

DOI: 10.7652/xjtuxb201502002

采用模糊层次聚类的社会网络重叠社区检测算法

李刘强^{1,2}, 桂小林^{1,2}, 安健^{1,2}, 孙雨¹

(1. 西安交通大学电子与信息工程学院, 710049, 西安;

2. 西安交通大学陕西省计算机网络重点实验室, 710049, 西安)

摘要: 为了能够发现社会网络中的重叠社区以及解决重叠社区之间关系的模糊性和层次性,提出了一种基于模糊层次聚类的重叠社区检测算法(CDHC)。算法中引入了距离加权因子来计算社区间的相似度,通过模糊层次聚类来合并相似度高的社区;针对合并生成的原始社区计算社区中节点的隶属度,再将隶属度小于阈值的节点从社区中移除,从而形成最终的网络重叠社区结构。该算法不仅可以发现重叠的社区结构,还可以处理孤立节点。在 Lancichinetti 基准网络和真实网络上将 CDHC 算法与具有代表性的重叠社区发现算法 CMP 和 LFM 进行了比较,结果表明:影响社区检测精度的主要因素是社区间的混合程度,而网络规模和网络中社区的规模的影响并不显著;CDHC 算法在小社区网络上的社区检测精度优于 LFM,在大社区网络上的社区检测精度优于 CMP。CDHC 算法在保持社区检测质量的同时,还具有较好的稳定性,是一种有效的社会网络重叠社区检测算法。

关键词: 社会网络;相似度;模糊层次聚类;重叠社区检测

中图分类号: TP393 **文献标志码:** A **文章编号:** 0253-987X(2015)02-0006-08

Overlapping Community Detection Algorithm Based on Fuzzy Hierarchical Clustering in Social Network

LI Liuqiang^{1,2}, GUI Xiaolin^{1,2}, AN Jian^{1,2}, SUN Yu¹

(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;

2. Shaanxi Province Key Laboratory of Computer Network, Xi'an Jiaotong University, Xi'an 710049, China;)

Abstract: A detection algorithm for overlapping communities based on fuzzy hierarchical clustering, CDHC, is proposed to detect the overlapping communities and to solve the fuzzy and hierarchical relationships among communities in social networks. The algorithm first utilizes the distance weighting factors to calculate the similarity among communities, and the communities with similarity larger than a given threshold are then merged together. The membership grade of each node for the merged community is computed and nodes with membership grades less than a given threshold are removed from the community to form a structure of the final overlapping community. The algorithm can not only detect the overlapping communities, but also detect the isolated nodes. The effectiveness of the proposed algorithm is tested through comparing it with two existing overlapping community detection algorithms, CMP and LFM, on the Lancichinetti synthetic network and real network datasets. Results show that the size of network and size of

收稿日期: 2014-07-12。 作者简介: 李刘强(1984—),男,博士生;桂小林(通信作者),男,教授,博士生导师。 基金项目: 国家自然科学基金资助项目(61172090);陕西省科技统筹创新工程资助项目(2013SZS16-Z01/P01/K01);中央高校基本科研业务费专项资金资助项目(XJJ2014049,XKJC2014008);陕西省自然科学基金资助项目(2012K06-30, 2014JQ8322)。

网络出版时间: 2015-01-07

网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20150107.1647.001.html>

communities have little effect on accuracy of detecting communities, and the main factor to affect the accuracy is the mixed degree among communities. The detection accuracy of the CDHC on social networks with small communities is higher than that of LFM, and it is better than CMP on networks with large communities. The CDHC algorithm improves the detection accuracy while its stability is good. Therefore, it can be concluded that the CDHC is an effective overlapping community detection algorithm for social networks.

Keywords: social network; similarity; fuzzy hierarchical clustering; overlapping community detection

社会网络是指社会个体成员之间因为互动而形成的相对稳定的关系体系。社会网络通常可抽象为节点以及节点间的边组成的图,图中的节点代表个体,节点间的边表示个体之间的联系。社会网络中蕴含着一些潜在的社区结构,同一社区内的节点具有较高的相似性,而不同社区间的节点相似性较低。了解社会网络中的社区结构有助于了解社会网络的本质,掌握社会网络中的社区结构已经成为近年来社会网络研究的热点。

2002年,Girven等提出了社区挖掘的概念^[1]。传统的社区挖掘算法是将网络划分为若干个独立的社区,每个节点只属于一个社区,代表性算法有基于划分的挖掘算法^[1-3]、基于模块性优化的算法^[4-7]、基于标签传播的算法^[8-9]和基于信息论的算法^[10]等。最近人们发现,社会网络中的社区之间并不是独立的,即一些节点可以同时隶属于多个社区,例如实际的人际关系中,每个人可以属于家庭、单位、朋友圈等不同的社会团体。因此,社会网络中的重叠社区结构更符合现实。

目前,社会网络中重叠社区的挖掘已经引起越来越多的关注,代表性的算法有Palla等提出的团渗算法(clique percolation method, CPM),其通过团过滤算法来发现网络中的社区结构^[11]。Evans等首次对网络中的链接进行了划分,以产生结点的重叠社区结构,即通过“用边表示节点,用节点形成边”的方法,将原始网络转化为线图,然后选择已有的节点划分算法获取链接社区结构^[12]。基于链接聚类的代表算法有LINK^[13]、DBLINK^[14]等。Lancichinetti等提出了基于局部扩展的重叠社区挖掘算法(local maxima for the fitness function, LFM)^[15],其从随机选择的一个种子节点出发,通过不断向外扩张来构建社区,直至社区目标优化函数达到局部最优为止。索勃等提出了基于信息流动分析的动态社区发现方法,通过对社交网络、微博等互联网应用的信息流动分析,来发现联系紧密、兴趣相近的节点集

合,从而实现动态社区发现^[16]。

针对社会网络中社区结构的重叠性和社区间的层次性关系,本文提出了一种基于模糊层次聚类的重叠社区检测算法(overlapping community detection based on fuzzy hierarchical clustering, CDHC)。在该算法中,通过引入距离加权因子来计算社区间的相似度,采用模糊层次聚类算法对相似度高的社区进行合并;计算合并生成的社区中节点对该社区的隶属度,并将隶属度小于隶属度阈值的节点从社区中移除。在Lancichinetti基准网络和真实网络上的实验表明,本文算法可有效发现社会网络中的重叠社区结构。

1 算法定义

基于模糊层次聚类的重叠社区检测算法定义如下。

定义1 网络图。通常,社会网络可以由加权的网络图 $G=(V,E,W)$ 来表示,其中 $V=\{v_1,v_2,\dots,v_n\}$ 表示节点的集合, $E=\{e_1,e_2,\dots,e_m\}$ 表示边的集合, $W=\{w_1,w_2,\dots,w_m\}$ 是对应于每条边的权重集合, n 和 m 分别表示网络中节点和边的个数。任意边 $e_i=(u,v)$ 表示 e_i 连接的2个节点分别为 u 和 v , e_i 对应的权重 w_i 在本文中表示节点之间的距离,权重越大,表示对应边连接的2个节点之间的距离越远,无权图的网络节点间的距离可记为1个单位长度。图1是一个简单的网络示意图,图中包含2个孤立节点 I_3 和 S_2 。

定义2 网络中的社区。社区 $C_i=\{v_1,v_2,\dots,v_k\}$ 是网络中节点的集合,代表网络中特定的结构, k 是社区 C_i 中包含节点的个数, k 越大,社区 C_i 的规模也越大。同一社区内节点之间连接紧密,不同社区的节点之间连接稀疏。图2给出了网络社区结构示意图,图中包含了4个非重叠社区,分别是 $\{I_3\}$, $\{S_2\}$, $\{S_4,W_6,W_8,W_7,W_9\}$, $\{S_1,W_1,W_2,W_3,W_4,W_5,I_1\}$ 。

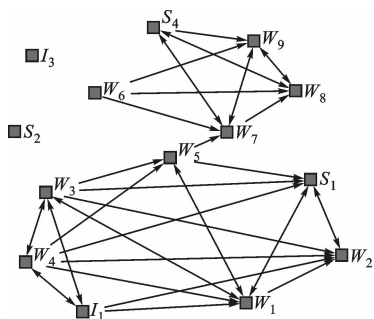


图1 网络示意图

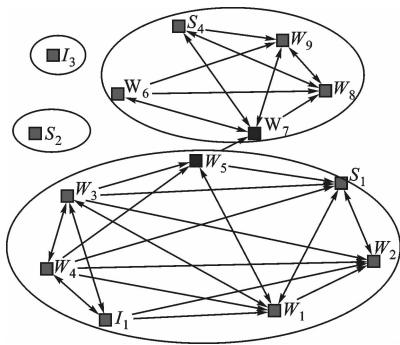


图2 网络社区结构示意图

定义3 社区间桥接边集合。社区 C_i 和 C_j 间的桥接边集合 $E_b(C_i, C_j)$ 是边 $e_b(v_i, v_j)$ 的集合, 其中 v_i 是社区 C_i 内的节点, v_j 是社区 C_j 内的节点, 则有

$$E_b(C_i, C_j) = \{e_b(v_i, v_j) \mid v_i \in C_i \wedge v_j \in C_j\} \quad (1)$$

社区间的桥接边集合表示社区间连接的紧密程度, 两个社区间的桥接边越多, 即 $|E_b(C_i, C_j)|$ 越大, 表示这两个社区间的联系越紧密。

定义4 社区间距离加权因子。社区 C_i 和 C_j 间的距离加权因子 $W_d(C_i, C_j)$ 是社区 C_i 和 C_j 间所有桥接边对应权重倒数的和, 即

$$W_d(C_i, C_j) = \begin{cases} |E_b(C_i, C_j)|, & \text{无权图} \\ |E_b(C_i, C_j)| \sum_{e_b \in E_b} 1/\omega(e_b), & \text{加权图} \end{cases} \quad (2)$$

$W_d(C_i, C_j)$ 与 $|E_b(C_i, C_j)|$ 、 $\sum_{e_b \in E_b} 1/\omega(e_b)$ 成正比。用最大-最小归一化方法将 $W_d(C_i, C_j)$ 的取值范围转换到 $[0, 1]$ 区间, 得

$$W_d^n(C_i, C_j) = \frac{W_d(C_i, C_j) - W_d^{\min}(C_i, C_j)}{W_d^{\max}(C_i, C_j) - W_d^{\min}(C_i, C_j)} \quad (3)$$

定义5 社区的邻居节点集合。社区 C_i 的邻居 $N(C_i)$ 是指所有和 C_i 中节点有边连接、且不属于 C_i 的节点的集合, 即

$$N(C_i) = \{v \mid e(v, v_i) \in G \wedge v_i \in C_i \wedge v \notin C_i\} \quad (4)$$

定义6 社区间相似度。社区 C_i 和 C_j 间的相似度

$$S(C_i, C_j) = \frac{|N(C_i) \cap N(C_j)|}{|N(C_i) \cup N(C_j)|} W_d^n(C_i, C_j) \quad (5)$$

式中: $N(C_i)$ 和 $N(C_j)$ 分别表示社区 C_i 和 C_j 的邻居节点集合; $W_d^n(C_i, C_j)$ 是社区 C_i 和 C_j 间的规范化距离加权因子。给定阈值 α , 若社区 C_i 和 C_j 的相似度大于 α , 则认为 C_i 和 C_j 相似。

根据定义1~6, 算法的第一步是比较两个社区间的相似度, 将相似度大于阈值 α 的两个社区进行合并。

定义7 节点的内部边集合与外部边集合。社区 C_i 中节点 v_i 连接的边的类型分为两种: 内部边集合与外部边集合。内部边集合是指和节点 v_i 相连且另一端节点 v 在 C_i 内的边的集合, 即

$$E_{C_i}^{\text{in}}(v_i) = \{e^{\text{in}} \mid e(v_i, v) \wedge v \in C_i\} \quad (6)$$

外部边集合是指和节点 v_i 相连且另一端节点 v 不在 C_i 内的边的集合, 即

$$E_{C_i}^{\text{out}}(v_i) = \{e^{\text{out}} \mid e(v_i, v) \wedge v \notin C_i\} \quad (7)$$

定义8 社区中节点的隶属度。节点 v_i 对社区 C_i 的隶属度定义为

$$\mu_{C_i}(v_i) = \frac{W(E_{C_i}^{\text{in}}(v_i))}{W(E_{C_i}^{\text{in}}(v_i)) + W(E_{C_i}^{\text{out}}(v_i))} = \frac{\sum_k \omega_k(e^{\text{in}})}{\sum_k \omega_k(e^{\text{in}}) + \sum_l \omega_l(e^{\text{out}})} \quad (8)$$

式中: $W(E_{C_i}^{\text{in}}(v_i))$ 表示节点 v_i 内部边集合中所有边的权重之和; $W(E_{C_i}^{\text{out}}(v_i))$ 表示节点 v_i 外部边集合中所有边的权重之和。

由式(8)可知, $W(E_{C_i}^{\text{in}}(v_i)) + W(E_{C_i}^{\text{out}}(v_i))$ 表示与节点 v_i 连接的所有边的权重之和, 则有 $0 \leq \mu_{C_i}(v_i) \leq 1$ 。

算法的第二步根据定义7和定义8给定隶属度阈值, 计算社区中每个节点对该社区的隶属度, 将小于隶属度阈值的节点从该社区中移除。

2 算法描述

2.1 算法伪代码

初始将所有单个节点作为一个独立的社区 C_i , 则初始社区集合 (C_s) 为 $C_s = \{C_1, C_2, \dots, C_n\}$, 其中 $C_1, C_2, \dots, C_n$ 分别为 $v_1, v_2, \dots, v_n$ 。

算法1 基于模糊层次聚类重叠社区检测算法。

输入 $G = (V, E, W), \alpha, \mu$

输出 节点社区集合 $C_s = \{C_1, C_2, \dots, C_n\}$

```

 $C_{S,temp} = \Phi$ 
do
  for  $\forall C_i, C_j \in C_S$ 
    if  $|E_b(C_i, C_j)| = 0$ 
      continue;
    endif
    if  $S(C_i, C_j) \geq \alpha$ 
       $C_k = \text{Combine}(C_i, C_j)$ 
      for each  $v_l \in C_k$ 
        if  $\mu_{C_k}(v_l) < \mu$ 
           $C_k = C_k - v_l$ 
        endif
      endfor
      if  $C_k \neq \Phi \wedge C_k \notin C_{S,temp}$ 
         $C_{S,temp} = C_{S,temp} \cup \{C_k\}$ 
      endif
    endif
  endfor
  if  $\exists C_m \in C_S \wedge \text{IsCombined}(C_m) = \text{false}$ 
     $C_{S,temp} = C_{S,temp} \cup \{C_m\}$ 
  endif
   $C_S = C_{S,temp}$ 
   $C_{S,temp} = \Phi$ 
while  $|C_S| = 1$ 
or  $\forall C_i, C_j \in C_S \wedge S(C_i, C_j) < \alpha$ 

```

算法 1 列出了 CDHC 的伪代码。可以看出, CDHC 的主程序包含 2 个阶段:第 1 阶段先合并相似度大于阈值 α 的社区;第 2 阶段对合并生成的原始社区进行裁剪,将社区中隶属度小于阈值 μ 的节点从该社区中移除。在第 1 阶段中,首先判断两个社区间是否有桥接边,若社区间没有桥接边,则两个社区没有交互关系,无需再计算相似度;在社区间有桥接边的情况下,计算社区间的相似度,若相似度大于阈值 α ,则合并这两个社区。然后,对第 1 阶段合并生成的社区中节点计算其隶属度,若节点的隶属度小于阈值 μ ,则将该节点从社区中移除;若经过裁剪后的社区非空,则将该社区放入临时社区集合 $C_{S,temp}$ 中。合并过程中可能出现孤立社区,即某社区与社区集合中的所有其他社区间的相似度都小于阈值 α ,该社区也放入临时社区集合 $C_{S,temp}$ 中,在下一轮的循环中继续参与合并。当所有节点都隶属于一个社区,或任意两个社区间的相似度都小于阈值 α 时,算法终止。

使用 CDHC 算法挖掘图 1 中的网络社区,结果

如图 3 所示,图中包含的 4 个社区分别为 $\{I_3\}$, $\{S_2\}$, $\{S_4, W_6, W_8, W_5, W_7, W_9\}$, $\{S_1, W_1, W_2, W_3, W_4, W_5, W_7, I_1\}$, 其中 W_5 和 W_7 是重叠节点。由此可见,CDHC 算法不仅可以发现重叠的社区结构,还可以处理孤立节点。

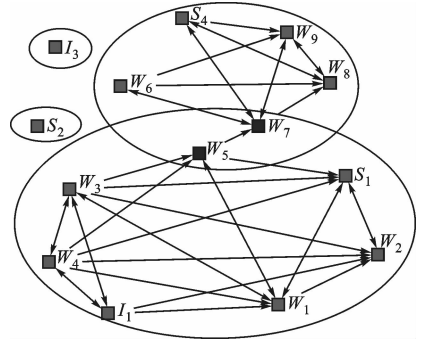


图 3 CDHC 算法的划分结果

2.2 时间复杂度

假设网络 G 包含 n 个节点和 m 条边,算法初始设置每个节点为一个独立的社区,并计算每对社区间的相似度,则 for 循环的时间复杂度为 $O(n^2)$ 。全连通网络中 $m = n^2$,实际网络中有 $m \ll n^2$,因此算法只需计算 m 次相似度即可,对应的进行 m 次社区的合并。对合并后的社区进行裁剪时,由于 $|C_k| \leq n$,因此裁剪过程的时间复杂度为 $O(n)$ 。选择合适的相似度阈值 α 和隶属度阈值 μ ,每轮循环结束后 $|C_S| \leq n$,假设需要进行 k 轮的合并和裁剪,则算法的整体时间复杂度为 $kO(n^2)$ 。

3 实验与结果分析

为了验证算法的有效性,将 CDHC 算法与具有代表性的重叠社区发现算法 CMP 和 LFM 进行了比较。实验环境:处理器为 Intel® Core™ i5-3470 CPU,主频为 3.20 GHz,内存为 4 GB,操作系统为 Windows7,运行环境为 JVM1.7。

3.1 实验数据

LFR 基准网络由 Lancichinetti 等在 2008 年提出^[17],该网络假设节点的度和社区的规模都服从幂率分布特性,且可模拟真实世界的网络。LFR 基准网络包含的参数及其含义见表 1。

Lancichinetti 基准网络由 Lancichinetti 等在 2009 年提出^[15],该网络在 LFR 基准网络的基础上增加了社区重叠的特性,其特有的参数及其含义见表 2,其他参数同表 1。

本文选取了 3 个真实网络数据进行实验,详细信息见表 3。

表 1 LFR 基准网络参数

参数	含义
N	网络中的节点总数
γ	节点的度的分布参数
β	社区规模的分布参数
K	节点的平均度
$K_{\min}$	节点的最小度
$K_{\max}$	节点的最大度
$s_{\min}$	最小社区包含的节点数
$s_{\max}$	最大社区包含的节点数
μ	混合参数,节点与社区外节点连接的概率

表 2 Lancichinetti 基准网络参数

参数	含义
N_0	重叠节点数
τ_1	节点的度的分布参数(同表 1 中的 γ)
τ_2	社区规模的分布参数(同表 1 中的 β)
μ_t	拓扑混合参数(节点与社区外节点连接的概率)

表 3 真实网络数据

ID	网络名称	节点数	边数
R1	Email	1 133	5 451
R2	Blogs	3 982	6 803
R3	PGP	10 680	24 316

3.2 评价指标

(1)基准网络评价指标 ENMI。

对于模拟生成的基准网络,由于网络中社区的结构信息是已知的,因此采用文献[15]中给出的扩展的规范化互信息(extended normalized mutual information,ENMI)(设为 ξ)来度量重叠社区挖掘算法的精度。假设基准网络中的社区集合为 C'_s ,挖掘算法发现的社区集合为 C''_s 。ENMI 指标度量 C''_s 和 C'_s 的一致性程度,挖掘算法发现的社区集合 C''_s 和基准网络的社区集合 C'_s 越接近,ENMI 指标就越大,挖掘算法的精度越高。

基准网络社区集合 C'_s 中的社区之间具有重叠性,节点 i 对社区的隶属关系可以用一个长度为 $|C'_s|$ 的向量 x_i 来表示, $(x_i)_k$ 为 1 表示节点 i 属于第 k 个社区, $(x_i)_k$ 为 0 表示节点 i 不属于第 k 个社区。将所有节点的二元向量的第 k 个元素组成的元组看作一个随机变量 X_k 。使用挖掘算法发现的社区集合记为 C''_s ,则所有节点的第 l 个元素组成的元组看作一个随机变量 Y_l 。定义 X_k 在 Y_l 上的条

件熵

$$H(X_k | Y_l) = H(X_k, Y_l) - H(Y_l) \quad (9)$$

C''_s 和 C'_s 越接近, $H(X_k | Y_l)$ 就越小,当 C''_s 和 C'_s 相同时, $H(X_k | Y_l)$ 为 0,因此定义 X_k 在 Y 上的条件熵为

$$H(X_k | Y) = \min_{l \in \{1,2,\dots,|C''_s|\}} H(X_k | Y_l) \quad (10)$$

对该条件熵进行规范化,有

$$H(X_k | Y)_n = H(X_k | Y) / H(X_k) \quad (11)$$

根据式(10)和式(11),定义 X 在 Y 上的规范化条件熵为

$$H(X | Y)_n = 1 / [|C'_s| \sum_k H(X_k | Y) / H(X_k)] \quad (12)$$

同理可以得到 Y 在 X 上的规范化信息熵,最后根据式(12)可以计算得到两个社区集合的扩展规范化互信息

$$\xi(X | Y) = 1 - [H(X | Y)_n + H(Y | X)_n] / 2 \quad (13)$$

(2)真实网络评价指标。

对于真实网络,由于网络中社区的结构信息是未知的,因此采用文献[18]中提出的具有重叠结构的模块度 Q_{ov} 来衡量重叠社区挖掘算法的精度。 Q_{ov} 的值越大,算法发现的社区结构和网络实际的社区结构越接近。

节点 i 和 j 对社区 C_i 的隶属度分别记为 μ_{i,C_i} 和 μ_{j,C_i} ,定义函数

$$F(\mu_{i,C_i}, \mu_{j,C_i}) = 1 / [(1 + e^{-f(\mu_{i,C_i})})(1 + e^{-f(\mu_{j,C_i})})] \quad (14)$$

式中: $f(x) = 40x - 20$ 。 $F(\mu_{i,C_i}, \mu_{j,C_i})$ 表示连接节点 i 和 j 的边 $e(i, j)$ 隶属于社区 C_i 的系数,令 $\beta_{e,c} = F(\mu_{i,C_i}, \mu_{j,C_i})$,则有

$$Q_{ov} = 1/m \sum_{C_i \in C_s} \sum_{i,j \in V} [\beta_{e,c} (A_{ij} - K_i K_j / m)] \quad (15)$$

式中: m 表示边数; V 表示节点集合; K_i 和 K_j 分别表示节点 i 和 j 的度; A_{ij} 表示网络中节点间的邻接矩阵。

3.3 基准网络社区检测结果

根据文献[17]中的参数设置,本文中 Lancichinetti 基准网络的参数如下:所有网络共享参数 $K = 20$, $K_{\max} = 50$, $\tau_1 = 2$, $\tau_2 = 1$;网络规模(网络中的节点总数) N 为 1 000 或 5 000, $s_{\min}$ 为 10 或 20, $s_{\max}$ 为 50 或 100, μ_t 为 0.1 或 0.3,重叠节点所占比率(N_0 / N)从 0 增加至 0.5,间隔 0.05。通过以上参数的设置可生成 8 种不同的基准网络,见表 4。

表 4 Lancichinetti 基准模拟网络信息

ID	N	K	$K_{\max}$	$s_{\min}$	$s_{\max}$	μ_t
S1	1 000	20	50	10	50	0.1
S2	1 000	20	50	10	50	0.3
S3	1 000	20	50	20	100	0.1
S4	1 000	20	50	20	100	0.3
S5	5 000	20	50	10	50	0.1
S6	5 000	20	50	10	50	0.3
S7	5 000	20	50	20	100	0.1
S8	5 000	20	50	20	100	0.3

CMP 中的 K 为 4; LFM 中的 α 为 1.2; CDHC 中相似度阈值 α 为 0.3, 隶属度阈值 μ 为 0.2。图 4 显示了算法 CMP、LFM 和 CDHC 在 4 个具有重叠社区的 Lancichinetti 基准网络上的社区检测结果。

由图 4 可以得到如下结果。

(1)随着网络中重叠节点所占比例的增加,网络中社区间的重叠性更高,社区检测的难度也更大, CMP、LFM 和 CDHC 3 种算法的社区检测精度都随着重叠节点所占比例的增加而有所下降。

(2)对比模拟网络 S1 和 S2、S3 和 S4、S5 和 S6、S7 和 S8 上的检测结果,发现在网络规模、社区规模和节点度分布相同的条件下,在低混合网络($\mu_t=0.1$)上,各个算法的社区检测结果对应的 ENMI 值优于对应的高混合网络($\mu_t=0.3$)的 ENMI 值。

(3)对比模拟网络 S1 和 S5、S2 和 S6、S3 和 S7、S4 和 S8 上的检测结果,发现 3 种算法在小规模网络($N=1\ 000$)上的社区检测精度略高于大规模网络($N=5\ 000$)。

(4)对比网络 S1 和 S3、S2 和 S4、S5 和 S7、S6 和 S8 上的检测结果,发现 3 种算法在小社区网络($s_{\min}=10, s_{\max}=50$)上的社区检测精度略高于大社区网络($s_{\min}=20, s_{\max}=100$)。

由式(1)~(4)可知,网络规模和网络中社区规模对社区的检测精度有一定的影响,但影响不显著,影响社区发现质量的主要因素是社区间的混合程度。

(5)在小社区网络中, CMP 算法的性能最好, CDHC 次之, LFM 最差;在大社区网络中, LFM 算法的性能最好, CDHC 次之, CMP 最差。由此可见, CDHC 算法在小社区网络上的社区检测精度优于 LFM,在大社区网络上的社区检测精度优于 CMP。

综上,在各种不同网络上, CDHC 算法在保持网络社区检测精度的同时,还具有稳定的性能。

3.4 真实网络社区检测结果

真实网络上的社区结构检测精度用 Q_{ov} 来衡量, Q_{ov} 越大,表明算法发现的社区结构与真实网络实际的社区结构越接近。

表 5 给出了各算法的参数设置及其在 3 个真实网络上获得的 Q_{ov} 。 LFM 算法中 α 控制社区的大小, α 越小,社区结构就越大, α 越大,社区结构就越小。当 $\alpha<0.5$ 时,算法只能检测到一个社区,即网络自身;当 $\alpha>2$ 时,检测到的每个节点被认为是一个社区^[15]。 CDHC 算法中 α 是社区间相似度阈值,其作用类似于 LFM 算法中的 α 。 CDHC 算法中 α 主要影响 CDHC 算法的合并过程,进而影响检测结果中社区的大小, α 越大,算法检测结果中社区结构就越小。当 $\alpha=0$ 时, CDHC 算法只能检测到一个社区,即网络自身;当 $\alpha>1$ 时,检测到的每个节点被认为是一个社区。

在真实网络 R1 上, CDHC 算法取得了最好的值 Q_{ov} 值, CMP 算法次之, LFM 算法的效果最差;在 R2 和 R3 上,通过选择合适的 α 值, LFM 算法取得了最好的 Q_{ov} 值, CDHC 在 R2 和 R3 上尽管不是最好的,但 Q_{ov} 的值和 LFM 非常接近。由图 4 可知, LFM 算法在大社区网络上的效果最好,因此 R2 和 R3 可能是大社区网络,而 R1 应该是一个小社区网络。

综上,在真实网络上, CDHC 也是一种有效的社区检测算法。

表 5 真实网络社区检测结果

ID	CMP		LFM		CDHC		
	K	Q_{ov}	α	Q_{ov}	α	μ	Q_{ov}
R1	4	0.462	1.4	0.125	0.3	0.2	0.471
R2	4	0.375	0.8	0.797	0.3	0.2	0.738
R3	4	0.568	0.9	0.653	0.3	0.2	0.634

注:加粗的数字表示网络上 Q_{ov} 的最大值。

4 结 论

本文提出了一种基于模糊层次聚类的重叠社区检测算法 CDHC。算法中通过引入距离加权因子来计算社区间的相似度,并通过模糊层次聚类来合并相似度高的社区;利用节点对社区的隶属度阈值将合并生成的社区进行裁剪,形成最终的 网络重叠社区结构划分。通过实验,在 8 个基准网络和 3 个真实网络上,将 CDHC 算法的检测结果与 CMP 和 LFM 算法进行了比较。结果表明, CDHC 在保持社

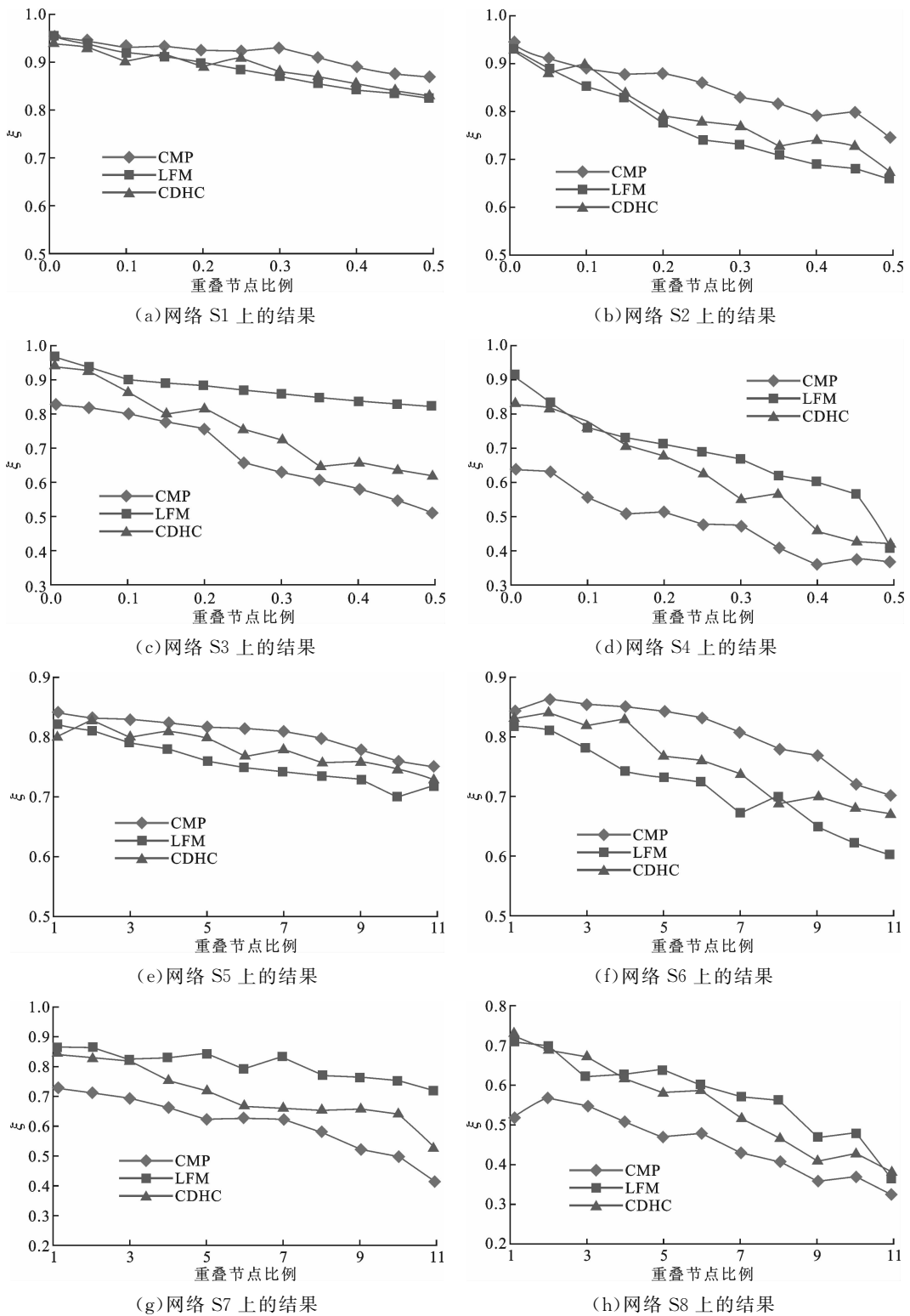


图 4 各算法的聚类精度随重叠节点所占比例的变化趋势

区检测质量的同时,具有较好的稳定性,是一种有效的社会网络重叠社区检测算法。

在今后的工作中,将从以下几个方面继续进行深入研究:①在更多的真实网络数据上进行实验,研究网络类型与相似度阈值、隶属度阈值之间的关系,

为不同类型的网络选择更合理的参数;②综合考虑层次与重叠结构之间的内在关系,研究合理度量层次化重叠社区结构的量化标准,实际网络结构的重叠社区检测算法。

参考文献:

- [1] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. *Proceedings of National Academy of Science*, 2002, 9(12): 7821-7826.
- [2] TYLER J R, WILKINSON D M, HUBERMAN B A. Email as spectroscopy: automated discovery of community structure within organizations [C] // *Proceedings of First International Conference on Communities and Technologies*. Dordrecht, Holland: Kluwer, 2003:81-96.
- [3] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks [J]. *Proceedings of National Academy of Science*, 2004, 101(9):2658-2663.
- [4] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. *Physical Review: E*, 2004, 69(2): 026113.
- [5] NEWMAN M E J. Modularity and community structure in networks [J]. *Proceedings of National Academy of Science*, 2006,103(23): 8577-8582.
- [6] LEE J, GROSS S P, LEE J. Modularity optimization by conformational space annealing [J]. *Physical Review: E*, 2012, 85(5):056702.
- [7] SHANG Ronghua, BAI Jing, JIAO Licheng, et al. Community detection based on modularity and an improved genetic algorithm [J]. *Physica: A Statistical Mechanics and Its Applications*, 2013, 392(5):1215-1231.
- [8] 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.
- [9] SUBELJ L, BAJEC M. Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction [J]. *Physical Review: E*, 2011, 83(3):036103.
- [10] 邓小龙,王柏,吴斌,等. 基于信息熵的复杂网络社团划分建模和验证 [J]. *计算机研究与发展*, 2012, 49(4): 725-734.
DENG Xiaolong, WANG Bai, WU Bin, et al. Modularity modeling and evaluation in community detecting of complex network based on information entropy [J]. *Journal of Computer Research and Development*, 2012, 9(4): 725-734.
- [11] PALLA G, DERENYI I, FARKAS I, et al. Uncovering the overlapping community structures of complex networks in nature and society [J]. *Nature*, 2005, 435(7043): 814-818.
- [12] EVANS T S, LAMBIOTTE R. Line graphs, link partitions and overlapping communities [J]. *Physics Review: E*, 2009, 80(1): 016105.
- [13] AHM Y Y, BAGROW J P, LEHMANN S. Link communities reveal multiscale complexity in networks [J]. *Nature*, 2010, 466(7303): 761-764.
- [14] 朱牧,孟凡荣,周勇. 基于链接密度聚类的重叠社区发现算法 [J]. *计算机研究与发展*, 2013, 50(12): 2520-2530.
ZHU Mu, MENG Fanrong, ZHOU Yong. Density-based link clustering algorithm for overlapping community detection [J]. *Journal of Computer Research and Development*, 2013, 50(12): 2520-2530.
- [15] LANCICHINETTI A, FORTUNATO S, KERTESZ J. Detecting the overlapping and hierarchical community structure in complex networks [J]. *New Journal of Physics*, 2009, 11(3): 033015.
- [16] 索勃,李战怀,陈群,等. 基于信息流动分析的动态社区发现方法 [J]. *软件学报*, 2014, 26(3): 547-559.
SUO Bo, LI Zhanhuai, CHEN Qun, et al. Dynamic community detection based on information flow analysis [J]. *Journal of Software*, 2014, 26(3): 547-559.
- [17] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms [J]. *Physics Review: E*, 2008, 78(4): 046110.
- [18] NICASIA V, MANGIONI G, CARCHIOLO V, et al. Extending the definition of modularity to directed graphs with overlapping communities [EB/OL]. (2009-03-16) [2013-12-13]. <http://iopscience.iop.org/1742-5468/2009/03/p03024>.

[本刊相关文献链接]

李长路,王劲林,郭志川,等. 两阶段密度意识子空间聚类模型. 2014,48(10):108-114. [doi:10.7652/xjtuxb201410017]
 杜辉,王宇平,董晓盼. 采用万有引力定律自动确定类数的 K 均值算法. 2014, 48(10): 115-119. [doi: 10.7652/xjtuxb201410018]
 董哲,伊鹏. 采用链路聚类的动态网络社团发现算法. 2014, 48(8):73-79. [doi:10.7652/xjtuxb201408013]
 张永斌,陆寅,张艳宁. 域名请求行为特征与构成特征相结合的域名变换检测. 2013,47(8):54-60. [doi:10.7652/xjtuxb201308010]

(编辑 武红江)