一种新的基于标签传播的重叠社区发现算法

沈海燕，李星毅

（江苏大学 计算机科学与通信工程学院，江苏 镇江 212013）

摘　要：发现高质量的社区是社区网络问题的研究热点。目前，社区发现算法大多针对非重叠社区，重叠社区发现算法较少。基于标签传播的算法是现有重叠社区发现算法中的一类，其中 COPRA 为典型算法。尽管该算法具有接近线性的时间复杂度，但存在随机因素，结果不稳定，产生的社区结构存在一定差异。为此，提出一种新的基于标签传播的社区发现算法，实验表明该算法在复杂度相近的情况下能明显提高所发现社区的质量，且具有较好的稳定性。

关键词：社区发现；重叠社区；标签传播；稳定性

DOI：10.11907/rjdk.1431038

中图分类号：TP312　　**文献标识码：**A　　**文章编号：**1672-7800(2015)004-0059-04

1　重叠社区及其发现算法

近年来，复杂网络研究受到广泛关注，主要涉及系统科学、统计物理学、社会科学、生物学等多个领域[1]。随着通信和互联网技术的快速发展，人们发现众多网络都存在社区结构[2]这一特征。所谓社区结构，简单来说，就是网络中的节点存在分组，一般组内的边连接比较稠密，而组间的边连接比较稀疏。社区结构在一定程度上可以反映出真实网络的拓扑关系。社区发现可以帮助更好地理解网络拓扑结构及其功能，从而更好地利用和改造网络，例如挖掘网络中的未知功能、控制疾病传播等。社区发现在某些特定应用环境中也有现实意义，例如发现恐怖分子、寻找犯罪团体等。

然而，真实网络存在一些同时属于多个社区的节点，这些节点叫作“重叠节点”，与其它社区存在重叠节点的社区就叫作“重叠社区”。例如在社会关系网络中，小王既参加了台球俱乐部，又参加了乒乓球俱乐部，那么小王就是这两个社区的重叠节点，台球俱乐部和乒乓球俱乐部是两个重叠社区。重叠社区较之非重叠社区具有更好的现实意义：一方面，重叠节点是网络中关键点，重叠社区因此而产生联系；另一方面，重叠社区反映了更加真实的网络结构。因此，研究重叠社区更符合真实网络的结构。

图 1 为非重叠社区结构，图 2 为接近真实网络的重叠社区结构。目前，能够发现重叠社区结构的算法主要包括以下 3 类：

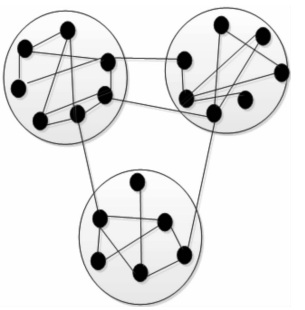

图 1　具有社区结构的小型网络示意图

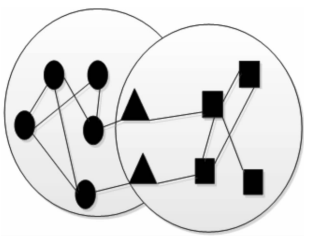

图 2　具有重叠社区结构的网络示意图

（1）基于 clique 的方法。典型算法有 CPM 算法[3]、EAGLE 算法[6]和 GCE 算法[7]。

（2）基于合并社区核心和扩展社区的方法[8]。

（3）基于标签传播的方法典型算法有 LPA 算法[4]和 COPRA 算法[5]。

基金项目：国家自然科学基金项目(10972027)；江苏大学校基金项目(11JDG064)

作者简介：沈海燕(1989－)，女，江苏南通人，江苏大学计算机科学与通信工程学院硕士研究生，研究方向为复杂网络；李星毅(1969－)，男，江苏镇江人，博士后，江苏大学计算机科学与通信工程学院教授、硕士生导师，研究方向为交通信息系统、复杂系统信息处理及模式识别与智能系统。

CPM 算法是一种派系过滤算法，主要通过寻找 K-clique 派系社区对社区进行划分。虽然 CPM 能够发现重叠点，但该算法在实际应用中依赖参数 K 的选取，不同 K 值导致划分出来的社区结构有很大差别。因此，在实际应用中存在一定局限性。同样地，EAGLE 和 GCE 算法是 CPM 的改进算法，存在同样的局限。

基于合并社区核心和扩展社区的算法需要人为指定两个参数，这两个参数需要先验知识，和具体网络相关并影响社区发现结果的好坏。

基于标签传播的算法是一类具有接近线性时间复杂度的算法，该算法的优点是计算过程简单，计算速度快，而它的缺点是算法稳定较差，每次运行的结果可能都不一样。本文提出一种新的标签算法，该算法对 COPRA 算法的初始化过程和随机选择过程作出了改进，从而大大提高算法的稳定性。

2 标签传播算法

本文重点对标签传播算法中的 COPRA 算法进行改进，故简要介绍标签传播思想和 LPA 算法，并分析 COPRA 算法与 LPA 的区别。

2.1 算法思想

标签传播算法最早由 zhu 等[9]于 2002 年提出，是基于图的半监督学习方法，其基本思想是通过标记节点的标签信息预测还未标记节点的标签情况。节点之间的标签传播主要依照标签相似度来进行，在传播过程中，未标记的节点根据邻接点的标签情况来迭代更新自身的标签信息，如果其邻接点与其相似度越相近，则表示对其所标注的影响权值就越大，邻接点的标签就更容易进行传播。

图 3 为标签传播过程。每个顶点都有一个唯一的标签作为社区标识，对图中所有的顶点进行标签迭代更新。在迭代过程中，每个顶点标签为其邻接点中出现次数最多的节点的标签。如果多个标签的数量都是最大值，则随机选择一个作为该顶点的标签。经过若干次迭代，最终形成一个完全连通图(代表一个社区)，并且该图内所有顶点都拥有相同标签。

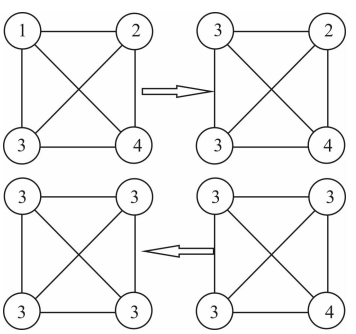

图 3 标签传播过程

2.2 LPA 算法

LPA 算法基于标签传播算法思想来发现社区结构。该算法由 Raghavan[4]提出，根据每个节点邻接点的社区情况来选择所要加入的社区。其主要思想是起初每个节点拥有独立的标签，每次迭代中对于每个节点将其标签更改为其邻接点中出现次数最多的标签，如果这样的标签有多个，则随机选择一个。通过迭代，直到每个节点的标签与其邻接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即属于同一个社区。

LPA 算法的执行步骤：

(1)初始化过程即为网络中的每个结点分配一个唯一的标签，对于节点 x，有 $G_x(0) = x$ 。

(2)设置 t=1。

(3)将网络中的节点随机排序，假设顺序为 x。

(4)对 x 中每个节点 x，将 x 的标签设置为新的标签 $C_x(t) = f(C_{x_1}(t),..,C_{x_m},C_{x_{m+1}}(t-1))$，f 函数返回节点 x 的邻接点中出现次数最多的标签。如果有多个，则随机选择一个。

(5)如果每个节点的标签都与其邻接点中出现次数最多的标签相同，则算法结束，否则设置 t=t+1，返回步骤 3 继续执行。

2.3 COPRA 算法

LPA 算法虽然有很多优势，但无法挖掘出重叠社区结构。对此，Steve[5]基于原先的算法，引入了新的标签结构(c,b)，对每个节点 x，$x \in G(x)$，都拥有这样的标签。其中，c 表示社区标识符，b 表示节点 x 在社区 c 中的从属系数，且 $0 \leqslant b \leqslant 1$ 。

COPRA 算法的执行过程如下：

(1)初始化，对任意的节点 x，$x \in G(x)$ 分配一个唯一的标签 $(c_x,1)$ 。

(2)节点 x 根据其邻接点集的标签情况更新自己的标签。如果有多个可选标签，算法会随机选取其中的 v 个标签作为结果。其中，$0 < |C(x)| \leqslant v$ 。节点的从属系数符合如下表达式：$\sum_{c_i \in C(x)} b_t(c_x,x) = 1$，其中 $C(x)$ 表示节点 x 可能属于的社区集合。

(3)如果每个节点的标签都与其邻接点中出现次数最多的标签相同，则算法结束，否则返回步骤 2 继续执行。

(4)将具有相同社区标签节点合并为同一社区。

3 社区质量评价指标

对于社区发现算法所产生的社区，需要通过量化指标衡量其质量好坏，从而进一步评估各种社区发现算法的优劣。Newman 和 Girvan[12]提出了一个评价社区质量的指标，称之为模块化度量 Q。考虑某种划分形式，将网络划

分为 K 个社区。令 e_{ij} 为网络中连接社区 i 到社区 j 的顶点之间边的一半，对角线元素为 e_{ii}，则对角线上的各元素之和为 $Tre = \sum_i e_{ii}$，表示网络中连接某一个社区内部各个节点的边在所有边的数目中的比例。$a_i = \sum_j e_{ij}$ 为每行（或者每列）中各元素之和，表示与第 i 个社区中的节点相连的边在所有边中的比例。在此基础上，用下式来定义模块性的衡量标准：

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \| e^2 \| \tag{1}$$

其中，$\| \chi^2 \|$ 表示矩阵 χ 中所有元素之和。式(1) 表示在同样社区结构下，网络中连接两个同类型的节点的边的比例减去任意连接这两个节点的边的比例的期望值。若社区内部边的比例小于或者等于任意连接时的值，则 Q 为 0。Q 的上限为 1，Q 越接近 1，说明社区结构越明显。实际上，该值通常为 0.3 ～ 0.7。

Q 函数只是用来评价非重叠社区结构的指标。Shen 等[10]对 Q 函数作了扩展，得到一种可以用来评价重叠社区结构的指标—EQ 函数，其公式为：

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

其中，O_v 是节点 v 可以同时从属的社区数目，A 表示由网络转换而成的邻接矩阵。

4　算法改进

根据上述 LPA 算法和 COPRA 算法描述，可以看出算法标签初始阶段是给网络中的所有节点分配唯一的标签，此后标签会不断更新，网络规模越大，更新标签所需要的资源消耗就越大。在标签更新阶段，如果存在多个标签可选，则进行一次随机选择，这样导致算法的结果非常不稳定。鉴于上述缺点，本文对 COPRA 算法标签初始化和标签选择进行改进。

4.1　标签初始阶段

对标签的初始化主要是借鉴 CPM 算法中 Clique 思想。Clique 代表网络中完全子图，用完全子图来代替大量的节点。这样就不需要对所有的节点进行处理，只要对每个完全子图分配标签。

标签预处理算法步骤如下：

(1)初始化节点数据。

(2)遍历节点，根据邻接点依次寻找网络中 K-Clique，在此过程中，对从属于完全子图的节点进行标注，以防完全子图出现重叠节点。

(3)对所发现的完全子图按综合度数排序，按照排序在标签传播过程中优先处理。

(4)为每个 Clique 和剩余节点分配一个唯一的标签。

4.2　标签选择过程改进

COPRA 算法第二步涉及随机选择，可以将此随机选择进行弱化，本文主要引入标签影响值来进行弱化。

对每一个标签节点进行标签更新时，若其邻接点都没有标签，则不进行更新。若其邻接点中有标签存在，则在所有邻接点上的标签组成的集合中，选择其中一个作为 x 的标签。影响 x 的因素有：每个标签存在的个数；所在边的权重以及其所在邻接点的度。综合考虑这些因素的影响力及其主次关系，最终提出将平均权重作为影响标签选择的主要因素，邻接点的度作为次要因素。

对 x 邻接点中出现的每一个标签 l 进行标签影响值 $influence(l)$ 的计算，其定义如下：

$$influence(l) = \frac{\sum w(l_i)}{C(l)} + (1 - e^{-\sum d(l_i)}) \tag{3}$$

其中，$l \in L$，$w(l_i)$ 表示每一个标签为 l 的顶点所在边的权重值；$C(l)$ 表示标签 l 的总个数；$\sum d(l_i)$ 表示所有标签为 l 的顶点度之和。

假设边的权重为正整数，则 $\frac{\sum w(l_i)}{C(l)} > 1$，而 $1 - e^{-\sum d(l_i)} < 1$。因此，比较两个标签时，若平均权重的差值大于 1，则平均权重起主要决定作用；若值小于 1，则由平均权重和顶点之和共同决定。故对于一个顶点 x，其标签为：

$$Label(x) = \text{argmax} \tag{4}$$

$$influence(l) = \text{argmax}[\frac{\sum w(l_i)}{C(l)} + (1 - e^{-\sum d(l_i)})] \tag{5}$$

5　实验验证与分析

本文实验硬件环境为：Pentinum(R)双核 2.7GHZ 处理器，4G 内存，算法具体实现语言为 JAVA。

5.1　基准测试集

为验证算法的有效性，将该算法应用到 Zachary's Karate Club Network[13] 和 Dolphins Social Network[14] 两个真实网络中。

5.1.1　Zachary's Karate Club Network

在复杂网络社区结构研究中，该网络经常被使用，它反映一所大学空手道俱乐部成员之间的关系。该数据集包含 34 个节点，78 条边，其中节点表示俱乐部的成员，边表示成员之间的关系。

从表 1 可以看出，由于网络社区规模小，因而基于标签传播的重叠社区算法并不能表现出优势。不仅如此，在速度上反而还要低于 CFinder 算法，而且这 3 种算法挖掘出的社区质量也没有明显的差别。

表 1　Karate 数据集测试结果

	CPM	COPRA	本文算法
EQ	0.265	0.459	0.481
时间(s)	0.098	0.168	0.159

5.1.2 Dolphin social network

该数据集为 Lusseau 等对新西兰海域一个有 62 个成员的宽吻海豚社会网络进行长达 7 年(1995 年到 2001 年)观测后给出的宽吻海豚社会网络。该网络具有 62 个节点,159 条边。

通过表 2 可以看出,本文新算法发现社区的模块化度量值比原始算法有所提高,新算法所发现的社区质量有所提高。当网络越复杂时,该优越性表现越明显。

表 2 基准测试集中的模块化度量 EQ

数据集	原始算法均值	本文算法均值	EQ 值提高率/%
空手道	0.3653	0.3715	1.7
海豚	0.4770	0.4972	4.0

5.2 抓取数据集

抓取到的数据集为 Leskovecsc 等[15]从 Arxiv 网站上抓取的关于广义相对论和量子宇宙学主题的论文集合,论文作者为图的顶点,作者之间的合作关系为边。该数据集总共有 5 242 个顶点和 14 484 条边。

通过表 3 可以看出,在处理大规模数据集时,本文算法无论是在执行效率上还是挖掘精度上较之原有的 COPRA 有明显的优势。

表 3 Arxiv 数据集上的测试结果

	CPM	COPRA	本文算法
EQ	0.535	0.722	0.739
时间(s)	123.8	31.5	27.1

6 结语

本文主要采用标签传播的思想从标签初始化和标签选择两个方面针对原有 COPRA 算法进行改进。通过在初始阶段对标签的预处理过程来减少初始化的数目,从而提高算法执行效率;在标签选择过程中引入标签影响值来弱化随机选择,使得算法更加稳定。实验结果表明,改进后的算法提高了算法稳定性,不仅能够成功挖掘出具有较高质量的社区,而且还能挖掘出重叠社区结构。

本文仅仅从标签初始化和标签选择两个方面对算法作了改进。对于标签传播的方式却未作出考虑。将从这方面对算法作进一步改进,以得到更具优势的算法。原有 COPRA 算法采用同步的方式,较之异步方式存在震荡的问题。可以考虑将这两种方式进行综合得到新的传播方式。

此外,在重叠社区结构发现方面,可以考虑将现有的层次发现算法和重叠社区算法相融合,得到层次性重叠社区发现算法,较之单一的重叠社区算法所挖掘出的社区更加真实。

参考文献:

[1] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006,162-191.

[2] TRAUD A L,KELSIC E D,MUCHA P J,PORTER M A. Comparing community structure to characteristics in online collegiate social networks[J]. SIAM Rev,2011,53(3):526-543.

[3] PALLA G ,DERENYI I ,FARKAS I ,et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature,2005,435:814-818.

[4] RAGHAVAN U N,REKA A,SOUNDAR K. Near liner time algorithm to detect community structures in large-scale networks [J]. Physical Review E,2007,76:36-106.

[5] GREGORY S. Finding overlapping communities using disjoint community detection algorithms [J]. Studies in Computational Intelligence,2009,207:47-62.

[6] HUAWEI SHEN,et al. Detect overlapping and hierarchical community structure in networks [J]. Physical A:Statistical Mechanics and its Applications Volume 388,Issue 8,15 April 2009: 1706-1712.

[7] C LEE,F REID,A MCDAID,et al. Detecting highly overlapping community structure by greedy clique expansion[C]. Tech. Rep. arXiv,2010.

[8] S MING-SHENG,C DUAN-BING Z,TAO. Detecting overlapping communities based on community cores in complex networks[J]. Chinese Physics Letters,2010(27):58-901.

[9] ZHU X JIAO-JIN,GHAHRAMANI Z. Learning from labeled and unlabeled data with label propagation, CMU-CALD[R]. Pittsburghers:Carnegie Mellon University,2002:2-107.

[10] XIE JIE-RUI ,SZYMANSKI B. Community detection using a neighborhood strength driven label propagation algorithm[C]. Proc of IEEE Network Science Workshop,2011:188-195.

[11] KIM Y,JEONG H. The map equation for link community [J]. Physical Review E,2011,84(2):26-110.

[12] NEWMAN M,EJ GIRVAN M. Finding and evaluating community structure in Networks[J]. Physical Review,2004(69):26-113.

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

[14] DAVID LUESSEAU, KARSTEN SCHNEIDER, OLIVER J BOISSEAU,et al . The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations [J]. Behavioral Ecology and Sociobiogy,2003(54):396-405.

[15] LESKOVEC J,KLEINBERG J,FALOUTSOS C. Graph evolution:densification and shrinking diameters[J]. ACM Trans on Knowledge Discovery from Data(ACM TKDD),2007,1(1) :1-40.

(责任编辑:陈福时)