ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 2,Mar. 2018,pp. 288－298
DOI:10.16337/j.1004-9037.2018.02.011
© 2018 by Journal of Data Acquisition and Processing
http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于多标签传播的重叠社区发现优化算法

杜长江　王志晓　邢贞明

(中国矿业大学计算机科学与技术学院,徐州,221116)

摘　要: 标签传播算法是一种被广泛应用的社区发现算法,该算法为网络中的每个节点分配一个初始标签,然后通过传播标签来发现复杂网络中的潜在社区,具有时间复杂度低的特点。当前基于标签传播的重叠社区发现算法存在忽略节点重要性差异、需要人为设置参数等不足。针对该类算法在重叠社区发现方面的缺陷,提出一种基于多标签传播的重叠社区发现优化算法。该算法使用 K-核分解方法找出若干个社区核心节点,以这些节点为种子节点,逐层向外传播标签;在进行标签选择的时候以邻居节点标签的种类来决定重叠节点的标签个数。实验表明,该算法明显改善了社区发现的性能,提高了划分结果的稳定性和准确性。

关键词: 复杂网络;重叠社区;标签传播;K-核分解

中图分类号: TP391　　**文献标志码:** A

Overlapping Community Detection Algorithm Based on Improved Multi-label Propagation

Du Changjiang, Wang Zhixiao, Xing Zhenming

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China)

Abstract: Label propagation is a widely used community detection method with low complexity. It assigns an initial label for each node in the network, and then propagates the labels to discover the potential community structure in complex networks. However traditional label propagation is faced with some inadequacies, such as ignoring the difference between nodes and input parameters demanding. To overcome those defects, this paper puts forward an overlapping community detection algorithm based on the improved multi-label propagation. It uses K-shell decomposition method to identify core nodes of the network firstly, and then updates labels outward layer by layer. The number of labels of overlapping nodes is determined by the types of neighbor node when choosing label for a node. Experiment results show that this algorithm makes the community detection results more accurate and stable.

Key words: complex network; overlapping community; label propagation; K-shell decomposition

引　　言

随着社交网络的兴起,社会网络引起了研究人员的重视。社区结构不仅反映了网络的拓扑结构,同

基金项目: 国家自然科学基金(61402482)资助项目;中国博士后基金(2015T80555)资助项目;江苏省博士后基金(1501012A)资助项目。

收稿日期: 2016-09-10;**修订日期:** 2016-10-14

时也被认为是能体现真实社会网络功能的一个重要属性[1-2],因此社区发现算法的研究有很重要的理论价值和现实意义[1,3-4]。比如针对互联网用户进行社区划分,可根据用户所在社区提供特定服务,有很大的商业价值。

近年来,研究人员提出了很多富有成效的社区发现算法。经典的基于图划分的 KL (Kernighan-Lin)算法[5]和谱平分算法[6],前者基于贪婪原理[5,7],后者则利用拉普拉斯矩阵[6,8]的特征向量,都可将网络划分成两个社区。但是这两种方法有很大的局限性,适用于社区数目已知的网络。基于层次聚类的 GN(Girvan-Newman)算法[9],Newman 和 Girvanc 采用分裂思想,构造出分层树,可将网络划分为多个社区,一定程度上克服了 KL 算法[5]和谱平分算法[6]的缺陷,但是该算法复杂度较高[10],而且社区数量不确定,依赖于在分层树上选取的划分位置[1]。2014 年,Jiang 等人[7]提出了基于优化的贪心算法的社区发现算法(Algorithm based on greedy surprise optimization, AGSO)及其改进版本 FAGSO(Fast-AGSO)算法,该算法在实验中展现出了比社区发现的几种重要算法更好的效果,包括 BGLL(Blondel-Guillaume-Lambiotte-Lefebvre)[11]、CNM (Clauset-Newman-Moore)[12]、OSLOM(Order statistics local optimization method)[13]等。不同于传统的算法,Raghavan 等人[14]首次将标签传播思想引入到社区发现领域,提出标签传播算法(Label propagation algorithm,LPA),综合考虑了网络结构和网络本身的传播特性,且具有较高的效率,适用于大规模网络的社区发现[15]。

然而上述算法均是针对非重叠社区,即网络中的一个节点只能属于一个社区[1]。但很显然,社会网络中一个节点属于多个社区的情况是普遍存在的,比如,一个人在不同的领域扮演不同的角色、在论坛中参与不同社区的话题讨论等。所以,重叠社区更符合复杂网络的实际意义[1,16]。2005 年,Palla 等人[17]提出了基于团过滤的重叠社区发现算法(Clique percolation method,CPM)。该方法假设社区由一系列相互连接的团(完全子图)构成,通过搜索融合相邻的团来划分社区。由于方法需要设置团的初始规模 K,因此划分结果受到参数 K 的影响。作为经典重叠社区发现算法,该算法常被当作基准算法与其他算法比较[18]。2007 年,Steve 等人[19]在 GN 算法的基础上进行改进给出了 CONGA(Cluster-overlap Newman Grivan algorithm)算法使其可用于重叠社区发现,并在其基础是进一步引入局部中介度的概念,进一步提出了 CONGO 算法(Cluster-overlap Newman Grivan optimization)[20]。

随着现实网络规模的不断增加,这些传统的方法由于时间复杂度高,在实际的应用中有很大的限制[21]。考虑到标签传播算法 LPA 是一种高效率的社区发现算法[1,21],很多学者对其进行了改进以应用于重叠社区发现。2010 年,Steve 等人[22]在单一标签的基础上引入了多标签的概念,提出了基于标签传播的重叠社区发现算法 (Community overlap propagation algorithm,COPRA),该算法不仅能发现重叠社区,且保留了 LPA 算法时间复杂度低的特点,其在大规模网络上的实验表明 COPRA 效率远高于 CPM 和 CNOGO 算法; 2012 年,Wu 等人[23]对其进行改进并提出了基于均衡多标签传播的社区发现算法(Balanced multi-label propagation algorithm, BMLPA) ,该算法重新设计了 COPRA 算法中的标签更新策略,一定程度上增加了算法的稳定性,但是也使得算法的适应性有所下降;2013 年,王庚等人[24]在 COPRA 算法的基础上提出了一种平衡重叠社区挖掘算法 BOCLP(Balanced overlapping community detecting algorithm by label propagation),进一步提高了算法的稳定性,其在人工网络上的实验表明该算法在社区结构变模糊时效果优于 COPRA。2015 年,Sun 等人[25]提出了一种优势标签传播算法(Dominant label propagation algorithm, DLPA),该算法引入优势标签的概念,认为节点倾向于处于优势标签所代表的社区中,并可根据输入参数控制社区重叠率,当然这也导致其结果不够稳定。随后 Liu 等人[26]经过进一步研究提出了 DLPAE(DLPA expansion)算法,提高了算法的稳定性和划分的准确度。大量的实验表明[22,24],虽然在处理社区结构清晰的小型网络时 CPM 有很好的效果,但是在大规模复杂网络中,基于标签传播的算法更有优势。

尽管基于标签传播的算法具备简单快速的优点,适合目前的大规模社会网络,但是普遍存在稳定性

差的问题，每次划分结果可能不一致，且需要人为设置参数来限制节点所属社区个数，这在很大程度上影响了算法在大规模网络数据集上社区划分的准确率。基于 COPRA 的改进算法也还存在需要设置参数，适应性降低等缺陷。此外，传统的多标签传播算法在标签初始化及标签传播过程中忽略了节点之间的差异，导致算法有较强的随机性。可见，基于多标签传播的社区发现算法仍有改进的空间。

1 MOLPA 算法

针对 COPRA 算法及其改进算法存在的缺陷，本文提出一种基于多标签传播的重叠社区发现优化算法(Multi-label propagation optimization algorithm, MOLPA)，并从标签的初始化、标签传播顺序、标签选择 3 个方面来对 COPRA 算法进行改进。

1.1 标签初始化

COPRA 算法初始化时给每一个节点分配唯一的标签，然后通过标签迭代更新来完成社区发现。这种方式随机性较强，每个节点之间都是平等的，容易导致社区发现结果不稳定，降低准确率。实际网络中，不同节点的重要性和影响力往往是不一样的，而且与节点在网络中的位置有很大关系。

Kitsak 等人[27]提出 K-核分解算法来确定节点在网络中的重要性，K 核是指所有节点度值均不小于 K 的最大子网络，其中度值等于 K 小于 $K+1$ 的那部分成为 K-shell，简称 K_s。其基本过程是：删除网络中所有度值为 1 的节点，然后更新网络中节点的度值，继续删除度值为 1 的节点，重复此步骤直到剩余节点度值均大于 1，此时被删的部分 $K_s=1$；同样的方式，继续删除度值最小的节点，直到所有节点都被删除，完成 K 核分解。该方法是一种全局的评估方法，在判断节点重要度方面有较高的准确度，而且时间复杂度为 $O(n)$，比较适用于大规模网络。

MOLPA 算法在标签初始化时，先通过 K-核分解方法来寻找网络影响力比较大的节点，作为社区初始核心，并逐渐向外传播标签，当多个核心节点往外扩张到一定程度时，以它们为中心所形成的社区很可能会产生交集，就是要寻找的重叠社区。如图 1 所示，核心节点 1 和 2 在往外传播标签的过程中产生的重叠部分的节点就是两个社区间的重叠节点。

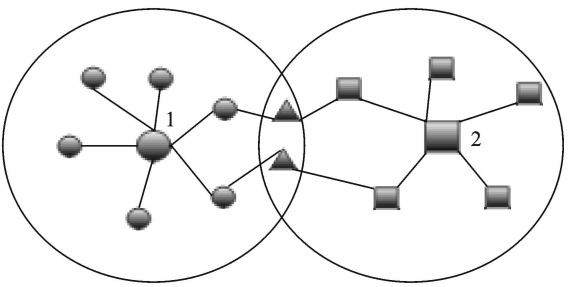


图 1 标签初始化示意图

Fig. 1 Initialized labels of nodes

1.2 标签传播顺序

COPRA 算法在进行标签传播的时候，首先将所有节点随机排列，然后按照随机序列的顺序来更新每一个节点的标签。这样的更新方式忽略了节点之间的差异，而且不同的排列顺序下所产生的标签传播结果也不同，导致社区划分的稳定性差、准确率不高。

为了提高社区划分结果的质量，减少随机性，MOLPA 对 COPRA 算法的标签更新顺序进行了改进。将标签初始化时找到的网络核心节点按照 K_s 值进行降序排列得到序列 T，首先更新序列 T 中的第一个核心节点的第一层邻居节点，然后更新第二个核心节点的第一层邻居节点，直到把所有核心节点的第一层邻居节点都更新完毕；下一次迭代再从第一个核心节点开始更新其第二层邻居节点，依次进行，每次迭代都按顺序以所有核心节点为中心向外传播一层，直到所有节点的标签都达到稳定状态，两个社区交界处的重叠部分就是要找的重叠社区。这样的标签更新策略，以节点的重要度为指导，使得标签传播过程更加有序，社区划分结果更稳定。并且由于只从核心节点向外层传播标签，减少了许多无意义社区的产生，使得其在提高社区划分准确度的同时，算法效率不至

于损失过大。

1.3 标签选择

COPRA 算法在标签更新的过程中,通过设置参数 v 来决定节点的标签个数,筛选出隶属度大于 $1/v$ 的标签。显然 v 值的设置很关键。然而在一个陌生的网络中,事先很难得知一个节点最多属于几个社区,而且每个节点的情况都不一样,如果用一个统一的参数去衡量它们,必定会影响到最终结果的准确性。

针对此问题,MOLPA 算法采用一种新的标签选择策略,由节点邻接点的标签种类来决定节点的标签个数。如图 2 所示,通过更新节点 1 的标签比较两种更新策略。

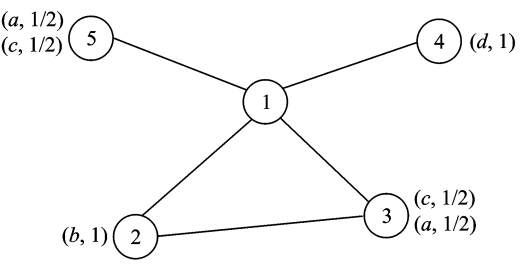


图 2 多标签传播过程示例

Fig. 2 Process of multi-label propagation

使用 COPRA 算法来更新节点 1 的标签,首先计算它的邻居节点标签的和 a:1/2+1/2=1;b:1;c:1/2+1/2=1;d:1。属于标签隶属度相同的情况,就随机选择两个(假设算法初始化的时候设置参数 $v=2$),比如选择 a,b,然后归一化处理,使得节点更新后的标签隶属度和为 1。因此,节点 1 更新后的标签为 $(a,1/2)$ 和 $(b,1/2)$。本来节点 1 是 a,b,c,d 这 4 个社区的重叠节点,但由于人为地设置参数 v 和随机选择,最终节点 1 只属于 a 和 b 这两个社区,这跟实际情况显然有出入。显然参数 v 的设置对 COPRA 划分结果的影响很大,而 MOLPA 采用新的标签选择策略能够避免标签选择时的随机性,有助于提高算法的稳定性和划分的准确度。

使用 MOLPA 算法的标签选择策略,就不需要设置 v 的值,而是根据节点 1 的邻接点标签种类个数来确定。因此,此时 v 等于 4,这 4 个标签值都大于 1/4,所以都保留下来,经过归一化,最终节点 1 的标签为 $(a,1/4)$,$(b,1/4)$,$(c,1/4)$,$(d,1/4)$,节点 1 同时属于这 4 个社区。很显然,与 COPRA 算法的结果相比,这种标签选择更加符合实际需求。

1.4 算法描述和流程

MOLPA 算法具体过程如下。

输入:$G=(V,E)$。

输出:社区划分结果。

Step 1 利用 K-核分解计算所有节点的 K_s 值,找到所有初始核心节点并按照降序排列为序列 $T=\{\mathrm{core}_i\}$,$i=1,\cdots,k$,为每一个核心节点 core_i 赋予唯一的标签,标签的格式为 $(\mathrm{core}_i,1)$。

Step 2 首先更新序列 T 中第一个社区核心节点的第一层邻居节点,接着更新序列 T 中第二个社区核心节点的第一层邻居节点,直到所有社区核心节点的第一层邻居都更新完毕;然后再从序列 T 中的第一个社区核心节点开始更新它的第二层邻居节点,依次进行。

Step 3 当所有节点标签不再变化,结束标签更新。

Step 4 将标签相同的节点合并为同一个社区。

MOLPA 算法的时间复杂度包括两部分:第一部分为使用 K-核分解法找出核心节点,时间复杂度为 $O(n+n\log n)$;第二部分为标签传播,时间复杂度为 $O(v_{\mathrm{avg}}m\log(v_{\mathrm{avg}}m/n))$,$v_{\mathrm{avg}}$ 为节点平均所属社区个数。因此算法总的时间复杂度为 $O(n+n\log n+v_{\mathrm{avg}}m\log(v_{\mathrm{avg}}m/n))$,在稀疏网络上,$v_{\mathrm{avg}}$ 值很小,此时算法的时间复杂度接近 $O(n+n\log n)$。MOLPA 算法按照核心节点的重要度大小来规定标签更新顺序,避免了许多不必要的迭代,实际应用中算法的效率很高。算法流程见图 3。

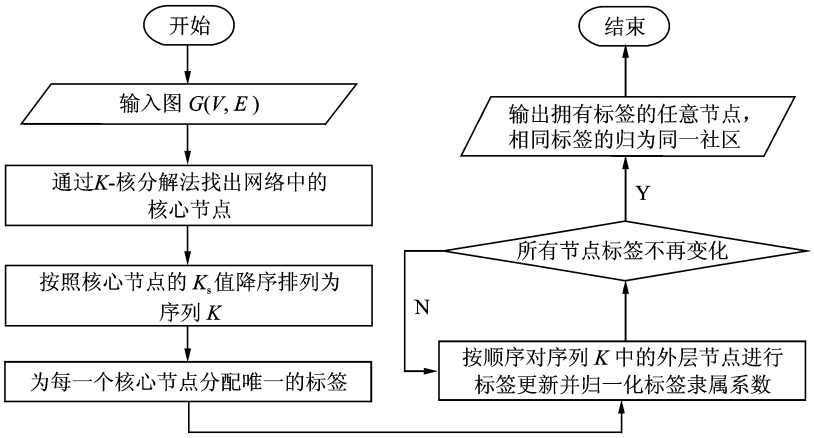


图 3 基于多标签传播的重叠社区优化算法流程图

Fig. 3 Process of MOLPA

2 实验结果和分析

为进一步验证算法的有效性，将 MOLPA 算法和两种标签传播重叠社区发现算法 COPRA 与 BO-CLP 算法以及传统的基于团过滤的重叠社区算法 CPM 算法进行实验对比，由于标签传播算法具有不稳定性，每一次的结果会有差异，因此以下实验都是进行 10 次并取平均值。

2.1 实验数据

(1)LFR 基准网络数据集

LFR 基准网络[28]是人工数据集，被许多研究者用来检测社区发现算法的功能，它能够生成用户指定分布的网络，包括网络节点度的分布和各个社区中节点数的分布。

LFR 基准网络主要包含以下参数：N 表示网络的节点数目；min$k$ 和 max$k$ 分别表示节点最小度数和最大度数；min$c$ 和 max$c$ 表示社区内节点个数的最小值和最大值；μ 为混合参数，表示社区结构清晰程度，它的取值范围为 0 到 1，μ 的值越大说明网络社区结构越模糊；o_n 指的是重叠节点个数；o_m 代表重叠节点所属社区的数目。可以对这些参数设置不同的值来控制节点数目、边密度、社区重叠情况等。实验所用的 LFR 数据集参数如表 1 所示。

表 1 LFR 基准网络数据

Tab. 1 Data of LFR networkks

网络	N	min$k$	max$k$	min$c$	max$c$	μ	o_n	o_m
b_1	3 000	20	50	10	50	0.1 ～0.8	100	4
b_2	5 000	40	100	20	100	0.1 ～0.8	300	4

(2)真实网络数据集

为验证算法在真实网络中的有效性，选取了 7 组真实网络数据，Zachary's Karate Club、Dolphin Social Network、American College Football 以及 PGP、DBLP 等[29]。这些真实网络的数据来自 http://www-personal.umich.edu/～mejn/netdata/和 http://snap.stanford.edu/data/index.html，具体参数如表 2 所示。

表 2 真实网络数据集

Tab. 2 Data sets of real networks

网络	网络规模
Karate	34
Dolphins	62
Football	115
C. elegans metabolic network	453
PGP network	10 680
Collaboration network	65 276
DBLP collaboration network	317 080

2.2 评价标准

实验用到两个评价指标:标准化互信息(Normalized mutual information,NMI)[1]和模块度函数 Q(Modularity)[9]。

(1)标准化互信息 NMI

2009 年,Lancichinetti 等人[30]提出了基于重叠社区结构的 NMI。假设 C 表示网络实际的社区集合,$|C|$ 表示社区的个数。则可以使用二元向量 x_i 来表示节点 i 属于哪一个社区,x_i 的长度为 $|C|$,$(x_i)_k$ 取值是 0 或 1,表示节点是否属于社区 k。将 x_i 的第 k 个元素当成一个随机变量 X_k,它的概率分布为 $P(X_k=1)=n_k/n$,$P(X_k=0)=1-n_k/n$,其中 n_k 指的是社区 k 的节点数目,n 指网络节点总数。类似地,在社区划分集合 C' 中,Y_l 代表节点属于社区 l 的概率分布。X_k 在 Y_l 上的条件熵定义为 $H(X_k|Y_l)=H(X_k,Y_l)-H(Y_l)$,根据 $H(X_k|Y)$,X_k 在 Y(所有 Y_l 构成的集合)上的条件熵 $H(X_k|Y)$ 为

$$H(X_k \mid Y)=\min_{l\in\{1,2,\cdots,|C'|\}} H(X_k|Y_t) \tag{1}$$

X(所有 X_k 构成的集合)在 Y 上的规范化条件熵为

$$H(X \mid Y)=\frac{1}{|C|}\sum_k \frac{H(X_k \mid Y)}{H(X_k)} \tag{2}$$

类似地,可以计算 Y 在 X 上的规范化条件熵 。最终根据式(1)计算 2 个社区集合的规范化互信息

$$\mathrm{NMI}(X \mid Y)=1-[H(X \mid Y)+H(Y \mid X)]/2 \tag{3}$$

(2)模块度函数 EQ

模块度 Q 指标针对的是非重叠社区结构,Shen 等[31]将模块度进一步扩展,给出可以衡量重叠社区结构的 EQ 函数。表达式如下

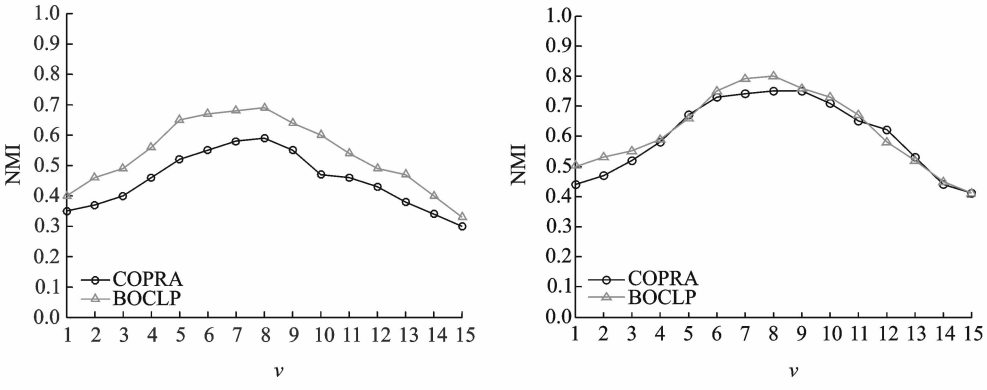
$$EQ=\frac{1}{2m}\sum_i \sum_{v\in c_i,w\in c_i} \frac{1}{O_v O_w}\left[A_{vw}-\frac{k_v k_w}{2m}\right] \tag{4}$$

式中:O_v 表示节点 v 从属的社区个数,A 表示网络所对应的邻接矩阵,A_{vw} 表示节点 i 和节点 j 之间的边数,m 表示网络中的边数,k_v 和 k_w 分别表示节点 v 和 w 的度。用 EQ 函数评价非重叠社区,EQ 的值越大则表示重叠社区的模块化程度越高。

2.3 测试实验参数

考虑到算法存在的不稳定性,为了更加客观地对比 MOLPA 与 COPRA、BOCLP 以及 CPM 4 种算法的性能,选取这些算法的最优结果与 MOLPA 算法结果进行对比。由于 COPRA 算法和 BOCLP 算法在标签传播过程中,需要设置参数 v(节点最多属于 v 个社区)来确定节点的标签个数,为了得到一个最优的 v 值,在 b_1 和 b_2 两个不同规模的人工网络上对 COPRA 算法和 BOCLP 算法进行实验,其中混合参数 $\mu=0.4$,实验结果如图 4 所示。由图 4 可知,当 v 取 7 或者 8 的时候,COPRA 算法和 BOCLP 算法的效果最好,在后面的人工数据实验当中这两种算法的 v 值统一取 8。

同理,由于 CPM 算法通过寻找完全子图的方式来挖掘社区结构,需要确定初始团的规模 K。Bron 和 Kerbosch[32]在测试团生成算法的时候,发现 K 值取 4~6 都能取得较为理想的社区发现结果。因此实验中选取 $K=4$。BOCLP 算法也需要寻找完全子图来确定网络初始社区核心,同样设置 $K=4$。



(a) b_1网络NMI随v值变化情况　(b) b_2网络NMI随v值变化情况

(a) NMI results of b_1　(b) NMI results of b_2

图 4　算法参数测试结果

Fig. 4　NMI results of COPRA and BOCLP of b_1 and b_2

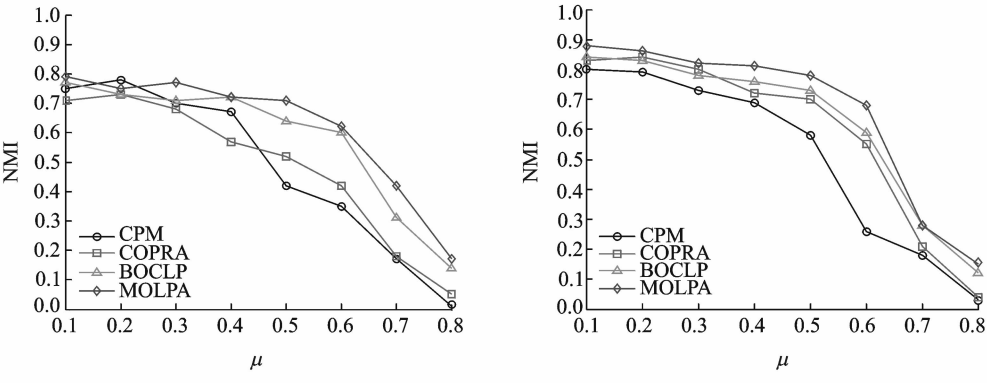
2.4　结果分析

2.4.1　LFR 基准网络数据集

实验选取了两种不同的人工网络，实验结果如下。

(1) NMI normalized mutual information 结果

从图 5 可以看出，在社区划分的准确度方面 MOLPA 算法要优于其他几种算法。在 μ 值较小时，所有的算法都能取得不错的结果。但是随着 μ 值的增大，社区结构越来越模糊，虽然社区发现的难度增大了，但 MOLPA 算法一直保持不错的效果。由于 CPM 算法和 BOCLP 算法都要通过寻找网络中相互连通的团(完全子图)，这对于网络结构的要求比较严格，不适用于完全子图较少的网络；而且 CORPA 算法和 BOCLP 算法还需要设置参数 v 来控制节点的标签个数，所以社区发现效果不够理想。这一实验表明了 MOLPA 初始化方法和标签传播顺序以及标签选择策略的有效性。



(a) b_1网络的NMI结果　(b) b_2网络的NMI结果

(a) NMI results of b_1　(b) NMI results of b_2

图 5　LFR 基准数据集上的 NMI 结果对比

Fig. 5　NMI results of LFP Networks

(2)模块度

LFR 基准数据集上模块度实验结果如图 6 所示。从图 6 明显看出,当 μ 值比较小的时候,这 4 种算法的模块度值较大,社区发现的结果质量较高;当 μ 逐渐增大的时候,即社区结构越来越模糊,这时几种算法的模块度都有不同程度的降低,但是本文提出的 MOLPA 算法的效果相比其他算法表现较好,说明通过 K-核分解方法以及新的标签传播策略来改进 COPRA 算法能够很大程度提高社区模块化程度,提升社区发现结果的质量。

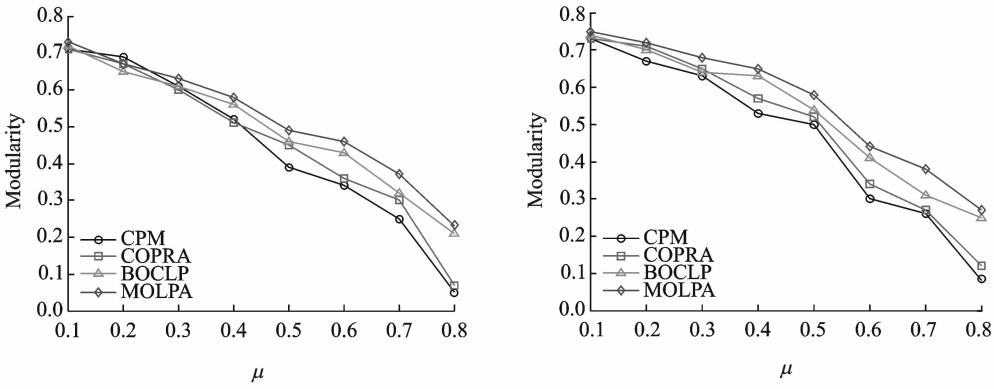


图 6 LFR 基准数据集上模块度实验结果

Fig. 6 Modularity test results of LFR networks

(3)算法效率

接下来在人工网络数据上测试这几种算法的效率。表 3 列出了实验所用参数。节点个数 N 为 1 000～5 000,网络规模不断增大,min$k$,max$k$,min$c$,max$c$,o_n,o_m 等也会增大。

表 3 人工网络实验数据

Tab. 3 Datas of LFR networks

参数	N	min$k$	max$k$	min$c$	max$c$	μ	o_n	o_m
值	1 000～5 000	10～40	20～100	40～200	10～100	0.4	50～300	4

不同算法的运行效率对比如图 7 所示。从图 7 中可以看出,MOLPA 算法的效率稍微逊色于 COPRA,但是比其他两种算法的性能都要好。CPM 和 BOCLP 算法都要寻找完全子图,这个过程耗费的时间比较多;而 MOLPA 算法通过使用 K-核分解方法寻找核心节点,K-核分解方法自身的效率就很高,而且固定的标签传播顺序也能减少迭代次数,社区结构很快就稳定,所以算法总体的时间复杂度相对于 COPRA 算法并没有提高太多。

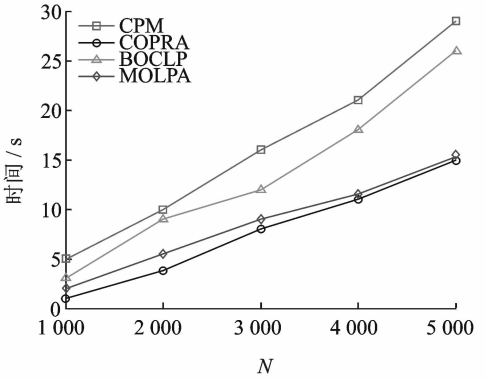


图 7 算法运行效率对比

Fig. 7 Comparison of efficiency of the four methods

2.4.2 真实网络数据集

(1)与 COPRA 的对比试验

在真实网络上,为了更加充分地了解 K-核分

解方法确定社区初始核心和标签传播顺序以及新提出的标签选择策略的优势，分别比较 3 个多标签传播算法：COPRA 算法，*K*-COPRA 算法和 MOLPA 算法。其中 *K*-COPRA 算法代表利用 *K*-核分解方法确定社区初始核心和标签传播顺序的 COPRA 算法，节点的标签个数仍由参数确定。比较 COPRA 算法和 *K*-COPRA 算法，可以验证 *K*-核分解方法的有效性，进一步比较 *K*-COPRA 算法和 MOLPA 算法可以评估新的标签选择策略的效果。实验结果如表 4 所示，其中 v 值均取算法最优值。

表 4 3 种多标签传播算法在真实网络中的模块度对比

Tab. 4 Comparison of modularity of three methods based on muti-label propagation in real networks

网络	COPRA		K-COPRA		MOLPA
	v	*EQ*	v	*EQ*	*EQ*
Karate	2	0.559	2	0.619	0.652
Dolphins	3	0.597	2	0.636	0.678
Football	2	0.480	4	0.629	0.664
C. elegans metabolic network	4	0.533	5	0.571	0.619
PGP network	7	0.547	8	0.565	0.634
Collaboration network	8	0.459	8	0.519	0.572
DBLP collaboration network	10	0.432	11	0.475	0.514

从表 4 中可以看出，在 Karate 网络上，COPRA 和 *K*-COPRA 算法取得最优值时参数都为 2，但是 *K*-COPRA 的社区模块度更高，说明 *K*-COPRA 算法提高了社区发现的质量，同时也验证了 *K*-核分解方法确定社区初始核心和标签传播顺序的有效性；通过对比 *K*-COPRA 算法和 MOLPA 算法，发现 MOLPA 算法的社区模块性能更好，这表明 MOLPA 算法的标签选择策略有效。其他两个网络数据上的实验结果也证实了 MOLPA 算法在真实网络上也能取得不错的效果。

(2)不同的重叠社区发现算法对比

在真实网络上测试 CPM、BOCLP 和 MOLPA 这 3 种重叠社区发现算法的表现。如表 5 所示，无论在规模比较小的空手道数据集上还是规模较大的足球比赛数据集上，MOLPA 算法得到的社区发现的模块度都高于其他 3 个算法，表明其社区划分的质量较好。

表 5 3 种重叠社区发现算法在真实网络上的模块度对比

Tab. 5 Comparison of modularity of three overlapping community detection algorithms in real networks

网络	CPM	BOCLP	MOLPA
Karate	0.560	0.581	0.652
Dolphins	0.607	0.633	0.678
Football	0.650	0.652	0.664
C. elegans metabolic network	0.581	0.569	0.619
PGP network	0.592	0.573	0.634
Collaboration network	0.563	0.545	0.572
DBLP collaboration network	0.493	0.508	0.514

3 结束语

本文介绍了一种基于多标签传播的重叠社区发现算法 MOLPA，并与传统的团过滤算法 CPM、标

签传播算法 COPRA 及其改进算法 BOCLP 进行了对比。CPM 算法是一种经典的重叠社区发现算法,但和标签传播算法相比效率不够高。COPRA 是标签传播算法的核心算法之一,但存在随机性强,社区发现结果不稳定等缺陷;BOCLP 对 COPRA 进行了改进一定程度上提高了稳定性,但划分结果仍受到参数设置的影响。本文提出的 MOLPA 算法首先利用 K-核分解法找出网络社区核心节点,给这些节点赋予唯一的标签,并且设置合理的标签迭代顺序,避免了传播过程中的随机性,提高了算法的稳定性;在标签选择的时候由标签种类来决定节点的标签个数而不是人为设置参数,避免了结果受到参数的影响。实验表明 MOLPA 算法能在很大程度上减少无意义社区的生成,改善传统标签传播算法的不稳定性问题,提高重叠社区发现的质量。

参考文献:

[1] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state-of-the-art and comparative study [J]. ACM Computing Surveys, 2011,45(4):115-123.

[2] Coscia M, Giannotti F, Pedreschi D. A classification for community discovery methods in complex networks [J]. Statistical Analysis & Data Mining, 2011,4(5):512-546.

[3] Yang J, Mcauley J, Leskovec J. Community detection in networks with node attributes[C]// 2013 IEEE 13th International Conference on Data Mining. [S. l.]: IEEE Computer Society, 2014:1151-1156.

[4] 李亚芳,贾彩燕,于剑, 等. 一种新的社区/动态社区优化方法 [J]. 数据采集与处理, 2015,30(6):1215-1224.
Li Yafang, Jia CaiYan, Yu Jian, et al. Novel community/dynamic community optimization algorithm[J]. Journal of Data Acquisition and Processing, 2015,30(6):1215-1224.

[5] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell Labs Technical Journal, 1970,49(2):291-307.

[6] Barnes E R. An algorithm for partitioning the nodes of a graph [J]. SIAM Journal on Algebraic & Discrete Methods, 1981,3(4):303-304.

[7] Jiang Y, Jia C, Yu J. An efficient community detection algorithm using greedy surprise maximization [J]. Journal of Physics A Mathematical & Theoretical, 2014,47(16):1644-1649.

[8] Jia H, Ding S, Xu X, et al. The latest research progress on spectral clustering [J]. Neural Computing & Applications, 2014,24(7/8):1477-1486.

[9] 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.

[10] 张海燕, 梁循, 周小平. 针对有向图的局部扩展的重叠社区发现算法 [J]. 数据采集与处理, 2015,30(6):683-693.
Zhang Haiyan, Liang Xun, Zhou Xiaoping. Overlapping commuity detection from local extension in directed graphs[J]. Journal of Data Acquisition and Processing, 2015,30(6):683-693.

[11] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics Theory & Experiment, 2008,2008(10):155-168.

[12] Clauset A, Newman M E, Moore C. Finding community structure in very large networks [J]. Physical Review E, 2005,70(6 Pt 2):264-277.

[13] Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks[J]. Plos One, 2011, 6(4):336-338.

[14] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007,76(3 Pt 2):036106.

[15] 骆志刚, 丁凡, 蒋晓舟, 等.复杂网络社团发现算法研究新进展[J]. 国防科技大学学报,2011,33(1):47-52.
Luo Zhigang, Ding Fan, Jiang Xiaozhou, et al. New progress on community detectionin complex networks [J]. Journal of National University of Defense Technology, 2011, 33(1):47-52.

[16] Chakraborty T. Leveraging disjoint communities for detecting overlapping community structure[J]. Journal of Statistical Mechanics Theory & Experiment, 2015(5):P05017.

[17] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043):814-818.

[18] Leskovec J, Lang K J, Dasgupta A, et al. Statistical properties of community structure in large social and information net-

works[C]// The International Conference of World Wide Web. 2008:695-704.

[19] Kim P, Kim S. Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering [J]. Physica A Statistical Mechanics & Its Applications, 2015,417(C):46-56.

[20] Abrahao B, Soundarajan S, Hopcroft J, et al. A separability framework for analyzing community structure [J]. ACM Transactions on Knowledge Discovery from Data, 2014,8(1):101-129.

[21] 王庚, 宋传超, 盛玉晓, 等. 基于标签传播的社区挖掘算法研究综述 [J]. 计算机技术与发展, 2013(12):69-73.
Wang Geng, Song Chuanchao, Sheng Yuxiao, et al. Research summary on communities mining algorithm based on label propagation [J]. Computer Technology and Development, 2013(12):69-73.

[22] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2009,12(10): 2011-2024.

[23] Wu Z H, Lin Y F, Gregory S, et al. Balanced multi-label propagation for overlapping community detection in social networks [J]. Journal of Computer Science and Technology, 2012,27(3):468-479.

[24] 王庚. 社会网络中基于标签传播的重叠社区挖掘研究[D]. 济南:山东建筑大学, 2013.
Wang Geng. Research to stable detecting overlapping communities by label propagation on social networks[D]. Jinan: Shandong Jianzhu University, 2013.

[25] Sun H L, Huang J B, Tian Y Q, et al. Detecting overlapping communities in networks via dominant label propagation [J]. Chinese Physics B, 2015, 24(1):551-559.

[26] Liu K, Huang J, Sun H, et al. Label propagation based evolutionary clustering for detecting overlapping and non-overlapping communities in dynamic networks [J]. Knowledge-Based Systems, 2015, 89(C):487-496.

[27] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks [J]. Nat Phys, 2010, 6 (11):888-893.

[28] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms [J]. Physical Review E, 2008, 78(4):561-570.

[29] Wang Z, Zhao Y, Xi J, et al. Fast ranking influential nodes in complex networks using a k-shell iteration factor [J]. Physica A Statistical Mechanics & Its Applications, 2016(461):171-181.

[30] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks [J]. New Journal of Physics, 2008, 11(3):19-44.

[31] Shen Huawei. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and Its Applications, 2009, 388(8):1706-1712.

[32] Bron C, Kerbosch J. Finding all cliques of an undirected graph (algorithm 457)[J]. Communications of the ACM, 1973, 16 (9):575-576.

作者简介:

杜长江(1991-),男,硕士研究生,研究方向:复杂网络社区发现,E-mail:1543258253@qq.com。

王志晓(1979-),男,博士,副教授,研究方向:社交网络分析,E-mail:zhxwang@cumt.edu.cn。

邢贞明(1991-),男,硕士研究生,研究方向:复杂网络社区发现。

(编辑:夏道家)