广西科学院学报
Journal of Guangxi Academy of Sciences

2017,33(1):12～18
Vol. 33, No. 1 February 2017

网络优先数字出版时间:2017-02-27 DOI:10.13657/j.cnki.gxkxyxb.20170227.002
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20170227.1538.004.html

# 一种改进的基于相干邻居亲近度的标签传播算法*

# An Improved Label Propagation Algorithm Based on Coherence Neighborhood Propinquity

张 超,武先强,董荣胜
ZHANG Chao, WU Xianqiang, DONG Rongsheng

(桂林电子科技大学计算机与信息安全学院,广西桂林 541004)
(School of Computer Science and Information Security of the Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China)

**摘要:【目的】**提高现有的基于相干邻居亲近度(Coherence neighborhood propinquity)的标签传播算法(Label propagation algorithm,LPA)社区发现的准确性,并减少标签传播过程花费的时间。**【方法】**在CNP-LPA算法基础上,引入节点间依赖度,提出一种改进的CNP-LPA+算法,在预处理阶段结合相干邻居亲近度与节点间依赖度,将依赖度高的节点并入本区域内的核心节点,并在得到的核心CNP网络基础上传播标签,显著提高了社区发现的质量。选取CNP-LPA算法使用的6组社交网络数据集,采用模块度$Q$评估LPA、CNP-LPA、CNP-LPA+ 3种算法的划分结果。**【结果】**CNP-LPA+算法在所有数据集上均取得了最高的$Q$值,有效提高了算法的准确性,并减少了标签传播过程花费的时间。**【结论】**CNP-LPA+算法是有效的。

**关键词:**社区发现 标签传播算法 相干邻居亲近度 节点间依赖度 核心CNP网络
**中图分类号:**TP311 **文献标识码:**A **文章编号:**1002-7378(2017)01-0012-07

**Abstract:【Objective】**To improve the accuracy of community detection by CNP-LPA and reduce the time it takes for the label propagation process. **【Methods】**In this paper, an improved CNP-LPA+ algorithm is proposed. In the preprocessing stage, the nodes with high dependency are integrated into the core nodes of the local region according to the coherent neighborhood propinquity and dependency. The quality of communities is significantly improved by spreading labels on the core CNP network. Six groups of social network data sets are selected, the modularity measure $Q$ is used to evaluate the results of detecting by LPA, CNP-LPA, CNP-LPA+ algorithms. **【Results】**Experiments show that the CNP-LPA+ algorithm achieves the highest $Q$ value on all data sets, which improves the accuracy of the algorithm and reduces the time spent on the label propagation process. **【Conclusion】**CNP-LPA+ algorithm is effective.

**Key words:** community detection, label propagation algorithm, coherent neighborhood propinquity, coherent neighborhood dependence, core CNP network

## 0 引言

**【研究意义】**以节点表示实体,边表示实体之间的关系,就可以将现实世界中的很多系统抽象为复杂网络,如人际关系网、论文合作网、电影明星合作网、博客引用网、电话通讯网等[1]。这些复杂网络往往具有潜在的社区结构,比如同一社区内的节点连接紧密,而社区之间的节点连接则较稀疏。对复杂

**收稿日期:**2016-12-20
**作者简介:**张 超(1991－),男,硕士研究生,主要从事社区发现和大规模图数据压缩研究,E-mail:guet_zhangchao@foxmail.com。
*广西可信软件重点实验室基金项目(kx201623)资助。

网络中的社区结构进行发现与分析,可以很好地理解其结构和行为,具有重要的研究价值与意义[2-3]。**【前人研究进展】**已有许多学者从不同角度对如何发现网络中的社团结构问题进行了研究,比较经典的社区挖掘算法有 GN 算法[4]、谱分析思想的算法[5]、层次距离算法[6]、边集聚系数法[7]等,Fortunato[8]也对复杂网络社区挖掘算法进行了详细的介绍和研究。传统的社团发现算法大多存在算法复杂度较高的问题,难以有效处理包含大量节点的社交网络。有的算法需要预先确定网络中的社区数目以及社区的大致规模,限制了算法的实际应用效率。Raghavan 等[9]提出了一种基于标签传播的社区发现算法,即标签传播算法(Label propagation algorithm, LPA),该算法能够在接近线性时间内查找出网络中的社区结构,在处理大规模的网络时具有很好的时间效率,且不需要事先知道网络中有多少个社区、社区规模如何,因此受到越来越多的关注。

LPA 算法是一种基于标签传播机制的启发式算法,其主要思想在于初始化时为网络中每个节点赋予一个独特的标签,根据每个节点的邻居节点集合的标签分布对该节点的标签进行迭代更新,通过多次迭代,网络中各节点的标签会趋于稳定,那些具有相同标签的节点则组合成同一社区。在现实的社交网络图中,节点与节点之间的关系不仅仅是存在边或者不存在边这两种状态,还有联系紧密程度的区分。LPA 算法在更新节点标签时,标签选择的策略仅考虑了邻接点中相同标签的个数,忽略了邻接点联系的紧密程度,有很大的局限性,Zhang 等[10]提出了相干邻居亲近度(Coherent neighborhood propinquity)的概念来度量网络中任意节点对间的亲近程度。Lou 等[11]将相干邻居亲近度引入标签传播算法,提出了 CNP-LPA 算法,该算法并不直接在原始网络上传播标签,而是对网络进行预处理,计算网络中任意节点对的 CNP 值,基于得到的 CNP 网络传播标签。在更新标签时,使用节点对的 CNP 值对标签进行加权,选择权值最大的标签进行更新。然而,CNP-LPA 算法进行预处理时只考虑到了网络节点间的亲密程度,没有考虑到节点间依赖度。**【本研究切入点】**在实际的社交网络中,影响力较小的节点所属的社区往往依赖于本区域内影响力较大的节点(核心节点)。此外,通过预处理得到的 CNP 网络,边的数目相比原始网络往往会大大增加,这就使得后续标签传播过程花费的时间明显增加。**【拟解决的关键问题】**针对上述问题,本研究在 CNP-LPA 算法基础上,引入节点间依赖度,提出一种改进算法 CNP-LPA+,其核心思想是在预处理阶段合并所有依赖度高的节点(冗余节点),仅仅保留网络中的核心节点,根据得到的核心 CNP 网络传播标签,传播过程完成后,将有相同标签的核心节点归入同一社区,冗余节点所属社区与本区域内核心节点所属社区保持一致。选取空手道关系网络、美国政治书籍网络、科学家引用网络等 6 组社交网络作为测试数据集,使用模块度 $Q$ 作为评估 LPA、CNP-LPA、CNP-LPA+ 3 种算法发现质量的评价函数。模块度 $Q$ 是最广泛使用的衡量社区发现质量的函数,由 Newman 和 Girvan[12]提出。实验结果证明 CNP-LPA+算法在 6 组数据集上均取得了最高的 $Q$ 值,显著提高了社区发现的质量,并减少了标签传播过程花费的时间。

## 1　朴素标签传播算法(LPA)及其改进

### 1.1　朴素 LPA

以节点表示实体,边表示实体之间的关系,就可以将社交网络抽象成一个无向简单图,形式化描述为 $G=(V,E)$,其中 $V(v\in V)$ 表示节点集,$E(e\in E)$ 表示边集,且满足 $E\subset V\times V$。

$G$ 中任意的节点 $v(v\in V)$,LPA 算法初始时都为其分配一个唯一的标签 $c_v$,用于表示节点 $v$ 所属的社区,$N(v)$ 表示节点 $v$ 的邻居集合,更新规则如式(1)所示:

$$c_v=\arg\max_l |N^l(v)|, \tag{1}$$

$|N^l(v)|$ 表示 $v$ 邻居节点集合中标签为 $l$ 的节点个数,每个节点的标签被更新为其数量最多的邻居节点所拥有的标签。LPA 算法的时间复杂度为 $O(t|E|)$,其中 $t$ 用来表示迭代次数,$|E|$ 表示网络中的边数,算法时间复杂度接近线性,在处理大规模的社交网络时具有很好的效率。LPA 算法标签的更新策略忽略了网络中节点与节点之间联系的紧密程度,Lou 等[11]将相干邻居亲近度引入 LPA 算法,提出了 CNP-LPA 算法。

### 1.2　CNP-LPA

#### 1.2.1　相干邻居亲近度

Zhang 等[10]提出了相干邻居亲近度(Coherent Neighborhood Propinquity)的概念,用来度量网络中任意节点对 $(v_1,v_2)$ 之间的亲近程度,如式(2)所示:

$$P(v_1,v_2)=|E(v_1,v_2)|+|N(v_1)\cap N(v_2)|+|E(G[N(v_1)\cap N(v_2)])|, \tag{2}$$

$P(v_1,v_2)$ 表示网络中任意节点对 $(v_1,v_2)$ 的 CNP 值，它由 3 个部分组成。$|E(v_1,v_2)|$ 表示节点 $v_1$ 与 $v_2$ 直接相连的边的数目。$|N(v_1)\cap N(v_2)|$ 表示 $v_1$ 与 $v_2$ 共享的邻居节点数，它的值等于 $v_1$ 邻居节点集合与 $v_2$ 邻居节点集合的交集大小。$|E(G[N(v_1)\cap N(v_2)])|$ 表示包含在 $v_1$ 与 $v_2$ 共享邻居节点集合的交集节点间连接的边的数目。$P(v_1,v_2)$ 值越大，说明节点 $v_1,v_2$ 之间联系越紧密，也越可能同属于一个社区。

图 1a 中 $v_A$ 与 $v_B$ 共享的邻居节点有 3 个，如图 1b 所示，$|N(v_A)\cap N(v_B)|=3$。$v_A$ 与 $v_B$ 共享邻居节点 C、D、E 之间连接的边有 2 条，如图 1c 所示，$|E(G[N(v_A)\cap N(v_B)])|=2$。图 1a 中，$v_A$ 与 $v_B$ 间存在边，$|E(v_A,v_B)|=1$。因此，对于图 1a 中节点对 $(v_A,v_B)$，$P(v_A,v_B)=1+3+2=6$。

同理对于节点对 $(v_D,v_E)$，$|N(v_D)\cap N(v_E)|=3$，$|E(G[N(v_D)\cap N(v_E)])|=3$。而在图 1a 中，$v_D$ 与 $v_E$ 并不存在边，$|E(v_D,v_E)|=0$，$P(v_D,v_E)=0+3+3=6$。求出网络中所有节点对的 CNP 值后，就可以将原始网络转换为相应的 CNP 网络。图 1d 是图 1a 生成的 CNP 网络，边上的数字表示相关节点对的 CNP 值。图 1 中原始网络 $G$ 中边的数目 $|E|=9$，而相应的 CNP 网络 $Gp$ 中边的数目 $|Ep|=10$。

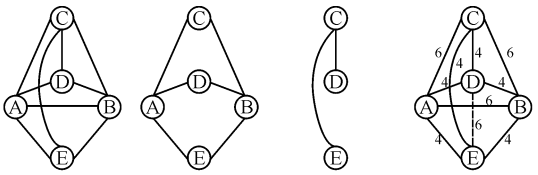


图 1 相干邻居亲近度

Fig. 1 Coherent Neighborhood Propinquity

### 1.2.2 CNP-LPA 算法

CNP 网络形式化描述为 $Gp=(V,Ep,P)$，$V(v\in V)$ 表示节点集，$Ep(ep\in Ep)$ 表示边集。$\forall v_1,v_2\in V$，且 $ep(v_1,v_2)\in Ep$，则 $P(v_1,v_2)$ 表示节点对 $(v_1,v_2)$ 的 CNP 值。

CNP-LPA 算法在预处理阶段将原始网络 $G$ 转换为 CNP 网络 $Gp$，并基于 $Gp$ 传播标签。更新标签时，根据节点对的 CNP 值对标签进行加权。两个节点之间的 CNP 值越大，表示它们联系越紧密，它们的标签对对方的影响也就越大，选择总的 CNP 值最大的标签进行更新，更新规则如式(3)所示：

$$c_v=\arg\max_l\sum_{s\in v^l}P(v,s)\ 。\tag{3}$$

在实际的社交网络中，影响力较小的节点所属的社区往往依赖于本区域内影响力较大的节点。可以发现图 2 中 $v_{12}$ 所属的社区一定和 $v_1$ 所属社区保持一致，同样 $v_{16}$ 所属社区依赖于 $v_{33}$ 和 $v_{34}$，本研究只考虑单个节点之间的依赖度，将节点间依赖度引入 CNP-LPA 算法，提出了 CNP-LPA+ 算法。

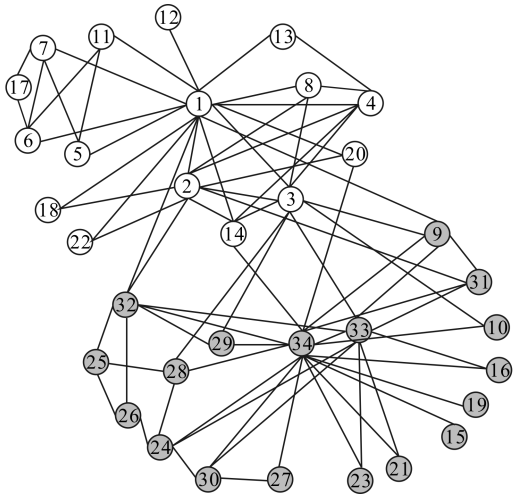


图 2 空手道关系网络图

Fig. 2 Zachary's karate club network

## 2 LPA-CNP+

### 2.1 相关定义

**节点间依赖度定义：**无向简单网络 $G=(V,E)$，$\forall u,v\in V$，且 $e(u,v)\in E$，则 $D(u,v)$ 表示节点 $u$ 依赖于节点 $v$ 的程度，$D(u,v)$ 值越大，节点 $u$ 越依赖于节点 $v$，节点 $u$ 所属的社区越可能与节点 $v$ 所属社区保持一致。

给出一种直观的方法来度量节点间依赖度，节点 $u$ 的覆盖范围包含在节点 $v$ 覆盖范围之中的程度就是节点 $u$ 依赖于节点 $v$ 的程度。

$$D(u,v)=\frac{|N(u)\cap N(v)|+1}{|N(u)|}。\tag{4}$$

排除没有邻居节点的孤立节点，则 $D(u,v)\in(0,1]$，当 $D(u,v)=1$ 时，表示节点 $u$ 在网络中的覆盖范围完全包含在节点 $v$ 的覆盖范围之中。

**冗余节点定义：**$u\in V,\exists v\in N(u)$，使 $D(u,v)\geqslant c$（$c$ 为给定的阈值），说明节点 $u$ 覆盖范围大部分包含在节点 $v$ 的覆盖范围中，确定节点 $v$ 的社区就可以确定节点 $u$ 所属的社区，称节点 $u$ 为冗余节点。

**核心节点定义：**$u\in V,\forall v\in N(u)$，使 $D(u,v)<c$，即节点 $u$ 的任意邻居节点 $v$ 的覆盖范围都不能大部分包含节点 $u$ 的覆盖范围，称节点 $u$ 为核心节点。

### 2.2　核心 CNP 网络

核心 CNP 网络形式化描述为 $Gc=(Vc,Ec,Pc)$,$Vc(v_c\in Vc)$ 表示核心节点集,$Ec(ec\in Ec)$ 表示核心边集,且满足 $Ec\subset Vc\times Vc$。$\forall v_1,v_2\in Vc$,且 $ec(v_1,v_2)\in Ec$,则 $Pc(v_1,v_2)$ 表示核心节点对 $(v_1,v_2)$ 的 CNP 总值。

图 3a 是空手道关系网络的一部分,图 3b 是其对应的 CNP 网络,可以看出后者相比前者增加了 5 条边。为了避免这个问题,作如下约定:原始网络 $G=(V,E)$ 中的节点对 $(v_1,v_2)$,如果 $v_1,v_2$ 间存在边,则保留该节点对的 CNP 值;如果不存在边,则将该节点对的 CNP 值清零。图 3a 中节点对 $(v_1,v_{17})$,$(v_5,v_6)$,$(v_5,v_{17})$,$(v_7,v_{11})$,$(v_{11},v_{17})$ 间均不存在边,因此忽略这 5 对节点的 CNP 值,图 3c 就是图 3a 对应简化的 CNP 网络。在图 3a 中,$D(v_{17},v_6)=D(v_{17},v_7)=1$,$v_{17}$ 的覆盖范围被完全包含在 $v_6$ 或者 $v_7$ 的覆盖范围中,$v_{17}$ 所属的社区可以与 $v_6$ 保持一致,也可以并入 $v_7$ 所属的社区。$D(v_5,v_1)=D(v_{11},v_1)=1$,$v_5$ 与 $v_{11}$ 的覆盖范围完全包含在 $v_1$ 的覆盖范围中,则 $v_5$ 与 $v_{11}$ 可以直接并入 $v_1$ 所属的社区。在后续标签传播过程中不再考虑 $v_5,v_{11},v_{17}$。

CNP-LPA+算法的主要思想就是在预处理阶段时,首先将原始网络转换为简化的 CNP 网络,在简化 CNP 网络的基础上,合并冗余节点到相应的核心节点,冗余节点与第三方节点之间的 CNP 值继承到相应的核心节点上。接着根据预处理阶段得到的核心 CNP 网络传播标签,传播过程完成后,将有相同标签的核心节点归入同一社区,冗余节点所属社区与相应的核心节点保持一致。合并过程如下:

在图 3c 的基础上,将节点 17 并入节点 6,并更新节点间的 CNP 值:

a) 令 $P(v_6,v_{17})=0$;

b) $P(v_6,v_7)=P(v_6,v_7)+P(v_7,v_{17})=3+2=5$;

c) $P(v_7,v_{17})=0$;

更新后如图 3d 所示,接着在图 3d 基础上将节点 5 并入节点 1,并更新节点间的 CNP 值:

d) 令 $P(v_1,v_5)=0$;

e) $P(v_1,v_7)=P(v_1,v_7)+P(v_5,v_7)=3+2=5$;

f) $P(v_5,v_7)=0$;

g) $P(v_1,v_{11})=P(v_1,v_{11})+P(v_5,v_{11})=3+2=5$;

h) $P(v_5,v_{11})=0$;

更新后如图 3e 所示,接着在图 3e 基础上将节点 11 并入节点 1,并更新节点间的 CNP 值:

i) $P(v_1,v_{11})=0$;

j) $P(v_1,v_6)=P(v_1,v_6)+P(v_6,v_{11})=3+2=5$;

k) $P(v_6,v_{11})=0$。

到此为止,合并全部完成。图 3f 就是图 3a 对应的核心 CNP 网络,其边数仅仅只有 3 条。预处理

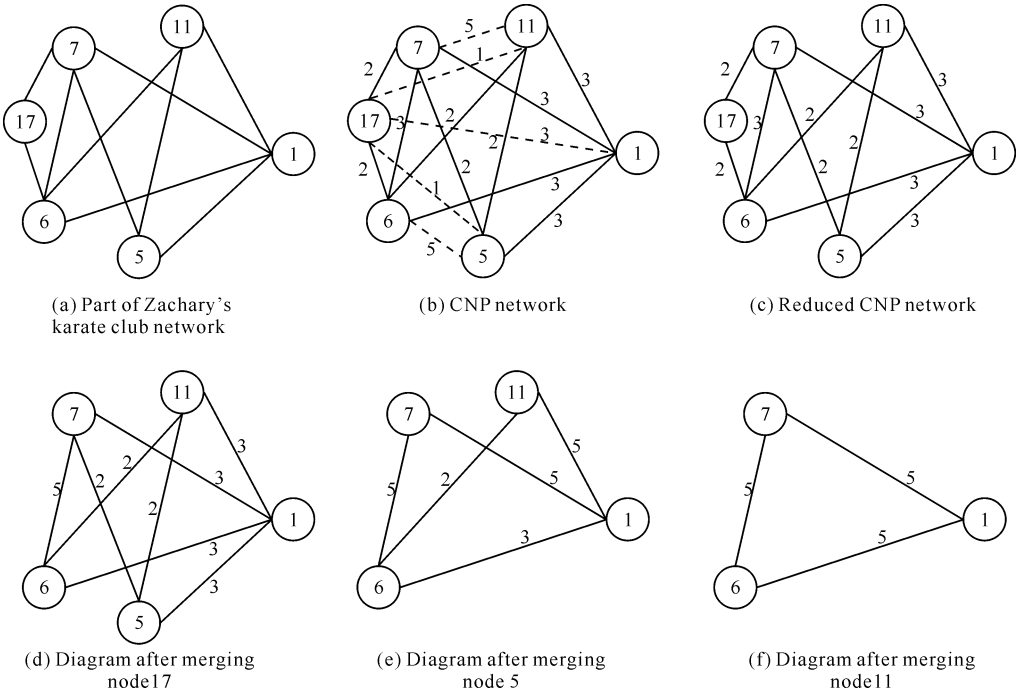


图 3　冗余节点合并过程

Fig. 3　Process of merging redundant nodes

阶段结束后，CNP-LPA＋算法在得到的CNP核心网络基础上传播标签，更新标签的规则如式(5)，其中 $v \in v_c$，$v_c^l$ 表示 $v$ 的邻居节点中标签为 $l$ 的核心节点集合，$Pc(v,s)$ 即核心节点对 $(v,s)$ 之间的CNP总值。

$$c_v = \arg\max_{l} \sum_{s \in v_c^l} Pc(v,s) 。 \quad (5)$$

### 2.3 CNP-LPA＋算法实现

这里给出CNP-LPA＋算法的伪代码：

Input：网络 $G(V,E)$，节点间依赖度阈值 $c$。

步骤1：计算每条边 $e(u,v)$ 对应节点对 $(u,v)$ 间CNP值，得到 $G$ 的简单CNP网络；

步骤2：根据给定的阈值 $c$，合并所有冗余节点，得到核心CNP网络 $Gc=(Vc,Ec,Pc)$；

步骤3：给每个核心节点 $v \in Vc$ 赋予唯一的标签；

步骤4：迭代次数 $t=1$；

步骤5：随机排序核心节点集 $Vc$；

步骤6：基于 $Gc=(Vc,Ec,Pc)$ 传播标签，根据式(5)更新标签，采用异步更新的方式来避免出现标签震荡；

步骤7：若 $t$ 达到设置的最大迭代次数或每个核心节点标签CNP权值达到最大，算法结束；否则 $t=t+1$，返回步骤5。

算法运行完成后，将具有相同标签的核心节点归入同一社区，冗余节点所属社区与相应的核心节点保持一致。CNP-LPA＋算法同时考虑到了节点间亲近度与节点间依赖度，在预处理阶段就将冗余节点并入本区域内的核心节点，并基于核心CNP网络传播标签，不仅减少了标签传播花费的时间，更重要的是在社区发现准确性方面有了显著的提高。

### 2.4 CNP-LPA＋算法时间复杂度分析

步骤1中 $\forall e(u,v) \in E$，需要计算节点对 $(u,v)$ 的CNP值 $P(u,v)$，时间复杂度为 $O(|E|)$；

步骤2中需要合并所有冗余节点，$\forall v \in V$，$v$ 是否为冗余节点，只需要在其邻居节点范围内寻找是否存在节点包含了 $v$ 的大部分覆盖范围。如果是则合并冗余节点 $v$，时间复杂度为 $O(|E|)$；

步骤3中初始化核心节点标签时间为 $O(|V|)$；

步骤5中随机排序核心节点序列 $O(|V|)$；

步骤6中一次标签传播时间为 $O(|E|)$；

步骤7中生成社区的时间为 $O(|V|)$。

可见CNP-LPA＋算法与标准LPA算法同样具有接近线性的时间复杂度。

## 3 实例验证

本研究选取6组真实的社交网络数据集分别对LPA算法、CNP-LPA算法、CNP-LPA＋算法进行测试分析。表1是这些数据集的介绍。由于这些算法的结果都具有一定的随机性，因此所有的实验都进行10次，结果取平均值。3种算法均使用C＋＋语言实现，实验环境为Windows10，Inter(R) Core(TM) i5-4690 CPU@3.50GHz，8 GB内存。

**表1 真实社交网络数据集**

**Table 1 Real-world networks with community structure**

| Network | Description | Nodes | Edges | Edges/Nodes |
|---|---|---|---|---|
| karate[13] | Zachary's karate club | 34 | 78 | 2.29 |
| books[14] | Books about US politics | 105 | 441 | 4.2 |
| netscience[15] | Network scientists | 1 589 | 2 742 | 1.73 |
| astro-ph[16] | Astrophysics collaborations | 16 706 | 121 251 | 7.26 |
| cond-mat[16] | Condensed matter collaborations | 16 726 | 47 594 | 2.85 |
| dblp[17] | Computer science bibliography | 326 186 | 1 615 400 | 4.95 |

CNP-LPA算法在预处理阶段将原始网络 $G$ 转换为CNP网络 $Gp$，而CNP-LPA＋算法则需转换为核心CNP网络 $Gc$（注意给定不同的阈值参数 $c$，$G$ 所得到的核心CNP网络 $Gc$ 也不同，实验中阈值 $c$ 默认取0.8），$G$、$Gp$、$Gc$ 边数的不同，后续标签传播过程花费时间也不同。图4中蓝色柱条代表CNP网络，红色柱条代表核心CNP网络，柱条上的值等于两者边数比上原始网络边数的比值。karate数据集原始网络有78条边，其对应的CNP网络则有343条边（343/78＝4.4），而当阈值 $c$ 取0.8时，它的核心CNP网络仅有32条边（32/78＝0.41）。

从图4可以看出6组数据集的CNP网络 $Gp$ 边数均远远超过了原始网络 $G$ 的边数，其中astro-ph数据集 $Gp$ 边数是 $G$ 边数的14.7倍，是6组数据集中最高的。与此同时所有数据集的核心CNP网络 $Gc$ 的边数都小于 $G$ 的边数，其中netscience数据集 $Gc$ 的边数仅是 $G$ 边数的1/20，只保留了原始网络中最关键的边。

预处理阶段结束后，CNP-LPA算法基于CNP网络传播标签，而CNP-LPA＋算法基于核心CNP网络传播标签，显然边数越少标签传播越快。设置最大迭代次数为20次，6组数据集在CNP-LPA算

法与 CNP-LPA+算法下传播标签平均所花费时间如表 2 所示。

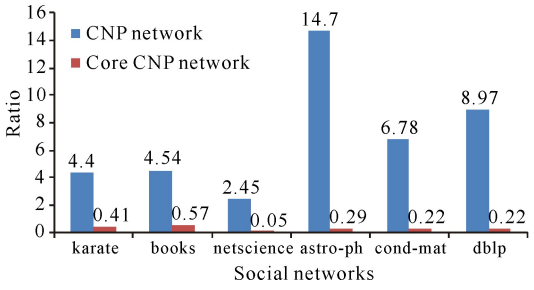


图 4　CNP 网络与核心 CNP 网络边数

Fig. 4　Edges of CNP network and core CNP network

**表 2　标签传播时间**

**Table 2　Time of spreading label**

| Network | Time (s) | |
|---|---|---|
| | CNP-LPA | CNP-LPA+ |
| karate | <0.01 | <0.01 |
| books | <0.01 | <0.01 |
| netscience | 0.243 | <0.01 |
| astro-ph | 29.0 | 0.257 |
| cond-mat | 3.143 | 0.083 |
| dblp | >129 | 2.056 |

由表 2 可以看出,当网络数据集中节点与边数较少时,CNP-LPA 算法传播时间与 CNP-LPA+传播时间的差距尚处于可以容忍的范围内,但当节点数增大时,传播时间的差距就变得非常明显。如 astro-ph 数据集,CNP-LPA 算法标签传播平均需要 29 s,而 CNP-LPA+算法平均仅需要 0.257 s;特别是 dblp 数据集,在 20 次的迭代范围内,CNP-LPA 算法标签传播并没有结束,而此时所花费时间已经超过 129 s,而 CNP-LPA+算法则正常运行结束,平均花费时间只有 2.056 s,明显优于前者。

使用模块度 $Q$ 作为评价社区质量的指标,分别对 LPA、CNP-LPA、CNP-LPA+ 3 种算法进行测试分析,结果如表 3 所示。

**表 3　模块度 $Q$ 对比**

**Table 3　Average modularity of LPA, CNP-LPA, CNP-LPA+**

| Network | Modularity | | |
|---|---|---|---|
| | LPA | CNP-LPA | CNP-LPA+ |
| karate | 0.356 | 0.276 | 0.373 |
| books | 0.495 | 0.451 | 0.509 |
| netscience | 0.895 | 0.936 | 0.956 |
| astro-ph | 0.621 | 0.625 | 0.680 |
| cond-mat | 0.710 | 0.767 | 0.795 |
| dblp | 0.761 | 0.785 | 0.812 |

由表 3 可以看出,除了前 2 个数据集,CNP-LPA 算法全部优于 LPA 算法,而 CNP-LPA+算法在所有数据集上优于 LPA 和 CNP-LPA 算法。特别是在 karate,books,astro-ph 数据集上,在 CNP-LPA 算法相比 LPA 算法并无提升的情况下,CNP-LPA+算法获得了很好的效果,达到或者明显超过了 LPA 算法划分的准确性。而在 netscience,cond-mat,dblp 3 组数据集上,在 CNP-LPA 算法相比 LPA 算法已经获得了明显的提升的基础上,CNP-LPA+算法相比 CNP-LPA 算法获得了进一步的提升,这充分说明了 CNP-LPA+算法的有效性。

综上所述,CNP-LPA+算法相比较 CNP-LPA 算法不仅减少了标签传播的时间,最重要的是在社区发现准确性方面有了显著的提高。

## 4　结论

本研究在分析 CNP-LPA 算法的特点和不足的基础上,提出了一种新的基于相干邻居亲近度的标签传播算法 CNP-LPA+,该算法同时考虑到相干邻居亲近度与节点间依赖度,并基于核心 CNP 网络传播标签,相比 CNP-LPA 算法不仅减少了标签传播过程所花费的时间,进一步提高了社区发现的准确性。CNP-LPA+算法仍然保持近似线性的时间复杂度。在接下来的工作中,将考虑更多节点间依赖度的度量方法,比较不同方法对标签传播算法的影响,还可以将节点间依赖度应用到具有重叠现象的社区发现等问题中。

**参考文献:**

[1] 林友芳,王天宇,唐锐,等. 一种有效的社会网络社区发现模型和算法[J]. 计算机研究与发展,2012,49(2):337-345.
LIN Y F, WANG T Y, TANG R, et al. An effective model and algorithm for community detection in social networks[J]. Journal of Computer Research and Development, 2012, 49(2): 337-345.

[2] 刘大有,金弟,何东晓,等. 复杂网络社区挖掘综述[J]. 计算机研究与发展,2013,50(10):2140-2154.
LIU D Y, JIN D, HE D X, et al. Community mining in complex networks[J]. Journal of Computer Research and Development, 2013, 50(10): 2140-2154.

[3] SERRANO M Á, BOGUÑÁ M, SAGUÉS F. Uncovering the hidden geometry behind metabolic networks [J]. Molecular Biosystems, 2011, 8(3): 843-850.

[4] 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.

[5] CAPOCCI A,SERVEDIO V D P,CALDARELLI G,et al. Detecting communities in large networks[J]. Physica A:Statistical Mechanics and Its Applications,2005,352(2/3/4):669-676.

[6] BOCCALETTI S,LATORA V,MORENO Y,et al. Complex networks:Structure and dynamics[J]. Physics Reports,2006,424(4/5):175-308.

[7] 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.

[8] FORTUNATO S. Community detection in graphs[J]. Physics Reports,2010,486(3/4/5):75-174.

[9] RAGHAVAN U N,ALBERT R,KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E,Statistical Nonlinear,and Soft Matter Physics,2007,76(3):036106.

[10] ZHANG Y Z,WANG J Y,WANG Y,et al. Parallel community detection on large networks with propinquity dynamics[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. York,NY,USA:ACM,2009.

[11] LOU H,LI S H,ZHAO Y X. Detecting community structure using label propagation with weighted coherent neighborhood propinquity[M]//BRESNAN J (ed.). The mental representation of grammatical relations. Massachusetts:MIT Press,1982:3095-3105.

[12] NEWMAN M E J,GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2004,69(2):026113.

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

[14] NEWMAN M E. Modularity and community structure in networks[J]. P Natl Acad Sci USA,2006,103(23):8577-8582.

[15] FIEDLER M. Algebraic connectivity of graphs[J]. Czech Math J,1973,23(98):298-305.

[16] NEWMAN M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences of the United States of America,2001,98(2):404-409.

[17] Laboratory for Web algorithmics[EB/OL]. [2016-11-12]. http://law.di.unimi.it/datasets.php.

(责任编辑:陆 雁)