

基于标签传播的两阶段社区检测算法

孙学良¹, 王巍², 黄俊恒², 辛国栋², 王佰玲²

(1. 哈尔滨工业大学计算学部, 黑龙江 哈尔滨 150001;

2. 哈尔滨工业大学(威海)计算机科学与技术学院, 山东 威海 264209)

摘 要: 社区检测是复杂网络分析的重要研究任务之一, 其检测结果有助于人们深入理解复杂网络的社区结构, 同时为下游任务提供支持, 如内容推荐、链路检测等。针对复杂网络的社区检测问题, 提出了一种基于标签传播的两阶段社区检测算法——TS-LPA。TS-LPA采用扩展邻域的思想来量化节点的传播能力, 并在此基础上, 利用节点信息和网络中边的权重等信息, 提出了新的评价指标来衡量节点的中心性和节点之间的影响力。所提算法在计算节点中心性的基础上确定了节点标签更新的顺序和种子节点的选择策略, 消除了算法在更新过程中的不稳定。在节点标签更新的过程中, 为了更好地利用邻居节点标签类别来进行标签更新, TS-LPA采用广度优先传播的思想, 提出了第二阶段标签传播方式。当标签开始传播的时候, 待更新节点的所有邻居节点都对该节点的类别标签产生影响, 同时, 为了减轻周围邻居节点对待更新节点的支配程度, 除邻居节点的影响外, 加入附近种子节点对待更新节点的影响, 共同完成节点的标签更新。在不同的真实数据集和人工合成数据集的实验结果分析表明, TS-LPA在消除随机性、表现出较强稳定性的同时, 有效提高了社区检测的质量。

关键词: 社区检测; 标签传播; 节点中心性; 广度优先算法

中图分类号: TP311

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2021099

Two-stage community detection algorithm based on label propagation

SUN Xueliang¹, WANG Wei², HUANG Junheng², XIN Guodong², WANG Bailing²

1. Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China

2. School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China

收稿日期: 2021-08-09; **修回日期:** 2021-10-08

通信作者: 辛国栋, gdxin@hit.edu.cn

基金项目: 国家重点研发计划(2020YFB2009502); 中央高校基本科研业务费专项(HIT.NSRIF.2020098); 山东省重点研发计划(2017CXGC0706)

Foundation Items: The National Key R&D Program of China (2020YFB2009502), The Fundamental Research Funds for the Central Universities (HIT.NSRIF.2020098), The Key R&D Program of Shandong Province (2017CXGC0706)

引用格式: 孙学良, 王巍, 黄俊恒, 等. 基于标签传播的两阶段社区检测算法[J]. 网络与信息安全学报, 2022, 8(2): 139-149.

Citation Format: SUN X L, WANG W, HUANG J H, et al. Two-stage community detection algorithm based on label propagation[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 139-149.

Abstract: Community detection is an important research topic of complex network analysis. The detection results help to understand the community structure of complex networks and provide support for downstream tasks, such as content recommendation, link detection. Considering the challenge of community detection in complex networks, a two-stage community detection algorithm based on label propagation (TS-LPA) was proposed. The TS-LPA algorithm quantified the propagation capability of nodes based on the extended neighborhood. Then a new evaluation index was proposed to measure the probability of influence between nodes using the information of nodes and the weight of edges in the network. Based on the calculation of node centrality, the algorithm determined the updating sequence of node labels and the selection strategy of seed nodes, which eliminated the instability of the algorithm in the updating process. The TS-LPA algorithm used the breadth-first propagation idea and introduced the second-stage label propagation method to improve the quality of community detection. When label began to spread, all neighboring nodes had an influence on the label of the related node. Meanwhile, the influence of neighboring seed nodes was added to complete label updating, in order to reduce the dominance of neighboring nodes on the updated node. The experimental results of different real data sets and synthetic data sets show that TS-LPA algorithm can eliminate randomness and show strong stability while effectively improving the quality of community detection.

Keywords: community detection, label propagation, node centrality, breadth-first search

0 引言

网络是表示事物与事物之间联系的常用工具。现实生活中,各个组织随着时间推移所产生的复杂信息常以网络的形式呈现,如科研协作网、社交关系网、金融交易网络等。在这些网络中,通常存在很多自然形成的社区,社区内部成员与成员之间联系紧密,社区之间联系稀疏。如何有效地从网络中识别出这些社区是社区检测的基本任务,具有重要的现实意义。例如,从科研协作网中识别出具有相近研究领域的学者,或者从社交网络中识别出具有相似兴趣爱好的用户等。

随着复杂网络研究的深入,相当多的社区检测算法被相继提出,许多算法从不同角度来解决社区检测问题^[1]。最开始是分裂算法,它基于去除边的思想将社区分开,但这种全局的方法对计算要求很高。为了克服这个问题,研究人员后续提出了凝聚算法或其他优化算法进行改进。在优化算法方面,基于模块度的算法成为优化算法的主流,但基于模块度的算法不易发现小的社区,计算量过大,因此后续发展出其他的解决方案。

2007 年, Raghavan^[2]提出的标签传播算法(LPA, label propagation algorithm)就是其中的代

表方案, LPA 因其易于实现、概念简单广受欢迎,该算法受传染病模型的启发,通过节点标签的迭代更新直至收敛来检测社区,但算法存在随机性强、运行结果不稳定等问题。

为了使社区检测更加真实地反映节点和社区之间的实际关系,本文在标签传播算法的基础上,消除了原算法的不稳定性,并采用矢量表示和广度优先传播的思想,对标签的传播策略进行了改进,主要工作如下。

1) 提出了基于标签传播的两阶段社区检测算法(TS-LPA, two-stage community detection algorithm based on label propagation),重点优化了标签更新过程。

2) 结合网络局部拓扑结构信息,本文采用扩展邻域的思想对节点的中心性进行度量,并在此基础上提出一种新的评价指标来衡量节点之间的影响概率。

3) 利用广度优先传播的思想,提出了第二阶段标签传播方式,来提高社区检测的质量。

4) 在不同数据集上的实验结果表明, TS-LPA 表现出较强的稳定性,有效提高了社区检测的质量。

1 相关工作

国内外学者在定义、检测和识别现实网络的社区方面做了很多工作,提出了各种算法,根据

社区检测的“方向”，可以分为自上而下的方法和自下而上的方法^[3]。

自上而下的方法: 基于图或者边去除的思想, 将整个网络分成小组, 以检测社区。分裂算法属于该类别。2012年, Prat-Perez 等^[4]提出了一个加权社区聚类算法 (WCC), 并认为好的社区是在所有的社区节点之间形成大量的三角形社区, 因此, 为了定义社区, WCC 测量节点 x 与集合 S 中的节点形成的三角形的比率, 而不是 x 在整个图中形成的三角形的数量。2018年, Qiao 等^[5]提出了 Picaso 的自上而下策略, 根据社区的特征, 有些边的权重会消失, 有些边的权重像山峰一样升起。算法包含两个阶段 (山模型的构建和社区合并更新), 经过实验分析表明, Picaso 算法可以处理大型的网络, 并且具有良好的效率。

自下而上的方法: 从局部结构开始扩展到整个网络, 在这个过程中, 逐步形成各种社区^[6]。许多不同的想法实现了自下而上的检测方法, 如模块度优化^[7-9]、标签传播等。模块度优化是找到一个从指定节点开始的子图, 使子图增加一个节点或者删除一个节点都会降低模块度值, 最终找到的社区满足模块度值最大, 实现社区的检测^[10]。2004年, Newman 首先定义了模块度概念并提出模块度算法^[11], 之后提出 Louvain 算法, 该算法在模块度算法基础上做了进一步的改进^[12], 将层次聚类与模块度优化有效结合, 分为两个阶段进行迭代划分检测。Louvain 算法被认为是当前较为高效的社区检测算法之一; 标签传播的基本思想是利用样本间的关系建立完全图模型, 将图模型中每个节点作为中心节点进行迭代标签更新, 迭代过程中利用一阶邻居标签对中心节点的标签进行更新。2007年, Raghavan 等^[2]将标签传播思想引入社区检测任务, 提出了标签传播算法, 该算法仅以网络结构为指导, 每个节点都用一个唯一的标签初始化, 并且在每一步每个节点都获得了它的大多数邻居当前拥有的标签。通过不断迭代, 密集的节点组倾向于获得相同的标签, 从而形成一个社区。该算法具有线性的时间复杂度, 更适用于社区结构已知的网络。但是 LPA

算法也存在稳定性较差的问题, 其示意如图 1 所示。

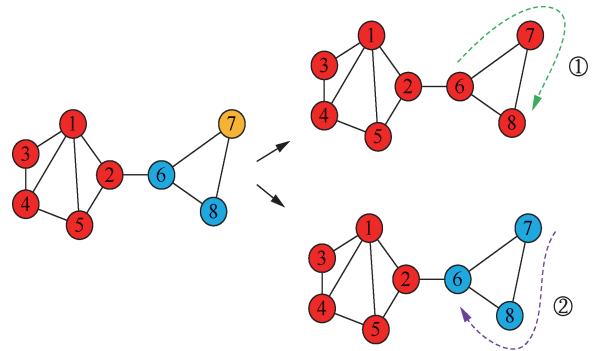


图1 标签传播算法示意

Figure 1 Schematic diagram of label propagation algorithm

从图 1 可知, 根据节点更新顺序的不同, 分别按照①号线条和②号线条更新时, 产生的社区检测结果不同。例如, 当标签按照①号线条传播时, 节点标签更新顺序为 6、7、8, 6 号节点邻居的标签分别为黄色、红色、蓝色节点各一个, 在不同类别邻居标签数量相同的情况下, 6 号节点随机选择了红色, 造成最终整个社区全都为红色节点。此外, 在更新策略方面, 同步更新可能引起标签震荡的问题, 也会造成算法的不稳定^[13], 其结果如图 2 所示。

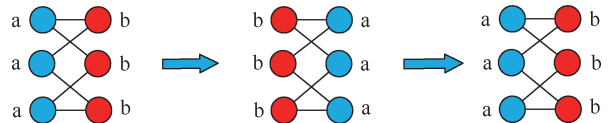


图2 人工网络同步更新的结果

Figure 2 Schematic diagram of manual network label synchronization update

后来, 研究人员在 LPA 的基础上做了不同程度的优化, 来适应更多的应用场景^[14-18]: 为了避免 LPA 将所有节点划分到同一社区, Barber 和 Clark^[19]提出了一种模块化标签传播算法 (LPAm), 将社区检测问题转化为求目标函数最大值问题, 但容易陷入局部最优解的情况。为了解决上述问题, Liu 等^[20]在 LPAm 的基础上, 融合多步贪婪凝聚算法, 提出了基于模块度最大化的标签传播 LPAm+, 将多对社区融合, 避免局部最大值的出现。Li^[21]探索了在 LPA 的基础上引入方向性的方法, 提出了约束定向的标签传播算法, 将边的方向转化为边的权重, 探索了定向化模块

的划分,但并未考虑边自身的权重信息。Zhang等^[22]提出了一种基于节点重要性和标签影响的算法,融合贝叶斯定理来计算节点的重要性,对节点的重要性进行了有效评估,但该方法只考虑了周围节点的信息,并没有融合边的信息,使标签影响能力的计算有所欠缺。

现有LPA的改进工作主要针对节点标签初始化、节点的标签选择策略、节点的标签更新顺序等方面进行改进,并没有考虑节点之间的影响。在网络中已有信息的利用上,也没有很好的策略。针对以上不足,本文提出了一种新的节点之间影响力计算方法,并引入广度优先传播的思想来有效利用已知信息。

2 TS-LPA

本文引入扩展邻域的概念来表示节点传播信息的能力(即节点的影响力),通过综合节点影响力和边的权重信息计算节点之间的影响概率,算法的基本思想概括如下。

1) 构建网络图,利用扩展邻域核心度的方法来计算节点的重要性,并根据网络的拓扑结构,综合利用邻居信息得到节点的影响力。

2) 结合节点影响力和边的信息计算单个节点对不同邻居的影响概率,并进行均一化表示,构建网络的节点影响概率矩阵。

3) 在节点的设置方面,根据节点的影响力确定种子节点选取策略,利用矢量来表示节点的类标签,提高迭代效率。

4) 标签迭代更新过程中,利用广度优先搜索方式搜索种子节点信息,根据搜索结果,对标签更新的类标签向量进行进一步更新。

2.1 节点影响力计算

现实中复杂网络的节点地位是不平等的,不同节点对网络结构和性能的影响可能会有很大的差异,网络的传播效应会受到一些重要节点的影响。例如,在传销的运作中,总经理、经理、主任、业务员等人员在组织运作过程中扮演着重要的角色。这种现实职位的重要性同样反映在交易网络中,总经理的账号往往周围邻居节点数更多或者边的交易金额、交易次数权重更大,更大幅度地影响着网络的结构和性能。

结合现实的分析,为了更有效地对节点的重要性进行衡量,本文采用节点的扩展邻域核心度^[23]来量化节点的传播能力,在此基础上,综合考虑节点与其邻居节点之间边的权重及邻居节点的度数,扩展得到节点综合影响力Node值。

定义 1 对于给定的网络 $G=(V,E)$, 其中表示网络的节点集, E 表示网络的边集。节点综合影响力Node值的计算如式(1)所示。

$$\text{Node}(i) = \frac{\text{ENcoreness}(i)}{d_i} + \sum_{j \in N(i)} \left(\frac{W_{ij}}{\sum_{Z \in N(i)} W_{ij}} \cdot \frac{\text{ENcoreness}(i)}{d_i} \right) \quad (1)$$

其中, $\text{ENcoreness}(i)$ ^[23] 表示节点 i 的扩展邻域核心度, W_{ij} 表示节点 i 与节点 j 之间边的权重, $N(i)$ 表示节点 i 的一阶邻居节点集合, d_j 代表节点 j 的度数。

2.2 节点之间影响概率的计算

LPA 的节点标签选择策略具有随机性,造成了社区检测的不稳定。此外,考虑周围邻居节点对更新节点的影响时并未涉及节点权重和边权重的影响。如图3所示, A、C、D与B进行业务往来, A与B进行了5次业务往来,涉及金额10 000元, C、D和B各进行了1次业务往来,涉及金额500元,根据标签传播算法,节点B应该划入C、D所属的社区,但实际上节点A对节点B的影响概率更大,这显然与实际不符。

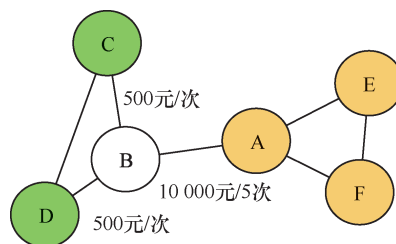


图3 一个人工生成的网络
Figure 3 An artificially generated network

为了解决此类问题,本文考虑邻居节点对待更新节点的影响,从节点的综合影响力、节点与邻居节点之间边的权重两方面来考虑,使待更新节点的所有邻居节点的标签都对其产生一定限度

的影响,从而使待更新节点的影响因素多样化,提高节点标签更新的准确性。考虑到在实际应用网络(如传销金融网络)中,若利用交易平均金额来表示边的权值,不同边的权重可能在几百到几万之间不等,那么待更新节点的不同邻居节点影响力与边的权重的乘积会相差过大,不利于标签影响概率的计算。因此,本文在两者相乘的基础上,引入 $\ln$ 函数来弱化这种影响,利用百分比来表示节点对不同邻居的影响程度,使节点的标签对各个邻居节点都有一定限度的影响概率。综上所述,节点之间影响概率 CL 如定义 2 所示。

定义 2 对于给定的网络 $G=(V,E)$, 节点之间影响概率 CL 值的计算如式(2)所示。

$$CL_{(j \rightarrow i)} = \frac{\ln(\text{Node}(j) \cdot W_{ij})}{\sum_{z \in N(j)} \ln(\text{Node}(z) \cdot W_{iz})} \quad (2)$$

其中, $CL_{(j \rightarrow i)}$ 代表节点 j 到节点 i 的传播概率, $N(i)$ 表示节点 i 的一阶邻居节点集合, W_{ij} 表示节点 i 与邻居节点 j 之间边的权重。

2.3 节点的设置

在种子节点的选取方面,由于网络中节点的影响力具有明显的差异性,如何有效地选取网络中有影响力节点的问题显得尤为重要,选择网络中影响力大的节点作为种子节点有利于社区的形成与划分,能够将信息更好地传播到网络中,相反,如果选择影响力较小的节点,则不利于社区的形成,从而增加迭代次数,浪费算法的执行时间。

因此,本文在计算得到的节点的综合影响力的基础上,将节点综合影响力从大到小排序,构建节点影响力的排序表 N ,根据排序表 N 选取一定比例数据组成种子节点集合 Seed 。

在节点标签的设置上,本文利用标签向量来更详细地描述节点的社区类别信息。一方面,通过矢量化对节点的社区标签类别进行表示,这样在每一次的迭代中保证节点可能的社区类别标签都能够得到更新。如图 4 所示,节点 3 处于淡蓝色社区和紫色社区重叠地方,且两个社区对它的影响相差不大时,传统的标签传播方法在每一次迭代时只选择淡蓝色或者紫色一种作为类别,具

有一定的绝对性,在使用矢量化表示后,3 号节点社区类别标签则可以表示为 $(0.45, 0.55)$ 的形式,在每一次迭代时都能够对两个社区类别标签同时进行更新。

此外,经过若干次迭代后,两个社区对更新节点的影响差距逐渐变大,使节点的社区类别的确定更加明确。另外,节点社区类别的矢量化表示有利于提高算法的执行效率。

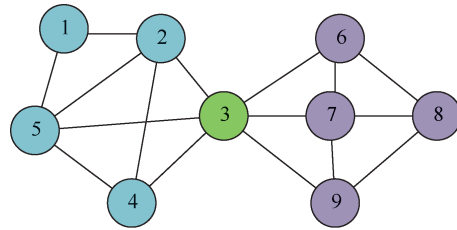


图 4 具有两个社区的网络示意
Figure 4 Schematic diagram of a network with two communities

综上,节点社区类别标签的设置策略如下:设网络节点的集合为 $V = \{v_i\}_{i=1}^n$, 其中, m ($m \leq n$) 个节点为已知标签节点, $(n-m)$ 个节点为未知标签节点,节点类别数 c 已知,定义一个 c 维的实值向量 $\mathbf{R}_i \in R^c$ 作为节点的标签向量,在标签向量初始化时,若节点 v_i 为已知标签的节点,且该节点属于第 p 类,则 $\mathbf{R}_i$ 的第 p 维数值置为 1,其余维度数值置为 0。若节点 v_i 为未知标签的节点,则 $\mathbf{R}_i$ 所有维度数值均置为 -1,于是,问题转化为如何利用 $(v_i, \mathbf{R}_i) (1 \leq i \leq m)$ 的数据预测 $\{v_{m+1}, \dots, v_n\}$ 的标签数据 $\{\mathbf{R}_{m+1}, \dots, \mathbf{R}_n\}$ 。

将输入的数据进行处理,计算节点更新顺序、影响概率、节点标签等信息,并利用矩阵向量的形式进行存储,TS-LPA 的数据预处理过程如下。

算法 1 TS-LPA 的数据预处理过程

输入 图 G 的邻接矩阵 A , 部分节点的标签 label

输出 节点之间影响概率值 CL, 节点种子集合 Seed , 排序后节点影响力 V_Sort , 节点的初始化标签向量 $\mathbf{R_init}$

1) $V_node = \text{Node}(A)$ //计算节点的影响力 Node 值

2) $V_Sort = \text{Sort}(V_node)$ //根据节点重要性对节点进行

```

3) Seed = seedChoice(label,V_Sort)//选择种子
节点
4) for  $\forall v \in V$ 
5) for  $\forall u \in N(v)$  //对于节点  $v$  的所有邻居
节点
6)  $CL(v,u) = \text{getCL}(V\_node,A)$ 
7) end for
8) end for //计算节点之间的传播概率
9) for  $\forall v \in V$ 
10) 设置节点的标签
11) end for

```

2.4 节点标签的分阶段更新策略

LPA 的不稳定性表现在算法更新过程中更新标签的选择,为了更好地利用邻居节点标签类别来进行标签更新,本文提出了两阶段更新策略。

第一阶段:网络中每一个节点都有一个标签向量 $\mathbf{R}_i$,代表节点对于各个类别的隶属度,当算法开始传播的时候,待更新节点的所有邻居节点都对该节点的标签产生影响,邻居节点的标签向量 $\mathbf{R}_j$ 与节点之间的影响概率 CL 共同影响待更新节点的标签。设 $\mathbf{L}_i$ 为第一阶段节点更新后的标签向量。第一阶段节点标签向量更新式如式(3)所示。

$$\mathbf{L}_i = \sum_{j \in N(v)} \mathbf{R}_j \cdot CL_{(j \rightarrow i)} \quad (3)$$

其中, $N(i)$ 表示待更新节点 i 的一阶邻居节点的集合, $CL_{(j \rightarrow i)}$ 表示节点 j 到节点 i 的传播概率, $\mathbf{R}_j$ 表示节点 j 的标签预测向量。

第二阶段:在标签更新时,待更新节点的邻居节点标签大部分是未知标签,若一个节点的社区类别是不确定的,那么对其邻居节点的标签影响说服力较弱。因此,为了减轻未知邻居节点对待更新节点的支配程度,除邻居节点的影响外,加入附近已知标签节点对待更新节点的影响,共同完成节点的标签更新。具体策略如下:以待更新节点为起点,应用广度优先搜索方式逐层向外搜索,直到出现已知标签节点所在层次,或查询超过 3 层时终止搜索。搜索到已知标签集合 K ,在集合 K 中标签数量最多的社区类别 $c1$,会在第一阶段标签更新的基础上修改 $\mathbf{L}_i$ 。因为在标签迭

代更新过程中,搜索到的社区类别 $c1$ 对待更新节点的影响较大,为了体现这种影响力,本文将 β 的取值设为 0.5~1。第一阶段节点标签预测向量更新如式(4)所示。

$$\mathbf{R}_{i(z)} = \begin{cases} \mathbf{L}_{i(z)} \cdot \beta, & z = c1 \\ \mathbf{L}_{i(z)} \cdot (1 - \beta), & z \neq c1 \end{cases} \quad (4)$$

其中, $\mathbf{R}_i$ 表示最终节点 i 的标签预测向量, $z=1, \dots, c$, $\mathbf{R}_{i(z)}$ 表示节点 i 对第 z 个社区的隶属度, $\mathbf{L}_i$ 表示第一阶段节点 i 的标签预测向量, c 表示社区类别的数量。

2.5 节点迭代传播目标函数设置

定义一个函数 Y_f ,该函数旨在停止算法的迭代,当函数 Y_f 的值变化很小时,节点标签趋于稳定,算法迭代停止。函数 Y_f 如式(5)所示。

$$Y_f = \sum_{i=1}^n \|\mathbf{R}_i(t) - \mathbf{R}_i(t-1)\| \quad (5)$$

$\mathbf{R}_i(t)$ 表示节点 i 当前第 t 次迭代的标签预测向量, $\mathbf{R}_i(t-1)$ 表示节点 i 第 $t-1$ 次迭代的标签预测向量, n 表示节点集合 V 中节点的数量。

利用算法预处理阶段得到的数据,进行算法的标签传播。标签迭代传播后,每个节点被分配到对应的标签,从而确定社区检测的结果。

TS-LPA 的标签传播过程如算法 2 所示。

算法 2 TS-LPA 的标签传播过程

输入 网络 $G=(V,E)$, 最大迭代次数 t_max ,

图 G 的邻接矩阵 A , 节点之间影响概率值 CL , 网络种子集合 $Seed$, 排序后节点影响力 V_Sort , 节点的初始化标签向量 $\mathbf{R_init}$

输出 网络中最终的社区划分结果 $P = \{P_1, P_2, \dots, P_n\}$

```

1) for  $\forall v \in V$ 
2) 计算影响节点  $v$  的标签类别  $M$ 
3) end for
4) iter = 0 //计算迭代的次数
5)  $\mathbf{R} = \mathbf{R\_init}$  //保留初始化的标签向量
6) while iter <  $t\_max$  &&  $Y_f > 1 \times 10^{-3}$  // 迭代

```

结束条件

```

7) for  $\forall v \in V\_Sort$ 
8)  $\mathbf{L}_v = \text{firstSpread}(\text{seed}, CL, \mathbf{R}, A)$ 

```

```

9)  $R_v = \text{secondSpread}(L_v, M)$ 
10)   for  $\forall u \in \text{Seed}$ 
11)    $R_u = R_{\text{init}_u}$ 
12)   end for
13) end for
14) end while
15) for  $\forall v \in V$  //迭代结束后, 计算得到未知节点的社区
16)  $P_v = \max(R_v)$ 
17) end for

```

2.6 算法复杂度分析

对于给定的网络 $G = (V, E)$, 其中 V 表示网络的节点集, E 表示网络的边集。 $|V|$ 大小为 n , $|E|$ 大小为 m , TS-LPA 的时间复杂度分析如下。

计算网络中节点重要性 $\text{Node}(i)$ 的时间复杂度是 $O(m+n)$, 节点重要性排序的时间复杂度为 $O(n \log n)$, 集合 V 的每个节点标签预测向量的初始化时间复杂度为 $O(n)$, 迭代一次的时间复杂度为 $O(n)$, 则迭代 t 次时间复杂度为 $O(nt)$, 有约束的广度优先搜索的时间复杂度为 $O(3nd_i)$, 其中 d_i 表示节点 i 的度数。综合的时间复杂度为 $O(m+n(3+t+3d_i)+n \log n)$ 。

3 实验与分析

为了验证算法的性能, 本文在真实数据集和人工数据集上分别进行了实验。实验环境是 Windows10 64 位版本, 处理器为 intel Core i5-8250U CPU @ 1.8 GHz, 内存容量为 8 GB, 编程语言采用 Python。

3.1 评价指标

标准化互信息^[24]是一种衡量社区检测算法结果与真实网络在结构上是否一致的方法。通过比较社区检测的结果与标准社区结构的相似度, 从而衡量社区检测的质量, 对于已知社区结构的社区检测, 可以衡量算法划分社区与标准社区之间的相似度。

对于给定的网络 $G = (V, E)$, 网络 G 划分为 k_i 个社区, 在每个社区 i 中, 每个节点 v 都被分配了标签 $l_{vp} = i$, 则真实划分 T 的熵计算如下:

$$H(T) = - \sum_{i=1}^{k_i} \frac{n_i^t}{n} \log \frac{n_i^t}{n} \quad (6)$$

其中, n_i^t 代表社区 i 中节点的个数, 因此, 真实社区检测 T 与预测社区检测 P 之间的互信息计算公式为

$$I(T, P) = \sum_{i=1}^{k_i} \sum_{j=1}^{k_p} \frac{n_{ij}^{tp}}{n} \log \left(\frac{\frac{n_{ij}^{tp}}{n}}{\frac{n_i^t}{n} \cdot \frac{n_j^p}{n}} \right) \quad (7)$$

其中, k_p 代表预测社区检测 P 的社区个数, n_{ij}^{tp} 代表真实社区 i 和预测社区 j 中相同节点的个数。经过最大值 $(H(T) + H(P)) / 2$ 正规化得到标准互信息公式为

$$\text{NMI} = \frac{-2 \sum_{i=1}^{k_i} \sum_{j=1}^{k_p} n_{ij}^{tp} \log \left(\frac{\frac{n_{ij}^{tp}}{n} \cdot n}{\frac{n_i^t}{n} \cdot \frac{n_j^p}{n}} \right)}{\sum_{i=1}^{k_i} n_i^t \log \frac{n_i^t}{n} + \sum_{j=1}^{k_p} n_j^p \log \frac{n_j^p}{n}} \quad (8)$$

3.2 基于真实数据集实验分析

本文实验采用 5 种标准的真实网络数据集对数据进行社区的划分, 选取的数据集分别是 Karate、Dolphins、Football、Political blogs 共 4 种公认的真实网络数据集和传销组织数据集 (MLM Organization)。其参数如表 1 所示。

表 1 真实网络数据集参数
Table 1 Real network data set parameters

数据集	节点	边
Karate	34	78
Dolphins	62	159
Football	115	616
Political blogs	1 490	19 090
MLM Organization	882	4 215

将 TS-LPA 与经典的 LPA、LPAM、LPAM+、半监督社区检测算法 (S_LPA^[25]、KLPA^[26]) 进行比较, 对于 LPA 和 S_LPA 两种不稳定的算法, 将算法运行 10 次后求其平均 NMI 值。通过表 2 可以看出, TS-LPA 与基于 LPA 的改进算法有更好的社区结构划分能力。在小规模的社区检测中, TS-LPA 的算法性能与最优的算法相差不大, 在涉

及节点数量较多的社区检测中, TS-LPA 表现出较强的社区检测能力, 从而能够有效地对节点所属社区进行判断, 该算法利用节点的重要性和种子节点等因素, 有效提高了社区节点划分的质量, 并且极大地消除了随机性, 具有更好的稳定性。

表 2 真实数据集实验数据比较
Table 2 Comparison of experimental data of real data sets

算法	Karate	Dolphins	Football	Political blogs
KLPA	92.44%	88.34%	84.79%	66.43%
LPA	79.12%	75.59%	84.35%	55.32%
S_LPA	88.82%	84.01%	86.44%	63.29%
LPAM	68.28%	66.37%	85.34%	44.12%
LPAM+	70.32%	70.24%	86.81%	51.45%
TS-LPA	100.00%	92.10%	90.10%	67.23%

TS-LPA 在 Dolphins 数据集、Football 数据集, 以及 MLM Organization 数据集上的社区划分结果如图 5 所示。在图 5(a)中, Dolphins 划分为两个社区, 社区的结构相对清晰, 社区成员划分与真实情况总体一致。在图 5(b)中, Football 数据集被划分为 12 个社区, 蓝色的 80、82、90 这 3 个节点为已知节点, 由于这 3 个节点所属社区属于一个管理组社区, 负责对其他球队社区进行组织与管理, 该社区成员与各个社区联系比较紧密,

所以社区检测的识别度较小, 无法形成小规模社区。但是除此之外其余社区检测结构比较清晰, 根据节点的标签可以判断, 实验结果各个社区的组成部分与真实情况社区的组成部分具有较高的一致性。在图 5(c)中, MLM Organization 数据集被划分为 10 个社区, 由于传销组织数据网络多为自我中心网络, 多数下线以头目为中心组成传销组织团伙, 所以易形成网络社区。非重要传销成员之间, 社区的结构比较清晰。

3.3 基于人工数据集实验分析

为了测试本文算法在不同网络中的适用性, 本文选取人工 LFR 基准网络进行实验分析, 并与半监督社区检测算法 (S_LPA、KLPA) 进行比较, 选取 NMI 值作为社区检测准确度的评价标准, 对于不稳定 S_LPA, 计算 10 次求平均值。

人工合成网络分别包含 1 000 个节点和 2 000 个节点, 具体参数如表 3 所示。其中, N 代表网络中节点的数量, K 代表网络的平均度值, μ 代表网络混合参数, 表示不同社区节点边数占网络总边数的比例, μ 取值为 0~1。 μ 值越小, 人工合成的社区结构越清晰; μ 值越大, 人工合成的社区结构越模糊。 $\max_k$ 代表网络中最大度值, $\min_c$ 和 $\max_c$ 分别表示网络中社区结构的最小值和最大值。在不同 μ 值下, 各个算法实验结果如图 6 和图 7 所示。

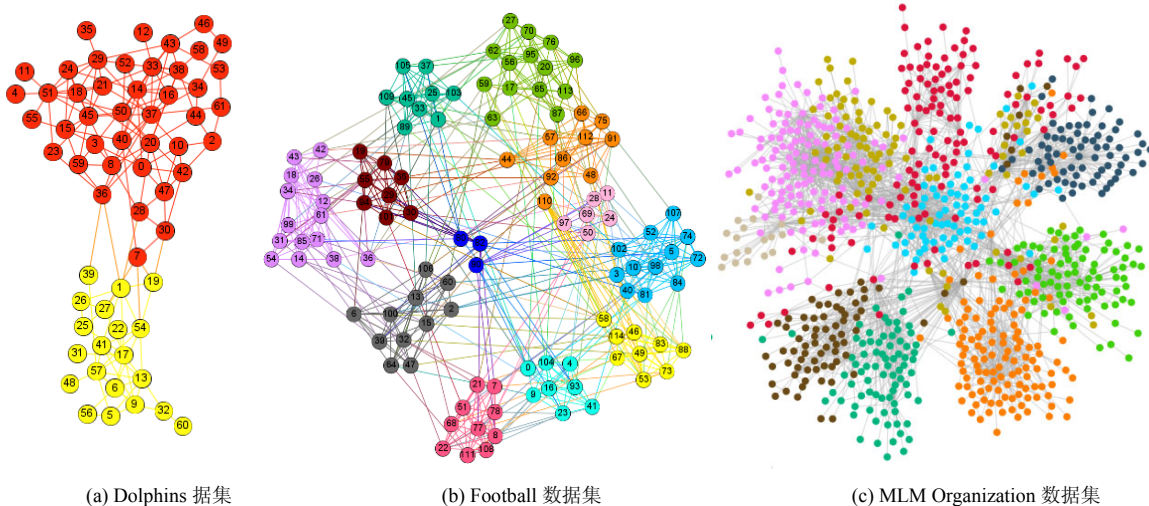


图 5 TS-LPA 在 3 个真实数据集划分结果
Figure 5 TS-LPA partition results in three real data sets

表3 人工合成网络参数
Table 3 Artificial synthesis network parameters

数据集	N	K	maxk	minc	maxc	μ
nc ₁	1 000	20	50	70	100	(0.2,0.7)
nc ₂	2 000	20	50	70	100	(0.2,0.7)

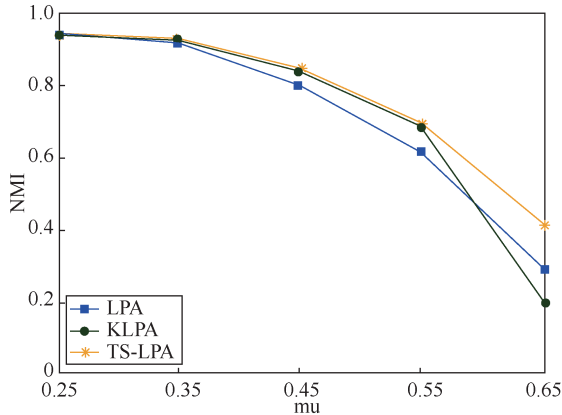


图6 当 $N=1\ 000$ 时不同算法数据比较
Figure 6 Comparison of data of different algorithms when $N=1\ 000$

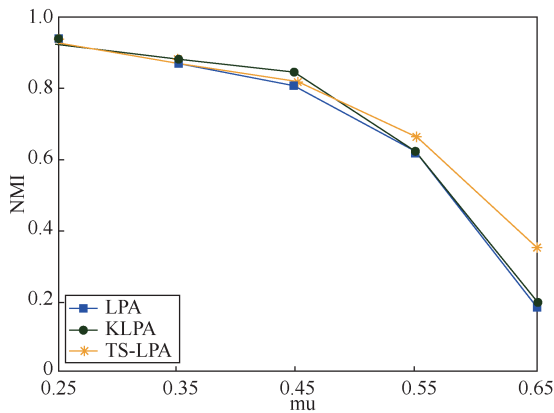


图7 当 $N=2\ 000$ 时不同算法数据比较
Figure 7 Comparison of data of different algorithms when $N=2\ 000$

由图6和图7可以看出,当 $\mu < 0.5$ 时,网络社区的结构比较清晰,3种算法对社区的划分有较高的准确性。当 $\mu > 0.5$ 时,网络社区的结构模糊,3种算法的社区检测的准确性开始下降。当 $N=1\ 000$, $\mu > 0.5$ 时,TS-LPA对社区检测的NMI值明显优于其余两种算法,当 $N=2\ 000$, $\mu > 0.55$ 时,TS-LPA整体高于其余两种算法的NMI值,但不明显。总体来看,TS-LPA在人工合成网络的性能优于KLPA和S_LPA。

3.4 参数的设置

TS-LPA引入参数 β 来确定种子节点在更新

过程中标签影响能力的大小,最终影响社区结构的划分,将TS-LPA在Political blogs数据集和nc₁($\mu = 0.2$)数据集选取不同的 β 参数进行实验。实验选用NMI值作为评价的标准。实验结果如图8所示,当 β 选取在特殊的值时,社区检测的NMI值相对较高。在不同的社交网络中,由于种子集合的选取和网络拓扑结构的不同,已知标签节点对未知标签节点的影响程度有所差异,因此,不同的网络结构需要不同的参数进行对应。

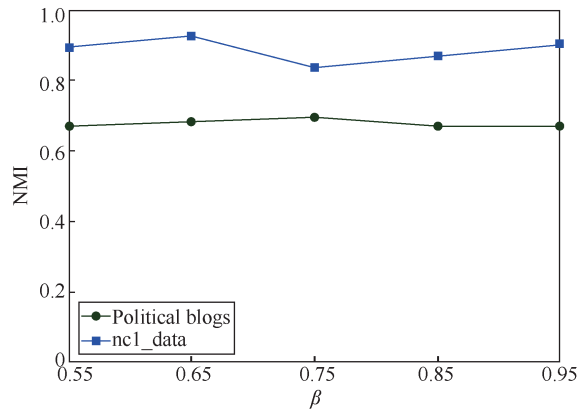


图8 不同数据集不同的 β 参数实验结果
Figure 8 Different β parameter detection results of different data sets

3.5 算法稳定性试验

LPA节点更新顺序的随机性和标签选择策略的随机性造成了算法的不稳定,多次运行的结果出现较大偏差。本文提出的TS-LPA在选择节点更新顺序时,通过对节点影响力进行排序来使更新顺序稳定;通过引入标签传播和广度优先的策略,确定了两阶段标签更新策略,消除了随机性。如图9所示,通过将TS-LPA在4个数据集上各进行15次运算,得到各个运行结果呈现一条直线,通过实验进行了验证。

4 结束语

本文在标签传播算法的基础上,提出了一种基于标签传播的两阶段社区检测算法,该算法综合考虑节点的局部影响力和全局影响力来计算节点在网络中的综合影响力,使节点影响力的量化更加具体。此外,所提算法引入节点影响力信息和边的权重信息来计算节点之间的影响概率,使其更接近真实网络。在传播迭代过

程中,通过未知节点邻域内已知节点的信息,实现对网络中节点的社区检测,实现标签传播过程中标签的两阶段调整,提高社区检测的准确度。实验结果证明,TS-LPA 能够有效提高社区检测的准确性和稳定性。

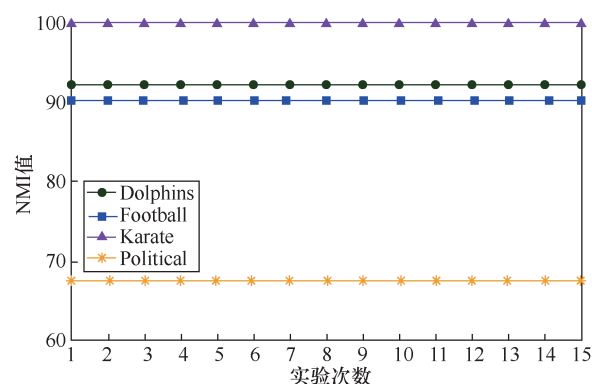


图 9 TS-LPA 在不同数据集多次运行结果
Figure 9 The results of multiple runs of the TS-LPA on different data sets

在接下来的工作中,将进一步考虑引入方向等信息和融合其他经典的算法来实现社区检测,提高社区检测的稳定性;对算法进行优化,减小算法运行的时间和空间复杂度。

参考文献:

- [1] SOURAVLAS S, SIFALERAS A, TSINTOGIANI M, et al. A classification of community detection methods in social networks: a survey[J]. International Journal of General Systems, 2021, 50(1): 63-91.
- [2] 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.
- [3] GARZA S E, SCHAEFFER S E. Community detection with the label propagation algorithm: a survey[J]. Physica A: Statistical Mechanics and Its Applications, 2019, 534: 122058.
- [4] PRAT-PÉREZ A, DOMINGUEZ-SAL D, BRUNAT J M, et al. Shaping communities out of triangles[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012: 1677-1681.
- [5] QIAO S J, HAN N, GAO Y J, et al. A fast parallel community discovery model on complex networks through approximate optimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1638-1651.
- [6] CHUNAIEV P. Community detection in node-attributed social networks: a survey[J]. Computer Science Review, 2020(37): 100286.
- [7] CAO J, BU Z, GAO G L, et al. Weighted modularity optimization for crisp and fuzzy community detection in large-scale networks[J]. Physica A: Statistical Mechanics and Its Applications, 2016(462): 386-395.
- [8] WEIR W H, EMMONS S, GIBSON R, et al. Post-processing partitions to identify domains of modularity optimization[J]. Algorithms, 2017, 10(3): 93.
- [9] LEE J, GROSS S P, LEE J. Modularity optimization by conformational space annealing[J]. Physical Review E, 2012, 85(5): 056702.
- [10] NEWMAN M E J. Community detection in networks: modularity optimization and maximum likelihood are equivalent[J]. arXiv preprint arXiv: 1606.02319, 2016.
- [11] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 70(6): 066111.
- [12] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.
- [13] BERAHMAND K, BOUYER A. A link-based similarity for improving community detection based on label propagation algorithm[J]. Journal of Systems Science and Complexity, 2019, 32(3): 737-758.
- [14] GDRZA S E, SCHAEFFER S E. Community detection with the Label Propagation Algorithm[J]. Physica A, 2019, 534(15): 122058.
- [15] 郑文萍, 车晨浩, 钱宇华, 等. 一种基于标签传播的两阶段社区发现算法[J]. 计算机研究与发展, 2018, 55(9): 1959-1971.
ZHENG W P, CHE C H, QIAN Y H, et al. A two-stage community detection algorithm based on label propagation[J]. Journal of Computer Research and Development, 2018, 55(9): 1959-1971.
- [16] CHEN N Y, LIU Y, CHEN H Q, et al. Detecting communities in social networks using label propagation with information entropy[J]. Physica A: Statistical Mechanics and Its Applications, 2017, 471: 788-798.
- [17] 於志勇, 陈基杰, 郭昆, 等. 基于影响力与种子扩展的重叠社区发现[J]. 电子学报, 2019, 47(1): 153-160.
YU Z Y, CHEN J J, GUO K, et al. Overlapping community detection based on influence and seeds extension[J]. Acta Electronica Sinica, 2019, 47(1): 153-160.
- [18] ZHOU H H. Research on community detection algorithm based on a small number of selected point[D]. Lanzhou: Lanzhou University, 2015.
- [19] BARBER M J, CLARK J W. Detecting network communities by propagating labels under constraints[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(2): 026129.
- [20] LIU X, 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.
- [21] LI X. Directed LPA: propagating labels in directed networks[J]. Physics Letters A, 2019, 383(8): 732-737.
- [22] ZHANG X K, REN J, SONG C, et al. Label propagation algorithm for community detection based on node importance and label influence[J]. Physics Letters A, 2017, 381(33): 2691-2698.
- [23] WU T, GUO Y X, CHEN L T, et al. Integrated structure investigation in complex networks by label propagation[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 448: 68-80.
- [24] 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.

- [25] 马秀. 基于节点影响力的标签传播算法的稳定性研究[D]. 太原: 中北大学, 2019.

MA X. Stability study of tag propagation algorithm based on node influence [D]. Taiyuan: North University of China, 2019.

- [26] WEI F F, YAN S J, YAN S K. Research on semi-supervised community discovery algorithm based on label propagation[J]. Software Guide, 2019(7): 92-95.



黄俊恒 (1966-), 河南新乡人, 哈尔滨工业大学(威海)副教授, 主要研究方向为社交网络、数据挖掘、人工智能。

[作者简介]



孙学良 (1999-), 男, 山东德州人, 哈尔滨工业大学硕士生, 主要研究方向为社交网络、机器学习。



辛国栋 (1976-), 男, 山东青岛人, 哈尔滨工业大学(威海)助理教授, 主要研究方向为金融安全、网络安全、社交网络。



王巍 (1975-), 女, 黑龙江齐齐哈尔人, 哈尔滨工业大学(威海)助理教授, 主要研究方向为社交网络, 数据挖掘, 机器学习和自然语言处理。



王佰玲 (1978-), 男, 黑龙江哈尔滨人, 哈尔滨工业大学(威海)教授、博士生导师, 主要研究方向为信息对抗、信息安全、信息搜索。