DegreeDiscount 算法的计算流程如下。

Step 1　初始化阶段。给定一个网络 $G(N,M)$，取关键节点组 $S=\phi$，关键节点组中节点数目为 k，计算网络中各节点的度值，并进行排序，选择度值最大的节点加入关键节点组 S。

Step 2　更新阶段。根据式(17)，对网络中不在关键节点组 S 的节点进行度折扣值更新，选择度折扣值最大的节点加入关键节点组。

Step 3　迭代阶段。重复 Step 2，迭代更新网络中剩余节点的度折扣值，直到选出 k 个节点加入关键节点组。

3.4.3　HC 算法

HC 算法[18]利用聚类算法将网络中的节点分成多个节点簇，从而解决网络中社区数量较少导致的种子节点数目不足的问题。利用相似性指数对节点进行聚类，保证了节点之间的距离相对分散，同时也考虑了一定的节点自身重要性。HC 算法的具体实现如下。

Step 1　建立相似度矩阵。使用局部路径（LP, local path）相似性指数来构建相似度矩阵，对网络中节点之间的相似性进行度量

$$\mathbf{Sim} = \boldsymbol{A}^2 + \lambda \boldsymbol{A}^3 \tag{18}$$

其中，**Sim** 表示节点的相似度矩阵；$\boldsymbol{A}$ 表示网络邻接矩阵，$(A^2)_{ij}$ 表示节点 i 与节点 j 之间长度为 2 的路径数目，也可以理解为节点 i 与节点 j 之间的公共邻居数目，同理，$(A^3)_{ij}$ 表示节点 i 与节点 j 之间长度为 3 的路径数目；λ 用来权衡不同路径长度对相似度的影响，λ 值越小，节点之间长度为 3 的路径数目对相似度影响越小。

Step 2　构建多个聚类。随机从网络中选取 k 个节点作为 k 个节点簇的中心，节点集合表示为 $S=\{v_1,v_2,\cdots,v_k\}$，对每个不在集合的节点 $v_i \notin S$，计算其与任意一个在集合中的节点之间的相似度 $\mathrm{Sim}_{v_iv_j}$，寻找相似度值 $\mathrm{Sim}_{v_iv_j}$ 最大的节点 v_j，将节点 v_i 分配给以节点 v_j 为中心的节点簇。遍历所有不在集合 S 的节点，将所有的节点分在不同的节点簇中，节点簇为 $C_1,C_2,\cdots,C_k$。

Step 3　节点簇中心更新。对于网络中的节点簇 C_t，定义节点 v_i 在节点簇 C_t 中的重要性值

$$B(v_i) = \sum_{v_j \in C_t} \mathrm{Sim}_{v_iv_j} \tag{19}$$

其中，$v_i,v_j \in C_t$，$\mathrm{Sim}_{v_iv_j}$ 表示两节点之间的相似度值。

在节点簇 C_t 中，选取在节点簇中具有最高重要性值的节点作为中心节点，即对集合 S 中的中心节点 v_t 进行更新；以此类推，遍历网络中的所有节点簇，完成对集合 S 中所有中心节点的更新。

Step 4　选取多个种子节点。重复 Step 2 和 Step 3，直到集合 S 中的中心节点不再发生变化，然后将 S 中的节点作为初始的 k 个传播节点。

3.4.4 NCA

NCA[14]利用节点的覆盖范围来解决网络的 Rich-club 问题，该算法将节点可能影响到的一阶邻居节点数目作为节点的覆盖范围，已经加入关键节点组的节点不受其他节点的影响，因此寻找种子节点时不会将已在关键节点组的节点加入计算，在每次挖掘种子节点时，通过考虑已加入节点组的节点和其他节点的共同覆盖范围，选择能覆盖到更多邻居节点的节点加入关键节点组中，最终使节点组能影响最多的节点。设 N_i 为节点 i 的邻居集合，节点 i 与节点 j 的共同覆盖范围计算式为

$$N_{ij} = N_i \cup N_j \tag{20}$$

以此类推，种子节点集合 S 的覆盖范围计算式为

$$N_S = N_{i_1} \cup N_{i_2} \cdots \cup N_{i_j} (j = |S|, i_1, i_2, \cdots, i_j \in S) \tag{21}$$

位于备选节点集合 R 中的节点 v 加入关键节点组 S 后的覆盖范围增益为

$$\text{gain}_v = |N_S \cup N_v \cap R| - |N_S| \tag{22}$$

其中，N_v 表示节点 v 的邻居集合，N_S 表示关键节点集合 S 的节点覆盖范围增益。

4 实验结果与分析

为了验证所提算法的有效性，本节选取上述介绍的 4 种算法（VoteRank 算法、DegreeDiscount 算法、HC 算法和 NCA）进行对比。

1) 不同规模初始节点组实验

基于 SIR 模型对得到的结果进行分析，随着种子节点的传播，将网络中最终的感染节点规模 $F(c)$ 作为判断算法有效性的重要评价指标。

以选取的节点作为初始感染节点，恢复概率为 γ，并以概率 β 对其邻居节点进行感染，等到网络中所有节点的状态不再变化时传播结束，此时将网络中被感染过的节点总数作为最终感染规模 $F(c)$。

如果网络中节点的感染率过小，则会导致最终感染规模得不到区分；如果网络中节点感染率过大，则会导致网络中几乎所有节点都被感染，因此需要选择合适的感染率来对算法结果进行有效的区分。其中，USAir、Netscience、C.Elegans、Email、GrQc、PGP 这 6 个数据集的传播概率 β 分别取值为 0.1、0.2、0.1、0.1、0.13、0.2。图 3 展示了在 6 个网络中种子节点传播范围随节点数目变化情况。整体来看，随着种子节点数目的增加，网络最终传播范围均得到扩大。

从图 3 中可以观察到，除了 Netscience 网络之外，VoteRank 算法在大部分网络中的影响范围总体均较低。HC 算法尽管在部分网络（如 C.Elegans 网络、PGP 网络）中具有较好的结果，但该算法的识别机制导致其在网络中不够稳定。DegreeDiscount 算法表现出良好的算法稳定性，但算法效果相较于 NCA 和 MDS 算法具有较大的差距。在图 3(d)和图 3(f)中，MDS 算法在种子节点较少时并未表现出最好的传播性能，而随着种子节点规模的增加，MDS 算法相较于其他算法表现出更好的性能。这可能是因为这 2 个网络中的模体结构密集存在于网络的核心结构之中，当种子节点较少时，本文算法所选的节点组即使由于覆盖范围的原因不是邻居节点，但也同处于网络的核心层次，从而导致影响力重叠问题。但总体来看，所提出的 MDS 算法相较于其他 4 种算法在 6 个网络中整体表现出更好的性能。

2) 不同传播概率实验

在现实复杂系统中，传播概率不是一成不变的，而是随着传播概率的变化，最终的传播结果也会发生很大的变化。这里通过固定关键节点组中的节点数目，改变传播概率来观察各种算法的泛化性能。图 4 为 6 个网络在不同传播概率下种子节点传播范围变化情况。当传播概率较小时，传播过程更多地依赖于节点邻居的直接传播，基于度指标的启发式算法具有更好的效果，如 DegreeDiscount 算法和 NCA。从图 4 可以观察到，在 C.Elegans 网络、Email 网络和 PGP 网络中，各种算法的差距并不明显，但依然可以观察到，MDS 算法随着感染率的增大均表现出最大的传播范围，而在 USAir 网络、Netscience 网络和 GrQc 网络中，MDS 算法相较于 4 种基准算法具有更明显的优势。特别是在传播概率较大时，MDS 算法表现出更大的传播优势。

3) 位置属性实验

节点组的平均距离反映了节点的分布情况。一般而言，所选节点之间的平均路径长度 L_s 越大，则意味着所选节点在网络中的分布越广泛，从而对富人俱乐部现象导致的影响力重叠有着越好的抑制作用。

图 5 展示了所提算法与 4 种基准算法在不同种子节点数目下节点平均最短路径长度变化情

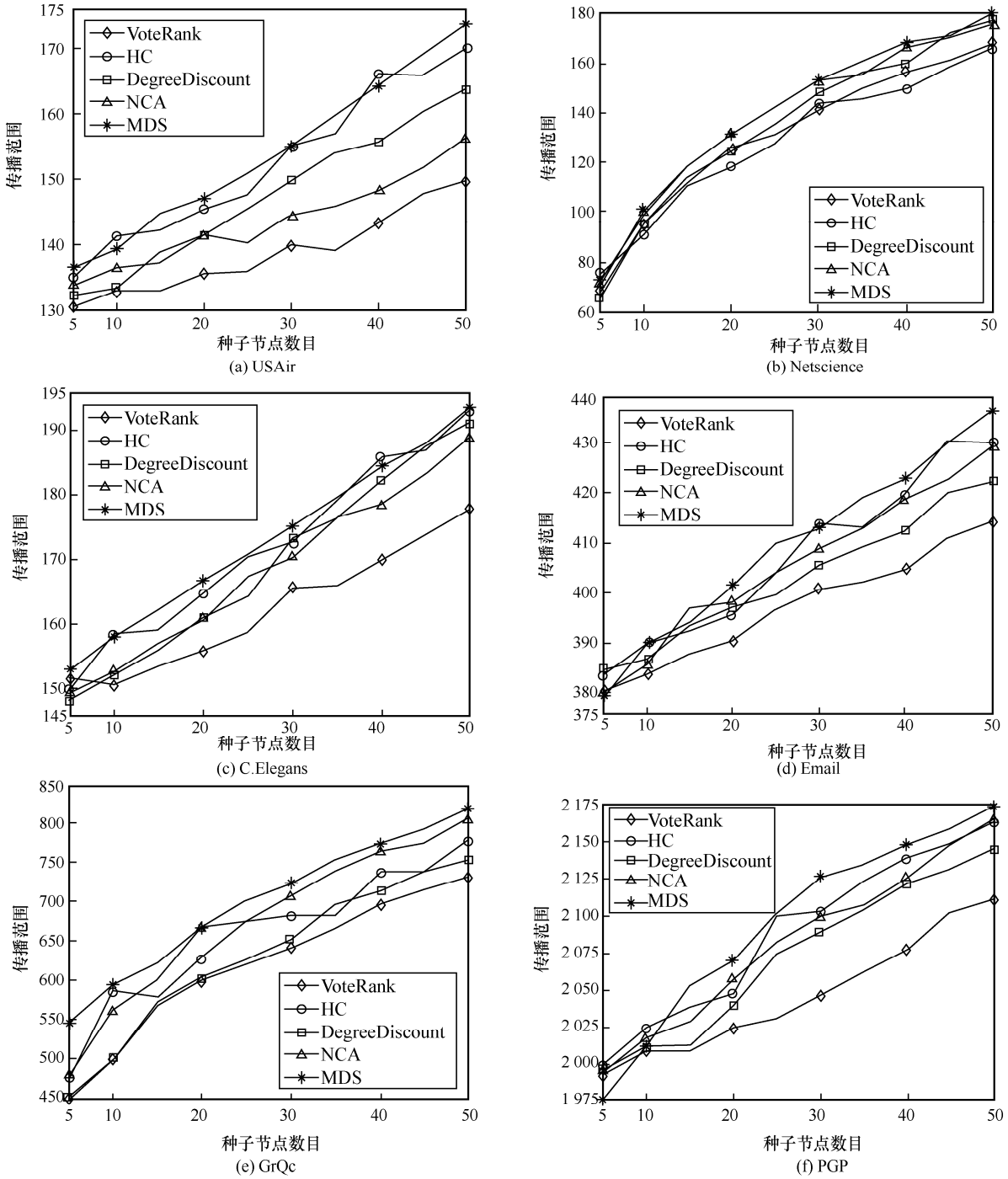


图 3　不同网络下种子节点传播范围随节点数目变化情况

况。其中，横坐标表示关键节点组中节点的数目，即初始感染节点的数目，纵坐标表示关键节点组中节点之间的平均最短路径长度 L_s。从图 5 可以看出，HC 算法在 6 个网络中均表现出较长的平均最短路径长度，特别是在图 5(d)和图 5(f)中表现出最高的平均路径长度，主要有以下两方面的原因。一方面，HC 算法是基于聚类的影响力节点挖掘算法，可根据聚类结果选择每个聚类中心加入关键节点组，因此不存在 2 个关键节点同属于一个聚类集合的问题，故关键节点组的最短路径长度性能较其他算法得到了提高；另一方面，HC 算法根据关键节点组中节点数目的不同，产生的种子节点结果也不同，即节点组数目 $k=5$ 时产生的结果与节点组数目 $k=10$ 时产生的结果不一定有重合，

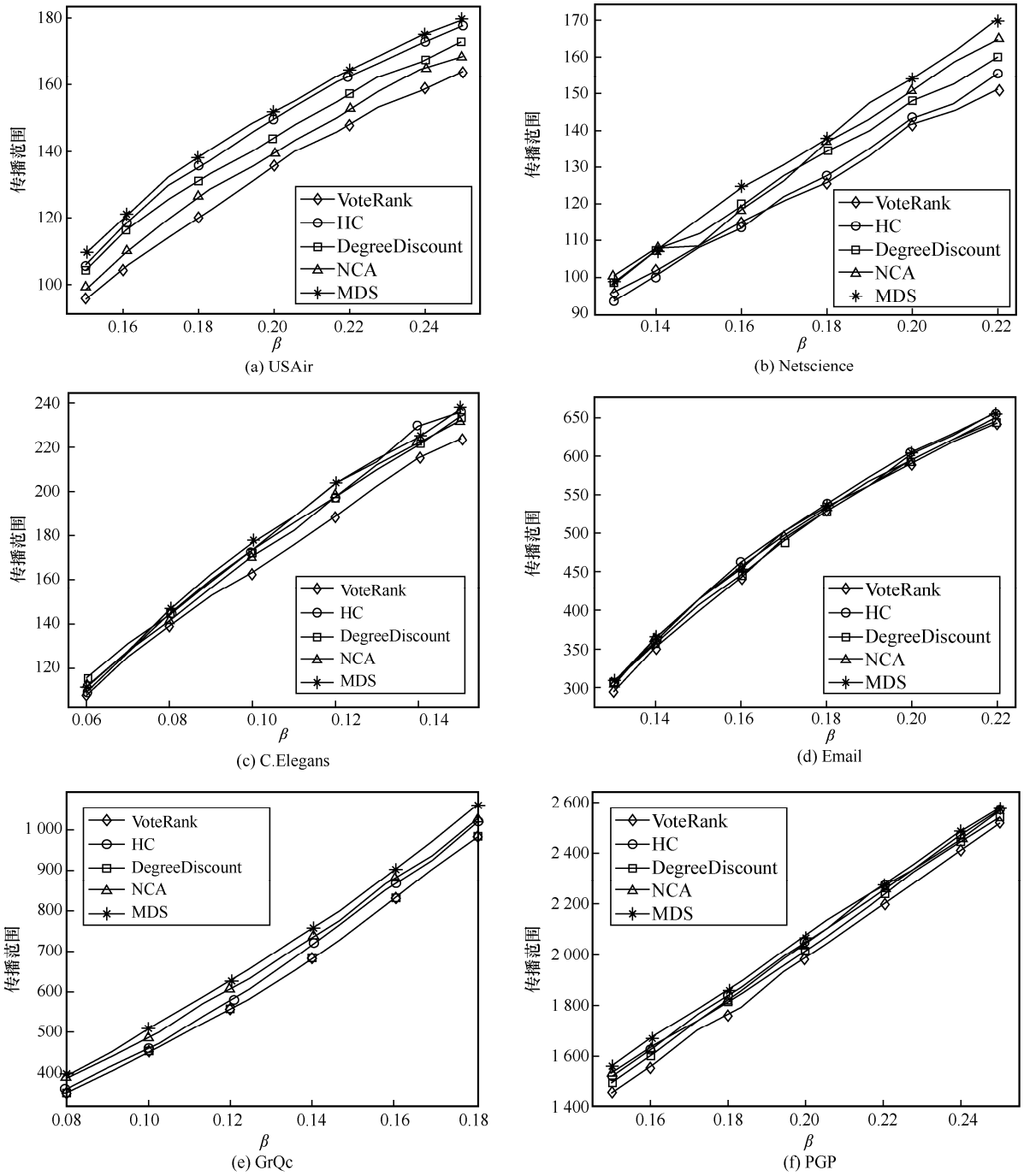


图 4　不同传播概率下种子节点传播范围变化情况

因此很大程度上保证了关键节点组中节点之间的最短路径长度。MDS 算法虽然在 Email 网络和 PGP 网络中没能表现出最优的结果，但相较于其他启发式算法，最短路径长度性能有了很大提高。另外，在 USAir 网络、Netscience 网络和 C.Elegans 网络中，MDS 算法表现出与 HC 算法结果相近的平均最短路径长度。这是因为，除 HC 算法以外，其他基准算法考虑的是关注节点的局部算法，而模体是网络中的中观结构，因此 MDS 算法也具有较高的平均最短路径长度。因此，本文所提算法吸收了聚类算法挖掘影响力节点的优点，相较于其他非聚类的启发式算法，识别的关键节点组中的节点更加分散，但并非单纯追求节点部分的广泛性，同时考虑了节点的综合传播影响力。

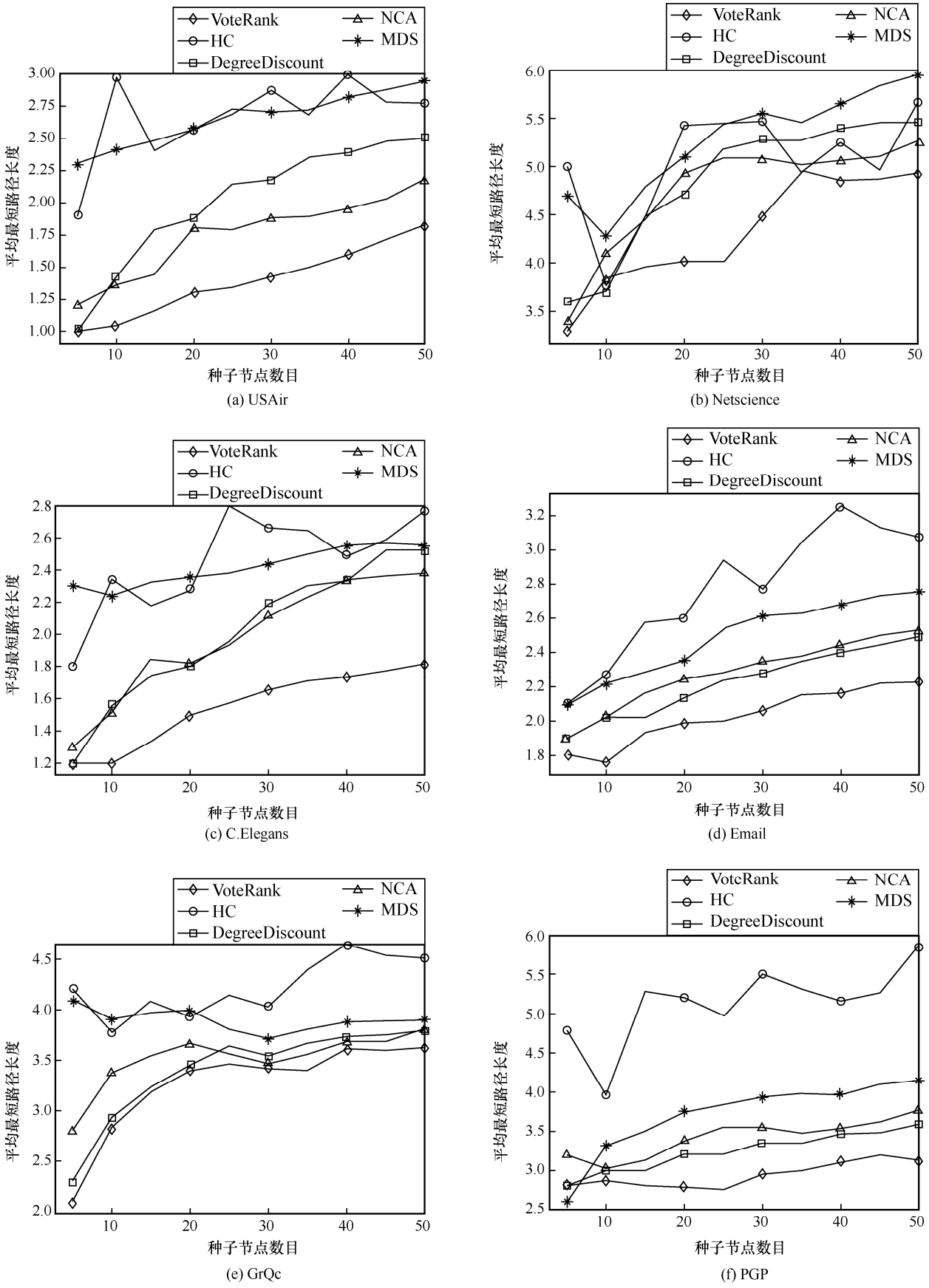


图 5　不同种子节点数目下节点平均最短路径长度变化情况

5 算法复杂度分析

本节对算法的复杂度进行分析，用 N 和 E 表示网络中节点和连边的数量，用 N_{sub} 和 N_M 表示网络中三阶子图和三角形模体的数量。

首先，需要对网络中的三角形模体进行识别，这一过程所需的时间复杂度均为 $O(N_{sub})$。然后，本文主要考虑在成功检测到模体之后挖掘关键节点组的过程所需的计算复杂度。在算法中，首先得到每个节点分别基于三角形模体和度信息的重要性评估，这一步骤需要的时间复杂度为 $O(N_M)+O(N)$，而后续使用 VIKOR 算法的计算复杂度为 $O(N^2)$，所以 MDS 算法的时间复杂度为 $O(N^3+N^2N_M)$。

6 结束语

本文以复杂网络理论为基础，探究网络模体结构对关键节点组识别的影响。利用 VIKOR 多属性决策方法，将其与对传播具有重要影响的度信息进行融合，最后基于种子排除法避免种子节点密集造成的影响力重叠问题，提出一种基于模体结构和度信息的关键节点组挖掘算法。基于 SIR 传播模型在 6 个不同的无向网络中将本文算法与 4 种基准算法进行比较，其中包括同样具备节点排除理论的 NCA 以及基于社区聚类的 HC 算法。实验结果表明，本文算法具有良好的性能，在提高准确性和稳定性的同时，对种子节点的选取也更加分散。本文仅使用在信息传播中具有重要作用的三角形模体进行分析，对于网络中的模体结构，同时利用多个模体结构进行信息挖掘可能会具有更好的效果，如何将多种模体以及更高阶模体进行融合，再对网络信息进行挖掘是一个需要考虑的问题，因此基于多个模体的关键节点组识别是下一步研究分析的重点。

参考文献：

[1] BADASHIAN A S, STROULIA E. Measuring user influence in GitHub: the million follower fallacy[C]//Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering. Piscataway: IEEE Press, 2016: 15-21.

[2] HOSNI A I E, LI K, AHMAD S. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors[J]. Information Sciences, 2020, 512: 1458-1480.

[3] REN J F, LIU M T, LIU Y, et al. Optimal resource allocation with spatiotemporal transmission discovery for effective disease control[J]. Infectious Diseases of Poverty, 2022, 11(1): 34.

[4] WANG J, LI C, XIA C Y. Improved centrality indicators to characterize the nodal spreading capability in complex networks[J]. Applied Mathematics and Computation, 2018, 334: 388-400.

[5] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 137-146.

[6] LESKOVEC J, KRAUSE A, GUESTRIN C, et al. Cost-effective outbreak detection in networks[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 420-429.

[7] GOYAL A, LU W, LAKSHMANAN L V S. CELF++: optimizing the greedy algorithm for influence maximization in social networks[C]//Proceedings of the 20th International Conference Companion on World Wide Web. New York: ACM Press, 2011: 47-48.

[8] BORGS C, BRAUTBAR M, CHAYES J, et al. Maximizing social influence in nearly optimal time[C]//Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. New York: ACM Press, 2014: 946-957.

[9] CHEN W, WANG Y J, YANG S Y. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 199-208.

[10] WANG J, MA X J, XIANG B B, et al. Maximizing influence in social networks by distinguishing the roles of seeds[J]. Physica A: Statistical Mechanics and Its Applications, 2022, 604: 127881.

[11] ZHAO X Y, HUANG B, TANG M, et al. Identifying effective multiple spreaders by coloring complex networks[J]. Europhysics Letters, 2014, 108(6): 68005.

[12] LIU D, JING Y, ZHAO J, et al. A fast and efficient algorithm for mining top-k nodes in complex networks[J]. Scientific Reports, 2017, 7: 43330.

[13] ZHANG J X, CHEN D B, DONG Q, et al. Identifying a set of influential spreaders in complex networks[J]. Scientific Reports, 2016, 6: 27823.

[14] 高菊远, 王志晓, 芮晓彬, 等. 基于节点覆盖范围的影响力最大化算法[J]. 计算机工程与设计, 2019, 40(8): 2211-2215, 2246.
GAO J Y, WANG Z X, RUI X B, et al. Node coverage based algorithm for influence maximization[J]. Computer Engineering and Design, 2019, 40(8): 2211-2215, 2246.

[15] LIU P F, LI L J, FANG S Y, et al. Identifying influential nodes in social networks: a voting approach[J]. Chaos Solitons and Fractals, 2021, 152: 111309.

[16] WANG B, ZHANG J K, DAI J Y, et al. Influential nodes identification using network local structural properties[J]. Scientific Reports, 2022, 12: 1833.

[17] SHANG J X, ZHOU S B, LI X, et al. CoFIM: a community-based framework for influence maximization on large-scale networks[J]. Knowledge-Based Systems, 2017, 117: 88-100.

[18] BAO Z K, LIU J G, ZHANG H F. Identifying multiple influential spreaders by a heuristic clustering algorithm[J]. Physics Letters A, 2017, 381(11): 976-983.

[19] BOZORGI A, SAMET S, KWISTHOUT J, et al. Community-based influence maximization in social networks under a competitive linear threshold model[J]. Knowledge-Based Systems, 2017, 134: 149-158.

[20] BENI H A, BOUYER A. TI-SC: top-k influential nodes selection

based on community detection and scoring criteria in social networks[J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(11): 4889-4908.

[21] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.

[22] 王兴隆, 石宗北, 陈仔燕. 空中交通网络模体识别及子图结构韧性评估[J]. 航空学报, 2021, 42(7): 551-561.
WANG X L, SHI Z B, CHEN Z Y. Air traffic network motif recognition and sub-graph structure resilience evaluation[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(7): 551-561.

[23] ZHAO H, XU X G, SONG Y Q, et al. Ranking users in social networks with motif-based PageRank[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(5): 2179-2192.

[24] 韩华, 刘婉璐, 吴翎燕. 基于模体的复杂网络测度量研究[J]. 物理学报, 2013, 62(16): 519-527.
HAN H, LIU W L, WU L Y. The measurement of complex network based on motif[J]. Acta Physica Sinica, 2013, 62(16): 519-527.

[25] OPRICOVIC S, TZENG G H. Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS[J]. European Journal of Operational Research, 2004, 156(2): 445-455.

[26] WANG Y, LI H Z, ZHANG L, et al. Identifying influential nodes in social networks: centripetal centrality and seed exclusion approach[J]. Chaos Solitons and Fractals, 2022, 162: 112513.

[27] ZENG A, LIU W. Enhancing network robustness against malicious attacks[J]. Physical Review E, 2012, 85(6): 066130.

[28] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74(3): 036104.

[29] BASSETT D S, PORTER M A, WYMBS N F, et al. Robust detection of dynamic community structure in networks[J]. Chaos, 2013, 23(1): 013142.

[30] DUCH J, ARENAS A. Community detection in complex networks using extremal optimization[J]. Physical Review E, 2005, 72(2): 027104.

[31] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graph evolution: densification and shrinking diameters[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 2.

[32] BOGUÑÁ M, PASTOR-SATORRAS R, DÍAZ-GUILERA A, et al. Models of social networks based on social distance attachment[J]. Physical Review E, 2004, 70(5): 056122.

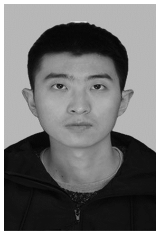
[33] NAMTIRTHA A, DUTTA B, DUTTA A. Semi-global triangular centrality measure for identifying the influential spreaders from undirected complex networks[J]. Expert Systems with Applications, 2022, 206: 117791.

[34] PASTOR-SATORRAS R, VESPIGNANI A. Epidemic spreading in scale-free networks[J]. Physical Review Letters, 2001, 86(14): 3200-3203.

[35] ZHOU F, SU C, XU S Q, et al. Influence fast or later: two types of influencers in social networks[J]. Chinese Physics B, 2022, 31(6): 068901.

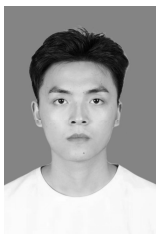
[作者简介]

杨云云（1987− ），女，山西吕梁人，博士，太原理工大学副教授，主要研究方向为复杂系统与复杂网络。



张辽（1997− ），男，山西河津人，太原理工大学硕士生，主要研究方向为节点重要性挖掘。

于海龙（2000− ），男，山西吕梁人，太原理工大学硕士生，主要研究方向为复杂网络。



王力（2000− ），男，四川达州人，太原理工大学硕士生，主要研究方向为网络动力学、图神经网络。