doi:10.11887/j.cn.201901025 http://journal.nudt.edu.cn

采用 PageRank 和节点聚类系数的标签传播重叠社区发现算法*

马 健,刘 峰,李红辉,樊建平
(北京交通大学 计算机与信息技术学院,北京 100044)

摘 要:基于标签传播的社区发现算法可以检测出复杂网络的重叠社区结构,因此提出了一种基于PageRank和节点聚类系数的重叠社区发现算法。该算法使用PageRank算法对节点的影响力进行排序,可以稳定社区发现结果,节点的聚类系数是一个与节点相关的值,使用节点聚类系数修改算法的参数并限制每个节点拥有最多标签的数量值,可以提高社区挖掘的质量。在人工网络和真实世界的网络上测试,实验验证了该算法能够有效地检测出重叠社区,并具有可接受的时间效率和算法复杂度。

关键词:社区发现;重叠社区;标签传播;聚类系数;PageRank算法;节点影响力
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2019)01-183-08

Overlapping community detection algorithm by label propagation using PageRank and node clustering coefficients

MA Jian, LIU Feng, LI Honghui, FAN Jianping
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Considering the fact that the community detection algorithm based on label propagation can detect overlapping community structures of complex networks, an overlapping community detection algorithm COPRAPC (community overlap propagation algorithm based on PageRank and clustering coefficient) was proposed. The algorithm used PageRank algorithm to rank the influence of nodes, which can stabilize the community finding results. The parameter of node clustering coefficient was a node-related parameter, which can be used to modify the parameters of the algorithm and limit the maximum number of labels each node, so as to improve the quality of community mining. Experiments on artificial networks and real-world networks show that the algorithm can effectively detect overlapping communities, and the algorithm has acceptable time efficiency and algorithm complexity.

Keywords: community detection; overlapping community; label propagation; clustering coefficient; PageRank algorithm; node influence

社区结构(community)是复杂网络所呈现的重要特征之一,从某种意义上说,整个网络由若干社区构成,具有社区内部节点之间连接比较密集,不同社区节点之间连接比较稀疏的特点[1]。复杂网络社区发现问题就是要揭示出复杂网络中存在的各个社区。从社区之间是否重叠的角度来看,社区结构是可以重叠的。例如人类社会关系网中的一个人可以拥有多个朋友圈,Internet中,一个路由器可以连接不同的局域网,在神经网络中,一个神经元可以属于不同的神经系统。

重叠社区的发现算法可以分为:基于派系过滤算法(Clique Percolation Method, CPM)的重叠社区发现算法,它将网络看作是完全连通子图(clique)的集合[2-3]。基于优化函数的重叠社区发现算法[4-5],其中,Lancichinetti 等[4]提出了LFM(Local Fitness Method)算法,该算法既可以发现层次社区又可以发现重叠社区。基于边的重叠社区发现算法[6-7],基于标签传播的社区发现算法,Gregory 等[8]提出了重叠社区发现算法(Community Overlap PRopagation Algorithm, COPRA)。Wu 等[9]提出了基于均衡多标签传播的重叠社团发现算法(Balanced Multi-Label Propagation Algorithm, BMLPA)算法,定义了粗糙核的概念对节点的标签进行初始化,通过更改标签更新时的参数改进算法。张昌理等[10]提出了多标签传播重叠社区发现算法COPRA-EP,算法利用节点信息熵和节点的局部相关性进行研究。Cui 等[11]从网络中发现所有最大子图,然后通过两个相邻最大子图的聚类系数来合并它们。Mohan 等[12]提出了一个分布式和可扩展的模型

* **收稿日期**:2018-02-22
基金项目:国家863计划资助项目(2015AA043701)
作者简介:马健(1985—),女,山西长治人,博士研究生,E-mail:13112083@bjtu.edu.cn;
樊建平(通信作者),男,教授,博士,博士生导师,E-mail:jp.fan@siat.ac.cn

来最大化信息的扩散,通过将图划分为社区并寻找每个社区中有影响的节点来检测社区,模型使用 PageRank 算法在每个社区中寻找有影响的节点。孙道平等[13]提出了一种算法计算网络节点的聚类系数,并选择具有最大聚类系数的节点来更新其在传播过程中的标签。Chen 等[14]提出了基于节点层次和标签传播增益的重叠社区发现算法,算法提出了新的多标签更新规则,设计节点之间标签传播增益得到重叠社区。Xie 等[15]提出了一种 SLPA 算法。

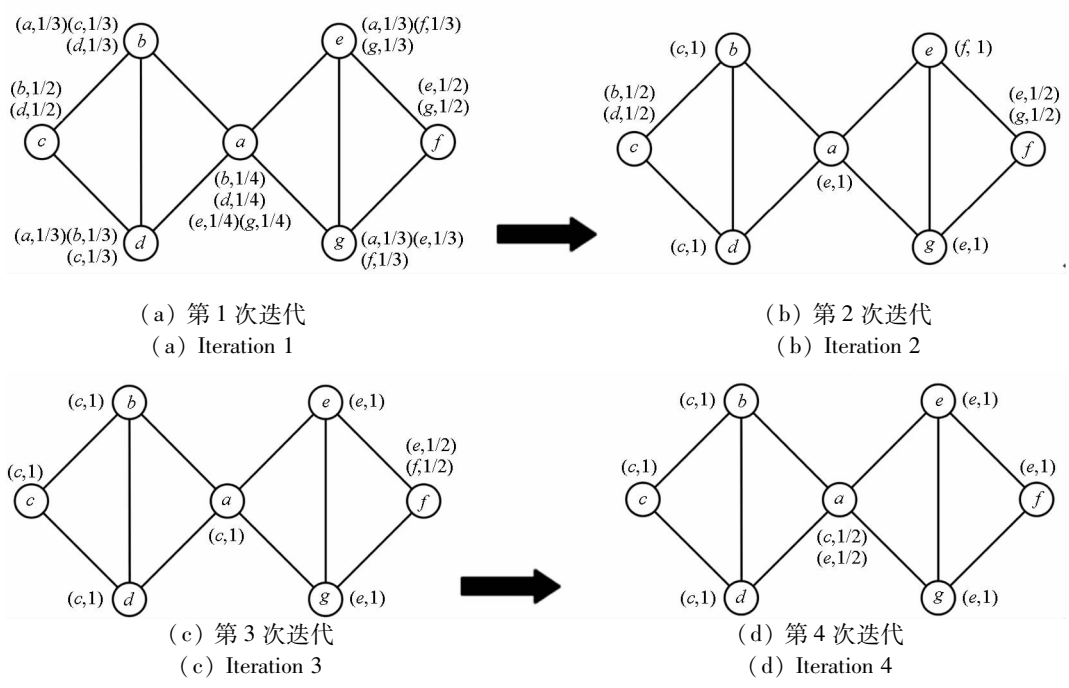
本文提出了一种基于 PageRank 和节点聚类系数的重叠标签传播算法(Community Overlap PRopagation Algorithm based on PageRank and Clustering coefficient,COPRAPC)。该算法在传播节点的社区标签时按节点影响力的大小进行排序,并且根据节点的聚类系数决定传播的阈值,最后通过一系列实验进行验证,实验结果表明所提出的重叠社区检测算法是可行且有效的。

1 COPRA 算法

1.1 COPRA 算法描述

COPRA 算法(见图 1)在标签传播算法(Label Propagation Algorithm, LPA)算法基础上改进解决重叠社区发现的问题,每个节点拥有多个标签(label)和隶属度(belonging coefficients),所有隶属度的和等于 1,标签采用同步更新的方式,用每个节点的邻居节点的标签更新该节点,属于相同社区标签的隶属度相加并标准化,如式(1)所示,隶属度小于 $1/v$ 的标签将从标签列表中删除,阈值 v 用来控制一个节点所属的社区数。COPRA 标签传播算法的时间复杂度为 $O(vm\log(vm/n))$,当 v 取较小值时,整个算法复杂度为线性时间复杂度。

$$b_t(c,x) = \frac{\sum_{y\in N(x)} b_{t-1}(c,y)}{|N(x)|} \tag{1}$$



(a) 第 1 次迭代
(a) Iteration 1

(b) 第 2 次迭代
(b) Iteration 2

(c) 第 3 次迭代
(c) Iteration 3

(d) 第 4 次迭代
(d) Iteration 4

图 1 COPRA 算法

Fig. 1 COPRA algorithm

1.2 评价函数

Qov 是 Nicosia 等将 Q 函数扩展提出的能够评价重叠社区的评价函数[16]。*Qov* 越大,社区结构越明显。

$$Qov = \frac{1}{m}\sum_{c\in C}\sum_{i,j\in V}\left[\delta(C_i,C_j,C)A_{ij} - \frac{\beta^{\text{out}}_{l(i,j),c}k^{\text{out}}_i\beta^{\text{in}}_{l(i,j),c}k^{\text{in}}_j}{m}\right] \tag{2}$$

其中,

$$\delta(C_i,C_j,C) = F(\alpha_{i,c},\alpha_{j,c})$$

$$F(\alpha_{i,c},\alpha_{j,c}) = \frac{1}{(1+e^{-f(\alpha_i,c)})(1+e^{-f(\alpha_j,c)})}$$

$$\beta^{\text{out}}_{l(i,j),C} = \frac{\sum_{j\in V}F(\alpha_{i,c},\alpha_{j,c})}{|V|}$$

$$\beta^{\text{in}}_{l(i,j),C} = \frac{\sum_{i\in V}F(\alpha_{i,c},\alpha_{j,c})}{|V|}$$

$$f(x) = 2px - p,\ p = 30$$

EQ 是评估重叠社区的另一个度量标准,*EQ* 值越大,社区的结构越明显。表达式如下[3]:

$$EQ = \frac{1}{2m}\sum_{i}\sum_{v,w\in C_i}\frac{1}{O_v O_w}\left[A_{vw} - \frac{d_v d_w}{2m}\right] \quad (3)$$

Lancichinetti 等将标准互信息(Normalized Mutual Information, NMI)扩展为能够评价重叠社区的一种评价指标[2]。

$$NMI(x|y) = 1 - \frac{1}{2}[H(x|y)_{\text{norm}} + H(y|x)_{\text{norm}}] \quad (4)$$

2　COPRAPC 算法

2.1　算法思想及描述

COPRAPC 算法和其他重叠标签传播算法的不同之处在于算法采用不同的阈值来限制每个节点拥有最多标签的数量值,每个节点可能出现在多个社区中,社区边缘节点拥有较小聚类系数一般是重叠节点,社区内部节点拥有大聚类系数往往是非重叠节点,社区边缘节点比社区内部属于更多的社区。小的聚类系数做阈值可以保留节点更多的标签,而大的聚类系数可以删除节点更多的标签,重叠的节点可以比非重叠的节点保留更多的标签。此外,复杂网络的小世界特性决定了节点具有较高聚类系数,从而避免了传播过程中节点的标签对过多的问题。

COPRAPC 算法采用同步更新策略。对于同步策略来说,可以比异步方式提供更稳定的结果,但是同步方案需要更多的迭代次数。

在 COPRA 算法中,每个节点都有一个标签,在传播过程中,节点标签在没有特定序列的情况下更新。这种随机策略导致运行结果不稳定。如图 1(c)中所示的 COPRA 算法中,在这次迭代之后,节点 a、b、c 和 d 已经形成了一个社区。如果算法选择节点 e 或 g,用节点 a 的标签替换它的标签,然后更新节点 f 的标签,那么,所有节点都使用 c 作为它们的标签,所有的节点将被划分为一个社区,结果导致整个网络在一个社区中。如果首先更新节点 f 的标签,将得到正确的分区结果。

图 2 显示了两个非重叠社区 C1 和 C2 迭代之后,如果节点 d, e, f 被分配到社区 C2,此时更新节点 g,将得到正确的社区分区,但如果更新节点 b 或 c 的标签并选择节点 d 和 e 作为它的标签,所有节点将拥有相同的标签,结果是所有节点被划分到相同的社区。

上述两个例子说明 COPRA 算法对节点更新序列很敏感,不是所有这些节点都具有相同的优先级。在社区之间的节点的影响力总是比社区内部的节点要低。PageRank 值反映了这个特性,本文使用排名方法 PageRank 算法[17]对节点进行排序。首先,将无向图转换为有向图,将无向图中的边转换为具有两个相反方向的弧。PageRank 的较小值往往位于社区中心,PageRank 值小的节点优先于 PageRank 值较大的节点,即

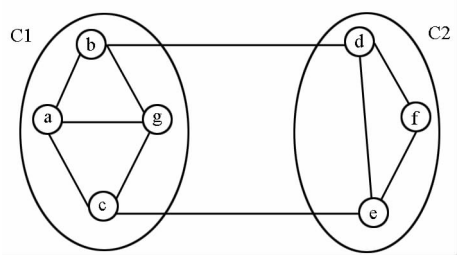


图 2　COPRAPC 算法实例

Fig. 2　Example of COPRAPC propagation

$$PR(v) = \frac{(1-\alpha)}{N} + \alpha\sum_{v'\in N(v)}\frac{PR(v')}{degress(v')} \quad (5)$$

其中,α 参数为 0.85,v' 是节点 v 的邻居,$degress(v')$为节点 v'的度。

节点的 PageRank 值:$PR(a) = 0.146\ 6$, $PR(b) = 0.147\ 6$, $PR(c) = 0.147\ 6$, $PR(d) = 0.152\ 0$, $PR(e) = 0.152\ 0$, $PR(f) = 0.107\ 6$, $PR(g) = 0.146\ 6$。节点在以下序列中更新它们的标签:f、a、g、b、c、d 和 e,可以很容易地得到正确的结果。

2.2　修改传播门限参数

在 COPRA 算法中,使用标签列表(c,b)表示每个节点拥有的标签,其中 c 表示一个标签(label),b 表示隶属度。如果 $b<1/v$,从节点标签列表中删除(c,b)。在重叠的社区中,每个节点可能属于多个社区。参数 v 控制重叠的社区数目。也就是说,参数 v 限制节点所属的社区最大数量。如果 v 太大,则可能导致网络中节点的标签对过少,节点容易被划分到一个社区,如果 v 太小,则会导致网络中节点的标签对过多。参数 v 的选择能够影响标签传播算法重叠社区发现的准确性,由于每个节点所属社区数目是不相同的,因此,本文提出了一种新的标签传播算法,根据不同的节点选择不同的参数。

在网络中,节点 v_i 的邻居节点之间实际存在的边数和可能存在最多边数的比值为节点 v_i 的聚类系数[18],每个节点 v_i 邻居节点之间可能最多是 $k_i\times(k_i-1)/2$ 条边。节点 v_i 的聚类系数如式(6)所示。

$$C_i = \frac{2E_i}{k_i(k_i-1)} \quad (6)$$

其中，E_i 表示连接到节点 v_i 的每个相邻节点的边数，k_i 表示节点 v_i 所有相邻节点的个数。聚类系数 C_i 总是在 0 至 1 之间。

COPRAPC 算法允许社区重叠，每个节点可能出现在多个社区中。社区边缘节点总是重叠节点，社区中间节点往往是非重叠节点，重叠节点聚类系数往往小于非重叠节点，重叠节点的聚类系数较小通常处在社区的边缘，非重叠社区的节点都在社区中间聚类系数较大，换句话说，一个社区边缘节点拥有小聚类系数可能比社区中间拥有大聚类系数属于更多的社区。更高的阈值可以删除节点更多的标签，重叠的节点可以比非重叠的节点保留更多的标签。

v 是一个与节点无关的参数。相反，聚类系数参数是一个与节点相关的参数，算法设置参数 μ，当节点的聚类系数大于 μ，阈值设为节点的聚类系数，反之，使用先前的值 $1/v$。这样做的目的是聚类系数大的节点一般在社区内部，往往不是重叠节点，这样的节点设置的门限高，可以去除较多的标签，而聚类系数小的节点一般在社区的边缘，容易成为重叠节点，设置的门限较低可以增加标签的数量。但对于聚类系数非常小的节点，阈值过低容易产生空的标签集合，这时就用原来的门限值 $1/v$。

COPRAPC 算法描述：

1）初始时，给每个节点赋予唯一的社区标签 $(c_x,1)$。

2）对网络中的节点按其 PageRank 值排序。

3）每个节点 x 通过最大邻居的数量来更新它的标签。属于相同社区的标签的隶属度相加并标准化，为了避免最后所有节点都划分到同一个社区，COPRAPC 算法使用节点的聚类系数来限制每个节点拥有最多标签的数量值。如果节点的聚类系数小于参数 μ，仍然使用原来的 $1/v$，如果一个节点的所有标签隶属度都小于传播参数，那么随机选择一个最大值邻居的标签作为该节点的标签。最后，当每个标签包含的节点数不变，算法迭代结束，否则，重复步骤 3。

4）将所有标签相同的节点划分为同一个社区。

5）将得到的社区进行删除其他社区的子集，不相连的社区分裂成更小的社区。

COPRAPC 算法的形式化语言如算法 1 和算法 2 所示。

算法 1　计算节点的聚类系数

Alg. 1　Calculate the value of node clustering coefficient

输入：网络的邻接矩阵 $\boldsymbol{A}$
存储节点 x 的邻居集合 *neighbours*
存储节点 x 的度数 *degress*
输出：节点的聚类系数集合 *cluster_coefficient*

1. 搜索节点 x 的邻居节点集合，加入到 *neighbours* 数组中
2. $count = 0$;
3. for i 是节点 x 的邻居
4. for j 是节点 x 的邻居
5. if 节点 i 和节点 j 之间存在边
6. $count = count + 1$
7. $cluster\_coefficient[x] = \dfrac{count}{degress[x] * (degress[x] - 1)}$
8. 返回 *cluster_coefficient*

算法 2　COPRAPC 算法

Alg. 2　COPRAPC algorithm

输入：网络 G
存储节点的 PageRank 值得到数组 *PR*
输出：网络的社区划分

1. 每个节点赋予唯一的社区标签
2. 每个节点新的社区标签设为空集
3. 将节点按 *PR* 值从小到大对其进行排序
4. for x 是 *PR* 中的节点
5. if $cluster\_coefficient[x] > \mu$
6. 用 $cluster\_coefficient[x]$ 代替传播门限 $1/v$
7. else
8. 用原先的门限值 $1/v$
9. end for
10. if 没有满足传播结束条件
11. 将新的社区标签集合赋给旧的社区标签集合
12. 转到第 4 步
13. end if
14. 删除其他社区子集的社区
15. 得到社区划分结果

2.3　复杂度分析

COPRAC 算法的时间复杂度，假设 n 为网络节点的数量。m 为边的数量，所有节点的聚类系数即网络的聚类系数为 $\bar{c}$，节点的平均度数为 $\bar{d}$。

1）初始化标签的时间复杂度为 $O(n)$。

2）设计算单个节点 PageRank 值的迭代次数为 t，矩阵相乘时间复杂度为 $O(n^2)$，所以计算单个节点的 PageRank 值的时间复杂度为 $O(n^2 \times t)$，计算网络中所有节点的 PageRank 值的时间复

杂度为 $O(n^3 \times t)$,根据节点排序更新序列的时间复杂度为 $O(n)$。

3)计算单个节点聚类系数的时间复杂度为 $O(\overline{d}^2)$,计算网络中所有节点的聚类系数需要的时间复杂度为 $O(n \times \overline{d}^2)$。

4)标签更新的时间复杂度为 $O(m/\bar{c}\log[m/\bar{c} \times n)]$。

5)迭代时,将节点加入社区 c,并且更新社区 c 的节点,所需时间复杂度 $O(n/\bar{c}^3)$。

6)社区的划分需要的时间复杂度为 $O[(m+n)/\bar{c}]$。

3　实验分析

为了验证所提出的 COPRAPC 算法,将它与几个重叠社区的发现算法进行比较:CPM, LFM, BMLPA 和 COPRA。

Lancichinetti 的 NMI 已被广泛用于发现重叠社区,因此本文在实验中采用扩展的 NMI 作为度量标准。实验独立运行 COPRA 算法 20 次实验数据,得到平均结果。COPRAPC 算法由于标签选择的随机性,本文独立运行了 10 次数据集取平均结果。CPM 算法,$k=4$, BMPLA 算法,$p=0.75$。

3.1　LFR 基准网络

LFR 基准网络,该网络由 Lancichinetti 等提出,是一类更接近真实网络的人工网络,该网络节点及社区模型度呈幂律分布,具有真实世界的网络特性。N 表示节点数目,d 表示节点平均度数,k 表示节点最大度,min$c$ 表示最小社区规模,max$c$ 表示最大社区规模,$t1$ 表示节点度的幂率分布指数,$t2$ 表示社区规模的幂率分布指数,混合参数 mu,o_n 表示重叠节点的个数,o_m 表示重叠节点所属的社区个数,表 1 和表 2 为 LFR 基准网络的参数。表 1 中的社区分别表示稠密小社区(Dense Small, DS),稠密大社区(Dense Large, DL),稀疏小社区(Sparse Small, SS),稀疏大社区(Sparse Large, SL)。表 1 中实验节点的聚类系数都较大,因此设置 $\mu=0$。

表 1　重叠社区的 LFR 基准网络参数

Tab. 1　LFR network parameters of overlapping community

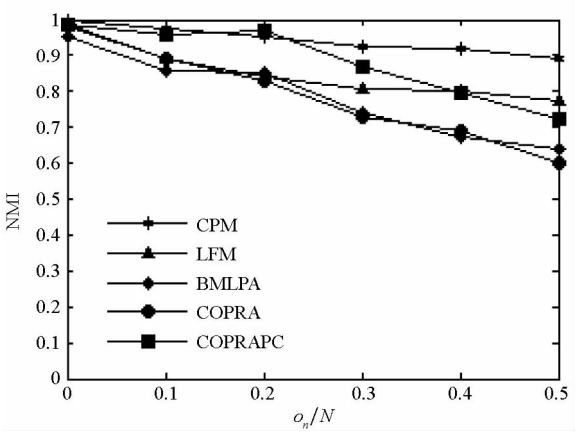
参数	DS	DL	SS	SL
N	1000	1000	1000	1000
k	20	20	20	20
$t1$	2	2	2	2
$t2$	1	1	1	1
min$c$	10	20	10	20
max$c$	50	100	50	100
$k_{\max}$	50	50	50	50
o_m	2	2	2	2
o_n/N	0 ~ 0.5	0 ~ 0.5	0 ~ 0.5	0 ~ 0.5
mu	0.1	0.1	0.3	0.3

表 2　非重叠社区的 LFR 基准网络参数

Tab. 2　LFR network parameters of non-overlapping community

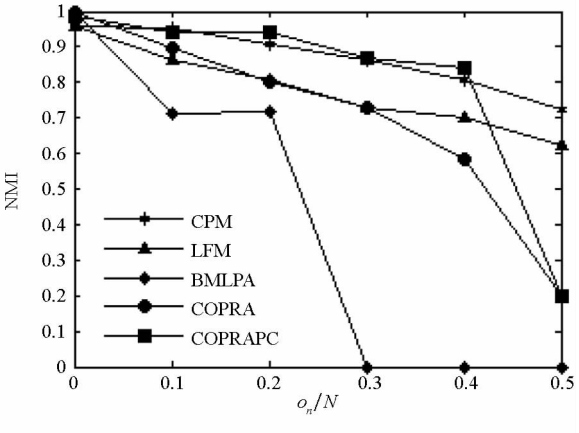
参数	Network a	Network b
N	1000	1000
k	20	20
$t1$	2	2
$t2$	1	1
min$c$	10	20
max$c$	50	100
$k_{\max}$	50	50
o_m	1	1
o_n/N	0	0
mu	0 ~ 0.5	0 ~ 0.5

图 3 显示该算法在实验中表现良好。随着参数 o_n 的增大,重叠节点的数量越来越多。COPRAPC 性能较优。



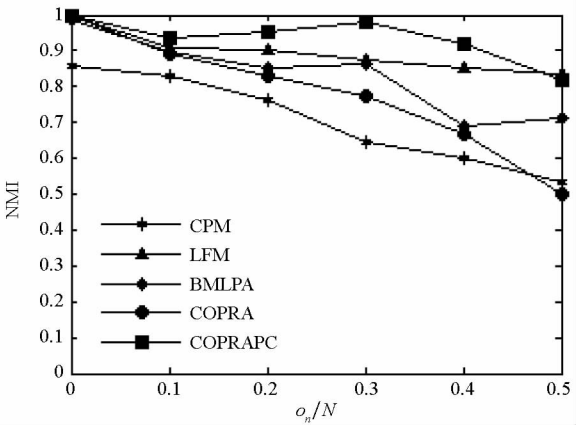
(a) DS 网络的 NMI 值

(a) NMI identified by DS networks



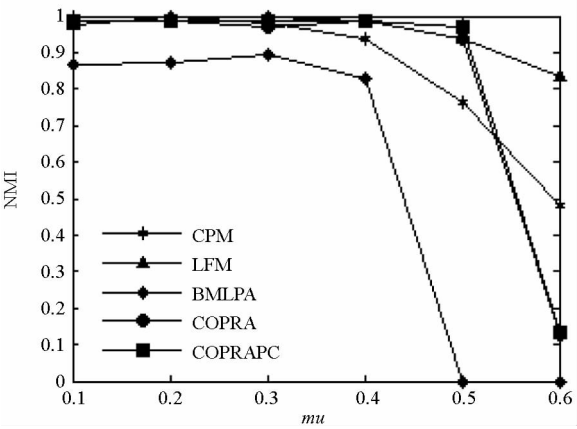
(b) SS 网络的 NMI 值

(b) NMI identified by SS networks



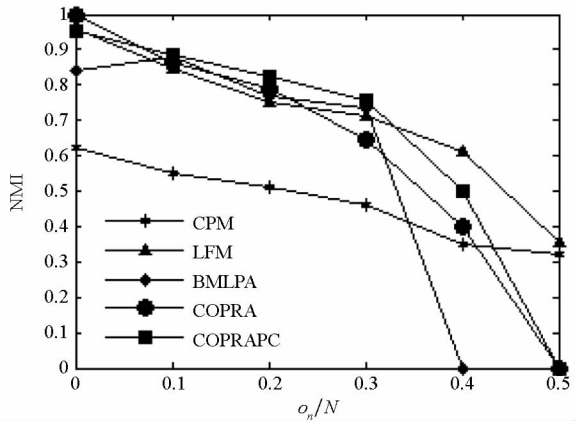
(c) DL 网络的 NMI 值

(c) NMI identified by DL networks



(a) Network a 的 NMI 值

(a) NMI identified by network a

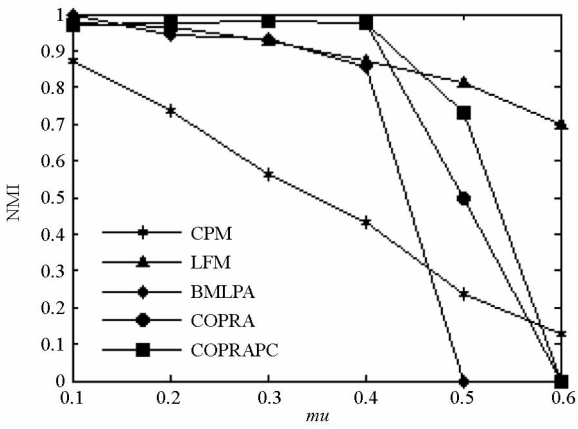


(d) SL 网络的 NMI 值

(d) NMI identified by SL networks

图3 CPM, LFM, COPRA, BMLPA and COPRAPC 算法的 NMI 对比结果

Fig. 3 The NMI identified by CPM, LFM, COPRA, BMLPA and COPRAPC



(b) Network b 的 NMI 值

(b) NMI identified by network b

图4 CPM, LFM, COPRA, BMLPA and COPRAPC 算法的 NMI 对比结果

Fig. 4 The NMI identified by CPM, LFM, COPRA, BMLPA and COPRAPC

表2 中的实验数据集节点的聚类系数较小，因此设 $\mu = 0.1$。

上述是一个具有非重叠节点的社区网络。随着参数 mu 的增加，当值 $mu = 0.4$ 时社区的结构变得模糊，图 4 显示该算法在这些网络中表现良好。

3.2 真实世界网络

本文在几个真实世界复杂网络上对算法进行测试，实验采用的数据分别是：Newman 提供的空手道俱乐部网络(Zachary's club network)、海豚网络(dolphin social network)、美国大学足球联盟网络(American college football)、爵士音乐网络(Jazz music)、政治书籍(political books)和邮件网络(the E-mail network of human interactions)。表3 为实验用到的真实网络参数。实验中，设置参数 $\mu = 0.2$。表4 和表5 分别给出了解情况种算法在个真实世界网络中的 Qov 和 EQ 值。

表3 真实世界网络参数

Tab. 3 Parameters of real-world networks

序号	网络	节点数	边数	社区数
1	Zachary's club	34	78	2
2	dolphins	62	159	2
3	football	115	613	12
4	Jazz	198	2742	—
5	political books	105	441	3
6	E-mail	1133	5451	—

在基准网络和真实网络的实验中，算法的社区发现结果是稳定的，基于此思想的算法相比其他标签传播重叠社区发现算法有明显的提高，和

其他重叠算法相比也有较好的挖掘结果,这也说明了基于 PageRank 和节点聚类系数的思路是可行的。

表 4　CPM LFM, COPRA, BMLPA and COPRAPC 算法的 *Qov* 值

Tab. 4　The *Qov* values identified by CPM, LFM, COPRA, and COPRAPC

Qov	CPM	LFM	COPRA	BMLPA	COPRAPC
Zachary's club	0.181 4	0.593 9	0.487 3	0.700 0	0.740 9
dolphins	0.355 8	0.686 8	0.702 1	0.770 0	0.604 5
football	0.600 1	0.506 5	0.623 0	0.657 4	0.671 6
Jazz	0.615 2	0.699 6	0.727 5	0.705 8	0.727 7
political books	0.677 9	0.833 6	0.780 2	0.758 3	0.863 8
E-mail	0.354 1	0.098	0.560 0	0.677 9	0.791 6

表 5　CPM, LFM, COPRA, BMLPA and COPRAC 算法的 *EQ* 值

Tab. 5　The *EQ* values identified by CPM, LFM, COPRA, and COPRAPC

EQ	CPM	LFM	COPRA	BMLPA	COPRAPC
Zachary's club	0.114 7	0.344 7	0.352 3	0.367 8	0.390 0
dolphins	0.288 5	0.389 4	0.505 9	0.420 6	0.497 4
football	0.559 3	0.536 2	0.578 8	0.538 5	0.589 7
Jazz	0.004 3	0.283 6	0.413 7	0.429 7	0.431 9
political books	0.430 8	0.454 1	0.513 5	0.505 4	0.518 9
E-mail	0.264 1	0.212 0	0.061 1	0.042 8	0.036 2

3.3　算法效率比较

改变 LFR 基准程序的参数,可以得到不同规模的人工网络,在这些网络上分别运行社区发现算法,可以得到算法运行时间的一般规律,从而能知道各个算法在效率上的差异。本节使用的 LFR 程序参数 $N = 1000 \sim 10\ 000$, $k = 20$, $k_{\max} = 50$, $c_{\min} = 20$, $c_{\max} = 100$, $t_1 = 2$, $t_2 = 2$, $o_n = 0$, $o_m = 2$。图 5 显示了 COPRA, BMLPA 和 COPRAPC 算法的执行时间,由于 CPM,LFM 算法的运行时间较长,本文不做比较。COPRAPC 算法具有合理的时间复杂度。此外,算法的时间效率和时间复杂度方面,标签传播算法具有明显的优势,本算法的时间在同类算法中也是可接受的。

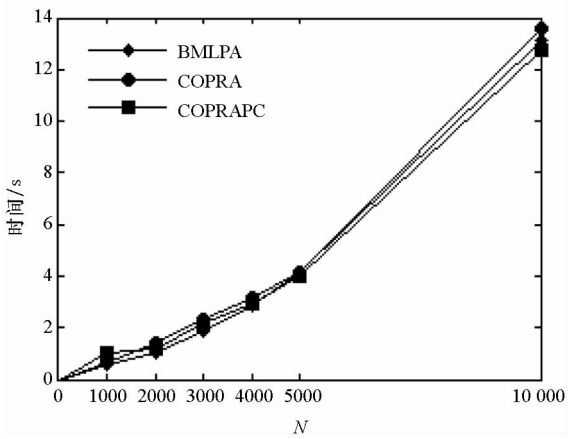


图 5　算法运行效率对比的实验结果

Fig. 5　Experiment results of runtime efficiency

4　结论

本文提出了一种改进的重叠社区发现算法,即基于节点的 PageRank 值排序和节点的聚类系数标签传播算法。节点标签的传播不按随机顺序,而是按照指定节点的 PageRank 值的顺序传播,设定节点聚类系数阈值对标签进行更新,进而实现对网络社区结构的划分和网络重叠节点的发现。在许多人工网络和真实世界网络的实验中,基于此思想的算法都取得了相比其他同类算法更好的效果,这也说明了该思路的可行性,同时该算法具有稳定的社区发现结果。并且该算法的时间效率和复杂度也在可接受的范围内。

参考文献(References)

[1]　水超, 陈洪辉, 陈涛, 等. 力导向模型的复杂网络社区挖掘算法[J]. 国防科技大学学报, 2014, 36(4): 163 – 168.
SHUI Chao, CHEN Honghui, CHEN Tao, et al. A community detect algorithm on force-directed model [J]. Journal of National University of Defense Technology, 2014, 36(4): 163 – 168. (in Chinese)

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

[3]　Shen H W, Cheng X Q, Cai K, et al. Detect overlapping and hierarchical community structure in networks [J]. Physica A: Statistical Mechanics & Its Applications, 2009, 388 (8): 1706 – 1712.

[4]　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.

[5]　张兴义, 郑雯, 王从涛, 等. 基于单步添加团的重叠社团检测算法[J]. 华南理工大学学报 (自然科学版), 2016, 44(9): 24 – 31.
ZHANG Xingyi, ZHENG Wen, WANG Congtao, et al. An

overlapping community detection algorithm based on addtion of a clique at each step [J]. Journal of South China University of Technology (Natural Science Edition), 2016, 44(9): 24 – 31. (in Chinese)

[6] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks [J]. Nature, 2010, 466(7307): 761 – 764.

[7] Jin D, Gabrys B, Dang J W. Combined node and link partitions method for finding overlapping communities in complex networks [J]. Scientific Reports, 2015, 5: 8600.

[8] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2010, 12(10): 103018.

[9] 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.

[10] 张昌理, 王一蕾, 吴英杰, 等. 基于信息熵和局部相关性的多标签传播重叠社区发现算法 [J]. 小型微型计算机系统, 2016, 37(8): 1645 – 1650.
ZHANG Changli, WANG Yilei, WU Yingjie, et al. Muti-label propagation algorithm for overlapping community discovery based on information entropy and local correlation[J]. Journal of Chinese Computer System, 2016, 37(8): 1645 – 1650. (in Chinese)

[11] Cui Y Z, Wang X Y, Li J Q. Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient [J]. Physica A: Statistical Mechanics & Its Applications, 2014, 405: 85 – 91.

[12] Mohan A, Kunnakadan S, Neelakantan B, et al. A scalable model for efficient information diffusion in large real world networks [C]//Proceedings of International Conference on Next Generation Intelligent Systems, Kottayam, 2017.

[13] 孙道平, 高原, 谢隽, 等. 一种用于中药方剂网络重叠社区发现的改进 COPRA 算法[J]. 南京大学学报(自然科学), 2013, 49(4): 483 – 490.
SUN Daoping, GAO Yuan, XIE Jun, et al. An improved COPRA algorithm applied to traditional Chinese medicine formula network [J]. Journal of Nanjing University (Natural Sciences), 2013, 49(4): 483 – 490. (in Chinese)

[14] 陈羽中, 施松, 陈国龙, 等. 基于节点层级与标签传播增益的重叠社区发现[J]. 模式识别与人工智能, 2015, 28(4): 289 – 298.
CHEN Yuzhong, SHI Song, Chen Guolong, et al. Overlapping community discovery based on node hierarchy and label propagation gain[J]. Pattern Recognition & Artificial Intelligence, 2015, 28(4): 289 – 298. (in Chinese)

[15] Xie J R, Szymanski B K. Towards linear time overlapping community detection in social networks [C]//Proceedings of Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2012: 25 – 36.

[16] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities [J]. Journal of Statistical Mechanics: Theory and Experiment, 2009(3): 03024.

[17] Naik A, Maeda H, Kanojia V, et al. Scalable twitter user clustering approach boosted by personalized PageRank [J]. Springer International Publishing AG, 2017: 472 – 485.

[18] Katzir L, Hardiman S J. Estimating clustering coefficients and size of social networks via random walk [C]//Proceedings of the 22nd International Conference on World Wide Web, 2013: 539 – 550.