基于链接相似性聚类的重叠社区识别

张桂杰[1,2],张健沛[1],杨 静[1],辛 宇[1]
(1.哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001;2.吉林师范大学计算机科学与技术学院,吉林四平 136000)

摘 要: 社区结构是社会网络最普遍和重要的拓扑属性之一,提出一种基于链接相似性聚类的重叠社区识别算法.该算法首先根据相邻链接的度分布状态,提出链接间的相似性度量方法;其次以链接相似性矩阵为输入,以链接社区的最优划分为目标,建立链接局部相似性聚类算法,实现了重叠社区的有效识别;然后对链接社区进行优化,解决了可能出现的过度重叠及孤立社区问题;最后在真实网络及人工合成网络上的实验验证了算法的高效性.

关键词: 社区识别; 链接社区; 局部链接相似性度量; 层次聚类; 重叠社区
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2015)07-1329-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.07.012

Overlapping Community Detection Based on Link Similarity Clustering

ZHANG Gui-jie[1,2], ZHANG Jian-pei[1], YANG Jing[1], XIN Yu[1]
(1. *College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*;
2. *College of Computer Science and Technology, Jilin Normal University, Siping, Jilin* 136000, *China*)

Abstract: Community structure is one of the most common and important social network topological properties. This paper proposes a link community detection algorithm based on hierarchical clustering. Firstly, the algorithm sets up similarity measure according to the degree distribution of links nearby; then sets up local link similarity clustering algorithm which takes the similarity matrix as input with the purpose of detecting the best link community; further more realizes link community detection effectively. And then, optimize the link community to solve the problem of excessive overlapping and isolated community. Experiment results based on real world and computer generated networks show that the algorithm is highly efficient.

Key words: community detection; link community; local link similarity metric; hierarchical clustering; overlapping community

1 引言

网络社区在信息传播与推荐、舆情预警、链接预测等领域有重要作用,因而社区识别成为当前的研究热点.目前该领域已有大量的研究成果,从社区识别结果上可分为硬社区识别和重叠社区识别.起初硬社区识别算法深受关注,最近研究者发现重叠现象是社会网络的重要特征,进而展开了重叠社区识别的研究,当前有代表性的重叠社区识别算法有基于派系的方法、局部扩展优化方法、模糊探测法等,Fortunato 在文献[1]中进行了详尽的分析和对比.

将网络抽象表达为图后,网络中的实体及实体间的联系表现为图中的节点(node)和链接(link).目前的研究往往过多关注节点信息,而忽视了链接信息,多数算法以节点的邻接矩阵为处理对象,而鲜见以链接为研究对象.最近学者们开始注意到构建链接图来揭示重叠社区结构的优势.Evans 等[2]开创性地对网络中的链接进行划分,将原始网络转换成线图,根据节点度分布的异构性提出了加权线图的构建方法,然后用基于随机游走的方式挖掘社区.Ahn 等[3]提出层次聚类方法划分链接,并给出划分密度函数来截取最优结果.Kim 等[4]将传统基于节点的 Infomap 方法扩展至基于链接的形式.Ball 等[5]提出用于探测链接社区结构的统计学方法.

尽管基于链接的社区识别算法取得了些成果,但仍存在以下问题:使用局部相似性度量对链接相似性的估计产生偏差,而使用全局相似性度量计算复杂度过高;将链接社区转换为节点社区时存在过度重叠或冗余孤立社区.

收稿日期:2014-04-08;修回日期:2015-01-09;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61370083, No.61073043, No.61073041);高等学校博士学科点专项科研基金(No.20112304110011, No.20122304110012);吉林省教育厅"十二五"科学技术研究基金(No.吉教科合字2013[207])

针对以上问题提出一种重叠社区识别算法 LinkCom.首先将研究主体进行转换,从节点的邻接矩阵出发,以节点-边的关联矩阵为过渡,无损转换为边-边邻接矩阵,构造出链接图.基于此提出局部链接相似性度量方法,构建链接相似度矩阵,采用层次聚类算法,建立链接聚类树,并设计有效的截断策略截取最优链接社区,最后还原为节点社区并通过设置重叠率阈值及孤立边消除进行优化.

本文主要贡献在于:(1)提出一种局部链接相似性度量方法,合理地刻画链接图中对象之间的相似性.(2)提出重叠社区识别算法 LinkCom,既克服了节点硬划分的问题,又避免了边社区识别中过度重叠及冗余现象的出现.

2 基本概念

设无权无向网络表示为图 $G(E,V)$,其中 $E(G)=\{l_{ij}\mid v_i,v_j\in V\}$ 为边集,$V(G)=\{v_i\mid i=1,\cdots,N\}$ 为节点集,$L=|E(G)|$ 为边数,$N=|V(G)|$ 为节点数.

定义 1 点邻接矩阵(node-node)

图 G 的节点间邻接矩阵 $\boldsymbol{A}$ 为 $N\times N$ 阶的对称方阵,其中 N 表示 G 中的节点个数,其元素 A_{ij} 取值为 0 或 1,当 v_i 与 v_j 有边相连时 $A_{ij}=1$,否则 $A_{ij}=0$.鉴于邻接矩阵的对称性,可提取其上(下)三角阵进行存储和运算,以节约存储空间并提高运算效率.

定义 2 关联矩阵(node-link)

图 G 的关联矩阵 $\boldsymbol{B}$ 为 $N\times L$ 的非对称矩阵,其元素表示为 $B_{i\alpha}$,其中 i 表示节点,α 表示链接.链接编号规则为:在点邻接矩阵 $\boldsymbol{A}$ 的上三角矩阵中,从第一个不为零的元素开始,按行优先顺序搜索,以自然数顺序对非零元素进行编号,得到链接序列.在关联矩阵 $\boldsymbol{B}$ 中,如果节点 i 与边 α 相关联,则 $B_{i\alpha}=1$,否则 $B_{i\alpha}=0$.关联矩阵表示图中节点与链接的关系,因此通过图 G 的关联矩阵可以推导出每个节点的度 k_i 和每条边所关联的节点数目 k_α 的关系,表达为

$$k_i=\sum_{\alpha=1}^{L}B_{i\alpha},k_\alpha=\sum_{i=1}^{N}B_{i\alpha} \tag{1}$$

定义 3 链接图

链接图 $L(G)$ 以图中的边为研究对象,通过边之间拥有公共节点的情况构造其拓扑结构.当 G 中两条边拥有公共节点时,链接图中两元素相邻.链接图满足 $V(L(G))=E(G)$,其元素记为 l_α,其中 $\alpha=1,\cdots,L$.

定义 4 链接图邻接矩阵(link-link)

图 G 的链接邻接矩阵 $\boldsymbol{E}$ 为 $L\times L$ 的对称矩阵,L 表示图中链接的数目,矩阵元素记为 $E_{\alpha\beta}$,其中 $\alpha,\beta=1,\cdots,L$.当图 G 中两条边拥有公共节点时 $E_{\alpha\beta}=1$,否则 $E_{\alpha\beta}=0$.

至此可得节点邻接矩阵 $\boldsymbol{A}$、关联矩阵 $\boldsymbol{B}$ 及链接图邻接矩阵 $\boldsymbol{E}$ 三者间的映射关系:

$$A_{ij}=\begin{cases}\sum_{\alpha=1}^{L}B_{i\alpha}B_{j\alpha}(i\neq j)\\0\qquad(i=j)\end{cases} \tag{2}$$

$$E_{\alpha\beta}=\begin{cases}\sum_{i=1}^{N}B_{i\alpha}B_{i\beta}(\alpha\neq\beta)\\0\qquad(\alpha=\beta)\end{cases} \tag{3}$$

定义 5 扩展邻接边

在链接图 $L(G)$ 中,边 l_α 的邻接边记为 $N(l_\alpha)$,表示与边 l_α 至少有一个公共节点但不包括 l_α 本身的边的集合,即 $N(l_\alpha)=\{l_\beta\mid E_{\alpha\beta}=1\}$.$l_\alpha$ 的扩展邻接边为由 l_α 及其邻接边构成,即 $N^+(l_\alpha)=N(l_\alpha)\cup\{l_\alpha\}$.

3 链接社区的局部相似性度量

链接的邻居的交集合及其度分布反映了链接之间联系的紧密程度.图 1 表达了相邻链接 l_α 和 l_β 的拓扑关系,其中 $N^+(l_\alpha)\cap N^+(l_\beta)$ 是以节点 v_i 为中心的星型结构,记作 v_i-*block* 社区.本文对 v_i-*block* 社区中链接 l_α 和 l_β 的相似性分析如下:

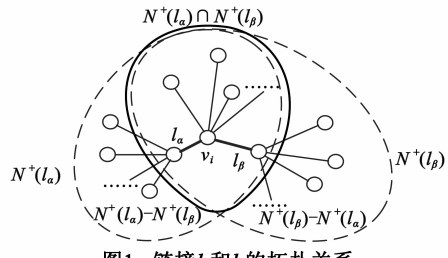


图1 链接l_α和l_β的拓扑关系

(1) v_i – *block* 社区中链接数量反映了星型结构的紧密程度,链接数量越多,l_α 与 l_β 的紧密度越高,相似性越强.

(2) v_i-*block* 社区的外部链接数量越多,v_i-*block* 的紧密度越低,l_α 与 l_β 的相似性越弱.由于 v_i-*block* 半径为 1,当 v_i-*block* 社区中的链接数量不变时,v_i-*block* 社区内某一链接度数越高,则其外部链接数量越多,该链接对 v_i-*block* 紧密度的贡献越低.

根据 l_α 和 l_β 的相似性分析可知,l_α 和 l_β 的相似性 $S(l_\alpha,l_\beta)$ 与 $N^+(l_\alpha)\cap N^+(l_\beta)$ 中链接的数量成正比,与 $N^+(l_\alpha)\cap N^+(l_\beta)$ 中链接的度数成反比.为此,本文对 l_α 和 l_β 的相似性 $S(l_\alpha,l_\beta)$ 初步建模如下:

$$S(l_\alpha,l_\beta)=\sum_{l\in N^+(l_\alpha)\cap N^+(l_\beta)}f(k_l) \tag{4}$$

其中 k_l 表示链接的度值,$f(k_l)$ 为 k_l 的减函数.式(4)兼备链接数量对 l_α 与 l_β 相似性的正向影响,及链接度数对 l_α 与 l_β 相似性的负向影响.根据社会网络局部相似性度

量方法[6],l_α和l_β的相似性$S(l_\alpha,l_\beta)$可归一化为:

$$S(l_\alpha,l_\beta)=\frac{\sum_{l\in N^+(l_\alpha)\cap N^+(l_\beta)} f(k_l)}{\sqrt{\sum_{l\in N^+(l_\alpha)} f(k_l)}\cdot\sqrt{\sum_{l\in N^+(l_\beta)} f(k_l)}} \tag{5}$$

其中,$f(k_l)$的构造需要满足:(1)$f(k_l)$为 k_l的减函数;(2)$f(k_l)$对于 k_l取值的差异性.为此,本文对以下 3 种常用的建模函数进行分析,其中 σ 为控制参数.

$$f(k_l)=k_l^{-\sigma},\sigma^{-k_l}(\sigma>1),\sigma k_l^{-1} \tag{6}$$

图 2 为 3 种建模函数对比图,其中函数 $f(k_l)=k_l^{-\sigma}$与$f(k_l)=\sigma k_l^{-1}$在 $k_l>5$ 时取值近似,与参数 σ 的取值无关;而函数 $f(k_l)=\sigma^{-k_l}$在参数σ 接近 1.1 时,取值保持了可区分性.为此,本文选择 $f(k_l)=\sigma^{-k_l}$作为建模函数,其中 σ 为控制参数且在区间$\sigma\in(1,1.5)$内最有效,由此,链接 l_α与 l_β的局部相似性度量可定义为:

$$S(l_\alpha,l_\beta\mid\sigma)=\frac{\sum_{l\in N^+(l_\alpha)\cap N^+(l_\beta)} \sigma^{-k_l}}{\sqrt{\sum_{l\in N^+(l_\alpha)} \sigma^{-k_l}}\cdot\sqrt{\sum_{l\in N^+(l_\beta)} \sigma^{-k_l}}} \tag{7}$$

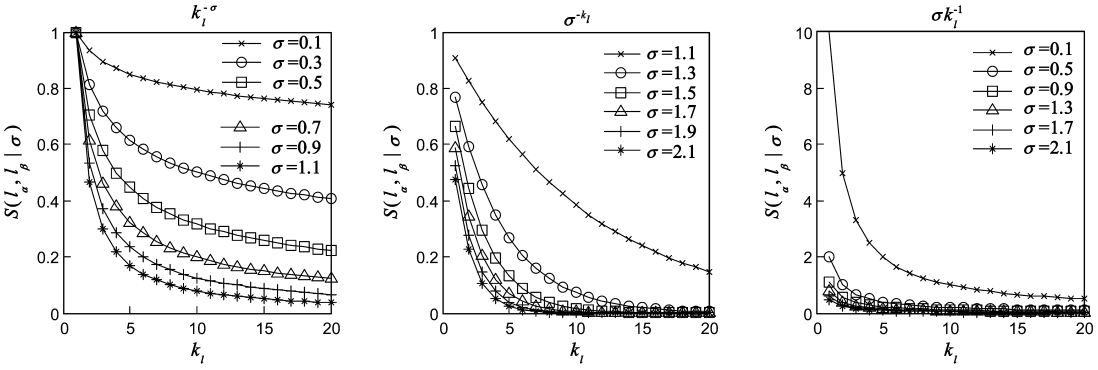


图2　3种建模函数对比

4　基于链接层次聚类的重叠社区识别算法

4.1　链接层次聚类

层次聚类算法又称为树聚类算法,是经典的社区识别技术.在网络中,通过对链接的连接强度进行度量,可以找出一些相互连接强度较大的对象集合,即为网络中的社区.具体步骤为:

(1)初始化.设置每个样本为一独立的社区;

(2)寻找最邻近社区;

(3)合并.将相似度最大的两个社区采用沃德法合并成一个新社区,同时更新相似度矩阵;

(4)若所有社区都合并成一个整体,则算法终止;否则,返回步骤 2.

伪代码描述如下:

算法 1　LinkCom 算法

输入:网络 $G=(V,E)$,link-link 邻接矩阵 $\boldsymbol{E}$,相似性区分度参数 σ.

输出:层次聚类树,每层对应一种划分 $C(n)$.

1. 初始化:$C=\{C_1,C_2,\cdots,C_l\}$,其中 $C_\alpha=\{l_\alpha\}$.
2. $C(n)\leftarrow C$.
3. 按照式(7)计算 link 间相似度,得到链接相似度矩阵 $\boldsymbol{S}$.
4. Do
5. 搜索矩阵 $\boldsymbol{S}$ 上三角部分最大值,$MaxS=S_{\alpha\beta}$;
6. For $k=1:(|C|-1)$
7. $S_{\beta k}\leftarrow S_{k\beta}\leftarrow(N_\alpha S_{ak}+N_\beta S_{\beta k})/(N_\alpha+N_\beta)$;
8. $S_{ak}\leftarrow S_{ka}\leftarrow 0$;
9. End for
10. $C_\beta'=C_\alpha\cup C_\beta$;
11. $C\leftarrow C\setminus\{C_\alpha,C_\beta\}\cup C_\beta'$;
12. $C(|C|)\leftarrow C$;
13. 删除 $\boldsymbol{S}$ 的α 行和β 列;
14. UNTIL $|C|=1$
15. $C(tree)\leftarrow\{C(n),\cdots,C(1)\}$;
16. 返回 $C(tree)$

在聚类过程中产生一个划分序列,图 3 为 Karate 网络经上述过程形成的层次聚类树(其链接的编号规则使用定义 2 中所述的方法).其中叶子节点代表网络中的链接,共 78 条链接.依次合并相似度最大的链接,并更新相似度矩阵,经过 77 次合并后聚类为一个社区.为了压缩整个聚类树的描述高度且能够清晰地表达树的聚合过程,每一次合并时在括号中标注操作的顺序号.

在层次聚类树中,每一层对应链接图的一种划分,即网络社区.为了从中截取最优划分,我们选择将 Newman 等针对节点集划分提出的模块度扩展至链接集合,表达为

$$Q(E)=\frac{1}{W}\sum_{C\in P}\sum_{\alpha,\beta\in C}\left[E_{\alpha\beta}-\frac{k_\alpha k_\beta}{W}\right] \tag{8}$$

链接型与节点型社区模块度函数具有相同的性质,模块度值越大,划分效果越好,即发现的社区结构

越合理.针对 Karate 网络,通过链接模块度的衡量,可得到每层划分所对应的模块度值,其中第 75 次合并前的划分得到最大模块度值 0.6183,此时网络划分为 4 个链接社区.由于链接社区和节点社区的不同特性可能导致社区间出现包含的情况,因此需要对聚类结果进行优化.

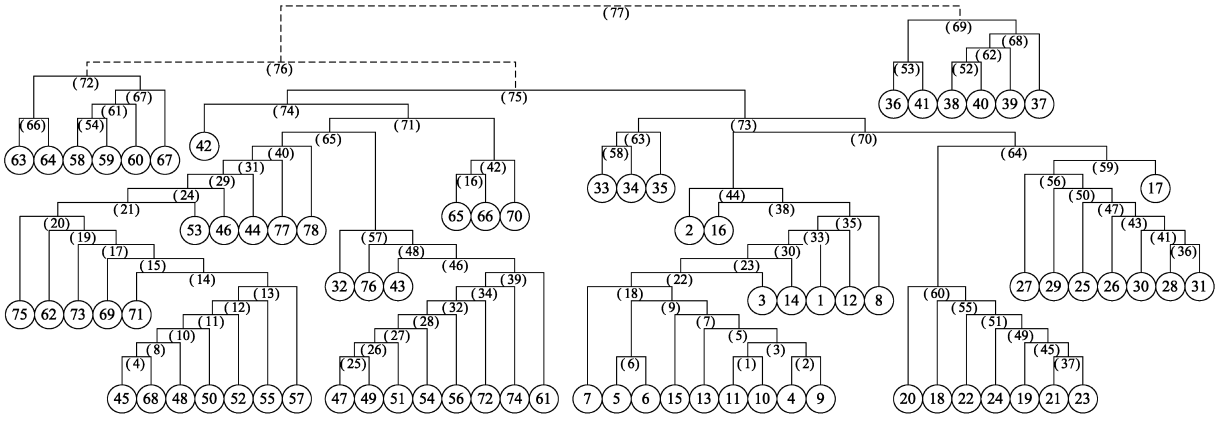


图3 Karate网络链接图的层次聚类树

4.2 链接型社区向节点型社区的转换及优化

链接型与节点型社区主要区别体现在:非重叠的链接社区对应着重叠的节点社区;链接社区可能造成节点社区的过度重叠;链接可以没有归属,而节点都有归属.因此,链接社区转换为节点社区得到的重叠结构需要优化.

本文针对社区的过度重叠及冗余问题,通过设定社区重叠率及消除单链接社区进行优化.社区重叠率指两个节点型社区间节点的重叠比例,通过计算两个社区交集的节点数与其中较小社区的节点数目的比值确定,表达为

$$O_v = |C_1 \cap C_2| / \min(|C_1|, |C_2|) \tag{9}$$

重叠率的限制阈值将根据具体网络结构及实际社区识别对重叠情况的要求进行调整.

链接型向节点型社区转换及优化的步骤为:

(1)根据链接社区与其原节点图的关系,将链接社区转变成对应的节点社区;

(2)查找仅由一条链接构成的单链接社区;

(3)若单链接社区关联的两个节点分别属于不同的社区,则将此单链接社区删除;

(4)根据式(8)计算重叠节点社区结构的 EQ 值;

(5)根据式(9)计算社区组间的重叠率,将超过一定阈值限制的社区组进行合并;

(6)若社区重叠率均小于设定的阈值,再没有社区需合并,则算法终止;否则,返回步骤 4;

(7)输出具有最大 EQ 的重叠社区.

5 实验

本节对 LinkCom 算法中的参数进行详细的分析,然后在真实及人工网络数据集上分别进行测试,并与经典算法的识别精度进行对比分析.

5.1 实验数据集

5.1.1 真实网络数据

为了评价本文算法的社区识别能力,选取表 1 所示的四个真实网络数据集进行实验分析.

表 1 真实网络数据集描述

networks	reference	nodes	edges
Karate	Zachary's karate club	34	78
Dolphin	Dolphin social networks	62	159
Polbook	Books about US politics	105	441
Football	American College football	115	613

5.1.2 人工网络数据集

LFR Benchmark[7]是当前社区识别研究中最常用的人工网络数据集,模型定义为

LFR model = $(N, d, d_{\max}, \gamma, b, c_{\min}, c_{\max}, on, om, \mu)$.

其中 N 表示节点的个数;d 和 $d_{\max}$分别表示网络中节点的平均度和最大度;γ 和 b 分别表示节点度和社区规模的幂率分布参数;$c_{\min}$和 $c_{\max}$分别表示最小和最大社区包含节点数;on 表示重叠节点数;om 表示重叠节点连接的社区个数;μ 为混合系数,表示节点与社区外部连接的概率,当 $\mu > 0.5$ 时,该随机网络的社区结构非常模糊.本文实验共生成 4 组 LFR 基准网络,共享参数 $d = 10, d_{\max} = 50, \gamma = -2, b = -1, om = 4$,其他参数如表 2 所示.

5.2 评价准则

为了评价算法在各数据集上的社区识别效果,本文采用五种常用的社区识别度量标准全面地对算法进行测试分析,分别为扩展模块度 EQ[8]、划分密度

PD[3]、基于信息论的 *Infomap*[4]、平均导电率 *AC*[9]和重叠社区模块度 *Qov*[10].

表 2　LFR Benchmark 参数设置

	N	$c_{\min}$	$c_{\max}$	μ	on
Data1	1000	10	50	0.1	100
Data2	2000	10	50	0.3	500
Data3	5000	20	50	0.1	100
Data4	8000	20	100	0.1	500

5.3　参数分析

5.3.1　参数 σ 取值分析

本实验从参数 σ 对特定链接组的影响入手,对参数 σ 的敏感性进行分析.

在由定义 2 所述方法构造的 Karate 网络的链接图中,选择部分度分布差异较大的链接.分别取 group1:(l_{74}, l_{75}),group2:(l_3, l_{35}),group3:(l_{37}, l_{39}),group4:(l_{43}, l_{78}),group5:(l_{58}, l_{59})共 5 组链接.在 σ 的有效取值区间(1,1.5)内,使用式(7)计算以上 5 组链接的相似度值,得到如图 4 所示的相似度变化曲线.当 $\sigma = 1$ 时,式(7)可表达为:

$$S(l_\alpha, l_\beta \mid \sigma = 1) = \frac{|N^+(l_\alpha) \bigcap N^+(l_\beta)|}{\sqrt{|N^+(l_\alpha)| \times |N^+(l_\beta)|}} \quad (10)$$

结合图 1,式(10)表达了 l_α与 l_β的拓扑结构差异性,即 $S(l_\alpha, l_\beta \mid \sigma = 1)$ 越大,l_α与 l_β的越相似.图 4 中随着 σ 的增大,各组间的相似度值越来越贴近,当 $\sigma > 1.5$ 以后,group1,group4 和 group5,及 group2 和 group3 的相似度十分接近,此时 $S(l_\alpha, l_\beta \mid \sigma > 1.5)$的差异性较小,无法有效区分链接相似度.以上分析表明参数 σ 的有效取值区间为$\sigma \in (1,1.5)$.

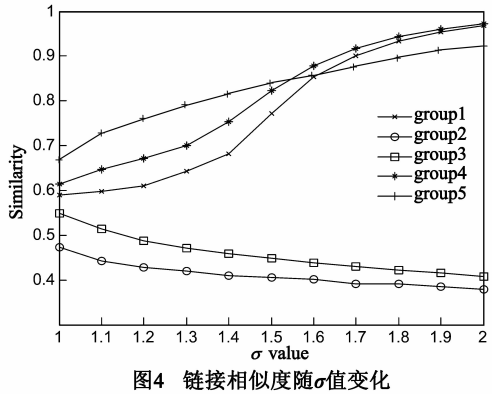


图4　链接相似度随σ值变化

5.3.2　重叠率参数 O_v分析

为直观表达参数 O_v对划分结果的影响,以 Karate 网络最优划分为目标,首先,在参数 σ 的最佳取值区间内选择三个不连续的值:1.25,1.2 和 1.05,使用本文算法进行链接层次聚类以得到链接聚类树;其次,使用 5 种评价标准截取聚类树,得到相应的最优链接社区结构;然后,在 $O_v \in (0.2,1)$连续区间的重叠率阈值限制下进行链接社区到节点社区的转换;最后,对节点社区进行评测,结果如图 5 所示,横轴表示重叠率,纵轴表示评价准则结果.

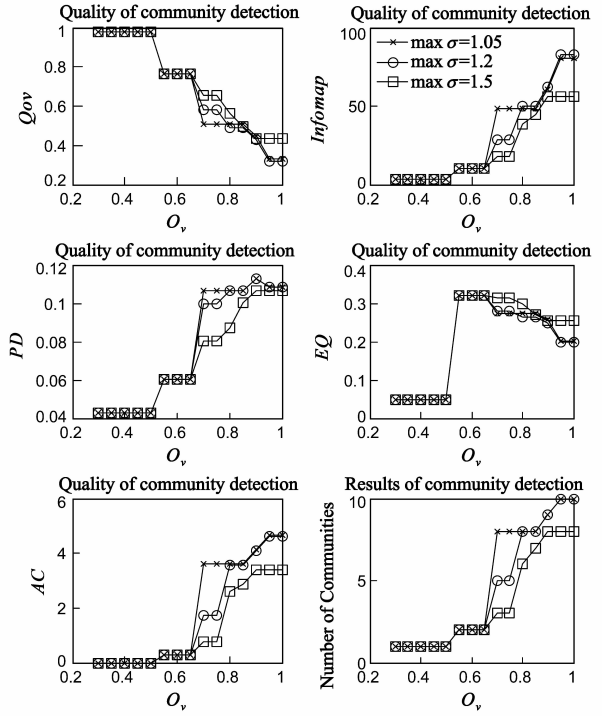


图5　Karate网络社区识别质量随参数σ及O_v的变化

其中,前 5 个子图为使用 5 种评价准则进行度量的结果.在各评价准则度量下,$0.3 < O_v < 0.7$ 的范围内,参数 σ 的变化不会引起评价准则值的波动,且均在 $O_v = 0.6$ 左右得到相对稳定的最优值.随着重叠率的增大,对比 σ 的不同取值情况下社区划分结果,$\sigma = 1.05$ 时 *Qov*, *EQ*, *Infomap* 和 *AC* 的值较优,而 *PD* 结果产生偏差,其原因在于 *PD* 标准是针对链接社区而设计,在衡量节点型社区时表现稍差.图 5 中社区数量对比表明,当 O_v在 0.6 左右时得到社区个数为 2,此时与真实网络社区结构吻合.通过 O_v的变化可知,若 O_v较高,则所发现的社区数目较多,社区规模较小;反之社区的重叠性较低,社区规模较大.本实验验证了 Karate 数据集 O_v的最优取值约为 0.6.

5.4　LFR benchmark 社区识别结果分析

为了测试 LinkCom 算法在 LFR benchmark 上的性能,本节实验选择 Link[3]、LFM[11]、Similarity[6] 和 CONRA[1]等四个典型算法进行对比,LinkCom 算法参数选取上文讨论中的最优值,$\sigma = 1.1$, $O_v = 0.6$.由于 LFR 网络真实社区结构已知,因此使用 *NMI*[12]评价算法质量.

$$NMI = \frac{-2\sum_{i,j} N_{ij} \log\left[\frac{N_{ij}N}{N_{i.}N_{.j}}\right]}{\sum_i N_{i.} \log[\frac{N_{i.}}{N}] + \sum_j N_{.j} \log[\frac{N_{.j}}{N}]} \quad (11)$$

其中 N_{ij}指社区 i,j 中公共的节点数,$N_{i.}$ 是 N 中第 i 行求和,$N_{.j}$是 N 中第 j 列求和. *NMI* 的取值在 0,1 之间,取 0 时表示两种结果完全不一致,取 1 则完全一致.

选取表 2 所示数据集,对比分析 5 种算法的性能,实验结果如图 6 所示.在各数据集上 *NMI* 值整体呈下降趋势,原因在于随着混合参数的增加,社区边界变得模糊,节点间会有更多的边链接于社区之间,因而 *NMI* 值降低.从 *NMI* 值的衰减速度分析,LinkCom 算法的 *NMI* 衰减速度低于其他算法,表现出一定的优势. COPRA 算法由于其随机性而造成 *NMI* 值的振荡;*Link* 算法由于对社区间的包含关系考虑不足而稍逊于本文算法;LFM 算法在社区间连边较少时效果显著,但当网络结构变得模糊时,社区识别结果较差;Similarity 算法是针对非重叠社区识别问题而设计,导致较低的重叠社区发现质量.

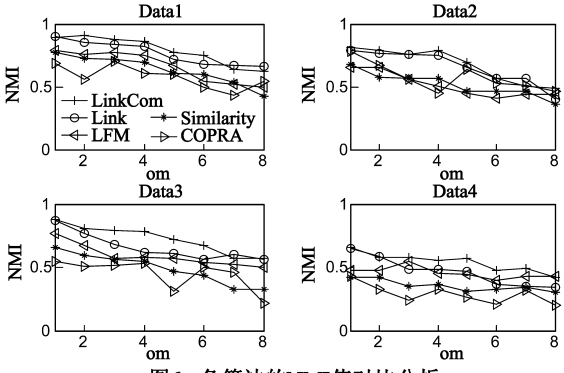


图6 各算法的NMI值对比分析

5.5 真实网络社区识别结果分析

社区识别的真正意义在于对网络社区结构的认识.通过在人工合成网络上验证了本文算法的高效性,还需要在真实网络上测试算法的性能.真实网络的社区识别更具有挑战性,由于不能预知其社区结构,只能从其返回的社区识别结果的各种评价指标进行比较.本文对表 1 中数据集进行了预处理,采用上节四种算法进行对分析,比较各算法在不同评价指标下的性能.本文 LinkCom 算法针对 Karate 网络,识别社区数目为 2,并取得 2 次最优社区评价结果,分别为在 *EQ* 和 *PD* 评价标准下的 0.4153 和 0.0915.针对 Football 网络,识别社区数目为 12,在 *PD* 标准下取得最优评价结果 0.2973.针对 Dolphin 网络,识别社区数目为 4,也在 *PD* 标准下取得最优评价结果 0.0745.在 Polbook 网络中,识别社区数目为 3,并在 3 个评价标准下取得了最优值,分别为 *EQ*:0.5873、*Infomap*:21.6529、*PD*:0.0844.总体上本文算法在各数据集及评价标准下,得到的社区数目与实际相符、取得最优评价的次数处于领先地位,虽然对于某些数据集,使用 *Qov* 和 *AC* 指标评价时 LinkCom 没有取得很好的划分精度,但是在针对链接型社区而设计的 *PD* 评价标准下,本文算法表现出明显优势.

6 结束语

本文以链接取代传统社区识别研究中的节点对象,利用链接社区特有的性质挖掘重叠节点社区结构,提出了基于链接的重叠社区识别算法.首先将描述网络拓扑性质的节点邻接矩阵无损转换为边关联矩阵;其次,提出一种链接相似性度量指标,该度量指标有效地避免了全局相似性指标具有较高的时间复杂度、传统局部相似性指标对已存在直接链接的对象间相似性的有偏估计问题;然后,通过链接相似度矩阵的更新、迭代,采用沃德法进行层次聚类,建立链接聚类树;最后,采取有效的截断策略截取最优链接社区,再通过重叠率阈值的限制及孤立链接的消除进行优化,解决了传统链接社区识别算法中过度重叠及冗余社区问题.

参考文献

[1] Fortunato S. Community detection in graphs[J]. Physics Reports, 2010, 486(3): 75 – 174.

[2] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities[J]. Physical Review E, 2009, 80(1): 016105.

[3] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 466(7307): 761 – 764.

[4] Kim Y, Jeong H. Map equation for link communities[J]. Physical Review E, 2011, 84(2): 026110.

[5] Ball B, Karrer B, Newman M E J. Efficient and principled method for detecting communities in networks[J]. Physical Review E, 2011, 84(3): 036103.

[6] 刘旭,易东云.基于局部相似性的复杂网络社区发现方法[J].自动化学报,2011,37(12):1520 – 1529.
LIU Xu, YI Dong-Yun. Complex network community detection by local similarity[J]. Acta Automatica Sinica, 2011, 37(12): 1520 – 1529. (in Chinese)

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

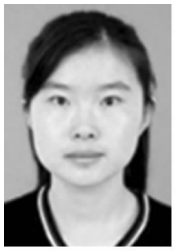
[8] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(8): 1706 – 1712.

[9] Leskovec J, Lang K J, Mahoney M. Empirical comparison of algorithms for network community detection[A]. Proceedings of the 19th International Conference on World Wide Web[C]. ACM, 2010. 631 – 640.

[10] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping com-

munities[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, (03): P03024.
[11] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.
[12] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005(09): P09008.

作者简介



张桂杰　女,1980 年生于吉林白山.哈尔滨工程大学计算机学院博士研究生,吉林师范大学讲师.主要研究方向为数据挖掘、社会网络社团分析.
E-mail: zhangguijie@hrbeu.edu.cn

张健沛(通讯作者)　男,1956 年生于黑龙江哈尔滨.哈尔滨工程大学计算机学院教授、博士生导师.主要研究方向为数据与知识工程、数据挖掘、社会网络等.
E-mail: zhangjianpei@hrbeu.edu.cn