基于标签传播概率的重叠社区发现算法

刘世超 朱福喜 甘 琳
(武汉大学计算机学院 武汉 430072)

摘 要 发现高质量的社区有助于理解真实的复杂网络,尤其是动态地分析社区重叠结构,对社区管理和演化具有重要意义.文中提出一种基于标签传播概率的LPPB(Label-Propagation-Probability-Based)重叠社区发现算法,该算法首先为每个结点赋予一个独立的标签,然后根据结点的影响力大小将结点进行排序;在标签传播的过程中,综合网络的结构传播特性和结点的属性特征计算标签传播的概率,同时利用结点的历史标签记录修正标签更新结果;最后将传播后具有相同标签的结点划分为同一社区,社区间的重叠结点构成了社区重叠结构.作者在基准数据集和带时间维度的C-DBLP网络上进行实验,结果验证了该算法具有较高的准确性和稳定性,并且通过对重叠结构的动态分析,揭示了社区重叠结点的行为特性和C-DBLP网络处于高"耦合度"的发展趋势.

关键词 重叠社区;标签传播概率;结点影响力;社区演化;社交网络;数据挖掘;社会媒体
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2016.00717

A Label-Propagation-Probability-Based Algorithm for Overlapping Community Detection

LIU Shi-Chao ZHU Fu-Xi GAN Lin
(*School of Computer, Wuhan University, Wuhan* 430072)

Abstract Finding high quality community helps the users to understand the real complex networks, especially the dynamical analysis of overlapping community structure has the vital significance for community management and evolution. This paper proposes a novel overlapping community detection algorithm called Label-Propagation-Probability-Based (LPPB) algorithm. Each node is assigned a unique label and determined the update order according to the value of node influence. Probability of label propagation depends on the structure propagation characteristics of complex networks and properties of the nodes, meanwhile revising the results from using history information of node label in the process of propagation. Finally nodes with the same tag are divided into one community after propagation, and the overlapping community structure consists of nodes which have more than one label. Experiment results from benchmark datasets and C-DBLP network with time dimension illustrate that LPPB is accurate and stable for overlapping community detection. The dynamic analysis of overlapping structure not only reveals the behavior characteristic of the community overlapping nodes, but also proves that C-DBLP network is undergoing the high "coupling" trend.

Keywords overlapping community; label propagation probability; node influence; community evolution; social networks; data mining; social media

收稿日期:2014-07-15;**在线出版日期**:2015-05-17.**本课题得到国家自然科学基金**(61272277)、**中央高校基本科研业务费专项基金**(274742)**资助**.**刘世超**,**男**,1989**年生**,**博士研究生**,**主要研究方向为社会网络分析、数据挖掘与智能计算**. E-mail: nani@whu.edu.cn.
朱福喜(**通信作者**),**男**,1957**年生**,**教授**,**博士生导师**,**主要研究领域为社会标签、数据挖掘与智能计算**. E-mail: fxzhu@whu.edu.cn.
甘 琳,**女**,1989**年生**,**博士研究生**,**主要研究方向为** Web **数据挖掘**.

1 引 言

社区结构是复杂网络的重要特性，在网络中发现社区就是把相似结点划分为一个集合，使得集合内结点之间的相互作用比它们与集合外结点的相互作用更强，即同一社区内部结点间的链接较为稠密，不同社区之间的链接较为稀疏[1]. 近年来社区发现被广泛应用于不同类型的网络，如万维网、社交网络和生物网络等. 现实的网络社区可能并不相互独立，如图 1 所示，网络中有阴影的结点可以同时属于两个社区，即社区是重叠的，这导致了网络结构更加复杂，为社区发现带来了新的挑战.

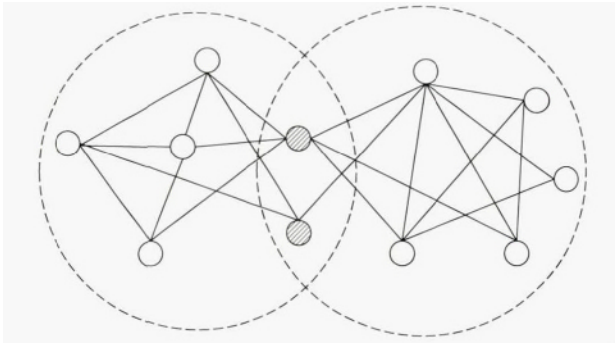

图 1 重叠社区的网络结构

社区的重叠结构由重叠结点组成，社区重叠结点是指网络中可以属于多个社区的结点，重叠结点越多表明社区的重叠度越高. 一方面，重叠结点具有“多面性”，对重叠结点的挖掘能更全面地了解结点的特性. 当前在线社会网络都是以用户为主体，分析用户的行为特性和喜好对商业推荐和用户管理具有重要的意义；另一方面，重叠结点是社区间的桥梁，对社区演化和社区间互通信息起到关键的作用. 当社区间的重叠度较高时，表明共同拥有大部分的成员，那么这些社区即将融合为一个大社区；反之，社区间的重叠度较低，说明网络的社区间相对独立. 因此，对社区重叠结构的研究有助于分析复杂网络的演化趋势.

然而，简单的社区聚类算法不能解决重叠社区和重叠结点的检测问题，尤其在大型网络面前束手无策，如何有效地、精准地挖掘出复杂网络的重叠社区是当前社会网络分析的重点，也是本文研究的核心问题.

本文的主要贡献如下：

(1) 提出了结点间的传播特性度量，以改进传统算法中仅仅基于邻域结构的传播方式.

(2) 在标签的传播过程中，根据结点影响力的大小进行更新，避免了不必要的标签更新和标签“逆流”的现象.

(3) 引入结点的属性特征来影响标签传播的概率，提高了算法的准确性，加快了算法的收敛.

(4) 动态地检测社区的重叠结构，能够分析和预测网络的发展趋势.

2 相关工作

目前社区发现算法大多是基于网络结构划分的，如基于图的划分算法[2-4]、边聚类算法[5-6]、种子扩散的方法[7-8]、随机游走[9]、层次聚类[10-11]和模块度[12]等. 当应用于大型社会网络分析时，几乎都遇到了算法复杂度高的问题，因此 Raghavan 等人[13]提出 RAK 算法，首次将标签传播应用于社区发现，显著地提高了算法的效率. 该算法的基本思想：为每一个结点赋予一个独特的标签，并随传播过程更新标签，计算邻居结点中最多的标签来更新结点的标签，最后把所有相同标签的结点都加入到同一社区. 为提高生成社区的质量，Barber 等人[14-15]提出基于模块度约束的标签传播算法，更加准确地划分社区. Xie 等人[16]结合了跳数衰减的思想扩展了对结点邻居的计算，提高了社区划分结果的精度. 算法[13-16]只允许结点保留一个标签，然而真实的网络社区可能是重叠的，例如一个研究员可以属于多个研究小组，一个学生可以加入不同的兴趣社团. 经典的团渗透算法(CPM)[2]以及边分割算法(Link)[5]虽然能够挖掘重叠社区，但都存在一定的局限性：CPM 依赖于网络中团的分布，而 Link 取决于划分密度 D，算法运行的时间和空间开销使得这两种算法无法应用于大型网络. 于是 Gregory[17]在 2010 年提出了 COPRA 算法，首次运用标签传播方法解决重叠社区挖掘的问题. 该算法每个结点都有一组属于某个社区的标签对(c,b)，c 表示社区名称，b 表示归属系数，任意结点的所有归属系数加起来等于 1. 更新过程中设置一个筛选阈值，阈值的大小取决于结点能够属于的最大社区数，结点的归属标签集中 b 小于阈值的标签对将被删除. 文献[18]对 COPRA 算法的筛选阈值的改进，是通过平衡阈值的选择来提高生成社区的质量.

以上所有算法仅利用结点的邻居结构来更新标签，忽略了结点自身的属性和历史标签信息对传播的影响. 尽管赵卓翔等人[19]引入结点的度来衡量标

签的影响传播概率，但是结点的度只能一定程度地反应其影响力值. 由于历史的标签信息对现在的更新决策有一定的影响作用，Qiang 等人[20]提出的 MLPAO 算法保留结点的历史标签信息，在选择标签时正比于其出现的次数. 然而，历史标签的影响会随着时间衰减，因此，MLPAO 算法将所有标签的影响等同处理，有一定的局限性.

本文总结了上述算法的优缺点，提出了基于标签传播概率(LPPB)的重叠社区发现算法，该算法在 COPRA 的基础上改进了网络的传播特性，根据结点影响力确定更新顺序，同时还综合计算了结点的属性特征和历史标签记录对传播的影响. 本文在第 3 节阐述 LPPB 的主要思想和具体步骤，第 4 节对该算法进行实验验证和分析，并与 COPRA 算法进行比较，第 5 节给出本文的工作总结和未来研究方向.

3　标签传播概率的社区发现算法

3.1　算法主要思想

本文拟从网络的传播特性、结点的更新顺序和结点的属性特征 3 个方面对经典的 COPRA 算法进行改进，如图 2 所示.

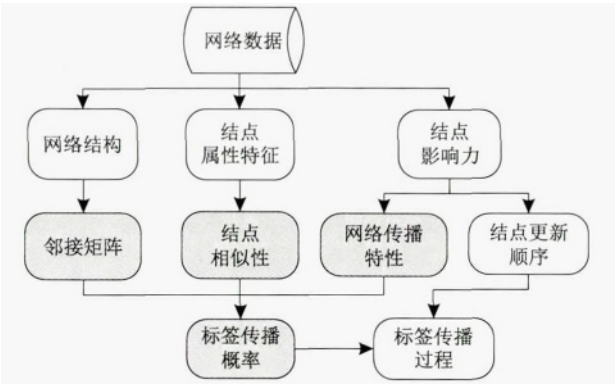

图 2　LPPB 算法流程图

首先计算网络所有结点的影响力值，根据影响力的大小对结点排序，并以此作为结点的更新顺序，可以减少算法不必要的标签更新和标签“逆流”的现象；然后利用网络的传播特性、结点属性相似性和邻接矩阵分析标签在结点间传播的概率，并依据标签传播概率进行传播迭代，这样不仅提高了算法的准确性，更加准确地探测重叠社区结构，而且加快了收敛速度.

定义 1.　结点影响力. 设网络 $G(V,E)$ 中每个结点 i 都拥有一个影响力值，用 Inf_i 表示. 由于大多网络并不是连通图，因此本文采用吕琳媛等人[21-22]提出的 LeaderRank 算法，计算结点的 Inf 值.

3.1.1　网络的传播特性 κ

目前的标签传播算法虽然考虑了社会网络的传播特性，即假定结点直接接受邻居结点的标签，然而这在真实的社会网络中并不成立. 每个结点都存在传播标签和接收标签两种能力，一般我们认为影响力大的结点传播标签的能力就强. 根据文献[23]的研究结果，有影响力的结点不易接受影响，那么影响力大的结点的接受能力较弱，因此做如下定义.

定义 2.　传播特性 κ. 定义 $\kappa_{i\leftarrow j}$ 为标签从结点 j 到结点 i 的传播特性度量值.

$$\kappa_{i\leftarrow j}=\frac{\log(1+Inf_j)}{\log((1+Inf_i)\times(1+Inf_j))},$$

由结点 i 和 j 的影响力决定，注意一般来说 $\kappa_{i\leftarrow j}\neq\kappa_{j\leftarrow i}$. 当 Inf_i 远小于 Inf_j 时，$\kappa_{i\leftarrow j}\approx 1$，说明由于 j 的影响力较大，结点 i 极易接受 j 的标签；反之，当 Inf_i 远大于 Inf_j 时，$\kappa_{i\leftarrow j}\approx 0$，说明 i 的影响力较大，结点 i 较难接受 j 的标签.

如图 3 所示，结点 1 的标签传播到结点 3 的 κ 值可以表示为

$$\kappa_{3\leftarrow 1}=\frac{\log(1+4)}{\log((1+0.6)\times(1+4))}\approx 0.774,$$

$$\kappa_{1\leftarrow 3}\approx 0.226,\ \kappa_{3\leftarrow 2}\approx 0.661,\ \kappa_{2\leftarrow 3}\approx 0.339.$$

由上述结果可知，在更新结点 3 的过程中，结点 1 对结点 3 的影响更大，标签更容易从结点 1 传播到 3.

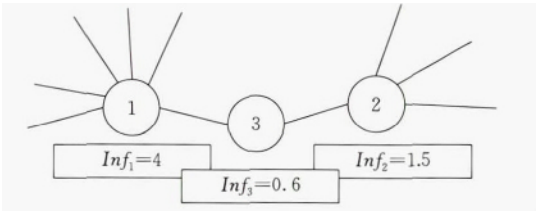

图 3　传播影响示意图

3.1.2　结点的更新顺序 U

由于先更新的结点影响传播的较远，很多 Inf 值较小的结点在传播过程中会反过来影响一些 Inf 值较大的结点，因此存在标签逆向传播的问题. 虽然算法在后续的迭代中可以修正结果，但耗费了大量的时间和更新操作，尤其是应用于大型社会网络时影响更为显著. 因此，根据结点的 Inf 值，由大到小排序作为结点的更新顺序(该排序用 U 表示)减少了算法不必要的标签更新和标签“逆流”的现象.

算法 1. 结点更新顺序.

输入:网络 $G(V,E)$,n 是结点总数

输出:结点更新顺序 U

```
1. U←{};
2. INF←{};
3. FOR EACH i∈V
4.    Inf_i=LeaderRank(i);
5.    INF←INF∪Inf_i;
6. ENDFOR
7. WHILE INF !=NULL
8.    U←U∪max(INF);
9.    INF←INF−max(INF);
10. ENDWHILE
```

3.1.3 结点的属性特征 S

现实的社会网络不仅具有拓扑结构特征,而且网络中结点的内在属性也容易获取,如 C-DBLP 中的学者记录都拥有研究方向、工作单位等信息,因此结点的属性特征 S 包含两部分:结构属性 St 和结点内在属性 In.

如图 4,学者 2、3、4 属于 A 社区,5、6、7 属于 B 社区,由于网络结构是对称的,此时学者 1 的候选标签集为 $\{A,B\}$,当学者 1 只能保留一个标签时,传统的标签传播算法在选择时存在较大的随机性,而实际上 1 和 2、3、4 同处于一个研究单位,如果算法随机地将学者 1 的标签更新为 B 是不恰当的,会影响社区划分的质量.

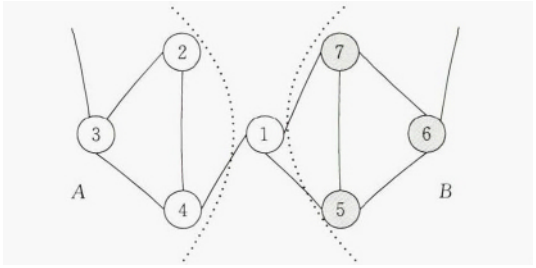

图 4 学者合著网络划分实例

为了解决上述随机性问题,本文综合结点的结构属性和内在属性共同计算结点间的相似性,即结点 i 和 j 的相似性,计算如下:

$$S_{i,j}=St_{i,j}+In_{i,j}.$$

采用常用的余弦公式计算结点 i 和 j 的结构相似性:

$$St_{i,j}=\frac{|\Gamma(i)\cap\Gamma(j)|}{\sqrt{|\Gamma(i)|\times|\Gamma(j)|}},$$

其中,$\Gamma(i)$表示结点 i 的所有邻居与结点 i 的集合. 定义 $In_i=\{in_{i1},in_{i2},\cdots,in_{iN}\}$ 为结点 i 的内在属性集合,in_{ik} 是结点 i 的第 k 个属性值,N 是内在属性个数. 那么结点 i 和 j 的内在属性相似性 $In_{i,j}$ 的计算公式为

$$In_{i,j}=\frac{1}{N}\sum_{k=1}^{N}\zeta(in_{ik},in_{jk}),$$

$$\zeta(in_{ik},in_{jk})=\begin{cases}1, & in_{ik}=in_{jk}\\ 0, & in_{ik}\neq in_{jk}\end{cases}.$$

例如给定结点的 3 个属性为{城市,性别,单位},结点 p 的值为{武汉,男,百度},结点 q 的值为{武汉,女,腾讯},那么 $In_{p,q}$ 值即为 1/3.

把结点间的相似性扩展到结点与社区的相似性就可以避免图 4 划分错误的问题,即结点 i 与社区 A 的相似性计算如下:

$$S_{i,A}=\frac{1}{M}\sum_{j\in A}S_{i,j}=\frac{1}{M}\sum_{j\in A}\frac{|\Gamma(i)\cap\Gamma(j)|}{\sqrt{|\Gamma(i)|\times|\Gamma(j)|}}+\frac{1}{M\times N}\sum_{j\in A}\sum_{k=1}^{N}\zeta(in_{ik},in_{jk}),$$

其中,M 为社区 A 的结点总数,在标签选择的过程中结点总是选择与自己相似性最大的社区作为标签.

本文将结点间的属性相似性值 S 作为网络关系的权重,并通过改进的网络传播特性 κ 共同决定标签传播概率.

3.2 标签传播概率

根据 3.1 节阐述的算法思想,本节给出在结点更新的过程中标签传播概率的定义.

定义 3. 标签传播概率. 结点 j 的标签以概率 $P(i\leftarrow j)$ 传播到结点 i,$P(i\leftarrow j)$ 取决于结点 i 和 j 的相似性度量 $S_{i,j}$、传播特性度量 $\kappa_{i\leftarrow j}$ 和邻接矩阵 $\boldsymbol{\delta}(i,j)$,即

$$P(i\leftarrow j)=S_{i,j}\times\kappa_{i\leftarrow j}\times\boldsymbol{\delta}(i,j) \tag{1}$$

其中 $S_{i,j}\in[0,1]$ 表示结点 i 和 j 的关系权重,即结点间的属性相似度,$\kappa_{i\leftarrow j}\in[0,1]$ 表示标签从结点 j 传播到结点 i 的度量值. $\boldsymbol{\delta}$ 是 $n\times n$ 的邻接矩阵,n 为结点总数,那么 $\boldsymbol{\delta}(i,j)$ 表示如下:

$$\boldsymbol{\delta}(i,j)=\begin{cases}1, & \text{结点 } i \text{ 和 } j \text{ 有连边}\\ 0, & \text{反之}\end{cases},$$

$P(i\leftarrow j)$ 的取值范围为 $[0,1]$,由 $S_{i,j}$、$\kappa_{i\leftarrow j}$ 和 $\boldsymbol{\delta}(i,j)$ 共同决定.

3.3 标签传播过程

定义 4. 结点的最大标签数 k. 根据输入网络,定义结点拥有的最大标签数 k,即为结点可以归属的最大社区数. 在 C-DBLP 的实验中证明,k 取值的波动会影响到社区划分的重叠度和模块度.

定义 5. 标签分布矩阵 $\boldsymbol{P}$. 每个结点拥有一组

标签，用 $1\times k$ 向量 $\boldsymbol{P}_i$ 表示结点 i 的标签分布：$\boldsymbol{P}_i=\{(\boldsymbol{P}_i(l_j),l_j)\mid j=1,2,\cdots,k\cap l_j\in C\}$，其中 $\boldsymbol{P}_i(l_j)$ 是结点 i 拥有标签 l_j 的概率，C 是全部标签的集合，且满足 $\sum_{j=1}^{k}\boldsymbol{P}_i(l_j)=1$. 所有结点的标签分布组成了标签分布矩阵 $\boldsymbol{P}$.

标签传播算法是一个迭代的过程，定义算法第 i 次迭代时的标签分布矩阵为 $\boldsymbol{P}^i$，图 5 给出本算法的迭代过程示意图. 每次迭代，根据结点的更新顺序 U 重新计算各个结点的标签分布.

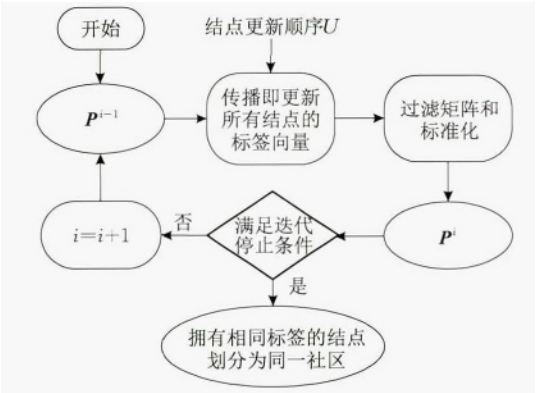

图 5　标签传播过程示意图

结点 i 除了概率地接受邻居结点的标签，还要考虑结点 i 的历史标签信息. 历史的标签信息对当前的更新决策有一定的影响，并且其影响随时间衰减，定义 $\sum_{k=1}^{t-1}\lambda^k P_i^k$ 为算法第 t 次迭代时结点 i 的历史标签信息对更新标签的辅助信息.

那么算法在第 t 次迭代时，结点 i 的标签向量 $\boldsymbol{P}_i^t$ 计算公式如下：

$$\boldsymbol{P}_i^t=\sum_{j\in nb_i}P_j^{t-1}\times P(i\leftarrow j)+\sum_{k=1}^{t-1}\lambda^k P_i^k \qquad (2)$$

等式(2)右侧第 1 项表示结点 i 概率地接受所有邻居结点的标签向量，第 2 项是结点 i 的历史标签集合，其影响经 λ 衰减后作为标签更新的辅助信息.

在更新结点标签时，文献[24]发现同步更新需要更多的迭代次数，但是输出结果更加稳定，这是由于异步更新在分别选择 t 与 $t-1$ 时刻邻居结点时的随机性引起的. 本文通过实验发现采用同步更新策略能取得更加准确和稳定的结果.

3.4　过滤矩阵和标准化

当分析的数据剧增时，如何缓解内存的压力是十分重要的问题，于是引入一个阈值 $\phi=1/k$，其中 k 为结点归属的最大社区数. 将标签分布矩阵 $\boldsymbol{P}$ 中概率值小于阈值的标签概率设置为 0. 如果结点拥有的标签概率都小于阈值，则计算结点与社区的相似性，选择具有最大相似值的社区作为该结点的标签，最后标准化分布矩阵 $\boldsymbol{P}$，即 $\boldsymbol{P}$ 中每行的概率和为 1.

算法 2.　过滤矩阵和标准化步骤.

输入：网络 $G(V,E)$，结点归属的最大社区数 k，S_{i,C_i} 为结点 i 与社区 C_i 的相似性

输出：标准化的分布矩阵 $\boldsymbol{P}$

---对每一个属于 V 的结点 i

1. FOR EACH $i\in V$

---过滤 $\boldsymbol{P}_i$ 中概率小于 ϕ 的标签

2.　FOR EACH $j\in[1,k]\wedge j\in Z$

3.　IF $\boldsymbol{P}_i(l_j)<\phi$

4.　SET $\boldsymbol{P}_i(l_j)=0$;

5.　ENDIF

6.　ENDFOR

---如果 $\boldsymbol{P}_i$ 的标签都被过滤，就选择与结点 i 相似性最大的标签；否则，标准化分布 $\boldsymbol{P}_i$

7.　IF $\sum_{j=1}^{k}\boldsymbol{P}_i(l_j)=0$

8.　$label(i)=\arg\max_{C_i\in C}S_{i,C_i}$;

9.　SET $\boldsymbol{P}_i=\{(label(i),1)\}$;

10.　ELSE

11.　$\boldsymbol{P}_i=\boldsymbol{P}_i\Big/\sum_{j=1}^{k}\boldsymbol{P}_i(l_j)$;

12.　ENDIF

13. ENDFOR

3.5　迭代停止条件

算法的主要部分是迭代执行标签传播过程，因此合理的选择迭代停止条件十分重要. 初始时，标签分布矩阵 $\boldsymbol{P}$ 的标签数量等于结点总数，并随着迭代过程而减少，最终得到最小值. 定义 i^t 为第 t 次迭代 $\boldsymbol{P}$ 的标签数量，也表示当前网络的社区数. 当 $i^t=i^{t-1}$ 时，算法停止，可能导致输出结果不准确，虽然此刻标签数量不变，但是很有可能在几次迭代后再次降低. 因此，终止条件除了满足 $i^t=i^{t-1}$，还需要观察结点拥有标签的变化情况. 第 t 次迭代时，假设网络中拥有标签 c 的结点总数为 i，那么

$$c^t=\Big\{(c,i):c\in V\wedge i=\sum_{x\in V,P_x(c)>0}1\Big\}.$$

拥有 c 的结点数量 i 在迭代过程中经常变化，于是本文观测拥有标签 c 的最小结点数的变化情况，当 $i^t=i^{t-1}$ 时

$$c_{\min}^t=\{(c,i):\exists p\,\exists q((c,p)\in c_{t-1}\wedge(c,q)\in c_t\wedge i=\min(p,q))\}$$

否则，$c_{\min}^t=c_t$. 一旦满足 $c_{\min}^t=c_{\min}^{t-1}$，算法即可停止迭代.

3.6 后处理

算法结束后根据标签分布矩阵 P，将具有相同标签的结点划分为同一社区. 然而得出的结果中可能存在某社区的子集或者非严格意义的子集(拥有几乎相同的结点)，因此需要删除掉这些重复的子集，这样就得出本文所挖掘的网络重叠社区结构，P 中拥有多标签的结点即为社区重叠结点.

3.7 LPPB 算法

这里给出 LPPB 算法的伪代码，如算法 3 所示.

算法 3. LPPB 算法描述.

输入：网络 $G(V,E)$，n 是结点总数，参数 k 是结点归属的最大社区数，邻接矩阵 δ，C 是全部标签的集合

输出：重叠社区 C_{ov}

1. 初始化，为每个结点赋予一个独立的标签，并生成 $n\times k$ 的初始标签分布矩阵 P；
2. 计算所有结点的影响力值(Inf)，由高到低排序作为结点更新的顺序 U；
3. $t=1$；
4. 根据顺序 U 更新每个结点的标签分布向量，任一结点 i 的计算方法如下：

$$P_i^t=\sum_{j\in nb_i}P_j^{t-1}\times S_{i,j}\times\kappa_{i\leftarrow j}\times\boldsymbol{\delta}(i,j)+\sum_{k=1}^{t-1}\lambda^k P_i^k$$

然后标准化分布矩阵 P；

5. 删除 P 中小于阈值 $\phi=1/k$ 的标签信息，如果结点 i 的标签概率都小于阈值，那么在标签集合 C 中选择与结点 i 的相似值最大的标签：

$$label(i)=\arg\max_{C_i\in C}S_{i,C_i}$$

然后标准化分布矩阵 P；

6. 判断是否满足设定的迭代停止条件

如果满足，则算法结束；

否则 $t=t+1$，进入下一次迭代，重复步 4～6；

7. 算法执行结束后，将拥有相同标签的结点划分为同一社区，输出 C_{ov}.

3.8 算法复杂度分析

假定网络有 n 个结点和 m 条边，k 是结点归属的最大社区数，以 m/n 表示结点的平均邻居数.

(1) 初始化标签分布矩阵需要 $O(kn)$.

(2) 计算结点的影响力 $O(n\log n)$.

(3) 计算结点相似性需要 $O(k^2n)$.

(4) 与 COPRA 算法一致，整个传播过程需要 $O(km\log(kn/m))$.

(5) 过滤矩阵和标准化需要 $O(kn)$.

(6) 最后划分社区需要 $O(k(m+n))$.

由于 k 一般取值较小，因此算法总的时间复杂度为 $O(n\log n)$，在真实网络中实验发现该算法能够快速地收敛. LPPB 完善了社会网络的传播特性，综合考虑了复杂网络的结构、结点的属性特征和历史标签记录对标签传播过程的影响，使得划分的社区结果更加准确和稳定.

4 实验分析

4.1 实验数据集和评估方法

我们选取 Zachary Karate Club[25] 数据集作为基准数据集，并将本算法得出的结果与标准结果进行比较，验证算法的准确性；另外选取真实的 C-DBLP 合著数据(来源于 WAMDM 实验室)来测试 LPPB 算法生成社区的质量和稳定性，该数据集带有时间维度，包含学者合著的时间，可用于动态分析社区重叠结构的变化和对社区演化的作用.

由于大部分的真实网络没有标准的结果作为对比，本实验采用改进的模块度 Q_{ov}[26]、社区重叠度 $Overlap$ 和 F 值作为分析指标来观测算法输出结果. 给定无向网络 $G(V,E)$，C 表示划分得到的社区数，m 是结点个数.

(1) 改进的模块度 Q_{ov}. Q_{ov} 取决于每个结点归属于各社区的概率和各社区成员的数量.

$$Q_{ov}=\frac{1}{2m}\sum_{c\in C}\sum_{i,j\in V}\left[r_{ijc}A_{ij}-w_{ijc}\frac{k_ik_j}{2m}\right],$$

其中，r_{ijc} 表示结点 i 和 j 同属于社区 c 的概率，$r_{ijc}=\ell(p_{i,c},p_{j,c})$，其中 $p_{i,c}$ 表示结点 i 属于社区 c 的概率，$\ell(p_{i,c},p_{j,c})$ 可表示为

$$\ell(p_{i,c},p_{j,c})=\frac{1}{(1+\mathrm{e}^{-f(p_{i,c})})(1+\mathrm{e}^{-f(p_{j,c})})}.$$

对函数 f 的定义，采用文献[26]的定义，令 $f(x)=60x-30$. w_{ijc} 表示结点 i 在社区 c 里或者结点 j 在社区 c 里的概率，表示如下：

$$w_{ijc}=\frac{\sum_{j\in V}\ell(p_{i,c},p_{j,c})}{|V|}\times\frac{\sum_{i\in V}\ell(p_{i,c},p_{j,c})}{|V|}.$$

(2) 社区重叠度 $Overlap$. 社区重叠结点的个数决定了社区重叠度 $Overlap$ 的值，是网络耦合度的具体体现，计算公式如下：

$$Overlap=\frac{1}{m}\sum_{c\in C}|c|,$$

其中 $|c|$ 表示社区 c 的结点个数，m 为网络结点总数.

(3) 综合指标 F 值. 一般情况下，重叠度较高的网络其模块度相对较低，两者呈负相关性，为了输出更加合适的社区结果，定义 F 值作为综合评估指标：

$$F=\frac{Q_{ov}\times Overlap\times 2}{Q_{ov}+Overlap},$$

F 值越大，表明算法输出结果越好.

4.2　参数 k 的选择

结点归属的最大社区数 k 作为 LPPB 算法的唯一参数，直接影响了算法输出的结果质量. 本文以综合指标 F 为效益函数，采用迭代局部搜索[27]的思想，在 k 取 $1\sim|C|$ 范围内寻找最优效益函数值，这里 C 是网络中全部标签的集合，$|C|$ 表示标签总数.

算法 4.　参数选择算法.

输入：参数 k 及对应的效益函数值 $F(k)$

输出：$k^*=\arg\max F(k)$

1. 初始化：随机生成初始解 k_0
2. 进行局部搜索：$k^*=LocalSearch(k_0)$
   ——$LocalSearch$ 采用爬山法
3. WHILE(未达到 CPU 预定运行时刻)
4.　对 k^* 进行随机扰动：
   $k'=Perturbation(k^*)$
5.　进行局部搜索：
   $k_l=LocalSearch(k')$
6.　IF $F(k_l)>F(k^*)$ THEN
7.　$k^*=k_l$
8.　ENDIF
9. ENDWHILE

4.3　基准数据集实验

Zachary Karate Club 包含 34 个成员，78 条关系，后来由于管理的分歧，该俱乐部主要分为 2 个社区，但是有一些成员处于重叠状态，如图 6 中浅灰色结点. 由于结点 33、34 和 1、2、3 分别作为两个社区的核心人物，尽管与其他社区重叠，但是根据本文提出的网络传播特性，这些核心人物属于其他社区的概率较低，不应该作为重叠结点被探测.

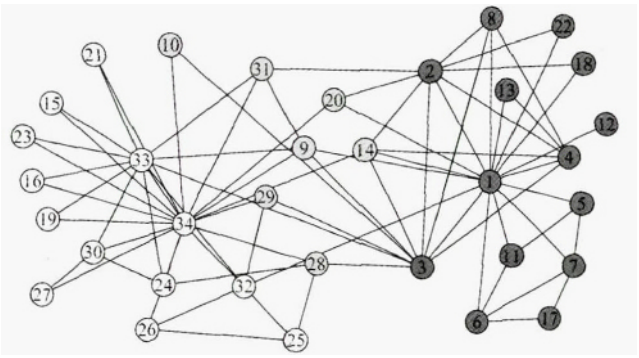

图 6　重叠的空手道网络

由于无法获取基准数据集实验中空手道网络的结点属性，因此本实验只能计算结点间的结构属性相似性，即取决于结点间的共同邻居数. LPPB 算法在该数据集上运行 100 次取平均值，在取相同的参数情况下与 COPRA 算法进行对比，实验得出的社区数稳定，能够探测准确的重叠点，且迭代次数明显减少，如表 1 所示. 与传统的 COPRA 不同，LPPB 算法能探测出部分重叠结点如 14、28 等，这些结点都与两个派系的核心结点相连，根据网络传播特性的影响，易接受高影响力结点的标签. 如果仅仅通过邻域结构计算标签传播会把这些结点只归为一个派系，因此本算法可以得出质量更高的社区划分结果.

表 1　算法 100 次运行结果分析

方法	平均社区数量	平均迭代次数	重叠结点
COPRA	3.7	10	3,9,10,20,31
LPPB	2.0	4	9,10,14,20,28,29,30,31

为了验证算法能够降低随机性造成模块度较低的影响，本实验没有选取最好的模块度进行比较，而是通过对比两种算法在不同数据集的模块度平均值 Avg，如表 2 所示，LPPB 算法方差较小，具有较好的稳定性，且显著提高了社区划分的质量.

表 2　平均模块化度量 Avg

数据集	k^*	COPRA 均值	LPPB 均值
Karate[25]	3	0.459±0.140	0.733±0.000
Dolphins[28]	4	0.654±0.041	0.702±0.018
Blogs[29]	8	0.726±0.012	0.756±0.002
PGP[30]	11	0.778±0.019	0.796±0.009

4.4　C-DBLP 数据集实验

本节实验采用 C-DBLP《计算机学报》、《软件学报》和《计算机研究与发展》3 个领域的学者合著关系网络，时间跨度为 1999～2009 年，共计 6908 篇文章，8257 个作者，19683 条合著关系. 本文抽取学者的合著次数、工作单位和研究方向作为结点的内在属性特征，把研究者作为网络的结点.

下面本实验将以静态和动态两种方式对 C-DBLP 数据进行分析，静态分析针对截止到 2009 年的 11 年合著数据，包括用户影响力分布、不同规模网络的执行时间和迭代次数、模块度与重叠度的综合分析、社区分布情况等；动态分析将 C-DBLP 分为 11 个时间段，即“1999～2000，1999～2001，…，1999～2009”，通过时间序列的递增，观测社区重叠结构（重叠点）的变化趋势及对社区演化的影响.

4.4.1　静态分析

由上述 11 年合著数据经过 LeaderRank 算法计算得出结点的影响力值作为更新顺序 U，结点的影响力服从厚尾分布，如图 7 所示. 当 Inf 值在 0.5～29 间波动时，κ 的取值范围为[0.11，0.89].

定义 6.　社区中心结点. 社区中具有最高拓扑

势的结点即为社区中心结点，结点 i 的拓扑势用如下公式计算[31]：

$$\varphi_i(\sigma)=\frac{1}{n}\sum_{j=1}^{h}n_j(v_i)\times e^{-\left(\frac{j}{\sigma}\right)^2},$$

其中，n 为社区结点个数，$n_j(v_i)$ 为结点 i 的第 j 跳邻居节点数，h 是计算结点影响的跳数，σ 表示结点影响的范围，文献[32]分析了网络拓扑势熵最小值存在的普遍性，那么 σ 值可由如下公式确定：

$$\sigma_{opt}=\arg\min H(\sigma),$$

其中 $H(\sigma)$ 为拓扑势熵，其定义为

$$H(\sigma)=-\sum_{i=1}^{n}\frac{\varphi_i(\sigma)}{Z(\sigma)}\ln\frac{\varphi_i(\sigma)}{Z(\sigma)},$$

$Z(\sigma)$ 作为标准化因子，定义为

$$Z(\sigma)=\sum_{i=1}^{n}\varphi_i(\sigma).$$

根据高斯函数的数学特性，对于给定的 σ 值，每个结点的影响范围近似为 $\lfloor 3\sigma/\sqrt{2}\rfloor$ 跳时，单位势函数很快衰减为 0，因此确定了 σ_{opt} 后，那么结点的影响跳数 h 取值为 $\lfloor 3\sigma_{opt}/\sqrt{2}\rfloor$.

一般情况下，在算法后期的迭代中会形成以影响力大的结点为中心的社区，本文选取影响力排名前 10%～20%的结点进行观测，根据 LPPB 算法输出的社区结果，分别计算各社区内部结点的拓扑势，选取最高拓扑势的结点最为该社区的中心结点，发现其中有 80%以上的观测结点是划分社区的中心结点，从而验证了本算法根据结点影响力值确定更新顺序的准确性.

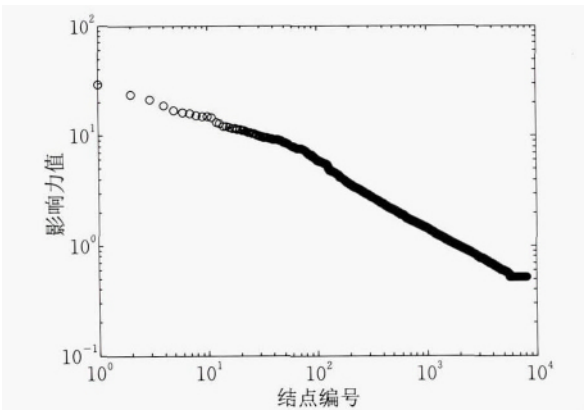

图 7 C-DBLP 网络结点影响力分布

本文依据 C-DBLP 数据集的时间特性，按照年份累积，构建了多个不同规模的网络，如表 3 所示，这 11 个网络的规模随着时间的推移稳步增长，可以用来测试算法的运行效率. LPPB 算法通过引入结点的属性特征相似性来减少算法的随机性，同时利用结点的历史标签信息指导当前的更新决策，因此算法可以快速收敛. 定义 LPPB* 表示去除结点属性信息影响的 LPPB 算法，可以验证在不引入结点属性信息的情况下，LPPB 算法是否仍然比原始的 COPRA 算法更优. 本实验通过综合指标 F 值和算法的运行效率（算法的运行时间和收敛的迭代次数）观测 3 种算法的性能，另外还进行了 LPPB 算法在模块度、重叠度和社区集中度等指标的分析.

表 3 不同规模网络数据

网络	累积截止年份	结点数	边数
t_1	1999	857	1293
t_2	2000	1602	2662
t_3	2001	2207	3899
t_4	2002	2910	5488
t_5	2003	3641	7122
t_6	2004	4386	8930
t_7	2005	5160	10954
t_8	2006	6196	13657
t_9	2007	7070	16247
t_{10}	2008	7790	18329
t_{11}	2009	8257	19683

(1) 综合指标 F 值分析

分析整体的 C-DBLP 数据集，采用迭代局部搜索算法得出的 k^* 值使得模块度和重叠度的综合指标 F 值最优，当 $k^*=13$ 时，LPPB 算法的 F 值最高；当 $k^*=12$ 时，LPPB* 算法的 F 值达到峰值；当 $k^*=7$ 时，COPRA 算法的 F 值达到最优. 通过试值计算，图 8 给出了 3 种算法的综合指标 F 值随 k 值的变化趋势，结果证明与搜索算法得出的参数 k^* 值一致. 从图中可以看出原始的 COPRA 算法随着参数 k 的变大，重叠度趋于稳定而模块度急剧降低，导致 F 值变得较小；LPPB 和 LPPB* 算法的 F 值波动相对平缓，显示了本文提出的网络传播特性对标签传播算法的改进效果；同时 LPPB 算法引入了结点的属性信息来影响标签传播的概率，输出效果较优于 LPPB* 算法.

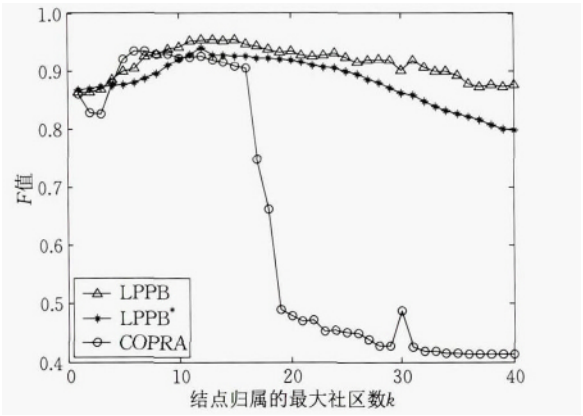

图 8 综合指标 F 值随 k 值变化

表 4 全面对比 3 个算法在 C-DBLP 数据集上运行的最优结果：LPPB 算法的平均模块度和重叠度相较于 COPRA 和 LPPB* 都有所提高；LPPB 和 LPPB* 的方差大大降低，说明了算法输出的结果更加稳定；对于综合指标 F 值，LPPB 算法在 COPRA 的基础上提高了 2.99%；LPPB 算法收敛的迭代次数和每次迭代的运行时间都比 COPRA 算法减少 10%以上.

表 4　最优 F 值的结果对比

指标	最高模块度	平均模块度	方差	平均迭代次数	总时间/s	重叠度	F 值
COPRA $k^*=7$	0.749	0.740	0.006	86.4	6.71	1.269	0.935
LPPB* $k^*=12$	0.756	0.753	0.002	79.9	5.43	1.291	0.951
LPPB $k^*=13$	0.767	0.764	0.002	76.6	5.11	1.302	0.963
分析	↑	↑	↓	↓	↓	↑	↑

(2) 算法的运行效率分析

分别在表 3 给定 C-DBLP 的不同规模网络上进行实验，然后计算 3 个算法的运行时间和迭代次数在参数 k 取最优时运行 100 次的均值，如图 9 和图 10 所示的实验结果，显示了 LPPB 算法比 LPPB* 和 COPRA 更有效. 通过比较 LPPB 和 LPPB* 两个算法的效果，验证了引入结点的属性特征相似性可以减少算法的随机性，降低了算法执行时间和迭代次数，在数据量增大的情况下依旧能够快速得出结果.

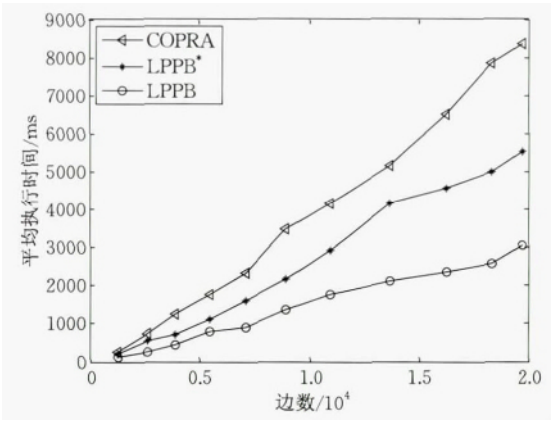

图 9　C-DBLP 不同规模网络的执行时间

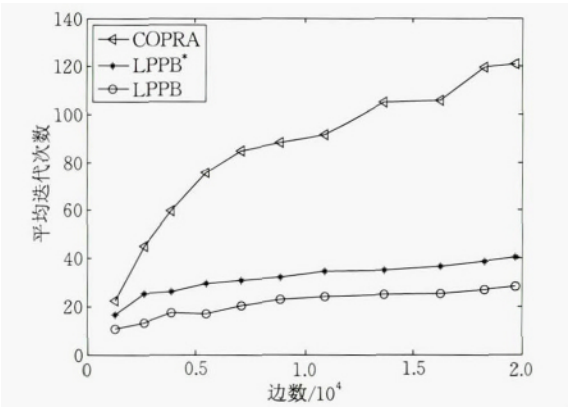

图 10　C-DBLP 不同规模网络的迭代次数

本节实验还分析了 LPPB 算法在 LFR[33] 网络上的运行效率，网络的平均度为 10，边数从 10000～200000 不断递增，其他参数为默认值. 如图 11 所示，运行结果验证了 LPPB 算法拥有接近线性的时间复杂度，能够较好地处理大型社会网络数据.

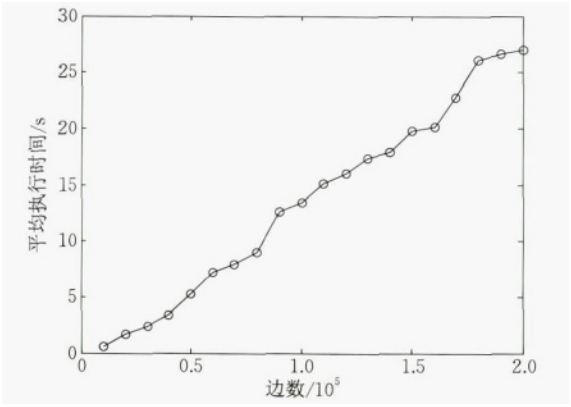

图 11　LFR 网络的执行时间

(3) 其他指标的分析

图 12 给出了算法的参数 k 对社区模块度 Q_{ov}（左侧纵坐标）和社区重叠度 $Overlap$（右侧纵坐标）的影响，输出结果在 k 较小时模块度较高，但此时网络的重叠度较低，只有小部分重叠结点被探测；当 k 值增大时，大部分的重叠结点被探测，重叠度渐渐的趋于某个极大值；当 k 大于 40，由于社区间重叠的部分较多，即算法输出的社区存在较多的子集和近似子集，删除这些子集，导致模块度和重叠度急剧降低.

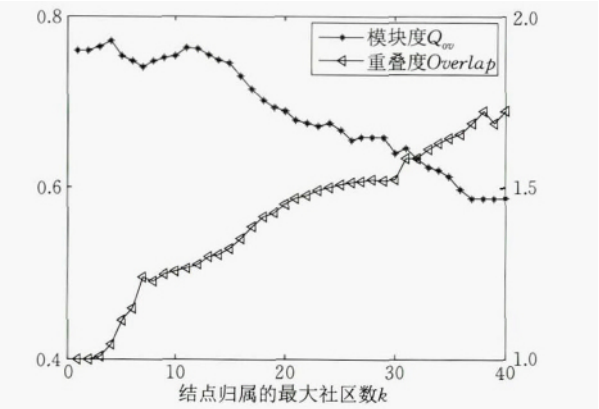

图 12　LPPB 算法的模块度和重叠度

定义 7.　社区集中度. 社区的分布情况是社区质量的重要衡量标准之一，用 HIM 表示社区的集

中度情况.

$$HIM=\sum_{i=1}^{N}\left(\frac{V_i}{V}\right)^2,$$

N 是社区的个数,V 是网络的结点总数,V_i 表示第 i 个社区包含的结点数.

HIM 值越大,表示划分的社区较集中,即容易出现超大型社区. 从图 13 中可以看出 LPPB 算法通过改进基于邻域结构的标签传播方式,使得社区划分的结果更加稳定、社区均匀分布,而 COPRA 算法在 k 值较大时,社区质量严重降低,此时出现较大社区,不利于检测出重叠结点.

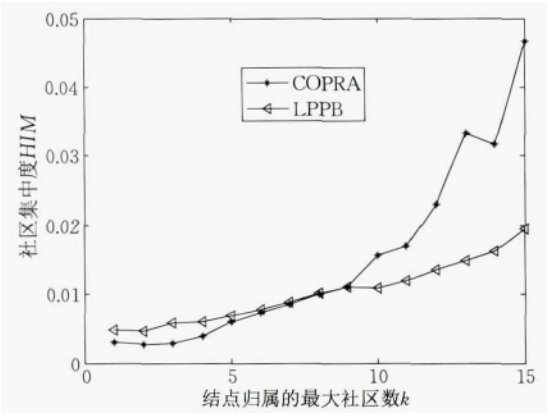

图 13 两种算法的 HIM 指数

另外,非单体社区是指结点总数大于 1 的社区,本实验将非单体社区数随参数 k 变化的波动性作为社区质量的衡量指标. 图 14 给出了 LPPB 算法和 COPRA 算法在 k 变化时非单体社区数的对比分析:当 k 取 2～4 时,此时网络的重叠度较低,COPRA 算法输出的非单体社区数较高;而后随着 k 增大,社区重叠部分增多,从而生成了部分社区的子集和近似子集,这部分子集被删除后导致非单体社区数急剧下降. LPPB 算法修正了网络的传播特性,减少了社区子集生成的概率,从而降低输出结果的波动性,提高了算法的稳定性.

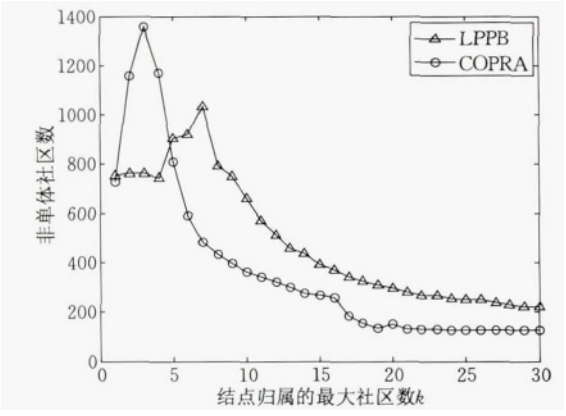

图 14 非单体社区数随 k 变化趋势

4.4.2 动态分析

为更好地了解 C-DBLP 网络,本节实验将网络划分成 11 个时间片段,动态地分析 LPPB 算法输出的结果.

(1) 社区重叠结点的行为分析

本节实验针对社区重叠结点的合著行为进行分析,有助于理解社区的演化过程.

定义 8. 合著转移概率 α. $t-1$ 时刻,算法得到重叠结点 i 归属社区 c 的概率为 $p_{i,c}^{t-1}$,C 是结点 i 归属的社区集合,显然 $\sum_{c\in C}p_{i,c}^{t-1}=1$;若 t 时刻,结点 i 存在合著行为,那么有 $1-\alpha$ 的概率结点 i 的合著者 j 还属于 C,那么 α 即为合著转移概率,如图 15 所示.

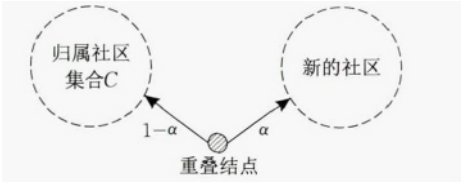

图 15 合著转移示例

根据年份将 C-DBLP 数据集分成 11 个时刻,同表 3 一致,动态地分析 LPPB 算法在每个时刻的输出结果,计算合著转移概率 α 的算法如下.

算法 5. 计算合著转移概率.

输入:$t-1$ 时刻重叠结点集合 on^{t-1} 及其归属社区集合 $\{C_i^{t-1}\mid i\in on^{t-1}\}$,$t$ 时刻网络结点的集合 V^t,t 时刻网络的邻接结点集合 $L^t=\{l_i^t\mid i\in V^t\}$,$t$ 的取值为 2～11

输出:合著转移概率 α

```
1.  sum=0; transfer=0;
2.  t=2;
3.  WHILE(t<=11)
4.    FOR EACH i∈on^{t-1}
5.      FOR EACH j∈l_i^t
6.        IF j∉C_i^{t-1}
7.          transfer=transfer+1;
8.        ENDIF
9.        sum=sum+1;
10.     ENDFOR
11.   ENDFOR
12.   t=t+1;
13. ENDWHILE
14. IF sum !=0
15.   α=transfer/sum;
16. ENDIF
```

算法 5 运行 20 次得出的结果如图 16 所示,算

法输出较为稳定，社区重叠结点的合著转移概率在均值 0.325 附近浮动，表明 C-DBLP 数据集中社区重叠结点仅以较小的概率进行合著转移，更多是融入已有的社区圈子.

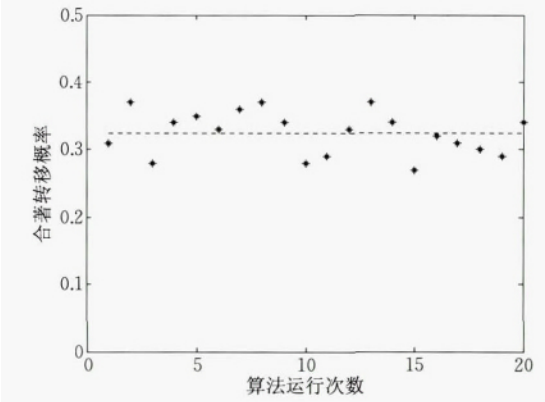

图 16　C-DBLP 合著转移概率

(2) 网络的演化趋势分析

本节实验计算了 k 分别取 1～30 时重叠结点占结点总数的比例，然后取该比例的平均值进行分析. 如图 17 所示，随着网络规模的扩大，重叠结点所占比例逐渐增加，表明网络的耦合度越来越高，即网络中结点的兴趣广泛，一般的结点都参与了多个社区. 很明显，实验选取的 C-DBLP 网络在这 11 年间处于高耦合度的发展趋势. 从图中还可以预测在 2009 年附近重叠度有可能趋于某个稳定值，如果重叠结点比例不变或降低，网络则处于内聚型发展，即结点专注于某一个社区内活动.

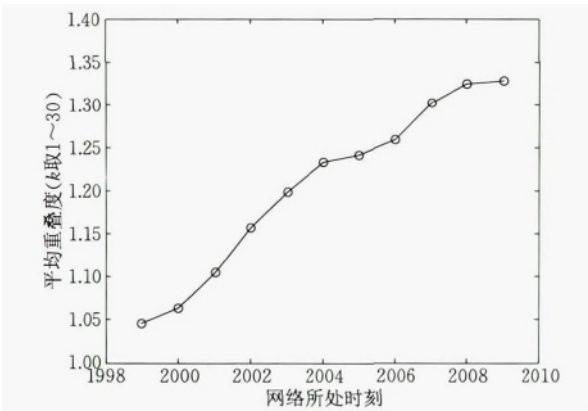

图 17　重叠结点所占比例随时间变化

本节还分析了算法输出的平均非单体社区数随时间的变化趋势，如图 18 所示. 平均非单体社区数呈近似线性增长趋势，表示随着网络规模的扩大，非单体社区数稳步增长，因此我们得出：在 C-DBLP 网络不断发展的过程中，社区的融合和分裂处于某种平衡状态.

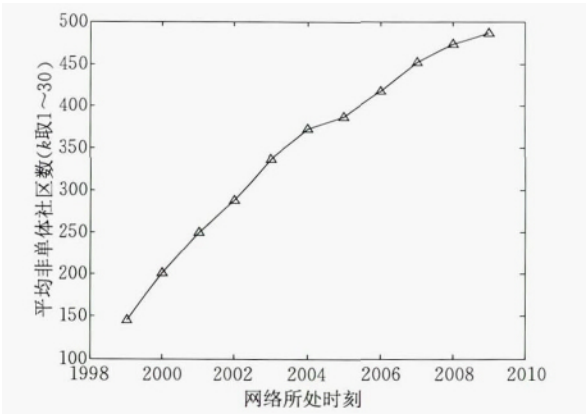

图 18　非单体社区数随时间变化

5　总结和展望

本文提出的基于标签传播概率的重叠社区发现算法，依据结点的影响力大小进行更新，综合网络的结构传播特性、结点的属性相似性和历史标签信息共同影响标签的传播过程，实验验证了本算法大大减少了运行时间和迭代次数，同时也保证了社区划分的准确性和稳定性，提高了社区划分结果的质量. 然而由实验结果可知，模块度与重叠度并不是正相关的，过分追求高模块度可能无法准确地探测出重叠结点；反之，追求高重叠度，而使得社区的模块度并不理想. 因此，文中给出两者的综合指标 F 值来选择算法最优的结果.

实验分析了社区重叠结点的行为特性，揭示了 C-DBLP 数据集中社区重叠结点仅以较小的概率进行合著转移，更多是融入已有的社区圈子；另外，还动态分析了社区的重叠结构，发现 C-DBLP 网络在 1999～2009 年间处于高“耦合度”的发展趋势，同时也预测了网络在 2009 年后可能会进入“内聚型”发展模式；此外，重叠结构对社区演化具有关键的作用，该作用会导致社区发生融合或者分裂，我们将在后续的工作中进行详细研究.

参　考　文　献

[1] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

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

[3] Hechter M. Principles of Group Solidarity. California, USA: University of California Press, 1988

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

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

[6] Shi C, Cai Y, Fu D, et al. A link clustering based overlapping community detection algorithm. Data & Knowledge Engineering, 2013, 87: 394-404

[7] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 2009, 11(3): 033015

[8] Whang J J, Gleich D F, Dhillon I S. Overlapping community detection using seed set expansion//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, USA, 2013: 2099-2108

[9] Pons P, Latapy M. Computing communities in large networks using random walks//Yolum P, Güngör T, Gürgen F, Özturan C eds. Computer and Information Sciences (ISCIS 2005). Springer Berlin Heidelberg, 2005: 284-293

[10] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113

[11] von Luxburg U. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395-416

[12] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111

[13] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007, 76(3): 036106

[14] Barber M J, Clark J W. Detecting network communities by propagating labels under constraints. Physical Review E, 2009, 80(2): 026129

[15] Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. Physica A: Statistical Mechanics and Its Applications, 2010, 389(7): 1493-1500

[16] Xie J, Szymanski B K. Community detection using a neighborhood strength driven label propagation algorithm// Proceedings of the 2011 IEEE Network Science Workshop (NSW). New York, USA, 2011: 188-195

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

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

[19] Zhao Zhuo-Xiang, Wang Yi-Tong, Tian Jia-Tang, Zhou Ze-Xue. A novel algorithm for community discovery in social networks based on label propagation. Journal of Computer Research and Development, 2011, 48(Suppl.): 8-15 (in Chinese)
(**赵卓翔，王轶彤，田家堂，周泽学. 社会网络中基于标签传播的社区发现新算法. 计算机研究与发展**，2011，48(Suppl.)：8-15)

[20] Qiang H, Yan G. A method of personalized recommendation based on multi-label propagation for overlapping community detection//Proceedings of the 2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM). Chengdu, China, 2012: 360-364

[21] Lü L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case. PloS One, 2011, 6(6): e21202

[22] Li Q, Zhou T, Lv L, et al. Identifying influential spreaders by weighted leaderrank. arXiv preprint arXiv: 1306.5042, 2013

[23] Aral S, Walker D. Identifying influential and susceptible members of social networks. Science, 2012, 337(6092): 337-341

[24] Leung I X Y, Hui P, Lio P, et al. Towards real-time community detection in large networks. Physical Review E, 2009, 79(6): 066107

[25] Zachary W. An information flow model for conflict and fission in small groups1. Journal of Anthropological Research, 1977, 33(4): 452-473

[26] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009(03): P03024

[27] Martin Q C, Stutzle T, Lourenço H R. Iterated local search. Handbook of Metaheuristics, 2003, 57: 321-353

[28] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003, 54(4): 396-405

[29] Gregory S. An algorithm to find overlapping community structure in networks//Kok J N, Koronacki J, Lopez de Mantaras R, et al, eds. Knowledge Discovery in Databases: PKDD 2007. Springer Berlin Heidelberg, 2007: 91-102

[30] Boguñá M, Pastor-Satorras R, Díaz-Guilera A, et al. Models of social networks based on social distance attachment. Physical Review E, 2004, 70(5): 056122

[31] Gan Wen-Yan, He Nan, Li De-Yi, et al. Community discovery method in networks based on topological potential. Journal of Software, 2009, 20(8): 2241-2254(in Chinese)
(**淦文燕，赫南，李德毅等. 一种基于拓扑势的网络社区发现方法. 软件学报**，2009，20(8)：2241-2254)

[32] Zhang Jian-Pei, Li Hong-Bo, Yang Jing, et al. Variable scale network overlapping community identification based on identity uncertainty. Acta Electronica Sinica, 2012, 40(12): 2512-2518(in Chinese)
(**张健沛，李泓波，杨静等. 基于归属不确定性的变规模网络重叠社区识别. 电子学报**，2012，40(12)：2512-2518)

[33] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4): 046110

LIU Shi-Chao, born in 1989, Ph. D. candidate. His research interests include social networks analysis, data mining and intelligence computing.

ZHU Fu-Xi, born in 1957, Ph. D. supervisor. His research interests include social tagging, data mining and intelligence computing.

GAN Lin, born in 1989, Ph. D. candidate. Her main research interest is web data mining.

Background

Community is an important structure in most complex networks, vertices in the same community are dense and intercommunity edges are sparse. Identifying communities especially the overlapping structure helps us to understand real large networks, and also has the vital significance for community management and commercial recommendation.

Label propagation is a new method applied to overlapping structure detection because of its approximate linear time complexity. Thus, We proposes a novel overlapping community detection algorithm called Label Propagation Probability Based Algorithm, in which probability of label propagation depends on structure propagation characteristics of complex networks, properties of nodes and the history label information in the process of propagation. Experiment results show that it outperforms the state of art algorithms in performance and efficiency. In addition, we conduct dynamic analysis of overlapping structure, finding that C-DBLP network is in the high "coupling" trend of development.

The work of this paper is supported by the National Natural Science Foundation of China under Grant No. 61272277; the Fundamental Research Funds for the Central Universities under Grant No. 274742. These projects aim to study the adaptive model of generation and optimization of social tags to match the user depend on his behaviors in real networks, users with the same tags may belong to the same group, also they can get the potential tags from others after label propagation process. This paper is a sub-project of these programs, whose purpose is to find the overlapping structure among user groups.