基于贡献函数的重叠社区划分算法

刘功申 孟 魁* 郭弘毅 苏 波 李建华
(上海交通大学电子信息与电气工程学院 上海 200240)

摘 要：现实世界中的网络结构呈现出重叠社区的特征。在研究经典的标签算法的基础上，该文提出基于贡献函数的重叠社区发现算法。算法将每个节点用三元组(阈值、标签、从属系数)集合来表示。节点的阈值是每次迭代过程中标签淘汰的依据，该值由多元线性方程自动计算而来。从属系数用于衡量当前节点与标签所标识社区的相关度，从属系数的值越大说明该节点与标签所标识社区的关联性越强。在每一次迭代的过程中，算法依据贡献函数计算每个节点的从属系数，并生成新的三元组集合。然后依据标签决策规则淘汰标签，进行从属系数规范化。通过对真实的复杂网络和 LFR(Lancichinetti Fortunato Radicchi)自动生成的网络进行测试可知，该算法的社区划分准确率高，而且划分结果稳定。

关键词：复杂网络；社区发现；重叠社区
中图分类号： TP309 **文献标识码**：A **文章编号**：1009-5896(2017)08-1964-08
DOI: 10.11999/JEIT161109

Overlapping-communities Recognition Algorithm Based on Contribution Function

LIU Gongshen MENG Kui GUO Hongyi SU Bo LI Jianhua
(*School of Electronic Information and Electrical Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: Overlapping is one of the most important characteristics of real-world networks. Based on the classic labeling algorithm, the overlapping-community orientated label propagation algorithm based on contribution function is proposed. In this algorithm, each node is indicated by a set of triples (threshold, label, and coefficient). The threshold value of every node is used as a metric for labels decision, which is calculated automatically by multiple linear regression equation. The dependent coefficient is used to measure the relevance of the current node with the correspondent community which is marked by the label. A greater value of dependent coefficient means a stronger association between the node and the community. During each iteration process, the dependent coefficients are calculated through Contribution Function (CF) of each node, and new triples are produced. Then the labels in terms of decision rules are selected, and the dependent coefficients of the node are normalized. According to the tests with real-world networks and automatic generation of LFR (Lancichinetti Fortunato Radicchi) test network, the algorithm can divide communication with high accuracy and robust result.

Key words: Complex networks; Communities detecting; Overlapping communities

1 引言

经典的社区发现算法假设某个节点仅仅属于一个特定社区，这种假设显然和现实不完全相符。在现实世界的复杂网络系统中，存在一些节点同时属于不同的社区，这便是所谓的重叠社区结构[1]。针对重叠社区的挖掘问题，Palla 等人[2]提出了针对重叠社区的派系过滤(Clique Percolation Method, CPM)算法，该算法将社区视作由一些互相连通的完全子图构成的集合。将节点数目为 k 的完全子图定义为 k-clique，当两个 k-clique 之间拥有 $k-1$ 个公共节点时，则认为这两个 k-clique 是相邻的。自从 Palla 提出 Clique 概念以来，许多研究者便不断尝试提出新的基于 Clique 思想的社区发现算法[3]。

Gregory[4]在 GN 算法的基础上提出了能发现重叠社区的 CONGA (Cluster-Overlap Newman Girvan Algorithm)算法，其主要思想是将边介数较高的节点再分裂成多个副本。对于分裂后的节点，采取经典的社区发现算法进行挖掘社区结构。由于

收稿日期：2016-10-18；改回日期：2017-04-24；网络出版：2017-05-26
*通信作者：孟魁 mengkui@sjtu.edu.cn
基金项目：国家 973 关键技术研究项目(2013CB329603)，国家自然科学基金(61472248)
Foundation Items: The National 973 Key Basic Research Program of China (2013CB329603), The National Natural Science Foundation of China (61472248)

某些节点在算法运行过程中分裂成多个副本，这些副本在后续运行过程中可能被划分到不同的社区，并最终实现一个节点可以同时属于多个社区的目的。

文献[5]采用了非负矩阵分解来发现重叠社区结构，采用非负矩阵分解在重叠社区发现任务上具有很好的准确性和良好的解释性。文献[6]提出了基于非负矩阵分解的图规范化方法，在该方法中使用了能反映节点间相似度的度量方法，实验证明这种方法能提高社区划分的准确度。文献[7]提出了一种基于非负矩阵的半监督社区划分方法，该方法利用了标签的先验知识进行训练，在划分工程中归并标签。

Raghavan 等人[8]提出的基于标签传播的社区发现算法(Label Propagation Algorithm, LPA)具有非常好的时间复杂度，这促使 Gregory[9]将 LPA 算法从非重叠社区拓展到了重叠社区，提出了基于标签传播的重叠社区发现算法。Gregory 的算法[9]继承了传统标签的优秀的时间复杂度指标，在不同的网络数据上具有较好的测试结果。但是，该算法也具有明显的劣势：(1)继承了传统标签传播算法的缺点—随机性。(2)算法执行时需要事先设置参数 v，而且 v 的不同取值对算法的结果影响非常大。但如何选择合适的 v 值是个难题。

本文在分析现有算法的基础上，提出基于贡献函数的重叠社区划分标签传播算法(Overlapping-Communities Recognition Algorithm based on Contribution Function, OCRA-CF)。通过引入贡献函数，克服了 Gregory 算法的随机行为，使算法结果更加稳定。同时能够自动计算各个节点的阈值，避免了 Gregory 所提算法中 v 值选择问题。此外，OCRA-CF 还具有较好的扩展性及并行计算优势。

2　算法设计

OCRA-CF 算法的创新改进在于：(1)为每个节点自动计算阈值，该阈值反映了某个节点属于多个社区的可能程度。(2)采用贡献函数计算从属系数。贡献函数反映了邻居节点的贡献情况。(3)采用较稳定的规则进行标签淘汰，最大程度地提高了算法结果的稳定性。

2.1 重要函数

在计算过程中，每一个节点 v 表示为一个集合 S，该集合的元素在整个计算过程中可能会一直发生变化。S 的每个元素是一个三元组 $(\mathrm{thr}, \mathrm{lab}, \mathrm{coe})$，其中，thr 表示该节点的阈值，参数 lab 表示社区划分的标签，而参数 coe 则称为从属系数。阈值 thr 在迭代过程中用作淘汰标签的标准，从属系数 coe 用于衡量节点 v 与社区 lab 的相关程度。coe 的值越大说明该节点 v 与社区 lab 的关联性越强。对于节点 v，它的所有从属系数的和保持为 1。

在每一次迭代的过程中，算法将决定每一个节点的所有邻接点的标签以及该标签对下一轮的贡献度。算法的核心由 4 个因素构成：阈值计算、贡献函数、从属系数和选择规则。在本文中，每个节点拥有独特的阈值，并且实现了阈值的自动计算。从属系数都经过规范化处理过程，某个节点的所有从属系数的值相加为 1。同样地，贡献函数的取值也经过规范化处理，某个节点的贡献函数的值相加也是 1。最后，根据选择规则来决策保留那些标签或者淘汰那些标签。

2.1.1 阈值计算　每个节点的阈值和该节点可能所属的社区数相关。事实上，节点的阈值 thr 就是该节点所属社区数的倒数。因此，估算社区数是阈值计算的核心任务。但在未执行完该社区划分算法之前，不可能知道每个节点实际属于几个社区，因此，只能采用估算的方法来预测节点可能属于几个社区。本文提出了为每个节点估算所属社区数量的方法，该方法主要包括估算社区度和计算社区数两个步骤。

(1)估算社区度：　所谓社区度就是衡量节点属于多个社区的可能程度。社区度是本文为每个节点计算的一个数值，该值越大说明该节点同时属于多个社区的可能性就越大，反之亦然。

本文把经典网络的标准划分结果作为先验知识，把每个节点的静态特征数据作为后验知识，通过多元线性回归模型拟合出估算社区数的方程。社区数估算函数为

$$\mathrm{ext} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \tag{1}$$

式中，ext 的取值区间为大于等于 0 的实数。为了使用式(1)为每个节点估算其所属的社区数，需要求出式(1)中的参数 $b_i (0 \le i \le n)$。求解参数的过程是经典的多元线性回归方法。公式拟合需要的先验数据从标准的 Karate 网络、Football 网络和 Dolphins 网络获得。式(1)中的 x_i 为每个节点的网络静态特征值(例如节点的度、介数、接近度、权威度等)。后文将给出详细的拟合过程。

(2)计算社区数：　计算社区数的工作就是根据已知的社区度估算出该节点可能同时属于的社区数量。根据美国认知科学家 George 的研究，人类短期记忆一般一次只能记住 5~9 个事物，也就是常说的“7 加减 2”原则。近年来对社区网络的数据统计分析的结论也证实，现实社会中的自然人或者网络中的虚拟人(即社会网络节点)尽管有较多的圈子存

在，但短期内频繁交往的圈子数约为 7 ± 2 个。基于此，本文把每个节点可能属于社区数的最大值定为 9，并按节点的社区度为每个节点赋予 $(0,9]$ 区间上的整数值。

假设 Y_0 取值为所有节点中社区度的最小值，Y_9 取值所有节点中社区度的最大值，且 $Y_0<Y_1<\cdots<Y_9$。则社区数函数表示为

$$f(\text{ext})=i,\ Y_{i-1}<\text{ext}\le Y_i,\ i\in(0,9] \tag{2}$$

令函数 $\text{count}(i)$ 表示社区数取值为 i 的节点总数。$\sum\text{count}(i),\ i\in(0,9]$ 就是社区节点总数。式(2)中 Y_i 的取值直接影响了 $\text{count}(i)$ 的值。确定 Y_i 的值时，有多种可选择策略：采用平均分段(本文举例部分采用该策略)、正态分布、偏态分布等。选择 Y_i 的最佳策略是使 $\text{count}(i)$ 在 $i\in(0,9]$ 区间上符合以 7 为峰值的负偏态分布。

2.1.2 贡献函数 对于某个节点 v，存在多个相邻节点 $N(v)$。把这些相邻节点的标签及从属系数，经过贡献函数计算加权后得到 v 的集合 S 的过程就是贡献计算。贡献函数的定义为

$$F(v,x)=\left\{f_y(x)\right\} \tag{3}$$

其中，$y\in N(v)$，x 是函数参数。

贡献函数 $F(v,x)$ 是针对节点 v 定义的一系列函数 $f_y(x)$ 的集合。$f_y(x)$ 表示 v 的第 y 个邻居节点对节点 v 的贡献度。$f_y(x)$ 的定义方法灵活多样，既可以根据节点的度、介数、接近度、权威度分别计算，也可以考虑把这些因素组合起来计算。$f_y(x)$ 的一般化定义为

$$f_y(\boldsymbol{X})=\frac{\sum\limits_{Y=y}\sum\limits_{i\in I}w_ix_i}{\sum\limits_{Y\in N(v)}\sum\limits_{i\in I}w_ix_i} \tag{4}$$

其中，x_i 是参与计算的所有节点网络静态特征(例如度、介数、接近度、权威度等)，i 标识了有多少项静态特征参与计算。相应地，w_i 就是 x_i 的权重值。在采用多个参数参与计算时，需要对 w 的取值进行估计。

2.1.3 从属系数计算 从属系数描述了当前节点 v 隶属于标签 lab 的程度，该值在每轮迭代中进行重新计算。从属系数计算是指，在第 t 次迭代中，节点 v 对于标签 lab 的从属系数计算方法。该计算方法的描述为

$$\text{coe}_t(\text{lab},v)=\sum_{y\in N(v)}\text{coe}_{t-1}(\text{lab},y)\cdot f_y(x) \tag{5}$$

其中，$N(v)$ 代表节点 v 的邻接点的集合。参数 t 和 $t-1$ 表示不同的迭代次数。

2.1.4 标签决策 标签决策是标签选择或淘汰的规则。在算法的每轮迭代中，对于每一个节点，需要按照一定的规则在集合 S 中选择标签或淘汰标签。

对于节点 v 及其对应的集合 S，标签处理规则如下：

(1)如果 S 中存在从属系数大于 v 的阈值的标签，则保留这些标签，删除其他标签，并对保留标签重新进行规范化。否则，进行(2)。

(2)如果 S 中存在多个并列最大值标签，则在 S 中多个并列最大值中随机选择标签，保留这个随机值，删除其他标签。并对保留标签重新进行规范化。否则，进入(3)。

(3)在 S 中随机选择标签，并保留这个随机值，删除其他标签。然后，对保留标签进行重新规范化。

2.2 参数拟合

参数拟合阶段的主要工作是分别求解式(1)参数 b_i 和式(4)中参数 w_i。(本部分工作参见本文实验部分)。经过参数拟合后，式(1)和式(4)完成了实例化，以备后续步骤中进行相关计算。

2.3 算法流程

算法的整个流程包括 4 个步骤：计算网络静态特征值、计算阈值、节点赋初值和迭代过程。其中，迭代过程是算法的核心部分，同时也是时间复杂度最高的部分。

(1)计算网络静态特征值： 网络静态特征值是指节点的度(degree)、介数(betweenness)[10]、接近度(closeness)[11]、权威度(authority)[12]等。本文的实验部分使用这几个参数完成了算法。作为对本文的扩展，在实际工程中还可以增加其他参数，或者仅仅使用其中的部分参数。

(2)计算节点阈值： 通过式(1)和式(2)，为每个节点估算出社区数，那么，该节点的阈值(thr)就是社区数的倒数。

(3)节点赋初始值： 迭代开始前为每个节点赋值一个标签，且从属系数为 1。每个节点初始状态包含如下两项内容：S (只包含一个三元组)和贡献函数表。

(4)迭代过程： 每一轮迭代主要包括 3 项任务，即从属系数计算、标签决策和终止条件判断。

3 实验与分析

3.1 实验数据

在社区发现算法的研究领域中，有两种方式对算法进行评价。一种是使用现实世界的网络数据集对算法进行测试。由于真实网络结构的复杂性，任何划分都不具有绝对的正确性，因此，需要通过一些社区评价指标作为评价标准。典型的评价指标包

括经典算法所提出的 Q 函数和 Mod 模块度[13]等。实验采用了许多具有代表性和研究价值的真实网络数据集。其中包括著名的空手道俱乐部数据集(Karate_Club，简称 Karate)、海豚网络数据集(Dolphin_Social，简称 Dolphin)、美国大学生职业足球联盟数据集(American_Football，简称 Football)、SNS 数据集(Social_Community，简称 S_C)、豆瓣数据集(Douban)、PGP 数据集(PGP)、安然邮件系统数据集(Eron_mail，简称 Eron)等。Z_follow 是匿名发布者在北邮人(sns.byr.edu.cn)上的所有 Follow 节点组成的 SNS 网络。Z_Friends 是所有 Follow 的 Follow 节点形成的 SNS 网络。根据数据提供者的说明，Z_Friends 网络的模块度是 0，也就是说，这个社区不存在小的社区结构。Bupt 是整个北邮人 SNS 网络的所有节点形成的社区。

另一种方式是利用人工构造数据集对算法进行评价。在利用人工构造数据集时，可以通过设定节点数、边数、每个节点的平均度数、重叠度等控制变量来生成结构固定的网络。由于生成的网络结构固定，因此可以有针对性地对算法进行评价。此外，还可以通过改变一到两个构造社区的参数变量，来针对性地测试算法的对不同网络的适应性。Lancichinetti 等人[14]提出的 LFR 基准程序是目前公认的构造人工网络的程序。

3.2 式(1)的参数估计

式(1)为典型的多元线性方程，本文采用多元线性回归模型拟合该方程的参数。参数的估计过程主要包括：数据准备和线性回归过程。

(1)数据准备：　国内外科研人员对 Karate, Football 和 Dolphins 网络进行了充分研究，给出了公认的划分结果[15]。本文基于这些公认的划分结果，每个节点准备一个标准值，并设计了一个算法来计算该值。该算法主要由以下 3 个部分组成：

(a)节点连接的社区数：　节点和社区的连接定义为："如果节点 v 和社区 A 中的任意节点有边相连，则称节点 v 和社区 A 有连接"。根据标准的社区划分，能方便地统计每个节点分别和多少社区有连接(记为 C)。

(b)度及其惩罚因子：　此处的"度"即传统的度(记为 d)。度非常高的节点倾向于作为社区的中心而不是跨社区的桥接点，因此，需要对特别高的度进行惩罚，记为 $1-\mathrm{e}^{-d/\overline{d}}$。其中，$\overline{d}$ 为网络中所有节点度的平均值。

(c)指向差异：　指向差异是指某个节点和多个社区连接情况的差异度量。对于节点 v 而言，如果该节点和 n 个社区有连接，且 v 分别有 $X_1, X_2, \cdots, X_n$ 条边连接到社区 $C_1, C_2, \cdots, C_n$。当 $X_1, X_2, \cdots, X_n$ 无明显差异时，v 更倾向于被当作 n 个社区的桥接点，也就是重叠节点。反之，当 $X_1, X_2, \cdots, X_n$ 差异较大时，v 更倾向于被划分到 X 值较大的那个社区，也就是说 v 不是重叠节点。图 1(a)图和图 1(c)的节点 v 都倾向于作为社区 C_1 的节点，而图 1(b)的节点 v 更倾向于作为 C_1 和 C_3 的桥接点。

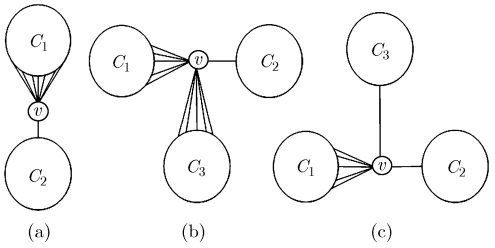

图 1 指向差异示意图

因此，指向差异的计算表示为

$$\mathrm{e}^{\left(1-\frac{\max_1(X)}{\max_2(X)}\right)} \tag{6}$$

其中，$\max_1(X)$ 表示取 $X_1, X_2, \cdots, X_n$ 中的最大值。$\max_2(X)$ 表示取 $X_1, X_2, \cdots, X_n$ 中的第 2 最大值。允许 $\max_1$ 和 $\max_2$ 的值相等。

最后，综合上述 3 部分指标，每个节点的参考值计算公式为

$$F = C + C\cdot \mathrm{e}^{\left(1-\frac{\max_1(X)}{\max_2(X)}\right)} - C\cdot\left(1-\mathrm{e}^{-d/\overline{d}}\right) \tag{7}$$

(2)多元线性回归：　Karate, Football 和 Dolphins 网络标准划分参见文献[15]。静态值的计算方法同 2.3 节。由于数据太多，表 1 仅给出了 3 个网络的部分参数(分别给出了 5 个节点的数据)，其他节点的数据略去。根据表 1 的数据，可以把式(1)的参数确定下来：

$$Y = -3.0467 - 0.2852x_1 + 4.8379x_2 + 17.6568x_3 + 3.9877x_4 \tag{8}$$

3.3 式(4)的参数估计

在式(4)中，x_i 是参与计算的节点的静态网络特征值，与此对应，w_i 就是 x_i 的权重值。本文实验使用了 4 项节点的静态参数，即度、介数、接近度、权威度等，因此，需要估算这 4 项参数相应的权重值 w 。

估算数据时，使用了人工生成网络作为标准,数据集是由 Lancichinetti 等人[14]开发的 LFR 基准程序生成的人工数据集。表 2 中 N 代表节点数；K 代表网络中节点的平均度数；$\max k$ 代表节点的最大度数；$\min c$ 代表最小的社区规模，$\max c$ 代表最大的社区规模；mu 代表构成这个网络的混合参数。mu

表 1 标准网络的静态特征值列表(部分)

网络	节点	度	介数	接近度	权威度	式(4)值
Karate	1	16	0.437635281	0.568965517	0.071412729	0.066131548
	2	9	0.053936688	0.485294118	0.053427231	0.283107528
	3	10	0.143656806	0.559322034	0.063719065	2.226199031
	4	6	0.011909271	0.464788732	0.042422737	0.270443437
	5	3	0.000631313	0.379310345	0.01526096	0.520041765
Dolphin	1	6	0.019082596	0.346590909	0.022833068	2.750868499
	2	8	0.213324436	0.371951220	0.007476274	0.650780668
	3	4	0.009072812	0.282407407	0.007064218	1.187597919
	4	3	0.002373797	0.308080808	0.014096546	1.850075725
	5	1	0	0.248979592	0.005203912	0.822860534
Football	1	12	0.032489949	0.423791822	0.010092324	2.032694610
	2	12	0.017621113	0.413043478	0.009132366	1.626829027
	3	12	0.013122497	0.407142857	0.011017018	1.539988579
	4	12	0.023070099	0.420664207	0.010068627	1.626829027
	5	11	0.010663869	0.402826855	0.009589377	1.753784351

表 2 4 类构造数据集参数情况

项目	小网络 小社区	小网络 大社区	大网络 小社区	大网络 大社区
N	1000	1000	5000	5000
K	20	20	20	20
max$k$	50	50	50	50
min$c$	10	20	10	20
max$c$	50	100	50	100
mu	0.1~0.9	0.1~0.9	0.1~0.9	0.1~0.9

值越大，网络内部的社区结构将越不明显。我们生成了 4 类人工网络用作参数估计：(1)小网络小社区；(2)小网络大社区；(3)大网络小社区；(4)大网络大社区。

图 2 的纵坐标是 NMI 值，横坐标是 mu 值。NMI (Normalized Mutual Information)主要用于计算两个网络之间的相似性。在实验中，NMI 取 4 类生成网络的平均值。从图 2 中可以知道，如果仅仅使用度作为参数，并不能获得较好的结果。同时，当 mu 值增大时，社区人为混合在一起，区分度越来越差，因此，划分的难度越来越大，算法很难完成社区的划分工作。

我们把式(4)中的 $\boldsymbol{X}$ 向量表示为(度，介数，接近度，权威度)，实验给出的最终估计值是 $w=(0.205,\ 0.274,\ 0.262,\ 0.259)$。该估计值用于后续计算从属函数。

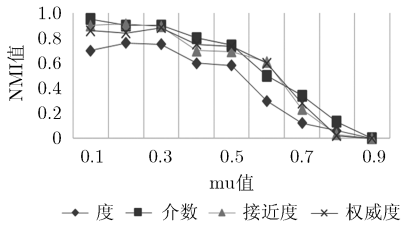

图 2 独立参数运行情况

3.4 真实网络数据集测试

(1)准确性评价： 对于真实网络构成的数据集，采用社区结构参数评价算法结果的好坏。本文主要采用适用于重叠社区发现结果的评价函数 Mod 进行算法结果的评估。

本文算法(OCRA-CF)是在传统标签算法的基础上改进而来，尽管通过贡献度函数大大降低了算法的不稳定性，但还是保留了一定的标签算法固有的不稳定性与随机性。在针对真实数据集测试的过程中，本文在同一个数据集上先后运行了 10 次算法。其中 Mod 值和社区数都是平均值。

经与 Karate, Dolphin 和 football 数据集公认的划分结果进行比对，OCRA-CF 算法的社区划分结果基本正确，几乎没有节点被错误划分。而对于大规模的网络数据集，从表 3 中的 Mod 一列可以看出 OCRA-CF 算法的运行结果也较为理想。Douban 的数据来自于互联网，节点之间的关联度非常低，其社区结构几乎为 0。Z_friends 的数据也是没有节

表 3　OCRA-CF 算法在不同数据集上的运行情况

数据集	节点数	边数	Mod	Mod 波动	平均社区数
Karate	34	78	0.544	±0.193	2.3
Dolphin	62	59	0.488	±0.105	5.1
Football	115	613	0.627	±0.024	10.8
S_C	6175	15969	0.543	±0.132	188.5
PGP	12189	83437	0.817	±0.187	24.5
Douban	115460	2235740	0.005	0	4
Eron	84429	325564	0.668	±0.004	165.5
Z_follower	161	439	0.769	±0.067	5.4
Z_friends	1077	40517	0	0	1
Bupt	23501	433635	0.760	±0.155	70.2

点之间连接的数据，其社区结构是 0，符合数据发布者对数据的特征说明。

(2)横向比较：　同样使用上述的测试数据集，我们将 OCRA-CF 算法与几个经典算法进行了对照测试，结果展示如表 4 和表 5 所示，其中，Gregory 为文献[9]的算法。

在表 4 中，OCRA-CF, CPM, Gregory 和 CFinder 算法采用 Mod 作为评价指标，而 GN，Newman 采用 Q 函数作为评价指标，LFM 采用 EQ 评价。表中的 OUT 表示该算法在 24 h 内无法完成该数据的计算过程。从表 4 的数据可以看出，本文提出的 OCRA-CF 算法在各个数据集上表现都非常好。

在表 5 中，主要进行运行速度的比较，时间参数为 s，从表中的结果可以看出，本文提出的 OCRA-CF 算法在各种数据集上的时间都较为理想。LFM 算法在部分数据集上也表现出了异常优越的效果，但也有些数据的效果非常差，非常不稳定。

总之，OCRA-CF 的优势主要有两点：(1)时间复杂度低。经典算法对于大型的数据集，都无法在实验允许的时间内(这里设为 24 h)得到社区发现结果。(2)社区划分的模块度适中。在各种数据集上都能获得较好的模块度评价指标。

3.5 人工数据集测试

用 LFR 基准程序所产生的人工网络拥有可控制的社区结构的，因此可以用来对算法的划分准确度进行测试。NMI 是评价标准的社区结构和算法输出的划分结构之间的相似度的定量指标。由于考虑到构造的数据集要有相当的规模才能具有测试的代表性，而 GN 或者 CPM 算法处理几千节点的数据集时要耗费很长时间，因此本文采用算法复杂度相对较低的 LFM 算法与 CFinder 算法作为比较对象。

在对比实验中，使用表 2 所列的参数生成了 4 组人工网络。在结果的展示中，OCRA-CF 算法、CFinder 和 LFM 算法都取 10 次重复实验的最佳结果进行比较。算法运行结果所得的 NMI 如图 3 所示。各个图中的纵坐标是 NMI 值，横坐标是 mu 的值。4 个图的网络属性分别为：(a)小网络/小社区；(b)小网络/大社区；(c)大网络/小社区；(d)大网络/大社区。

由图 3 可以得到如下结论：(1)对于 $N = 1000$ 的两个数据集，即图 3(a)和图 3(b)的两张折线图，OCRA-CF 算法在 mu 值较低时的 NMI 比 CFinder 要好，稍稍逊色于 LFM 算法。而对于 $N = 5000$ 的两个数据集，即图 3(c)和图 3(d)所表述的内容，OCRA-CF 算法的 NMI 都比其他两个算法要高。这说明 OCRA-CF 算法在较大规模的数据集的效果比较小规模的数据集的效果要好。(2)就划分结果的稳定性而言，OCRA-CF 算法要比 LFM 算法优秀，但对于较高的混合参数 mu 时，CFinder 算法稳定性最强。

表 4　不同算法划分结果的模块度比较

数据集	OCRA-CF	Gregory	GN_Q	LFM_EQ	Newman_Q	CPM_Mod(k)	CFinder_Mod(k)
Football	0.627	0.467	0.337	0.558	0.418	0.511(3)	0.543(4)
Dolphin	0.488	0.581	0.413	0.752	0.250	0.585(3)	0.588(3)
Karate	0.544	0.499	0.488	0.731	0.312	0.482(4)	0.555(3)
S_C	0.543	0.568	OUT	0.441	OUT	0.429(4)	0.405(5)
PGP	0.817	0.717	OUT	OUT	OUT	OUT	OUT
Douban	0.005	0.011	OUT	OUT	OUT	OUT	OUT
Enron	0.668	0.720	OUT	OUT	OUT	OUT	OUT
Z_follow	0.769	0.599	0.551	0.297	0.608	0.547(4)	0.573(4)
Z_friends	0	0	OUT	N/A	0.363	0.447(3)	OUT
Bupt	0.760	0.337	OUT	OUT	OUT	OUT	OUT

表 5 不同算法的速度比较(s)

数据集	OCRA-CF	Gregory	GN	LFM	Newman	CPM	CFinder
Football	7.8	6.2	489.6	1.926	0.135615	0.96	0.703
Dolphin	13.3	9.8	12.6	1.054	0.28507	0.14	0.062
Karate	3.4	3.0	3.9	0.202	0.149233	0.1	0.004
S_C	1129.3	19.6	OUT	677.411	OUT	3.7	1.330
PGP	1656.1	228.0	OUT	OUT	OUT	OUT	OUT
Douban	44479.5	6782	OUT	OUT	OUT	OUT	OUT
Enron	15973.5	5778	OUT	OUT	OUT	OUT	OUT
Z_follow	20	18.0	411.7	22.859	3.7	0.82	0.227
Z_friends	193.4	234.2	OUT	OUT	12.0744	755.4	OUT
Bupt	5330.1	633.8	OUT	OUT	OUT	OUT	OUT

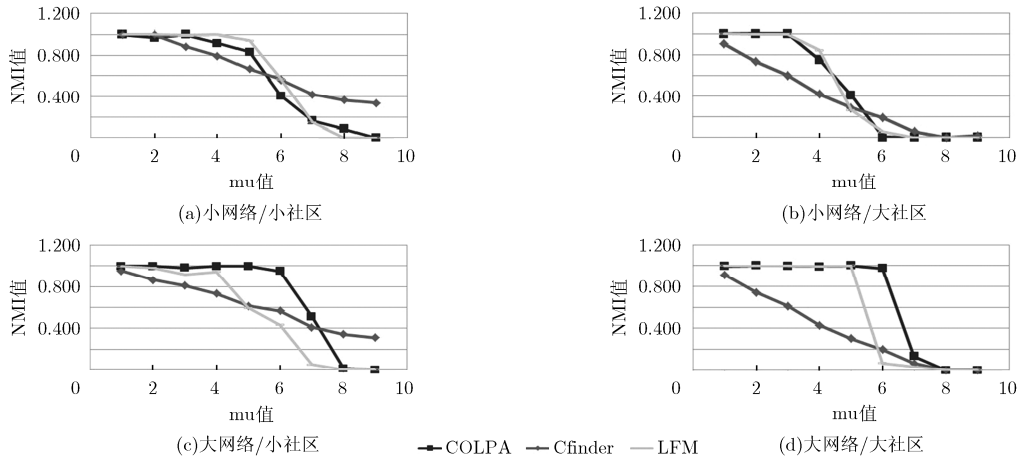

图 3 针对人工数据集的准确度比较

4 结论

在现实的社会网络中，社区结构重叠是普遍存在的现象，因此，面向重叠社区的自动发现算法具有重要的研究意义和使用价值。本文提出的基于贡献函数的重叠社区发现算法，即 OCRA-CF 算法，既继承了传统标签算法的速度优势，又能达到较好的划分效果。通过在各种人工构造数据集和真实数据集上的进行测试，可以看出 OCRA-CF 运行结果较为理想，达到了可用的目标。在实际工作中，本文算法具还有两大优势：(1)由于 OCRA-CF 是基于经典标签算法的改进，所以，该算法能比较方便地移植到 Hadoop 或 Spark 等并行计算平台，以适应社交网络的大数据需求。(2)OCRA-CF 算法中使用的贡献函数具有较强的扩展性，可以通过调整采用的网络参数多少以及参数对应的权重获得不同的贡献函数，适应不同的应用场景。

参 考 文 献

[1] WANG Xiaofeng, LIU Gongshen, PAN Li, *et al*. Uncovering fuzzy communities in networks with structural similarity[J]. *Neurocomputing*, 2016, 210(1): 26–33.

[2] PALLA G, DERENVI I, FARKAS I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society[J]. *Nature*, 2005, 435(7043): 814–818.

[3] LEE C, REID F, McDAID A, *et al*. Detecting highly overlapping community structure by greedy clique expansion [C]. ACM International Conference on Paper Presented at SNA-KDD Workshop, Washington DC, USA, 2010. arXiv: 1002.1827.

[4] GREGORY S. An algorithm to find overlapping community structure in networks[J]. LNCS, 2007, 4702(12): 91–102.

[5] SHI Xiaohua. Community detection in social network with pair wisely constrained symmetric non-negative matrix factorization[C]. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 2015: 541–546.

[6] LIU Xiao, WEI Yiming, WANG Jian, *et al*. Community detection enhancement using no-negative matrix

factorization with graph regularization[J]. *International Journal of Modern Physics B*, 2016, 30(20): 1650130.

[7] WANG Zhaoxian, WANG Wenjun, XUE Guixiang, *et al*. Semi-supervised community detection framework based on non-negative factorization using individual labels[C], The Sixth International Conference on Swarm Intelligence, Beijing, China, 2015, 349–359

[8] RAGHAVAN U N, ALBERT R, and KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 2007, 76(3 Pt 2): 036106.

[9] GREGORY S. Finding overlapping communities in networks by label propagation[J]. *New Journal of Physics*, 2009, 12(10): 2011–2024.

[10] ULRIK B. A faster algorithm for betweenness centrality[J]. *Journal of Mathematical Sociology*, 2001, 25(2): 163–177.

[11] EPPSTEIN D and WANG J. Fast approximation of gentrality[J]. *Journal of Graph Algorithms and Applications*, 2004, 8(1): 39–45.

[12] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. *Journal of the ACM* (*JACM*), 1999, 46(5): 604–632.

[13] NICOSIA V, MANGIONI G, CARCHIOLO V, *et al*. Extending the definition of modularity to directed graphs with overlapping communities[J]. *Journal of Statistical Mechanics Theory & Experiment*, 2009, 2009(3): 3166–3168.

[14] LANCICHINETTI A, FORTUNATO S, and RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 2008, 78(2): 046110.

[15] CAO Xiaochun, WANG Xiao, JIN Di, *et al*. Identifying overlapping communities as well as hubs and outliers via nonnegative matrix factorization[J]. *Scientific Reports*, 2013, 03: 2993. doi: 10.1038/srep02993.

刘功申：　男，1974 年生，副教授，研究方向为内容安全、自然语言理解.

孟　魁：　女，1973 年生，高级工程师，研究方向为移动安全、数据安全和社会网络.

郭弘毅：　男，1992 年生，硕士生，研究方向为推荐系统研究.

苏　波：　男，1971 年生，副研究员，研究方向为社会网络分析.

李建华：　男，1965 年生，教授、博士生导师，研究方向为电子与通信工程、信息安全.