计算机科学
COMPUTER SCIENCE

基于马尔可夫相似性增强和网络嵌入的社区发现

曾祥宇, 龙海霞, 杨旭华

引用本文

曾祥宇, 龙海霞, 杨旭华.基于马尔可夫相似性增强和网络嵌入的社区发现[J]. 计算机科学, 2023, 50(4): 56-62.

ZENG Xiangyu, LONG Haixia, YANG Xuhua. Community Detection Based on Markov Similarity Enhancement and Network Embedding [J]. Computer Science, 2023, 50(4): 56-62.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100155

基于马尔可夫相似性增强和网络嵌入的社区发现

曾祥宇　龙海霞　杨旭华
浙江工业大学计算机科学与技术学院　杭州 310023
(zxyu15947@163.com)

摘　要　社区结构普遍存在于自然界的各种复杂网络中,是网络结构的重要特征之一。社区发现算法可以识别网络中的有用信息,有助于分析网络的结构和功能,被广泛应用于社交网络、生物和医学等领域。文中针对目前基于局部相似性的复杂网络社区发现算法精确度不高的问题,提出了一种基于马尔可夫相似性增强和网络嵌入的社区发现算法。首先,受马尔可夫链思想启发,提出了一种马尔可夫相似性增强方法,通过对初始网络的马尔可夫迭代状态进行转移,来获取稳态的马尔可夫相似性增强矩阵,根据马尔可夫相似性指标对网络进行初始的社区划分。然后结合网络的拓扑结构和网络嵌入,提出了一种新的社区相似性指标,将初始社区结构中的小社区与其连接紧密的社区合并,得到网络社区结构。在7个真实网络和可变参数的人工网络上,通过与其他5个知名社区发现算法的比较,证明了所提算法具有良好的社区发现效果。

关键词:社区发现;复杂网络;马尔可夫相似性;网络嵌入;社区相似性
中图法分类号　TP391

Community Detection Based on Markov Similarity Enhancement and Network Embedding

ZENG Xiangyu, LONG Haixia and YANG Xuhua
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

Abstract　Community structure is ubiquitous in various complex networks in nature and is one of the important characteristics of network structure. Community detection can identify useful information in the network, and help to analyze the structure and function of the network. It is widely used in social networks, biology, medicine and other fields. Aiming at the low accuracy of the current community detection algorithm based on local similarity in complex networks, a community detection algorithm based on Markov similarity enhancement and network embedding is proposed. Firstly, inspired by the idea of Markov chain, a Markov similarity enhancement method is proposed, which obtains the steady-state Markov similarity enhancement matrix through the Markov iterative state transition of the initial network. According to the Markov similarity index, the network is divided into initial community structure. Then, a new community similarity index is proposed by combining the network topology and network embedding. The small community in the initial community structure is merged with its closely connected community to obtain the network community structure. On 7 real networks and artificial networks with variable parameters, compared with other 5 well-known community detection algorithms, it is proved that the proposed algorithm has a good community detection effect.

Keywords　Community detection, Complex network, Markov similarity, Network embedding, Community similarity

1　引言

自然界中许多不同领域的复杂系统都可以建模为网络或图,将其中的实体称为节点,实体之间的连接称为边。这些复杂的网络产生了大量的数据,如何分析这些数据,获取隐藏在其中的有用信息,是当前的研究热点。社区结构是复杂网络的一个重要特征,它与特定网络的组织结构和功能特征密切相关,有助于人们深入了解网络的功能和组织形式。一般而言,网络中社区内部节点之间的联系相对紧密,社区之间的节点的联系相对稀疏。例如,在社交网络[1]中,具有相似爱好的用户往往会聚集形成群体;在科学家网络[2]中,按照研究方向的归属,科学家们往往会形成不同的研究团队等。

社区发现算法旨在探测网络中潜在的社区结构,挖掘网络的隐藏信息,有助于获取网络的组织结构和预测网络的功能,是网络分析中重要的方法之一。例如,在社交网络中,通过社区发现能够找到拥有共同兴趣的用户,从而可以进行用户推荐;在蛋白质网络[3]中,通过社区发现能够找到拥有相似功能的蛋白质群,有助于发现蛋白质物质能量代谢的反应机制。因此,研究高效、准确的社区发现算法具有重要意义。目前学者们已经提出了许多社区发现算法[4],如模块度优化

到稿日期:2022-01-16　返修日期:2022-06-08
基金项目:国家自然科学基金(62176236,62106225)
This work was supported by the National Natural Science Foundation of China(62176236,62106225).
通信作者:杨旭华(xhyang@zjut.edu.cn)

算法[5]、标签传播算法[6]、层次聚类算法[7]、随机游走算法[8]等。Louvain 算法[9]是一种知名的模块度优化算法，它通过最大化社区合并之后的模块度增益大小来合并社区进行社区发现。标签传播算法（Label Propagation Algorithm）通过为每个节点分配一个唯一标签，然后根据邻居中数目最多的标签更新自身的标签，直到所有节点的标签与其大多数邻居的标签一致，该类算法简单快速，但结果不是很稳定，精度得不到保证。GN 算法[10]是一种知名的层次聚类算法，该算法通过删除网络中具有最大边介数值的连边，将网络划分为多个社区，然后通过构造网络的层次树来划分社区。Infomap 算法[11]是一种随机游走算法，引入信息熵来优化随机游走的过程，再通过最小化平均编码完成社区发现。

局部相似性能够很好地反映节点之间的相似程度，节点之间的相似性越高，属于同一个社区的概率就越大，反之越低。Eustace 等[12]发明了一种结合局部相似性和社区相似性的社区发现算法，提出了一种社区邻域比函数来衡量社区之间的联系，通过合并紧密联系的社区来形成社区结构。Liu 等[13]提出了一种基于模糊相似关系的局部社区发现算法，通过模糊关系来衡量两个节点之间的相似关系，采用最大连通子图算法获取节点所在社区。但这类算法的效果依赖于节点的相似性指标，不同的节点相似性指标划分的社区结果可能会不同。

网络嵌入算法能够将高维和稀疏的网络数据映射至低维稠密的向量空间，其学习到的特征向量可以用于分类、回归、聚类和社区发现等机器学习任务。目前已经提出许多结合网络嵌入的社区发现算法。Kumar 等[14]提出了一种基于网络嵌入和引力搜索的社区检测算法，利用网络嵌入算法和 k-means 聚类来寻找网络中社区的质心，然后采用引力搜索算法来改进簇质心的结果。Zhao 等[15]提出了一种结合社区嵌入和节点嵌入的社区发现算法，通过节点嵌入和社区嵌入将节点和社区映射为低维的向量表示，再结合节点相似性进行社区发现。He 等[16]提出了一种网络嵌入增强贝叶斯模型来识别社区，该模型将原始邻接矩阵表示与网络嵌入相结合，开发了一种新的贝叶斯概率模型来识别广义社区。

本文提出了一种基于马尔可夫相似性增强和网络嵌入的社区发现算法，提升了社区发现效果。本文的主要贡献如下：

（1）提出了一种马尔可夫相似性增强方法，对初始网络进行马尔可夫相似性迭代转移，获取稳态的马尔可夫相似性增强矩阵，在此基础上进行社区识别，相比传统的局部相似性算法，该方法识别社区会更加精确。

（2）结合网络拓扑结构和网络嵌入算法 SDNE[17]，提出了一种社区相似性指标进行社区合并，以完成社区划分。

2 相关工作

2.1 基于节点相似性的社区发现算法

基于节点相似性的社区发现算法大致可以分为两类：基于局部相似性指标的方法和基于中心性的方法。局部相似性指标通常用来衡量节点与其邻居之间的接近程度，目前常见的相似性指标有 Salton 相似性指标、Jaccard 相似性指标、Sørensen 相似性指标和 RA 相似性指标；中心性指标可以计算节点的重要性，通常用于发现社区的中心，常见的中心性指标有度中心性指标、介数中心性指标、Katz-Bonacich 中心性指标和亲密中心性指标等。节点相似性指标可以用于复杂网络的社区发现算法。Wang 等[18]同时考虑了一阶和二阶邻居信息，结合局部相似性 Jaccard 指标和聚类系数来划分社区。Zhang 等[19]提出了一种基于种子扩展的社区发现算法，揭示了节点相似性和社区融合在社区发现中的作用，并引入核心相似性来改进度中心性和适应度的定义。Verma 等[20]提出了一种基于标签传播算法的社区发现算法，通过标识出网络中不同社区的中心节点和传播中心节的影响力来发现网络中的社区。You 等[21]提出了一种基于局部和全局信息的三阶段社区发现算法，使用了一种新的指标来识别社区的中心节点。

2.2 网络嵌入

网络嵌入旨在用低维和稠密的向量表示高维、稀疏的网络，同时保留网络的拓扑结构。网络嵌入作为一种有效且高效的网络表示方式，目前已经在社交网络数据挖掘、链路预测和标签分类等方面取得了良好的效果。例如，DeepWalk[22]为网络的每个节点生成随机游走序列，并将节点序列视为 word2vec[23]中的句子，然后将这些序列作为自然语言处理的经典模型 Skip-Gram 的输入，以获取网络表示；Node2vec[24]通过定义两个参数来平衡深度优先采样（Depth First Sampling）和广度优先采样（Breadth First Sampling），修改随机游走的策略，从而使其可以捕获更全面的节点邻域信息；LINE[25]结合了节点的一阶、二阶邻域信息来优化目标函数，保留了网络的局部和全局结构，此外，通过边缘采样的方法解决了随机游走时梯度下降的问题；SDNE 通过半监督神经网络模型联合优化一阶和二阶的邻近度，用自编码器重构二阶邻近度维护网络的全局结构，以一阶邻近度来保留网络的局部结构，约束了节点的低维嵌入。

3 基于马尔可夫相似性增强和网络嵌入的社区发现（MSE）

本文提出了一种基于马尔可夫相似性增强和网络嵌入的社区发现算法。首先，研究组提出了一种基于马尔可夫相似性指标来获取网络的初始社区，进而基于网络拓扑结构和 SDNE 嵌入算法提出了一种社区相似性指标将小社区与大社区合并，从而获取网络的社区结构。

3.1 问题描述

针对一个无向无权的网络 $G=(V,E)$，其中 V 表示节点集合，E 表示节点间连边的集合。

目标是从网络中获取马尔可夫相似性增强矩阵，进而根据马尔可夫相似性指标获得网络的初始社区，再结合网络拓扑结构和 SDNE 网络嵌入算法提出新的社区相似性指标，进行初始社区合并以完成社区发现。

3.2 MSE 框架

如图 1 所示，本文算法主要分为 3 步：

（1）对初始网络进行马尔可夫相似性迭代转移，获取网络稳态下的马尔可夫相似性增强矩阵。

（2）通过马尔可夫相似性增强矩阵得到节点之间的马尔可夫相似性指标，获取每一个节点的最相似节点，

进而获得初始社区。

(3)利用 SDNE 网络嵌入算法将高维稀疏的网络映射成低维稠密的向量,提出了一种社区相似性指标对初始社区进行合并,完成社区划分。

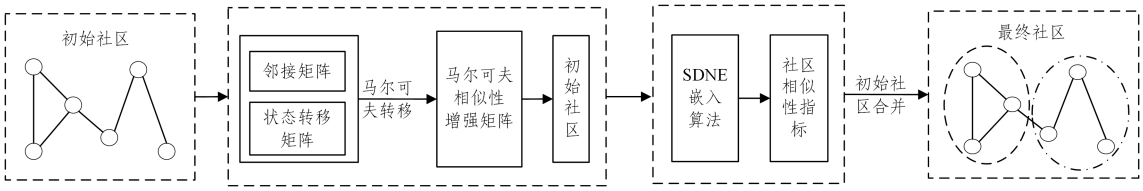


图 1 MSE 算法框架图

Fig. 1 Diagram of MSE algorithm framework

3.3 马尔可夫相似性增强矩阵

马尔可夫链是一个满足马尔可夫性质的随机过程。马尔可夫性质指未来状态只与当前状态相关,而与过去状态无关。过去的状态和未来的状态是条件独立的。也就是说,马尔可夫链任何时刻的状态值只依赖于前一时刻的状态,并且结合初始的状态分布和一个确定的状态转移矩阵,最后一定会达到一个稳定的状态。在马尔可夫相似性的增强部分,通过不断地将状态转移矩阵相乘进行马尔可夫相似性迭代转移,在此过程中,强相似性节点之间的相似性会变得更强,而弱相似性节点之间的相似性会变得更弱,直至状态转移矩阵收敛,得到最终的马尔可夫相似性增强矩阵。

本文以网络的邻接矩阵 $\boldsymbol{A}$ 为网络的初始状态矩阵,如式(1)所示:

$$\boldsymbol{A}=\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & s_{2,2} & \cdots & a_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N} \end{bmatrix} \tag{1}$$

其中,$a_{i,j}=1$,表示节点 v_i 和 v_j 之间存在连边,$a_{i,j}=0$,表示不存在连边。

我们使用 Jaccard 局部相似性指标来衡量节点之间的相似性,其中,$\Gamma(v_i)$ 表示节点 v_i 邻居的集合,$|\Gamma(v_i)|$ 意味着节点 v_i 邻居的数量,如式(2)所示:

$$sim(v_i,v_j)=\frac{|\Gamma(v_i)\bigcap\Gamma(v_j)|}{|\Gamma(v_i)\bigcup\Gamma(v_j)|},v_j\in\Gamma(v_i) \tag{2}$$

该网络的状态转移矩阵定义为:

$$\boldsymbol{S}=\begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,N} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ s_{N,1} & s_{N,2} & \cdots & s_{N,N} \end{bmatrix} \tag{3}$$

$$s_{i,j}=\begin{cases} 0, & v_j\notin\Gamma(v_i) \\ sim(v_i,v_j), & v_j\in\Gamma(v_i) \end{cases}$$

其中,$s_{i,j}$ 表示节点 v_i 和 v_j 的相似性,可以理解为节点 v_i 和 v_j 在同一个社区的概率,也可以认为从节点 v_i 状态转移到节点 v_j 状态的概率。由于每个状态转移到其他状态的概率之和为 1,因此需要对矩阵 $\boldsymbol{S}$ 做一个归一化操作:令 $s_{i,j}=s_{i,j}/\sum_{k=1}^{N}s_{i,k}$

马尔可夫相似性增强矩阵定义为:

$$\boldsymbol{Se}=\boldsymbol{AS}^n \tag{4}$$

其中,$n=\lceil|E|/N\rceil$,表示网络边的数量与网络节点的数量之比,N 表示网络中节点的数量,$|E|$ 表示网络中节点间连边的数量。

结合马尔可夫性质,通过对网络的初始状态进行马尔可夫相似性迭代转移,直至状态转移矩阵收敛,得到网络趋于稳态时的状态矩阵,即马尔可夫相似性增强矩阵。在此过程中使得网络中强相关节点之间的联系更加紧密,弱相关节点之间的联系更弱,这样更方便进行网络社区划分,使得到的网络社区结构更加清晰,识别的社区更加精确。

3.4 初始社区划分

定义 $Se_{i,j}$ 为节点 v_i 和 v_j 之间的相似性指标。对于任意节点 v_i,如果:

$$j=\text{argmax}(Se_{i,j}) \tag{5}$$

则节点 v_j 为它的最相似节点。

这样可以找到任意节点 v_i 的最相似节点 v_j,形成最相似性节点对 (v_i,v_j),遍历网络可以得到最相似性节点对集合 P,其中集合 P 中的每一列表示一个最相似节点对。因为节点对的相似性越高,二者在同一个社区的概率就越大,因此本文认为最相似节点对处于同一社区。在不考虑网络原有连边的情况下,我们连接集合 P 中每一个最相似节点,提取其中的连通成分,形成初始社区。例如,对于一个网络的最相似节点对集合 P,有:

$$P=\begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 \\ v_2 & v_1 & v_4 & v_6 & v_4 & v_5 & v_8 & v_9 & v_7 \end{pmatrix} \tag{6}$$

其中,每一列表示一个最相似节点对;把每一个最相似节点对相连,得到该网络的初始社区结构由 3 个社区 (v_1,v_2),(v_3,v_4,v_5,v_6),(v_7,v_8,v_9) 组成。

3.5 新社区相似性指标和初始社区合并

本文提出了一种社区合并策略,在初始社区结构中,节点数量小于阈值 λ 的社区被认为是小社区,大于或等于阈值 λ 的社区被认为是大社区。小社区应该要并入相邻而且联系最紧密的大社区中去,以便获得网络的最终社区结构。

为了找出与小社区最近且联系最紧密的大社区,本文提出了一种社区相似性指标。首先使用 SDNE 网络嵌入算法把高维稀疏的网络映射成为低维稠密的向量,获取网络中每个节点的低维向量,其中任意节点 v_i 的低维向量表示为 $\boldsymbol{Ls}_i$,网络中任意两节点 v_i 和 v_j 之间的距离 $dis(\boldsymbol{Ls}_i,\boldsymbol{Ls}_j)$ 为其对应的低维向量之间的欧氏距离。在合并社区这一步,只计算待合并小社区和其他社区之间连边所对应节点对之间的距离,而不是任意两节点之间的距离。对于初始社区集合 $C=\{C_0,C_1,\cdots,C_k\}$,定义任意社区 C_m 和 C_n 之间的社区相似性指标为:

$$Sc(C_m,C_n)=\sum_{v_i\in C_m,v_j\in C_n}\frac{sum(C_m)-dis(\boldsymbol{Ls}_i,\boldsymbol{Ls}_j)}{sum(C_m)}E(v_i,v_j) \tag{7}$$

其中，$E(v_i,v_j)$表示节点v_i和节点v_j是否有连边，如果有，则$E(v_i,v_j)=1$，否则$E(v_i,v_j)=0$；$sum(C_m)$表示社区C_m和其他社区之间连边所对应节点对之间的距离之和。该指标表明社区之间的连边越多，以及该连边对应节点对的距离越近，社区相似性就越大。

接下来根据新社区间的相似性，对于任意小社区，计算该小社区和其邻居社区的社区相似性，将该小社区和其中相似性最大的社区合并。遍历网络，将所有小社区与相似性最大的邻居社区合并，从而得到最终的社区结构。

3.6 模型实现步骤

算法 1 基于马尔可夫相似性增强和网络嵌入的社区发现算法

输入：Network G(V,E)

输出：社区集合 C={C_1,…,C_p}

1. for i=1 to N do
2. for j=1 to N do
3. S←Sim(v_i,v_j)
4. end
5. end
6. 归一化矩阵 **S**
7. **Se**=**AS**^{n}
8. for i=1 to N do
9. (v_i,v_j)←v_i的最相似节点v_j
10. 最相似节点对集合 P←(v_i,v_j)
11. end
12. 提取最相似节点对集合 P 的连通成分形成初始社区 C
13. for i=1 to |C| do
14. if|C_i|<λ
15. for j=1 to |C| do
16. if i!=j do
17. 根据式(7)，计算C_i的最相似社区C_j
18. end
19. end
20. 合并社区C_i和C_j
21. end
22. end

MSE 算法的整体步骤如下：

(1)根据网络的拓扑结构得到网络的初始状态矩阵。

(2)根据式(3)计算得到网络的状态转移矩阵。

(3)根据式(4)计算得到马尔可夫相似性增强矩阵。

(4)通过节点之间的马尔可夫相似性指标得到网络中每个节点的最相似节点，构成最相似节点对，提出它们之间的连通成分，得到初始社区。

(5)通过 SDNE 网络嵌入算法得到网络节点的低维向量表示。

(6)通过式(7)计算得到所有节点数小于阈值 λ 的小社区和其他社区的社区相似性，将其并入最相似的社区中以得到最终社区。

4 数值仿真

为了研究 MSE 算法的精确性，本文在真实网络和人工网络上进行了一系列数值仿真实验，将 MSE 算法与 5 种知名算法进行了比较，包括 Newman 提出的 FN 算法[26]、Leading eigenvector 算法[27]、Raghavan 等提出的标签传播算法(LPA)、Clauset 等提出的 CNM 算法[28]以及 Wang 等提出的基于局部相似性和度聚类信息的算法(BLI)。

4.1 评价指标

(1)模块度[29](*Q*)：模块度指标通常用于衡量社区发现的质量。*Q* 越大，表明得到的社区结构越合理，社区发现质量越高，反之，社区发现效果越差。模块度的具体计算式如式(8)所示：

$$Q=\frac{1}{2|E|}\sum_{i,j}\left[\boldsymbol{A}_{ij}-\frac{|v_i||v_j|}{2|E|}\right]\delta(c_i,c_j) \tag{8}$$

其中，$\boldsymbol{A}_{ij}$表示邻接矩阵，$|v_i|$表示节点v_i的度，$|E|$表示边的数目，$\delta(c_i,c_j)$表示节点v_i，v_j是否处于同一个社区，若处于，则为 1，否则为 0。

(2)归一化互信息[30](*NMI*)：*NMI* 通常用于评估划分社区和真实社区的接近程度，$NMI\in[0,1]$，*NMI* 值越接近 1，说明划分的社区与真实社区越接近。其计算式如式(9)所示：

$$NMI(X;Y)=\frac{2I(X;Y)}{H(X)+H(Y)} \tag{9}$$

其中，$H(X)$表示 X 的熵，$I(X;Y)$表示 X 和 Y 之间的互信息。

4.2 真实网络

本文使用 7 种知名的真实网络来验证 MSE 算法的精确性，包括 karate 网络[31]、dolphins 网络[32]、polbooks 网络[33]、football 网络[34]、lesmis 网络[35]、polblogs 网络[36]和 facebook 网络[37]。这 7 个真实网络的基本信息如表 1 所列。

表 1 7 个真实网络的基本信息

Table 1 Basic information of 7 real networks

Network	Node	Edge	Community-Truth
karate	34	78	2
dolphins	62	159	2
polbooks	105	441	3
football	115	613	12
lesmis	77	254	—
polblogs	1490	19090	—
facebook	4039	88234	—

经过多次实验，表 2 列出了 MSE 算法和其他 5 种算法在 7 个网络上模块度的大小。其中，7 个网络中设置的最佳阈值 λ 分别为 4,3,7,4,3,12,10。从表 2 可以看出，在 karate 网络上，MSE 算法取得了最高的 *Q* 值 0.417，远高于 BLI 算法；在 dolphins 网络上，MSE 算法取得了最高的 *Q* 值 0.514，Leading eigenvector 算法取得了最低的 *Q* 值 0.491；在 polbooks 网络上，MSE 算法取得了最高的 *Q* 值 0.519，远高于 Leading eigenvector 算法的 *Q* 值 0.467；在 football 网络上，MSE 算法取得了最高的 *Q* 值 0.6，远高于 Leading eigenvector 算法取得的 *Q* 值 0.492；在 lesmis 网络上，Leading eigenvector 算法取得了最高的 *Q* 值 0.532，MSE 算法的 *Q* 值为 0.472；在 polblogs 网络上，Leading eigenvector 算法和 LPA 算法取得了最高的 *Q* 值 0.521，BLI 算法取得了最低的 *Q* 值 0.39，MSE 算法取得的 *Q* 值为 0.512，略低于 Leading eigenvector 算法和 LPA 算法；在 facebook 网络上，MSE 算法取得了最佳的 *Q* 值 0.811，FN 算法和 CNM 算法取得了最低的 *Q* 值 0.774。

表 2 MSE 与 5 种知名算法在真实网络上的模块度比较

Table 2 Comparison of modularity between MSE and five well-known algorithms in real networks

Networks	Modularity/Number of communities					
	MSE	FN	CNM	Leading eigenvector	BLI	LPA
karate	**0.417/4**	0.380/3	0.380/3	0.393/4	0.370/3	0.375/4
dolphins	**0.514/6**	0.495/4	0.495/4	0.491/5	0.510/5	0.506/4
polbooks	**0.519/5**	0.501/4	0.501/4	0.467/4	0.511/6	0.496/4
football	**0.600/11**	0.588/10	0.549/6	0.492/8	0.550/11	0.582/11
lesmis	0.472/4	0.516/5	0.500/5	**0.532/8**	0.525/6	0.526/5
polblogs	0.512/5	0.502/25	0.516/20	**0.521/6**	0.390/97	**0.521/6**
facebook	**0.811/46**	0.777/13	0.777/13	0.799/18	0.801/98	0.796/53

经过多次实验，表 3 列出了 MSE 算法和其他 5 种算法在具有真实划分的 4 个网络上的 *NMI* 值，其中 MSE 算法在表 2 中的 4 个网络中设置的最佳阈值 λ 分别为 5，12，7，4。在 katate 网络上，MSE 算法取得了最高的 *NMI* 值 0.837，FN 算法取得了最低的 NMI 值 0.579；在 dolphins 网络上，MSE 算法取得了最高的 *NMI* 值 0.888，Leading eigenvector 算法取得了最低的 *NMI* 值 0.449；在 polbooks 上，MSE 算法和 LPA 算法取得了最高的 *NMI* 值 0.539，Leading eigenvector 算法取得了最低的 *NMI* 值 0.52；在 football 网络上，MSE 算法取得了最高的 *NMI* 值 0.921，CNM 算法取得了最低的 *NMI* 值 0.697。可以看出，MSE 算法在 4 个网络中都取得了最高的 *NMI* 值，并且在 karate 网络和 dolphins 网络中，MSE 算法取得的 *NMI* 值远高于其他 5 种算法。

表 3 MSG 和 5 种知名算法在真实网络上的 NMI 比较

Table 3 Comparison of NMI between MSE and five well-known algorithms on real networks

Networks	Modularity/Number of communities					
	MSE	FN	CNM	Leading eigenvector	BLI	LPA
karate	**0.837/2**	0.579/5	0.690/3	0.677/4	0.699/3	0.563/4
dolphins	**0.888/2**	0.572/4	0.572/4	0.449/5	0.510/5	0.552/4
polbooks	**0.539/5**	0.515/4	0.530/4	0.520/4	0.533/6	**0.539/4**
football	**0.915/11**	0.878/10	0.697/6	0.698/8	0.889/12	0.909/11

表 4 列出了 MSE 算法和 BLI 算法在进行马尔可夫相似性迭代转移和不进行马尔可夫相似性迭代转移时的性能。其中，NMSE 表示 MSE 算法不进行马尔可夫相似性迭代转移，直接使用局部相似性 Jacard 指标识别初始社区；MBLI 算法表示在 BLI 算法的基础上，对其中的局部相似性识别初始社区那一部分进行马尔可夫相似性迭代转移。经过多次实验，取其中的最优结果。

表 4 进行马尔可夫迭代转移和不进行马尔可夫迭代转移时 MSE 和 BLI 算法的模块度 Q 与 *NMI* 值的比较

Table 4 Comparison of modularity Q and *NMI* value of MSE and BLI algorithms with and without Markov iterative transfer

Networks	MSE		NMSE		MBLI		BLI	
	Q	*NMI*	Q	*NMI*	Q	*NMI*	Q	*NMI*
karate	**0.417**	**0.837**	0.371	**0.837**	**0.400**	**0.837**	0.370	0.699
dolphins	**0.514**	**0.888**	0.509	0.582	0.461	**0.776**	**0.510**	0.510
polbooks	**0.519**	**0.539**	0.435	0.477	**0.519**	**0.539**	0.511	0.533
football	**0.600**	**0.915**	0.578	0.904	**0.577**	**0.899**	0.550	0.889

从表 4 可以看出，MSE 算法在 4 个网络上的 Q 值和 *NMI* 值都高于 NMSE 算法，MBLI 算法除了在 dolphins 网络上的 Q 值低于 BLI 算法，其他情况下的性能均优于 BLI 算法。因此可以看出，进行马尔可夫迭代相似性转移之后的效果远远优于不进行马尔可夫迭代相似性转移的效果。

综上所述，在 7 个真实的网络数据集上，MSE 算法在 5 个网络数据集中取得了最优的模块度，在所有网络中都获得了最优的 *NMI* 值。因此，MSE 算法的整体性能最佳。

4.3 人工网络

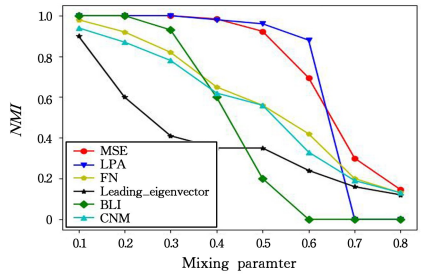
本文使用 LFR 基准网络测试来检验 MSE 算法。LFR 网络提供了一系列参数来控制生成网络的拓扑结构，具体参数如表 5 所列。

表 5 LFR 基准图参数描述

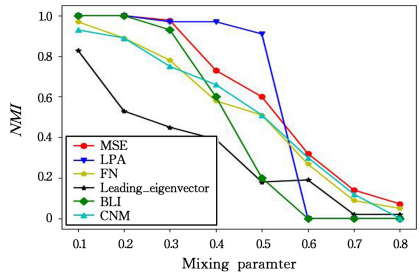
Table 5 LFR benchmark map parameter description

参数	描述
N	网络的节点数
k	网络中节点度的平均值
max k	网络中节点度的最大值
μ	混合参数
α	度分布的幂率指数
β	社区分布的幂率指数
min c	网络中最小社区的节点数
max c	网络中最大社区的节点数

图 2(a)、图 2(b)分别给出了 MSE 算法和其他 5 种算法在节点数为 1000 的 LFR 人工网络上改变网络中社区的最小规模、最大规模和混合参数的情况下的 *NMI* 对比结果。



(a)1000-S



(b)1000-B

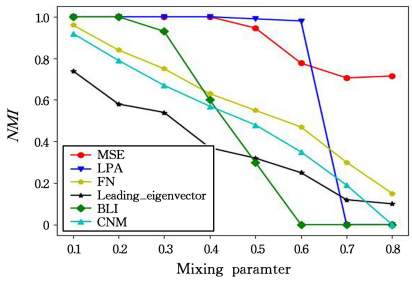
图 2 在 $N=1000$ 的 LFR 网络上各算法 *NMI* 的比较

Fig. 2 Comparison of *NMI* of various algorithms on LFR network with $N=1000$

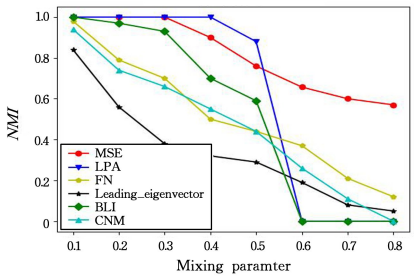
图 2(a)具体的参数如下：节点个数 $N=1000$，节点度的平均值 $k=15$，度的最大值 max $k=50$，度分布的幂率指数 $\alpha=2$，社区分布的幂率指数 $\beta=1$，最小社区的节点数 min $c=10$，最大社区的节点数 max $c=50$，混合参数 μ 设置为 0.1～0.8，基于此参数生成的网络称为 1000-S。图 2(b)具体的参数为：节点个数 $N=1000$，节点度的平均值 $k=15$，度的最大值max $k=50$，度分布的幂率指数 $\alpha=2$，社区分布的幂率指数 $\beta=1$，最小社区的节点数 min $c=20$，最大社区的节点数

$max\ c=100$，混合参数 μ 设置为 0.1～0.8，基于此参数生成的网络称为 1000-B。由图 2(a)可以看出，当 $\mu\in[0.1,0.4]$时，MSE 算法和 LPA 算法取得了最优的 *NMI* 值，远高于其他算法；当$\mu\in[0.5,0.6]$时，MSE 算法的性能只略逊于 LPA 算法，远优于其他算法；当 $\mu\in[0.7,0.8]$时，LPA 算法和 BLI 算法已经不能识别社区结构，MSE 算法取得了最优的 *NMI* 值。由图 2(b)可以看出，当 $\mu\in[0.1,0.3]$时，MSE 算法和 LPA 算法取得了最优的 *NMI* 值，远高于其他算法；当 $\mu\in[0.4,0.5]$时，MSE 算法的 *NMI* 值只低于 LPA 算法；当 $\mu\in[0.6,0.8]$时，可以看出，此时 LPA 算法和 BLI 算法已经不能识别社区结构，MSE 算法取得了最优的 *NMI* 值。综合考虑，MSE 算法的性能最优。

图 3(a)、图 3(b)分别给出了 MSE 算法和其他 5 种算法在节点数为 4000 的 LFR 人工网络上改变网络中社区的最小规模、最大规模和混合参数的情况下的 *NMI* 对比结果。



(a)4000-S



(b)4000-B

图 3 在 $N=4000$ 的 LFR 网络上各算法 *NMI* 值的比较

Fig. 3 Comparison of *NMI* values of various algorithms on LFR network with $N=4000$

图 3(a)具体的参数如下：节点个数 $N=4000$，节点度的平均值 $k=15$，度的最大值 $\max k=50$，度分布的幂率指数 $\alpha=2$，社区分布的幂率指数 $\beta=1$，最小社区的节点数 $\min c=10$，最大社区的节点数 $\max c=50$，混合参数 μ 设置为 0.1～0.8，基于此参数生成的网络称为 4000-S。图 3(b)的具体参数如下：节点个数 $N=4000$，节点度的平均值 $k=15$，度的最大值 $\max k=50$，度分布的幂率指数 $\alpha=2$，社区分布的幂率指数 $\beta=1$，最小社区的节点数 $\min c=20$，最大社区的节点数 $\max c=100$，混合参数 μ 设置为 0.1～0.8，基于此参数生成的网络称为 4000-B。从图 3(a)中可以看出，当 $\mu\in[0.1,0.4]$时，MSE 算法和 LPA 算法取得了最优的 *NMI* 值，远高于其他算法；$\mu\in[0.5,0.6]$时，MSE 算法的性能只略逊于 LPA 算法，远优于其他算法；$\mu\in[0.7,0.8]$时，LPA 算法和 BLI 算法已经不能识别社区结构，MSE 算法的 *NMI* 值还保持在 0.7 以上，远远高于其他算法。从图 3(b)可以看出，当 $\mu\in[0.1,0.3]$时，MSE 算法和 LPA 算法取得了最优的 *NMI* 值，远高于其他算法；当 $\mu\in[0.4,0.5]$时，MSE 算法的 *NMI* 值只低于 LPA 算法；$\mu\in[0.6,0.8]$时，LPA 算法和 BLI 算法已经不能识别社区结构，而 MSE 算法的 *NMI* 值还保持在 0.6 以上，远远高于其他算法。综合考虑，MSE 算法的性能最优。

结束语 本文受马尔可夫链思想的启发，提出了一种马尔可夫相似性增强方法，得到了稳态的马尔可夫相似性增强矩阵，根据马尔可夫相似性指标对网络进行初始社区的划分，并在此基础上结合网络拓扑结构和网络嵌入算法，提出了一种新的社区相似性指标来合并初始社区，从而得到网络的社区结构。在真实网络和人工网络上，对比其他知名的社区发现算法，证明了 MSE 算法能获得更优的社区结构。未来将研究马尔可夫相似性方法在属性网络和重叠社区的检测策略。该算法可以用于社交网络用户社区的划分、脑网络不同功能区域的发现、医院网络中规划医疗社区结构等领域，具有广阔的应用前景。

参 考 文 献

[1] LI M, LU S, ZHANG L, et al. A Community Detection Method for Social Network Based on Community Embedding[J]. IEEE Transactions on Computational Social Systems, 2021, 8(2): 308-318.

[2] ZHANG W, SHANG R, JIAO L. Complex Network Graph Embedding Method Based on Shortest Path and MOEA/D for Community Detection [J]. Applied Soft Computing, 2020, 97: 106764.

[3] ACMAN M, DORP L V, SANTINI J M, et al. Large-scale Network Analysis Captures Biological Features of Bacterial Plasmids[J]. Nature Communications, 2020, 11(1): 1-11.

[4] HAN N, QIAO S J, YUAN C A, et al. Fast Community Parallel Detection Algorithm in Mobile Social Networks[J]. Journal of Chongqing University of Technology: Natural Science, 2020, 34(1): 94-102.

[5] IMANE M, NADJET K. Community Detection UsingFireworks Optimization Algorithm[J]. International Journal of Artificial Intelligence Tools, 2019, 28(3): 1950010.

[6] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E, 2007, 76(3): 036106.

[7] CLAUSET A, NEWMAN M E J, MOORE C. Finding Community Structure in Very Large Networks[J]. Physical Review E, 2004, 70(6): 066111.

[8] HU F, ZHU Y, SHI Y, et al. An Algorithm Walktrap-SPM for Detecting Overlapping Community Structure[J]. International Journal of Modern Physics B, 2017, 31(15): 1750121.

[9] NEWMAN M E J, GIRVAN M. Finding and Evaluating Community Structure in Networks[J]. Physical Review E, 2004, 69(2): 026113.

[10] GIRVAN M, NEWMAN M E J. Community Structure in Social and Biological Networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826.

[11] ROSVALL M, BERGSTROM C T. Maps of RandomWalks on Complex Networks Reveal Community Structure[J]. Procee-

dings of the National Academy of Sciences, 2008, 105(4): 1118-1123.
[12] EUSTACE J, WANG X, CUI Y. Community Detection Using Local Neighborhood in Complex Networks[J]. Physica A: Statistical Mechanics and its Applications, 2015, 436: 665-667.
[13] LIU J L, WANG D L, ZHANG Y F, et al. Local Community DetectionApproach Based on Fuzzy Similarity Relation[J]. Journal of Software, 2020, 31(11): 3481-3491.
[14] KUMAR S, PANDA B S, AGGARWAL D. Community Detection in Complex Networks Using Network Embedding and Gravitational Search Algorithm[J]. Journal of Intelligent Information Systems, 2020, 57(1): 51-72.
[15] ZHAO X, LI X, ZHANG Z H, et al. Community Detection Algorithm Combining Community Embedding and Node Embedding [J]. Computer Science, 2020, 47(10): 121-125.
[16] HE D, WANG Y, CAO J, et al. A Network Embedding-enhanced Bayesian Model for Generalized Community Detection in Complex Networks[J]. Information Sciences, 2021, 575: 306-322.
[17] WANG D, CUI P, ZHU W. Structural Deep Network Embedding[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1225-1234.
[18] WANG T, YIN L Y, WANG X X. A community Detection Method Based on Local Similarity and Degree Clustering Information[J]. Physica A, 2018, 490: 1344-1354.
[19] ZHANG J, DING X, YANG J. Revealing the Role of Node Similarity and Community Merging in Community Detection[J]. Knowledge-Based Systems, 2018, 165: 407-419.
[20] VERMA P, GOYAL R. Influence Propagation Based Community Detection in Complex Networks[J]. Machine Learning with Applications, 2020, 12: 100019.
[21] YOU X, MA Y, LIU Z. A Three-stage Algorithm on Community Detection in Social Networks[J]. Knowledge-Based Systems, 2020, 187: 104822.
[22] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online Learning of Social Representations[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 701-710.
[23] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv: 1301.3781, 2013.
[24] GROVER A, LESKOVEC J. Node2vec: Scalable Feature Learning for Networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.
[25] TANG J, QU M, WANG M, et al. Line: Large-scale Information Network Embedding[C]// Proceedings of the 24th International Conference on World Wide Web. 2015: 1067-1077.
[26] NEWMAN M E J. Fast Algorithm for Detecting Community Structure in Networks[J]. Physical Review E, 2004, 69(6): 066133.
[27] NEWMAN M E J. Finding Community Structure in Networks Using the Eigenvectors of Matrices[J]. Physical Review E, 2006, 74(3): 036104.
[28] CLAUSET A, NEWMAN M E J, MOORE C. Finding Community Structure in Very Large Networks[J]. Physical Review E, 2004, 70(6): 066111.
[29] HESAMIPOUR S, BALAFAR M A. A New Method for Detecting Communities and their Centers Using the Adamic/Adar Index and Game Theory[J]. Physica A: Statistical Mechanics and its Applications, 2019, 535: 122354.
[30] SU Y S, LIU C L, NIU Y Y, et al. A Community Structure Enhancement-based Community Detection Algorithm for Complex Networks[J]. IEEE Transactions on Systems Man Cybernetics-Systems, 2021, 51(5): 2833-2846.
[31] SUN Z J, SUN Y N, CHANG X F, et al. Community Detection Based on the Matthew Effect[J]. Knowledge-Based Systems, 2020, 205: 106256.
[32] ZHANG L, PAN H B, SU Y S, et al. A Mixed Representation-Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection[J]. IEEE Transactions on Cybernetics, 2017, 47(9): 2703-2716.
[33] JIANG H, LIU Z, LIU C, et al. Community Detection in Complex Networks with an Ambiguous Structure Using Central Node Based Link Prediction[J]. Knowledge-Based Systems, 2020, 195: 105626.
[34] LIU H J, MA H F, ZHAO Q Q, et al. Target Community Detection with User Interest Preferences and Influence[J]. Journal of Computer Research and Development, 2021, 58(1): 70-82.
[35] LU H, SHEN Z, SANG X S, et al. Community Detection Method Using Improved Density Peak Clustering and Nonnegative Matrix Factorization[J]. Neorocomputing, 2020, 415: 247-257.
[36] CHEN Y Z, SHI S, ZHU W P, et al. An Incremental Community Detection Algorithm Based on Neighborhood Following Relationship[J]. Chinese Journal of Computers, 2017, 40(3): 570-583.
[37] NATH K, SHANMUGAM R, VARADARANJAN V. MaCODE: A Multi-phase Approach on Community Detection in Evolving Networks[J]. Information Sciences, 2021, 569: 326-343.
[38] SAID A, ABBASI R A, MAQBOOL O, et al. CC-GA: A Clustering Coefficient Based Genetic Algorithm for Detecting Communities in Social Networks[J]. Applied Soft Computing, 2018, 63: 59-70.

ZENG Xiangyu, born in 1998, postgraduate. His main research interests include community detection and graph neural network.

YANG Xuhua, born in 1971, Ph.D, professor, is a member of China Computer Federation senior. His main research interests include machine learning and network science.

(责任编辑:喻藜)