计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180277
55(9):1959-1971,2018

# 一种基于标签传播的两阶段社区发现算法

郑文萍[1,2,3] 车晨浩[2] 钱宇华[1,2,3] 王 杰[2]
[1](山西大学大数据科学与产业研究院 太原 030006)
[2](山西大学计算机与信息技术学院 太原 030006)
[3](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(wpzheng@sxu.edu.cn)

# A Two-Stage Community Detection Algorithm Based on Label Propagation

Zheng Wenping[1,2,3], Che Chenhao[2], Qian Yuhua[1,2,3], and Wang Jie[2]
[1](*Research Institute of Big Data Science and Industry*, *Shanxi University*, *Taiyuan* 030006)
[2](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[3](*Key Laboratory of Computational Intelligence and Chinese Information Processing* (*Shanxi University*), *Ministry of Education*, *Taiyuan* 030006)

**Abstract** Due to the random process in node selection and label propagation, the stability of the results of traditional LPA is poor. A two-stage community detection algorithm is proposed based on label propagation, abbreviated as LPA-TS. In the first step of LPA-TS, the labels of nodes are updated according to their participation coefficients in non-decreasing order, and the node label is determined according to the similarity of nodes in the process of node label updating. Some clusters found by Step1 might not satisfy the weak community condition. If a cluster is not a weak community, in the beginning of Step2, we will merge it with the cluster that has most connections with it. Next, we treat each community as a node, and the number of edges between two communities as their edge weights between corresponding nodes. We compute the participation coefficients of each node of the resulted network, and use similar process to get the final results of the communities. The proposed algorithm LPA-TS reduces the randomness in the process of node selection and label propagation; hence, we might obtain stable communities by LPA-TS. In addition, LPA-TS combines small scale communities with adjacent communities in the second phase of the algorithm to improve the quality of community detection. Compared with other classical community detection algorithms on some real networks and artificial networks, the proposed algorithm shows preferable performance on stability, NMI, ARI and modularity.

**Key words** complex network; community detection; label propagation; participation coefficient; weak community

**摘 要** 针对标签传播社区发现算法在节点更新顺序及标签传播过程中存在较大随机性而导致划分结果稳定性差的问题,提出一种基于标签传播的两阶段社区发现算法(a two-stage community detection

**收稿日期**:2018-04-10;**修回日期**:2018-07-12
**基金项目**:国家自然科学基金项目(61672332,61572005);国家自然科学基金优秀青年科学基金项目(61322211);山西省回国留学人员科研资助项目(2017-014)
This work was supported by the National Natural Science Foundation of China (61672332, 61572005), the National Natural Science Foundation of China for Excellent Young Scientists (61322211), and the Research Project of Shanxi Scholarship Council of China (2017-014).
**通信作者**:钱宇华(jinchengqyh@sxu.edu.cn)

algorithm based on label propagation, LPA-TS),通过参与系数确定节点更新顺序,并在标签传播过程中依据节点间相似性更新节点标签,得到初始社区划分.将社区看作节点,社区间连边数作为边权重,得到社区关系网络.按照参与系数由低到高的顺序合并社区关系网络中的节点,得到最终社区划分结果.算法 LPA-TS 减少了传统 LPA 方法在节点更新和标签传播过程的随机性;在第 2 阶段,将不符合弱社区定义的初始社区与连边最多的相邻社区合并,再按照社区参与系数由低到高的顺序合并初始社区提升社区发现质量.通过与一些经典算法在 8 个真实网络及不同参数下 LFR benchmark 人工网络数据集上的实验比较表明 LPA-TS 算法表现了良好的稳定性,在 NMI、ARI、模块性等方面表现良好.

**关键词** 复杂网络;社区发现;标签传播;参与系数;弱社区

**中图法分类号** TP181

复杂网络分析在社会学、传染病学和生物学等领域有着广泛的应用[1-3].通常可以用图 $G=(V,E)$ 表示一个复杂网络,其中 $V$ 表示网络中个体的集合,$E$ 表示个体间联系的集合.社区结构(community structure)是复杂网络的重要特征之一,即一个网络可以分成若干社区,社区内节点之间连接相对紧密,社区间节点连接相对稀疏.有效的社区发现算法可以发现社会网络中的社区结构、生物网络中的蛋白质功能模块等,有助于深入研究各种类型复杂网络的功能模块及其演化特征,对准确地理解并分析复杂系统的拓扑结构及动力学特性具有十分重要的理论意义和应用价值[4-5].

目前复杂网络中的社区发现方法主要有基于图划分的聚类算法[6]、基于谱分析的聚类算法[7]、基于层次的聚类算法[8]和基于密度的聚类算法[9-10]等.Newman 提出了一种基于贪心策略的快速社区发现算法(fast modularity maximization, FMM)[11],以最优化模块性为目标函数进行社区合并和更新.Blondel 等人提出了 BGLL 算法[12],随机选择一个节点作为初始社区,迭代选择使当前社区模块性增长最大节点加入当前社区完成社区扩展过程.由于现实网络包含大量的小规模社区,网络社区内部连接数不一定比社区之间的连接数多,导致模块性不能较好地度量社区划分质量.Bai 等人[13]基于互补熵理论提出了一种度量网络中社区发现质量的目标函数,该函数综合考虑社区内紧密程度和社区间稀疏程度对社区发现结果进行评价,并以公共邻居数度量节点相似性给出了一种图聚类算法 ISCD+.

Raghavan 等人提出了标签传播算法(label propagation algorithm, LPA)[14],该算法中起初每个节点拥有独立的类标签,每次迭代中对于每个节点将其标签更改为其邻居节点中出现次数最多的标签,通过迭代,直到每个节点的标签与其邻居节点中出现次数最多的标签相同,则达到稳定状态,算法结束.此时具有相同标签的节点属于同一个社区.LPA 算法有接近线性时间的复杂度,但划分过程中,节点更新顺序与标签传播过程存在很大的随机性,使划分结果表现了较强的不稳定性.Barber 等人对 LPA 算法进行改进提出算法 LPAm[15],参照随机连接定义节点标签更新方式,使得算法结果具有较高的模块性.然而,LPAm 算法可能陷入局部最优解且结果存在不稳定性.Liu 等人提出 LPAm+算法[16],在 LPAm 算法之后引入后处理步骤,合并一些小社区进一步提高划分结果的模块性.Li 等人基于 LPA 算法提出一种分阶段的社区发现算法 LPA-S[17],依据节点间的相似性更新节点标签,得到初始社区划分;再根据社区相似性进行社区合并,使得最终划分结果中社区内部连边具有较高的密度.

以上算法主要针对 LPA 节点标签更新过程的不确定性进行了改进,并对划分结果进行适当处理以缓解过早陷入局部极值的问题.尽管如此,这些算法没有处理节点标签更新顺序的随机性,使得划分结果仍然存在较大的不稳定性.按合理的顺序选择待更新节点可以提高算法性能并得到稳定的社区划分结果.针对此问题,本文提出了一种基于标签传播的两阶段社区发现算法(a two-stage community detection algorithm based on label propagation, LPA-TS),减少了传统标签传播方法在节点更新和标签传播过程的随机性,可以得到稳定的计算结果;通过与一些经典算法在 8 个真实网络以及不同参数情况下 LFR benchmark 人工网络数据集上的实验比较分析,结果表明 LPA-TS 算法社区发现结果表现了良好的稳定性,且在标准互信息、调整兰德系数、模块性等方面均表现出较好的性能.

## 1 背景知识

一个复杂网络可以用图 $G=(V,E)$来表示,其中节点集 $V=\{v_1,v_2,\cdots,v_n\}$表示网络个体的集合,边集 $E$ 代表网络个体间联系的集合,记作边 $e_{i,j}=(v_i,v_j)$. 令 $n=|V|$且 $m=|E|$. 除非特别声明,本文仅对无向简单图进行讨论. 在图 $G$ 中,节点 $v_i$ 的邻域 $N_G(v_i)$定义为 $N_G(v_i)=\{v_j|(v_i,v_j)\in E\}$,其中 $v_j\in N_G(v_i)$称为节点 $v_i$ 的邻居节点. 节点 $v_i$ 的度为 $d(v_i)=|N_G(v_i)|$,在不引起混淆的情况下,简记为 $d_i$. 假设 $\Omega=\{V_1,V_2,\cdots,V_k\}$是 $V$ 的一种划分,$V_r\in V$ 且 $|V_r|=n_r$,称 $k$ 为该划分中的社区个数. 令 $d_i(V_r)=|\{v_j|(v_i,v_j)\in E$ 且 $v_j\in V_r\}|$表示节点 $v_i$ 与社区 $V_r$ 内节点的连边数.

一个社区结构可以分为强社区和弱社区[18]. 令 $d_i^{\text{in}}(V_r)$ 表示节点 $v_i(\in V_r)$与社区 $V_r$ 内节点连接数,$d_i^{\text{out}}(V_r)$表示节点 $v_i\in V_r$ 与社区 $V_r$ 外部节点连接数. 一般情况下,强社区是指社区中任何一个节点与社区内部节点连接的度大于其与社区外部节点连接的度:

$$\alpha\times d_i^{\text{in}}(V_r)>d_i^{\text{out}}(V_r),\forall i\in V_r. \tag{1}$$

而弱社区是指社区中所有节点与社区内部节点的度数之和大于社区中所有节点与社区外部节点连接的度数之和:

$$\alpha\times\sum_{i\in V_r}d_i^{\text{in}}(V_r)>\sum_{i\in V_r}d_i^{\text{out}}(V_r). \tag{2}$$

默认取 $\alpha=2$. 通常一个社区应该至少表现弱社区的性质.

2017 年,Bertolero 等人定义了节点的参与系数[19],用来刻画节点与网络中不同社区连边的分布情况:

$$PC_i=1-\sum_{r=1}^{k}\left(\frac{d_i(V_r)}{d_i}\right)^2. \tag{3}$$

参与系数的值越高则表示节点与较多社区存在连边,该节点对某社区的归属度比较低;相反,值越低表示节点的连边情况越集中于少数社区,则该节点对某社区的归属度较高. 从具有明显社区归属的节点开始进行社区发现,有助于提高社区发现质量,并提高算法稳定性.

为了对社区划分结果进行度量,2004 年 Newman 等提出了模块性[20]的概念,它反映了网络社区内部连接的强弱,作为一种社区划分的评价标准得到了广泛使用. 将网络用邻接矩阵 $\boldsymbol{A}$ 来表示,若节点 $x$ 与 $y$ 直接相连,则 $A_{xy}=1$,否则 $A_{xy}=0$. 模块性定义为

$$Q=\frac{1}{2m}\sum_{ij}\left(\boldsymbol{A}_{ij}-\frac{d_id_j}{2m}\right)\delta_{i,j}, \tag{4}$$

其中,$m$ 是网络中边的总数,$\frac{d_id_j}{2m}$表示随机网络中节点 $i$ 和节点 $j$ 的连边数的期望,当节点 $i$ 和节点 $j$ 属于同一个社区时,$\delta_{i,j}=1$,否则 $\delta_{i,j}=0$. 在网络社区内部和社区之间呈现随机连接且社区内部连接较社区间连接稠密的情况下,尤其是网络中社区规模呈高斯分布时,模块性指标能够较好地度量网络社区划分的质量,模块性越高,社区划分质量越高. 然而,现实网络包含大量的小规模社区,网络社区内部连接数不一定比社区之间的连接数多,导致模块性在某些情况下不能较好地反映社区划分质量.

为了合理地对复杂网络中的节点进行社区发现,将节点间相似性作为衡量节点连接紧密程度的重要标准. 目前已经有一些基于网络拓扑特征的相似性度量函数[21-22]. 公共邻居数(common neighbors,CN)度量[21]认为 2 个节点间的公共邻居节点越多,则它们在结构上越相似,连接越紧密:

$$CN(x,y)=|NG_x\cap NG_y|. \tag{5}$$

## 2 一种基于标签传播的两阶段社区发现算法

基于节点参与系数与节点相似性,本文给出一种基于标签传播的两阶段社区发现算法(LPA-TS). 算法包括 2 个主要过程:1)根据节点参与系数定义节点的更新顺序,并更新节点标签为与其具有最高相似性的邻居节点标签,得到社区的初始划分结果;2)将当前社区进行合并,并在目标函数的监督下完成社区划分的过程. 第 1 阶段中首先根据节点的参与系数高低,从低到高确定节点的更新顺序;其次依据节点相似性,选择与当前遍历节点最相似的邻居节点的标签作为当前遍历节点的标签,得到第 1 阶段的划分结果. 第 2 阶段中,将社区作为节点计算其参与系数以确定社区合并顺序,最后在目标函数的监督下将社区进行合并得到最终的划分结果.

### 2.1 节点更新顺序

在 LPA 算法中,不同的节点更新顺序会使得最终的社区划分结果有很大差异. 如图 1 所示,可以看出,图 1 中 7 个节点应该被分成 2 个社区. 算法初始将每个节点看作 1 个单独社区,假设当前虚线框住的节点已经被赋予了相同的社区标签. 随后,若首先

选择节点 $v_4$ 进行标签更新,有很大可能将节点 $v_4$ 与节点$\{v_1,v_2,v_3\}$划分到同一社区,进而将所有节点划分为 1 个社区.而如果选择节点 $v_5$ 进行更新则会有很大可能得到正确的社区划分.这是由于节点 $v_4$ 位于 2 个社区的边界处,对社区的归属度不强,对其首先更新容易将 2 个社区合并为 1 个大社区.

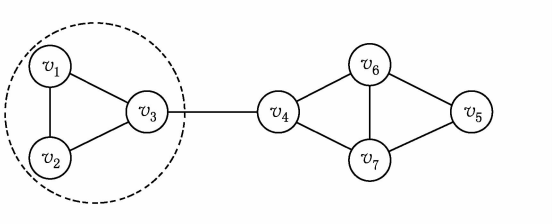



Fig. 1 An example network with two communities

图 1 具有 2 个社区的网络示例

可以看出,在节点标签更新的过程中,如果先更新归属度较强的社区内部节点的标签,会获得一个更符合实际社区结构且更稳定的划分结果.

可以用节点的参与系数 $PC_i$,见式(3),定义节点与社区的归属度,参与系数值越小,节点的社区归属度越高.如图 1 所示,此时,节点 $v_5$ 分别与社区$\{v_6\}$和$\{v_7\}$有边连接,故其参与系数 $PC_{v_5}=1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right]$,有 $PC_{v_5}=0.5$.而对节点 $v_4$,节点 $v_6$ 以及节点 $v_7$ 均与 3 个社区各存在 1 条连边,计算参与系数可得 $PC_{v_4}=PC_{v_6}=PC_{v_7}=0.667$.

从参与系数的定义看,一个节点的度越低,其社区归属程度越高;而一个节点的邻居节点社区归属越集中,其社区归属程度越高.优先选择参与系数低的节点进行更新,可以尽早得到更稳定的社区结构,进而得到更准确的社区划分结果.如图 1 最终得到 2 个社区划分$\{v_1,v_2,v_3\}$和$\{v_4,v_5,v_6,v_7\}$.

### 2.2 标签传播

LPA 算法在对节点的标签进行更新时,选取邻居节点的标签中出现次数最多的标签为自身标签,即认为其所有邻居节点的重要性相同,并没有考虑不同邻居节点的不同相似性.然而,在同一个社区中的个体往往具有较高的相似性.如在图 1 中,节点 $v_4$、节点 $v_6$ 和节点 $v_7$ 具有相同 $PC$ 值,对于节点 $v_4$,其邻居节点的标签出现次数均为 1,则有较大可能将 $v_4$ 划分入社区$\{v_1,v_2,v_3\}$,导致错误的划分结果.

实际上,对节点 $v_4$ 而言,社区$\{v_1,v_2,v_3\}$,$\{v_6\}$和$\{v_7\}$中各有一个节点与其关联,从图 1 中可以看出,由于 $v_4$ 与 $v_6$(或 $v_7$)有一个公共邻居节点,因此相较于节点 $v_3$,$v_4$ 更有可能与 $v_6$(或 $v_7$)属于同一社区.

因此,用 CN 对节点间的相似性进行度量,2 个节点间的公共邻居节点越多,则它们在结构上越相似,连接越紧密,在标签更新的过程中选择与其相似性最高的邻居节点的标签,可使最终划分在同一个社区中的节点具有较高的相似性,也更符合实际的社区分布.如在图 1 中,节点 $v_4$ 确定标签时,由于其与节点 $v_6$ 和节点 $v_7$ 的相似性高于其他节点,故标签会确定为节点 $v_6$ 或节点 $v_7$ 的标签,得到正确的划分.但是公共邻居数度量节点相似性在某些特殊情况下并不适用.例如若节点 $x$ 和节点 $y$ 之间存在连边,但无公共邻居节点,但它们之间的相似性应该大于 0;特别地,一个悬挂点与其相邻点之间无公共邻居节点,但通常与其相邻点位于同一社区.基于此,本文对节点相似性计算为

$$CN(x,y)=|NG_x\cap NG_y|+\frac{1}{d_x d_y}. \tag{6}$$

### 2.3 初始社区发现过程

根据节点更新顺序和标签传播过程,算法给出网络初始划分结果.首先将网络中的每个节点看作一个独立社区,赋予唯一社区标签;计算节点的参与系数 $PC_i(1\leqslant i\leqslant n)$,按从低到高的顺序依次更新节点标签.节点标签更新过程中,考虑当前节点与其邻居节点的公共邻居相似性,选择相似性最高的邻居节点标签作为当前节点的新标签.以上过程迭代进行,直到划分结果不再变化或者达到最大迭代次数.算法 1 给出 LPA-TS 算法的第 1 阶段即初始社区发现过程.

**算法 1.** 初始社区发现算法(LPA-TS 第 1 阶段).

输入:网络 $G=(V,E)$、最大迭代次数 $t_{\max}$,其中 $V=\{v_1,v_2,\cdots,v_n\}$,$\boldsymbol{A}$ 是图 $G$ 的邻接矩阵;

输出:网络的初始划分结果 $L(V)=\{l_1,l_2,\cdots,l_n\}$,其中,$l_i$ 表示节点 $v_i$ 的初始划分社区标签.

步骤 1. 根据

$$CN(V_i,V_j)=|NG_{v_i}\cap NG_{v_j}|+\frac{1}{d_i d_j},$$

计算网络中节点 $v_i$ 和 $v_j$ 间的相似性.

步骤 2. 令 $t=0$.令 $l_i^{(t)}=i$,为节点赋唯一初始社区标签.

步骤 3. 计算当前节点标签集合中不同的标签数 $k=|L(V)|$.

步骤 4. 根据

$$PC_i = 1 - \sum_{r=1}^{k} \left( \frac{d_i(V_r)}{d_i} \right)^2,$$

计算节点 $v_i$ 的参与系数.

步骤 5. 按参与系数对节点进行排序 $\{v_{i_1}, v_{i_2}, \cdots, v_{i_n}\}$,满足 $PC_{i_1} \leqslant PC_{i_2} \leqslant \cdots \leqslant PC_{i_n}$.

步骤 6. 对 $1 \leqslant x \leqslant n$,计算 $j = \arg\max_{v_j \in NG(v_{i_x})} CN(v_{i_x}, v_j)$,令 $l_{i_x}^{(t+1)} = l_j^{(t)}$.

步骤 7. $t = t+1$.

步骤 8. 若 $t > t_{\max}$ 或对所有的 $1 \leqslant j \leqslant n$,有 $l_j^{(t+1)} = l_j^t$,则算法结束;否则返回步骤 3.

图 2 给出了在 Karate 网络上的初始社区发现结果.其中节点被初始划分为 4 个社区(用不同形状表示).可以看出,算法 1 的初始社区划分结果中,度数较大节点由于与邻居节点的相似性较高,容易将自身标签传播给邻居节点,从而形成以大度节点为中心的较大社区.而位于社区边缘的一些节点,由于度数偏低且与邻居节点的相似性小,容易形成一些规模较小的社区,如图 2 中菱形节点和三角形节点构成的社区.

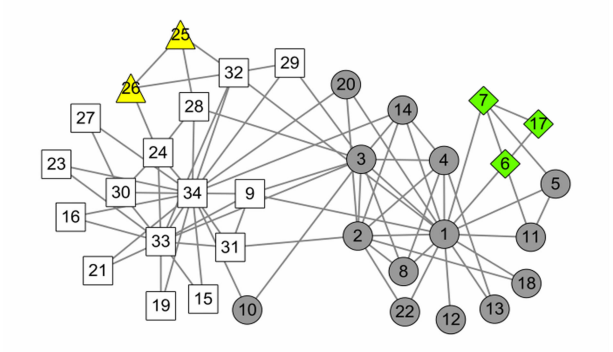



Fig. 2 Communities of the Karate network discovered by Step 1 of LPA-TS

图 2 LPA-TS 第 1 阶段在 Karate 网络的初始社区发现结果

随着网络规模的增大,特别是网络连接比较稀疏时,算法第 1 阶段会得到大量的特别小规模的社区,造成对原始网络的过划分.为了得到更准确的社区划分结果,还需要对初始社区划分结果进行社区合并.

## 2.4 社区合并过程

针对初始社区发现过程网络过度划分的问题,首先分析得到的初始社区划分结果中,根据式(2)依次判断各初始社区是否满足弱社区的定义.通常情况下,式(2)中的内部度系数 $\alpha = 2$,表示一个社区的内部度大于其外部连边数则为弱社区.当所分析网络中包含大量小规模社区时,特别是社区个数远大于单个社区规模时,一个社区的内部度可能会小于其外部连边.为更好地反映网络的社区组成,此时,需要根据网络类型适当提高式(2)中的内部度系数 $\alpha$.

将不满足弱社区定义的初始社区,与其相邻的且具有最多关联边数的社区合并为一个社区.在此为基础上继续进行 LPA-TS 算法第 2 阶段的标签传播过程.

将以上所得的每个社区看作一个节点,社区之间连边数作为 2 个社区对应节点连边的权重,给出该带权无向网络中节点参与系数的定义方式:

$$PC_{s_i} = 1 - \frac{\sum_{r=1}^{n'} (w_{ir})^2}{\left( \sum_{j=1}^{n'} w_{ij} \right)^2}. \tag{7}$$

其中,$n'$表示当前网络节点数,$w_{ij}$ 表示节点 $s_i$ 与节点 $s_j$ 的边权重,则 $w_{ij} = \sum_{v_x \in s_i, v_y \in s_j} A_{xy}$,令 $w_{ii} = 0$.节点 $s_i$ 的参与系数越高,表明该节点对应社区与较多的社区存在连边.相反,如果节点 $s_i$ 的参与系数低,表明该节点对应社区仅与少数社区存在连边,与其他社区合并成为更大社区的可能性更高.

因此,此处仍然按社区对应节点的 $PC$ 值从低到高的顺序进行标签更新.在节点 $s_i$ 的标签更新过程中,选择与其具有最高连边权重的邻居节点标签作为 $s_i$ 的新标签.每次节点标签的更新都意味着合并 2 个初始社区.

为了对节点标签传播过程进行有效控制,需要从社区内部连接紧密程度以及社区间连接的稀疏程度同时考虑社区发现质量,因此本文选择文献[13]提出的基于互补熵的评价函数,对当前社区合并结果进行评价:

$$F(\Omega) = \sum_{i=1}^{n'} \delta_i I_i. \tag{8}$$

其中,$\delta_i = \sqrt{\sum_{r=1}^{n'} \left( \frac{f_i(V_r)}{\sum_{j=1}^{n'} f_i(V_j)} \right)^2}$ 表示节点 $v_i$ 与其他社区的分离度.$f_i(V_r) = \frac{d_i(V_r)}{|V_r|}$,其中 $|V_r|$ 表示社区 $V_r$ 的节点个数.而 $I_i = \sum_{r=1}^{n'} d_i(V_r) \times f_i(V_r)$ 用来评价节点 $v_i$ 与其他社区的紧密程度.

算法 2 给出了 LPA-TS 第 2 阶段社区合并的具体过程.

**算法 2.** 社区合并过程(LPA-TS 第 2 阶段).

输入:网络 $G=(V,E)$,其中 $V=\{v_1,v_2,\cdots,v_n\}$,网络的初始划分结果 $L(V)=\{l_1,l_2,\cdots,l_n\}$;

输出:网络的最终划分结果 $\Omega=\{V_1,V_2,\cdots,V_k\}$.

步骤 1. 对于上一阶段划分出的社区中不符合弱社区定义的社区,选择与其连边数最多的社区进行合并,令 $\Omega_0=\{V_1^{(0)},V_2^{(0)},\cdots,V_{n'}^{(0)}\}$ 为合并后所得的社区集合,$n'=|\Omega_0|$,令 $t=0$.

步骤 2. 将每个 $V_i^{(t)}$ 看作一个节点 $s_i$,社区间的连边数作为边权重,得到一个新网络 $G'=(V',E',W)$,其中 $V'=\{s_1,s_2,\cdots,s_{n'}\}$. 若节点 $s_i$ 和 $s_j$ 对应社区之间存在连边,则边 $(s_i,s_j)\in E'$,且 $w_{ij}=\sum_{v_x\in s_i,v_y\in s_j}A_{xy}$. 计算当前网络划分的评价函数值 $F(\Omega_t)$.

步骤 3. 根据式(7)计算每个节点 $s_i$ 的参与系数. 令 $s_m=\arg\max_{s_i\in V'}PC_{s_i}$.

步骤 4. 对计算 $j=\arg\max_{s_j\in V'}w_{mj}$,合并社区 $V_m^{(t)}$ 与 $V_j^{(t)}$,令 $n'=n'-1$,得到新的社区划分结果 $\Omega_{t+1}=\{V_1^{(t+1)},V_2^{(t+1)},\cdots,V_{n'}^{(t+1)}\}$.

步骤 5. 计算当前网络划分的评价函数值 $F(\Omega_{t+1})$.

步骤 6. 若 $F(\Omega_{t+1})>F(\Omega_t)$,令 $t=t+1$,返回步骤 2;否则,算法 2 结束,返回 $\Omega_{t+1}$ 为最终结果.

### 2.5 时间复杂度分析

本文提出的基于标签传播的两阶段社区发现算法 LPA-TS 包括2 个主要过程:

1) 根据参与系数定义节点的更新顺序,并将节点标签更新为与其具有最高相似性的邻居节点标签,得到社区的初始划分结果;

2) 将初始划分结果中的小规模社区与其有最多连边的相邻社区进行合并;本文采用基于互补熵的评价函数 $F(\Omega)$ 作为目标函数对社区发现结果进行判断,得到 $F(\Omega)$ 值最大的社区发现结果作为算法最终输出.

算法 1 中,计算节点相似性的代价为 $O(n^2)$,计算节点参与系数的代价为 $O(n^2)$;在算法 2 中,假设当前网络中的社区数为 $n'$,计算各社区间的连边数的代价为 $O(n'^2)$;2 阶段的标签更新的代价为 $O(t\times(n+n'))$,$t$ 为算法的迭代次数,因此算法 LPA-TS 的总时间复杂度为 $O(n^2+n'^2+t\times(n+n'))$.

## 3 实验结果与分析

选择不同参数情况下的 LFR benchmark 人工网络[23]和 8 个经典真实网络用本文算法 LPA-TS 进行社区发现,并选择 LPA, LPAm, LPAm+, LPA-S, BGLL, ISCD+, FMM, Infomap[24] 等包括经典的标签传播算法和以模块性为优化目标的算法作对比进行性能比较.

### 3.1 评价指标

对于有标签的网络选择标准互信息($NMI$)[25]与调整兰德系数($ARI$)[26]评价算法聚类结果与网络原始标签的吻合程度. 对于一个有 $n$ 个节点的集合 $V$,$C=\{C_1,C_2,\cdots,C_k\}$ 表示一个网络划分结果,$P=\{P_1,P_2,\cdots,P_{k'}\}$ 表示网络原始划分,对集合 $C_i$ 和 $P_j$($1\leqslant i\leqslant k$,$1\leqslant j\leqslant k'$),令 $n_{ij}=|C_i\cap P_j|$,$b_i=\sum_{j=1}^{k'}n_{i,j}$,$d_j=\sum_{i=1}^{k}n_{i,j}$. $NMI$ 和 $ARI$ 的定义为

$$NMI=\frac{2\sum_i\sum_j n_{ij}\ \mathrm{lb}\left(\frac{n_{ij}n}{b_id_j}\right)}{-\sum_i b_i\ \mathrm{lb}\left(\frac{b_i}{n}\right)-\sum_j d_j\ \mathrm{lb}\left(\frac{d_j}{n}\right)},\tag{9}$$

$$ARI=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{b_i}{2}\sum_j\binom{d_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{b_i}{2}+\sum_j\binom{d_j}{2}\right]-\left[\sum_i\binom{b_i}{2}\sum_j\binom{d_j}{2}\right]\Big/\binom{n}{2}},\tag{10}$$

划分结果与原始划分的吻合程度越高,$NMI$ 和 $ARI$ 的值越高. 如果划分结果与网络随机划分结果相差越大,$ARI$ 的值更高.

为了对算法划分结果的稳定性进行评价,本文定义算法的稳定系数 $\sigma$. 对网络 $G$,若 $|V(G)|=n$,对第 $t$ 次计算结果 $\Omega_t$ 构造 $n\times n$ 阶的矩阵,其元素定义为

$$\delta_{i,j}^{(t)}=\begin{cases}1, & \text{节点 } i \text{ 和 } j \text{ 属于同一社区},\\ 0, & \text{否则}.\end{cases}\tag{11}$$

若算法运行 $T$ 次,可得到计算结果的方差矩阵 $\boldsymbol{S}$,其元素定义为

$$S_{i,j}=\frac{1}{T}\sum_{t=1}^{T}\left(\delta_{i,j}^{(t)}-\frac{1}{T}\sum_{x=1}^{T}\delta_{i,j}^{(x)}\right)^2,\tag{12}$$

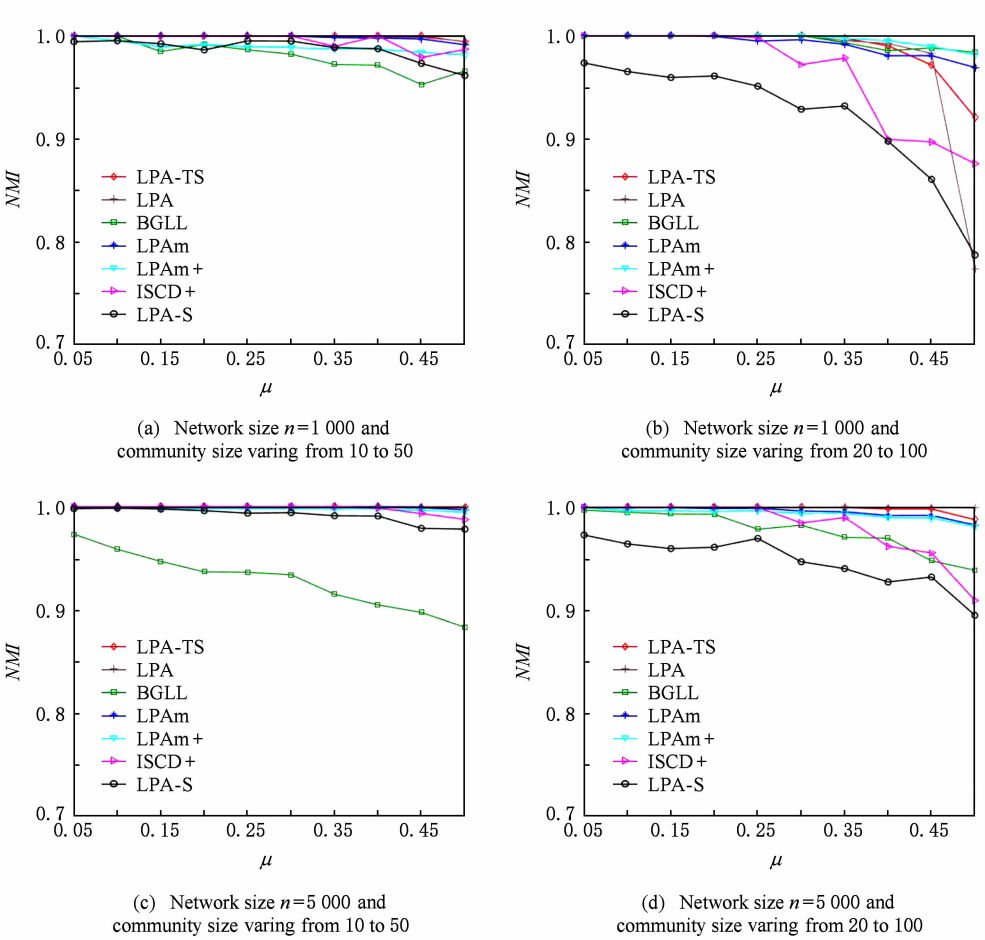
由此,将算法稳定系数 $\sigma$ 定义为

$$\sigma = \frac{1}{n^2} \sum_{i,j \in [1,n]} S_{i,j}. \tag{13}$$

一个算法多次运行结果的算法稳定系数越低,说明算法划分结果越稳定.

### 3.2 人工网络实验

人工网络实验采用在社区发现算法性能检测中广泛采用的 LFR benchmark 数据集上进行,分别考察网络规模 $n$ 为 1 000 或 5 000,社区规模区间为 10～50 或 20～100,混合参数 $\mu$ 为 0.05～0.5 的各种不同参数下 LFR 人工网络上本文算法 LPA-TS 与对比算法的性能比较. 所有实验网络节点平均度为 20,最大度为 50,节点度序列满足指数为 2 的幂律分布,社区规模序列满足指数为 1 的幂律分布. 取各算法在每个网络上执行 30 次的结果取平均值进行比较,结果如图 3 和图 4 所示. 可以看到,混合参数较低($\mu$ 为 0.05～0.3)时,网络的社区结构比较明显,各算法都取得了与实际网络中社区分布高度吻合的结果. 随着 $\mu$ 的增大,网络中社区结构明显性降低,由于在本文 LPA-TS 算法中,选择参与系数较低的节点进行标签传播,确保算法在保持社区发现质量的同时,减少了节点标签传播过程中的随机性,从而使得算法运行结果比较稳定.




(a) Network size $n$=1 000 and community size varing from 10 to 50

(b) Network size $n$=1 000 and community size varing from 20 to 100

(c) Network size $n$=5 000 and community size varing from 10 to 50

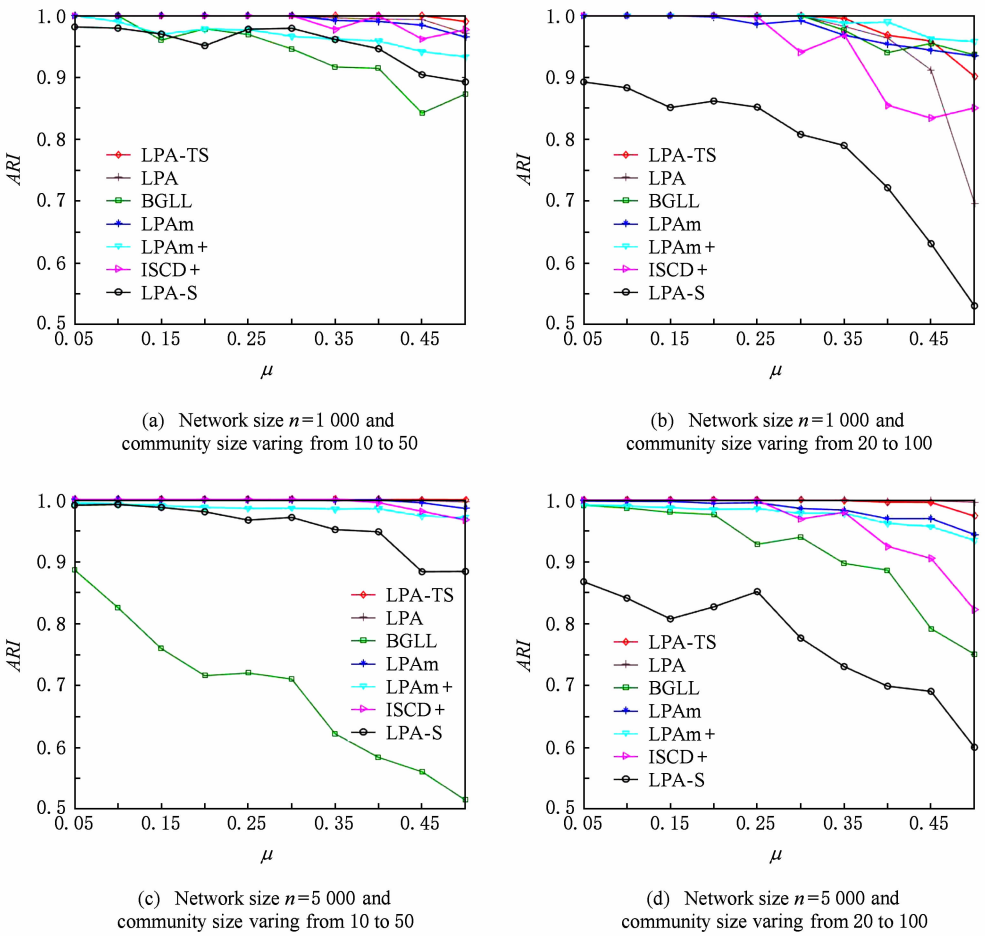
(d) Network size $n$=5 000 and community size varing from 20 to 100

Fig. 3 Comparison of *NMI* on LFR benchmark networks

图 3 各算法在 LFR benchmark 网络上的 *NMI* 比较结果

实际上,本文也与基于随机游走的社区发现算法 Infomap 进行了对比实验. Infomap 算法在本文实验的 LFR benchmark 网络上社区发现结果与初始生成的社区划分结果完全吻合. 这是由于 LFR benchmark 网络生成过程中,其社区内部连边和社区之间连边分别采用随机连接方式产生. 这导致网络中各社区内部连边分布比较均匀,2 节点之间的连边概率与其度数成正相关关系. 所以 LFR benchmark 网络中社区分布比较平衡,网络结构也相对稳定. 因此,在 LFR benchmark 网络上,本文所提算法 LPA-TS 和算法 LPA-S 在各类算法比较中的优势没有在真实网络的实验结果表现明显.

然而,真实网络的社区构成更加多样化,社区内部连接和社区间连接也体现了很大的非平衡性. 因此,我们进一步在经典的真实网络上对各算法性能进行了比较.




(a) Network size $n$=1 000 and community size varing from 10 to 50

(b) Network size $n$=1 000 and community size varing from 20 to 100

(c) Network size $n$=5 000 and community size varing from 10 to 50

(d) Network size $n$=5 000 and community size varing from 20 to 100

Fig. 4 Comparison of *ARI* on LFR benchmark networks

图 4 各算法在 LFR benchmark 网络上的 *ARI* 比较结果

### 3.3 真实网络比较实验

在本节中,我们通过在空手道俱乐部(Zachary's Karate Club)[27]、海豚社交网络(Dolphins Social Network)[28]、Polbooks[13]和大学生足球网络(American College Football Network)[29]四个带标签的真实网络以及 Les Misérables[30],NetScience[31],Email[32]和 Yeast[33]四个无标签的真实网络上进行实验以对算法进行评测. 数据基本情况如表 1 所示:

**Table 1 Real Network Datasets**

**表 1 真实网络数据集**

| Data Set | Number of Nodes | Number of Edges | Number of Communities |
|---|---|---|---|
| Karate | 34 | 78 | 2 |
| Dolphins | 62 | 159 | 2 |
| Polbooks | 105 | 441 | 3 |
| Football | 115 | 613 | 12 |
| Les Misérables | 77 | 254 | |
| NetScience | 379 | 914 | |
| Email | 1 133 | 5 451 | |
| Yeast | 2 375 | 11 693 | |

由于 LPA-TS 算法引入了弱社区判断,以适应复杂的真实网络中多样性的社区变化,使其在真实网络上表现良好. 算法比较结果如表 2 和表 3 所示,其中每个算法运行 30 次,取各项指标的平均值做比较. 评价指标 $k$ 表示算法最终发现的社区个数,$Q$ 是模块性指标,*time* 表示算法运行时间,$\sigma$ 是算法稳定系数.

表 2 给出了算法 LPA-TS 与 8 种经典社区发现算法 FMM, LPA, BGLL, LPAm, LPAm+, Infomap, ISCD+, LPA-S 在有标签网络数据集的实验比较结果. 可以看出,当网络规模较小时,各算法所发现的社区个数与实际社区数吻合得较好,算法稳定性也表现良好;随着网络规模的逐渐增大,本文算法 LPA-TS 在社区个数和结果稳定性方面表现突出.

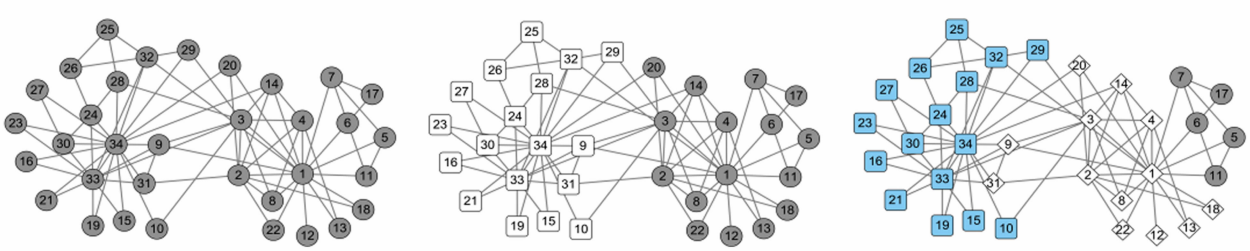
图 5 给出了经典 LPA 算法在 Karate 网络上的 3 种不同的划分结果,由于其在节点更新顺序上的随机性,LPA 对于网络的划分结果存在较大的不稳定性,甚至在划分过程中会将整个网络中的节点划分在一个社区中.

**Table 2 Comparison of Real Networks with Labels**

**表 2 带标签真实网络实验结果对比表**

| Data Set | Index | FMM | LPA | BGLL | LPAm | LPAm+ | Infomap | ISCD+ | LPA-S | LPA-TS |
|---|---|---|---|---|---|---|---|---|---|---|
| Karate | $k$ | **2** | 1-3 | 3-4 | 4-7 | 4 | 3 | **2** | 2-3 | **2** |
| | $ARI$ | 0.882 3 | 0.663 2 | 0.544 5 | 0.409 2 | 0.524 5 | 0.702 2 | **1.000 0** | 0.956 0 | 0.882 3 |
| | $NMI$ | 0.837 2 | 0.657 4 | 0.653 7 | 0.577 1 | 0.652 1 | 0.699 5 | **1.000 0** | 0.942 6 | 0.837 2 |
| | $Q$ | 0.371 8 | 0.314 7 | 0.407 4 | 0.371 0 | **0.416 4** | 0.402 0 | 0.371 5 | 0.369 0 | 0.371 6 |
| | $\sigma$ | | 0.121 4 | 0.048 7 | 0.043 8 | 0.025 6 | | | 0.013 1 | **0** |
| | Running Time/ms | 33 | 9 | 10 | 14 | 49 | 3 | 3 | 21 | 13 |
| Dolphins | $k$ | 3 | 2-5 | 4-5 | 8-11 | 4-6 | 6 | 4 | 2-6 | **2** |
| | $ARI$ | 0.479 5 | 0.510 6 | 0.425 9 | 0.230 8 | 0.335 5 | 0.361 4 | 0.345 8 | 0.644 2 | **0.948 1** |
| | $NMI$ | 0.605 8 | 0.634 9 | 0.519 5 | 0.443 7 | 0.508 6 | 0.527 0 | 0.470 8 | 0.682 0 | **0.969 8** |
| | $Q$ | 0.494 2 | 0.492 0 | 0.520 2 | 0.498 8 | **0.522 6** | 0.515 8 | 0.491 7 | 0.437 3 | 0.375 9 |
| | $\sigma$ | | 0.060 5 | 0.031 0 | 0.017 3 | 0.030 8 | | | 0.078 5 | **0.007 9** |
| | Running Time/ms | 40 | 20 | 13 | 22 | 73 | 3.4 | 3 | 39 | 35 |
| Polbooks | $k$ | **3** | 1-4 | 3-5 | 8-12 | 4-8 | 5 | **3** | 2-6 | 2 |
| | $ARI$ | 0.656 3 | 0.492 1 | 0.596 6 | 0.318 3 | 0.536 6 | 0.646 3 | 0.639 0 | 0.651 4 | **0.667 1** |
| | $NMI$ | 0.556 6 | 0.423 1 | 0.523 4 | 0.432 0 | 0.506 3 | 0.536 9 | 0.524 5 | 0.554 3 | **0.597 9** |
| | $Q$ | 0.499 3 | 0.380 1 | 0.521 0 | 0.489 0 | 0.519 7 | **0.526 7** | 0.497 3 | 0.497 1 | 0.456 9 |
| | $\sigma$ | | 0.122 8 | 0.025 9 | 0.036 7 | 0.044 1 | | | 0.037 2 | **0** |
| | Running Time/ms | 76 | 31 | 17 | 36 | 90 | 4.5 | 3 | 56 | 50 |
| Football | $k$ | 5 | 8-13 | 9-10 | 10-14 | 9-10 | 10 | 13 | 7-18 | **12** |
| | $ARI$ | 0.444 4 | 0.739 0 | 0.768 1 | 0.823 2 | 0.777 5 | 0.760 1 | 0.889 0 | 0.680 1 | **0.889 3** |
| | $NMI$ | 0.686 2 | 0.872 5 | 0.874 0 | 0.902 3 | 0.878 1 | 0.880 1 | 0.925 4 | 0.8406 | **0.926 9** |
| | $Q$ | 0.549 4 | 0.581 9 | **0.603 7** | 0.582 1 | 0.601 9 | 0.590 2 | 0.594 9 | 0.529 1 | 0.601 0 |
| | $\sigma$ | | 0.024 5 | 0.010 3 | 0.009 6 | 0.008 1 | | | 0.029 6 | **0** |
| | Running Time/ms | 66 | 36 | 17 | 46 | 112 | 5.2 | 12 | 108 | 59 |

Notes: Bolded part indicates the best result among the 9 algorithms.



(a) A result with only 1 community (b) A result with 2 communities (c) A result with 3 communities

Fig. 5 Different results of the LPA on Karate network

图 5 LPA 算法在 Karate 网络上的 3 种不同的划分结果

图 6 给出了本文算法 LPA-TS 在 Karate 网络上的划分结果,将该网络稳定地划分为 2 个社区.

图 7 给出了本文算法 LPA-TS 在 Dolphins 网络上的一种划分结果,该网络表示 62 只宽吻海豚之间相互联系的情况,节点表示海豚,若 2 只海豚间存在频繁联系则对应节点间存在边,Dolphins 网络中

有 2 个社区标签. LPA-TS 可以得到 2 种不同的划分结果,这 2 种划分结果的区别仅在于节点 40(图 7 中虚线圈出的节点)的社区归属.

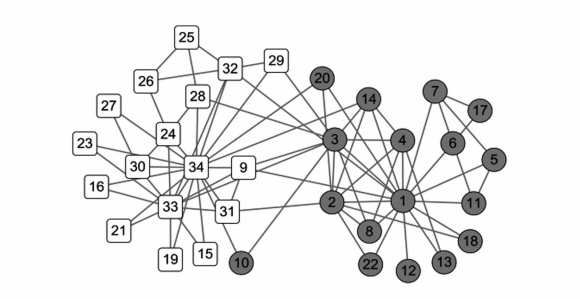


Fig. 6 Result of LPA-TS on Karate network

图 6 LPA-TS 在 Karate 网络上的划分结果

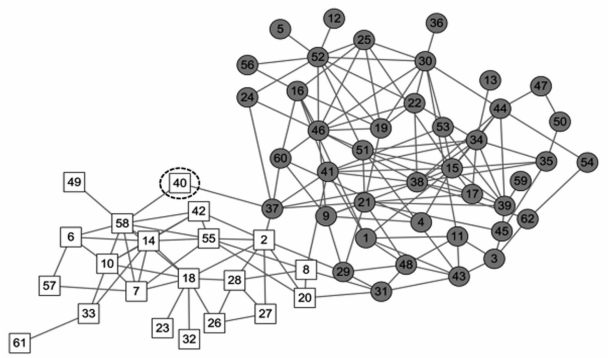


Fig. 7 Result of LPA-TS on Dolphins network

图 7 LPA-TS 在 Dolphins 网络上的划分结果

图 8 给出了 Polbooks 网络的原始划分情况. 该网络节点表示在 Amazon 在线书店上销售的与美国政治相关的图书,如 2 本图书曾被同一用户购买过,则对应节点之间存在一条无向边. 这些图书被分为

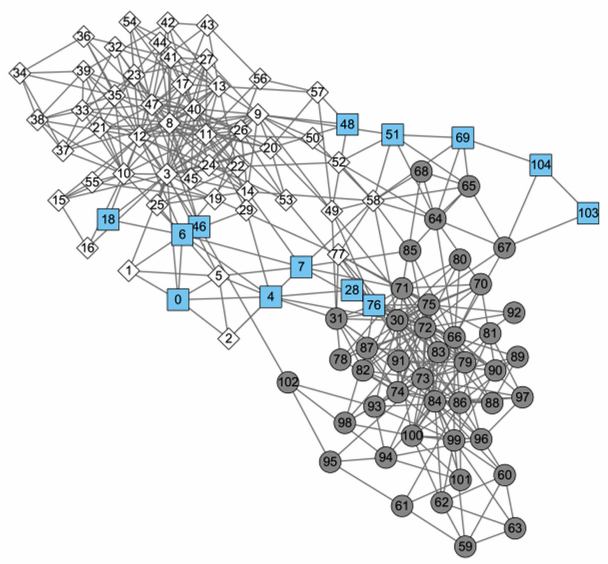


Fig. 8 Primitive division on Polbooks network

图 8 Polbooks 网络原始划分

3 类:“自由派”(圆形节点)、“保守派”(菱形节点)和“中间派”(方形节点). 可以看到,“自由派”和“保守派”这 2 个社区内部连接比较紧密,社区间连边比较稀疏. 而“中间派”节点代表的图书没有明显的政治倾向,其对应的社区结构也并不明显,这给社区发现算法的结果准确性和稳定性带来一定的困难. 从表 2 可以看出,本文算法 LPA-TS 在 *NMI*, *ARI* 等指标上表现更好;与 LPA, BGLL, LPAm, LPAm+, LPA-S 五种结果呈现随机性的算法相比,LPA-TS 的算法稳定系数更低,因此运行结果比较稳定. 图 9 给出了本文算法 LPA-TS 在 Polbooks 网络上所得的最终划分结果,将 Polbooks 网络划分为 2 个社区,且各社区内部的连边都比较稠密,社区间的连接比较稀疏,本文算法得到较为合理的划分.

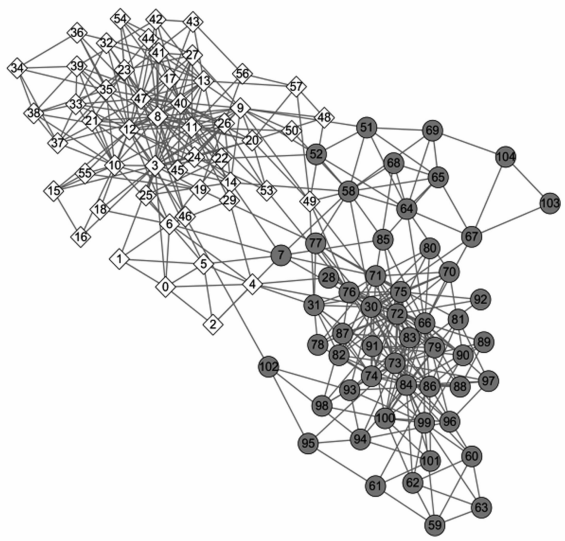


Fig. 9 Result of LPA-TS on Polbooks network

图 9 LPA-TS 在 Polbooks 网络上的划分结果

Football 网络是根据美国大学足球联赛 2000 年一个赛季的比赛赛程而建立的实际网络,共有 12 个足球联盟,每个球队都属于某一个联盟,因此包含 12 个社区标签. 若 2 支队伍之间进行过比赛,则对应节点间存在连边. 图 10 给出了 LPA-TS 在 Football 网络上的划分结果,与真实划分更为接近.

可以看出,在带标签的真实网络数据中,社区划分的模块性并不一定很高,这是因为真实网络社区结构更加多样化,模块性不能完全反映这种情况.

表 3 给出了算法 LPA-TS 与其他 8 种经典社区发现算法在 4 种无标签真实网络上的实验结果,分别从划分结果的模块性和稳定性2 个方面对各算法性能进行比较. 可以看出,本文算法 LPA-TS 的稳定系数较低,说明其具有良好的稳定性.

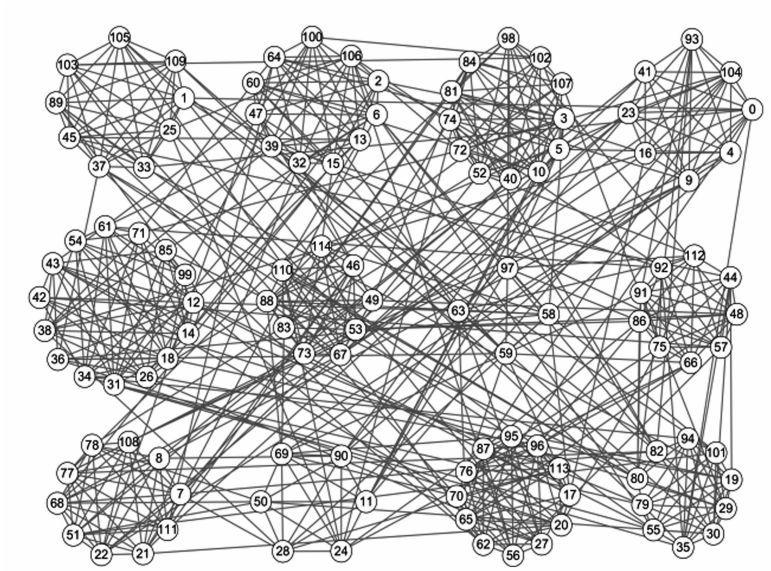


Fig. 10 Result of LPA-TS on Football network

图 10 LPA-TS 在 Football 网络上的划分结果

**Table 3 Comparison of Real Networks without Labels**

**表 3 无标签真实网络实验结果对比表**

| Data Set | Index | FMM | LPA | BGLL | LPAm | LPAm+ | Infomap | ISCD+ | LPA-S | LPA-TS |
|---|---|---|---|---|---|---|---|---|---|---|
| Les Misérables | $k$ | 4 | 2-5 | 5-6 | 10-14 | 6-10 | 8 | 2 | 3-23 | 6 |
| | $Q$ | 0.496 1 | 0.271 9 | 0.546 1 | 0.531 5 | **0.549 9** | 0.536 3 | 0.254 0 | 0.445 8 | 0.500 7 |
| | $\sigma$ | | 0.083 3 | 0.019 3 | 0.010 3 | 0.011 2 | | | 0.053 4 | **0** |
| | Running Time/ms | 51 | 24 | 12 | 25 | 83 | 2.9 | 4 | 54 | 35 |
| NetScience | $k$ | 18 | 31-40 | 18-20 | 68-74 | 64-70 | 7 | 6 | 4-60 | 15-23 |
| | $Q$ | 0.838 5 | 0.808 1 | **0.846 4** | 0.717 5 | 0.729 9 | 0.778 9 | 0.754 6 | 0.753 7 | 0.823 8 |
| | $\sigma$ | | 0.013 0 | 0.005 3 | **0.002 0** | 0.003 5 | | | 0.057 3 | 0.012 6 |
| | Running Time/ms | 477 | 197 | 76 | 501 | 3100 | 19 | 5 | 380 | 253 |
| Email | $k$ | 8 | 1-3 | 8-13 | 92-127 | 88-123 | 50 | 5 | 177-361 | 28-29 |
| | $Q$ | 0.346 1 | 0.000 2 | **0.552 2** | 0.476 3 | 0.484 1 | 0.530 9 | 0.456 3 | 0.345 8 | 0.526 4 |
| | $\sigma$ | | 0.001 5 | 0.051 1 | 0.015 7 | 0.018 1 | | | 0.003 1 | **0.001 1** |
| | Running Time/ms | 130 0 | 300 | 700 | 760 0 | 336 00 | 50 | 90 | 387 00 | 390 0 |
| Yeast | $k$ | 30 | 104-131 | 16-24 | 329-354 | 325-350 | 10 | 4 | 199-656 | 34-75 |
| | $Q$ | 0.702 1 | 0.663 9 | **0.729 1** | 0.642 9 | 0.645 0 | 0.526 9 | 0.519 2 | 0.572 3 | 0.691 8 |
| | $\sigma$ | | 0.025 8 | 0.029 4 | **0.002 1** | 0.002 3 | | | 0.004 5 | 0.015 3 |
| | Running Time/ms | 834 00 | 219 9 | 530 3 | 596 80 | 381 700 | 130 | 600 | 1 045 800 | 127 200 |

Notes: Bolded part indicates the best result among the 9 algorithms.

## 4 总 结

本文提出了一种基于标签传播的两阶段社区发现算法,通过节点的参与系数确定节点更新顺序,并在标签传播过程中依据节点间相似性更新节点标签,得到社区的初始划分结果.判断得到的初始社区是否满足弱社区定义,若不满足则将其与相邻连边最多的社区进行合并.将初始划分得到的社区看作节点,社区之间的连边数作为节点间的边权重,得

到社区关系网络,并按照参与系数由低到高的顺序合并社区关系网络中的节点,得到最终的社区划分结果.

通过与 FMM,LPA,BGLL,LPAm,LPAm+,ISCD+,LPA-S 等经典社区发现算法在不同参数情况下 LFR benchmark 人工网络数据集以及真实网络上的实验比较分析,结果表明:由于 LPA-TS 算法减少了在节点更新和标签传播过程的随机性,社区发现结果表现了良好的稳定性,且在标准互信息、调整兰德系数、模块性等方面均表现出较好的性能.

实际上,现实复杂网络中存在复杂多样的社区分布情况,如存在大量的小规模社区、社区规模分布呈现非平衡性、小社区内部连接数比社区间连接数少、网络结构的动态性等特点,这给复杂网络社区发现问题带来了多方面的挑战,未来将针对复杂网络中这些特殊的社区结构特点,研制有效的社区发现算法.

## 参 考 文 献


[1] Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network [J]. Cell, 2014, 159(5): 1212-1226

[2] Yang Bo, Liu Jiming, Feng Jianfeng. On the spectral characterization and scalable mining of network communities [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(2): 326-337

[3] Fortunato S, Barthélemy M. Resolution limit in community detection [J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(1): 36-41

[4] Li Huijia, Li Huiying, Li Aihua. Analysis of multi-scale stability in community structure [J]. Chinese Journal of Computers, 2015, 38(2): 301-312 (in Chinese)
(李慧嘉, 李慧颖, 李爱华. 多尺度的社团结构稳定性分析 [J]. 计算机学报, 2015, 38(2): 301-312)

[5] Chen Junyu, Zhou Gang, Nan Yu, et al. Semi-supervised local expansion method for overlapping community detection [J]. Journal of Computer Research and Development, 2016, 53(6): 1376-1388 (in Chinese)
(陈俊宇, 周刚, 南煜, 等. 一种半监督的局部扩展式重叠社区发现方法[J]. 计算机研究与发展, 2016, 53(6): 1376-1388)

[6] Newman M E J. Community detection and graph partitioning [J]. Europhysics Letters, 2013, 103(2): 28003

[7] Newman M E J. Spectral methods for community detection and graph partitioning [J]. Physical Review E, 2013, 88(4): 042822

[8] Lin Chunchen, Kang Jiarong, Chen Jyunyu. An integer programming approach and visual analysis for detecting hierarchical community structures in social networks [J]. Information Sciences, 2015, 299: 296-311

[9] Bai Xueying, Yang Peilin, Shi Xiaohu. An Overlapping community detection algorithm based on density peaks [J]. Neurocomputing, 2017, 226: 7-15

[10] Ren Jun, Wang Jianxin, Li Min, et al. Identifying protein complexes based on density and modularity in protein-protein interaction network [J]. BMC Systems Biology, 2013, 7(Suppl4): S12

[11] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 066133

[12] Blondel V D, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008

[13] Bai Liang, Cheng Xueqi, Liang Jiye, et al. Fast graph clustering with a new description model for community detection [J]. Information Sciences, 2017, 388-389: 37-47

[14] 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

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

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

[17] Li Wei, Huang Ce, Wang Miao, et al. Stepping community detection algorithm based on label propagation and similarity [J]. Physica A: Statistical Mechanics and Its Applications, 2017, 472: 145-155

[18] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658-2663

[19] Bertolero M A, Yeo B T T, D'Esposito M. The diverse club [J]. Nature Communications, 2017, 8(1): 1277

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

[21] Lü Linyuan, Zhou Tao. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170

[22] Wang Jie, Liang Jiye, Zheng Wenping. A graph clustering method for detecting protein complexes [J]. Journal of Computer Research and Development, 2015, 52(8): 1784-1793 (in Chinese)

(王杰,梁吉业,郑文萍. 一种面向蛋白质复合体检测的图聚类方法[J]. 计算机研究与发展,2015,52(8):1784-1793)

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

[24] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123

[25] Danon L, Díaz-Guilera A, Duch J, et al. Comparing community structure identification [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005 (9): P09008

[26] Rand W M. Objective criteria for the evaluation of clustering methods [J]. Journal of the American Statistical Association, 1971, 66(336): 846-850

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

[28] Lusseau D, Newman M E J. Identifying the role that animals play in their social networks [J]. Proceedings of the Royal Society B: Biological Sciences, 2004, 271(Suppl6): S477-S481

[29] 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

[30] Yang Gui, Zheng Wenping, Wang Wenjian, et al. Community detection algorithm based on weighted dense subgraphs [J]. Journal of Software, 2017, 28(11): 3103-3114 (in Chinese)
(杨贵,郑文萍,王文剑,等. 一种加权稠密子图社区发现算法[J]. 软件学报,2017,28(11):3103-3114)

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

[32] Guimerà R, Danon L, Díaz-Guilera A, et al. Self-similar community structure in a network of human interactions [J]. Physical Review E, 2003, 68(6): 065103

[33] Qian Yuhua, Li Yebin, Zhang Min, et al. Quantifying edge significance on maintaining global connectivity [J]. Scientific Reports, 2017, 7: 45380

**Zheng Wenping**, born in 1979. PhD. Associate professor. Member of CCF. Her main research interests include graph theory algorithms, bioinformatics.

**Che Chenhao**, born in 1994. Postgraduate. His main research interest is clustering algorithm (714181016@qq.com).

**Qian Yuhua**, born in 1976. Professor and PhD supervisor. Member of CCF. His main research interests include granular computing, social computing and machine learning for big data.

**Wang Jie**, born in 1988. PhD candidate. Student member of CCF. His main research interests include data mining, bioinformatics (xhcwj@sina.com).