

重叠社区发现的两段策略

陈端兵 尚明生 李 霞

(电子科技大学互联网科学中心 成都 611731)

摘 要 复杂网络中的社区特别是重叠社区在信息传播与推荐、舆情控制、商业营销等领域中具有重要作用。在实际的网络中,由于有些节点天然地属于多个功能团体,重叠社区的挖掘越来越受到重视。提出了一种重叠社区挖掘的两段策略算法:初始社区抽取与社区合并。在社区抽取阶段,选择网络中最大度节点及其紧密的邻居节点作为初始社区,将与该初始社区联系紧密的节点也一并加入;在社区合并阶段,如果两个社区合并之后使得模块度增加,则合并这两个社区。用包括大规模网络在内的3个实际网络对所提算法进行了测试,结果表明,该算法可有效挖掘网络中的重叠社区。

关键词 复杂网络,重叠社区,社区抽取和合并,连接度

中图法分类号 TP301.6 **文献标识码** A

Two-phase Strategy on Overlapping Communities Detection

CHEN Duan-bing SHANG Ming-sheng LI Xia

(Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract Communities, especial overlapping communities in complex networks are significant in many fields such as information spreading and recommending, public opinion controlling, and commercial marketing. Overlapping communities detecting is attracting increasing attentions since some nodes may naturally belong to several groups in real-world networks. This paper proposed an overlapping community detecting algorithm based on two phase strategies: initial community extracting and community merging. In extracting phase, a node with maximal degree and its tight neighbors are selected as an initial community, and nodes tight with the community are also included. In merging phase, two communities are merged if the modularity gets larger after merging. Three real-world complex networks including a large-scale one were used to evaluate the algorithm. Experimental results demonstrate that the proposed algorithm is efficient for detecting overlapping communities in complex networks.

Keywords Complex network, Overlapping community, Community extracting and merging, Connecting degree

1 引言

现实中的许多复杂系统可以用图或网络进行描述^[1-3]。网络中的社区结构对深入理解复杂网络有重要作用。早期的研究主要集中在非重叠社区,Newman等人^[4]在2004年提出了基于模块度的社区挖掘算法,引发了社区挖掘研究的热潮。在Newman的工作基础上,对模块度进行扩充,得到了各种变形的模块度及相应的社区挖掘算法^[5-7]。除了基于模块度的社区挖掘,对启发式社区挖掘算法的研究也较多,Chen等人^[8]提出了一种滚雪球式的社区挖掘算法,其基于节点和社区的连接紧密度不断向社区中加入节点以形成更大的社区,直到得到一个局部最优的社区。Li等人^[9]提出了一种基于事件驱动的社区挖掘算法——ECODE。

在真实的网络中,一个节点可能属于多个社区,重叠社区的挖掘在最近几年引起了极大的关注。Lancichinetti等人^[10]

提出了一种重叠社区挖掘的局部算法,算法的核心是设计一个适应值函数,然后对此函数进行优化,其局部最优值对应的划分就是社区划分的结果。Palla等人^[11]基于 k -派系提出了一种重叠社区挖掘算法,由于节点可能属于多个 k -派系,从而保证能够找到网络中的重叠社区。Shen等人^[12]在Newman提出的模块度基础上,从最大派系的视角出发,提出了一个重叠社区划分效果的评价指标,通过优化此评价指标,作者提出了一种重叠社区挖掘算法,基于最大派系和汇聚的思想框架,他们还提出了重叠社区的层次划分方法^[13]。Gregory^[14]提出了一种重叠社区挖掘的两阶段策略,首先利用节点介数对节点进行分割,将原始网络转为一个新的网络,然后在转换后的网络上使用分裂算法对社区进行划分。Chen等人^[15]提出了一种快速的重叠社区挖掘算法,算法从初始只有一个节点的初始社区出发,采用滚雪球方式对社区进行扩展。

本文提出的重叠社区挖掘算法包括两个主要阶段:从最

到稿日期:2012-03-18 返修日期:2012-07-28 本文受国家自然科学基金(60973069, 90924011, 60903073, 60973120), 华为高校合作基金(YBCB2011057)资助。

陈端兵 副教授,硕士生导师,主要研究方向为数据挖掘、社会计算、信息物理等;尚明生 教授,博士生导师,主要研究方向为社会网络、信息推荐、并行计算等;李 霞 硕士生,主要研究方向为复杂系统、现代经济管理等。

大度节点出发的社区抽取阶段和基于模块度的社区合并阶段。用俱乐部合作网络、科学家合作网络及互联网络对算法性能进行了评价。实验结果表明,本文提出的算法可有效挖掘出网络中的重叠社区。

2 基本概念

2.1 连接度

本文定义了两种连接度:节点和社区间的连接度,以及两个社区之间的连接度。Gregory^[16]在他最近的工作中也使用了这个概念,读者可参阅文献[16]以获得更多详细信息。

2.1.1 节点与社区之间的连接度

对给定的节点 u 和社区 c , u 和 c 之间的连接度 $C(u, c)$ 定义为:

$$C(u, c) = \frac{\sum_{v \in c} w_{uv}}{k_u} \quad (1)$$

式中, k_u 为节点 u 的度,如果 u 和 v 之间存在连边,那么 $w_{uv} = 1$, 否则, $w_{uv} = 0$ 。式(1)中的 $C(u, c)$ 反映了节点 u 和社区 c 的联系紧密程度。如果节点 u 的所有邻居节点在社区 c 中, $C(u, c) = 1$, 否则 $C(u, c) < 1$ 。例如,在图 1 中,如果社区 c 包含节点 1, 2, 3 和 4, 那么节点 5 和社区 c 的连接度 $C(5, c) = 3/4 = 0.75$, 同样地, 节点 6 和社区 c 的连接度 $C(6, c) = 2/4 = 0.5$ 。

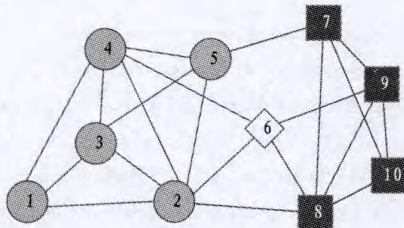


图 1 具有社区结构的网络示意图(图中有两个社区,其中节点 6 是正方形标记的社区和圆形标记的社区的公共点)

2.1.2 两个社区之间的连接度

两个社区 c_1 和 c_2 之间的连接度定义为:

$$C(c_1, c_2) = \frac{\sum_{u \in c_1, v \in c_2} w_{uv}}{\min(\sum_{u \in c_1} w_{uv}, \sum_{v \in c_2} w_{uv})} \quad (2)$$

式(2)反映了一个小社区和一个大社区联系的紧密程度。对给定的社区,其连接度越大,表明社区联系越紧密。例如,在图 1 中,圆圈标记的社区 c_1 和正方形标记的社区 c_2 之间有两条边,而社区内部分别有 9 条和 6 条边,因此,社区 c_1 和 c_2 之间的连接度为 $2/\min(9, 6) = 1/3$ 。

2.2 模块度

Newman 和 Girvan^[4]提出了模块度 Q 来衡量社区划分的效果,但文献[4]定义的模块度只适用于不重叠社区的情形。为了对重叠社区划分效果进行评价,一些研究者对 Newman 提出的模块度进行扩展而得到了各种改进的模块度。Lázár 等人^[17]提出了一种模块度指标用于评价社区划分的效果,并设计了一种通过优化此模块度而得到重叠社区划分的算法。Nepusz 等人^[18]提出了另外一种模块度指标用于评价模糊社区,其实就是重叠社区的划分效果。本文采用文献[19]提出的模块度指标方法评价重叠社区划分的效果,其定义为:

$$Q_o = \frac{1}{2m} \sum_{c \in P} \sum_{u, v \in c} \alpha_{cu} \alpha_{cv} (A_{uv} - \frac{k_u k_v}{2m}) \quad (3)$$

式中, A 是网络的邻接矩阵,如果 u 和 v 之间存在一条连边,那么 $A_{uv} = 1$, 否则 $A_{uv} = 0$; m 是网络的总边数; P 是对应划分的社区集合; k_u 和 k_v 分别是节点 u 和 v 的度; α_{cu} 是节点 u 隶属于社区的程度^[19]。

3 社区挖掘算法

本文提出的重叠社区划分算法包括两个阶段:(i)从一个标记为“F”的最大度节点出发,扩展而抽取出一个社区;(ii)基于式(3)定义的模块度合并社区最终得到一个更好的网络划分。在社区划分过程中,对已经找到的社区,将其中的节点标记为“T”,这些节点集合记为 V_T , 剩下的节点集合记为 $V_F = V - V_T$ 。初始时,所有节点都标记为“F”。在社区初始点选择时,从 V_F 中取一个度最大的节点,而在社区扩展过程中,从所有节点 V 中考察。在社区抽取阶段, Hu 等人^[20]和 Nandini 等人^[21]也提出了类似的思路。

3.1 社区抽取

对于给定的网络,从一个标记为“F”的最大度节点出发,可以抽取出一个社区,其具体步骤如算法 1 所示。

算法 1 社区抽取

输入:网络 $G(V, E)$; 阈值 C_L (本文 C_L 取为 0.45)

输出:社区集合

将网络中所有节点标记为“F”, List 为抽取出的社区集合,初始为 Φ ;
while 存在“F”节点 do

 选择标记为“F”的最大度节点以及它的邻居节点,这些节点形成一个初始社区 c ;

 for 社区 c 中每一个节点 v do

 计算节点与社区间的连接度 $C(v, c)$;

 end for

 从社区 c 中移除连接度 $C(v, c)$ 小于 C_L 的节点 v ;

 Flag = true;

 while Flag = true do

 Flag = false;

 找出社区 c 的所有邻居节点,记为 N_c ;

 for N_c 中每一个节点 v do

 计算节点与社区的连接度 $C(v, c)$;

 if $C(v, c) \geq C_L$ then

 将节点 v 加入社区 c , 得到新的社区,仍记为 c ;

 Flag = true;

 end if

 end for

 end while

 将社区 c 中的节点标记为“T”;

 List = List $\cup$ $\{c\}$;

end while

return List;

例如,在图 2 中,由于节点 9 的度最大,选择节点 9 及其邻居 {6, 7, 8, 10, 13, 14} 作为初始社区 {6, 7, 8, 9, 10, 13, 14}。由于节点 13 和初始社区的连接度 ($= 2/5 = 0.4$) 小于阈值 C_L ($= 0.45$), 将其从初始社区中移除, 同样, 节点 14 也需要从初始社区中移除, 其它节点不需要再移除。因而, 最终得到的初

始社区 c_1 为 $\{6, 7, 8, 9, 10\}$ 。现在,考察社区 c_1 的所有邻居节点 $\{2, 4, 11, 12, 13, 14, 15\}$,只有节点 15 和社区 c_1 的连接度 $C(15, c_1) = 2/3 > 0.45$,其它节点和社区 c_1 的连接度都小于 0.45,不能加入社区 c_1 ,因而,抽取出的第一个社区 c_1 为 $\{6, 7, 8, 9, 10, 15\}$,同时,将社区 c_1 中的节点标记为“T”。类似地,从剩下的“F”节点中挑选最大度节点,可抽取得到其它两个社区 $c_2 = \{1, 2, 3, 4, 5, 6\}$ 和 $c_3 = \{11, 12, 13, 14\}$ 。

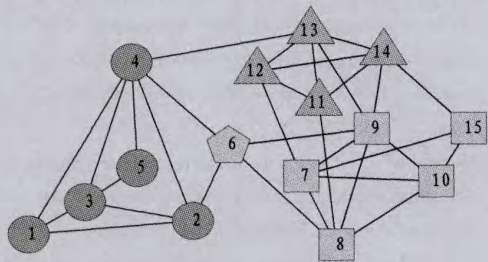


图2 社区抽取的例子

3.2 社区合并

通过社区抽取阶段,可以抽取网络中的初始社区。但对这些社区而言,有些社区之间联系可能非常紧密,需要进一步合并以获得更好的划分,其具体步骤如算法2所示。

算法2 社区合并

输入:初始的社区集合 P^{in}

输出:最终得到的社区划分集合 P^{out}

如果两个社区之间的连接度大于阈值 R_L (本文中 R_L 仍取为 0.45),则作为候选合并社区。合并的基本原则是:如果两个社区合并之后模块度增加,就真实地合并这两个社区。用 List 表示两个社区的二元组集合。

Flag=true;

while Flag=true do

List= Φ ;

C_{checked}= Φ ;

Flag=false;

for 集合 P^{in} 中每一个社区 c_1 do

C_{checked}←C_{checked} $\cup$ $\{c_1\}$;

for 集合 $P^{in} \setminus C_{checked}$ 中每一个社区 c_2 do

计算社区之间的连接度 $C(c_1, c_2)$;

if $C(c_1, c_2) > R_L$ then

将二元组 $\langle c_1, c_2 \rangle$ 加入集合 List;

end if

end for

end for

for 集合 List 中每一个二元组 $\langle c_1, c_2 \rangle$ do

暂时合并 c_1 和 c_2 ,并计算合并之后的模块度 Q_0 ;

if Q_0 大于合并之前的模块度 then

真实地合并 c_1 和 c_2 ;

Flag=true;

end if

end for

合并之后的社区集合仍记为 P^{in} ;

end while

$P^{out} \leftarrow P^{in}$

return 最终的社区挖掘结果 P^{out} ;

4 实验结果分析

对本文提出的算法,用 C# 编程语言实现,所有程序运行于主频 3.0GHz,内存 3GB 的酷睿 2 微型计算机。选取 3 个实际网络对本文提出的社区挖掘算法进行了验证,即空手道俱乐部网络^[22]、科学家合作网络^[7]以及互联网络。

4.1 空手道俱乐部网络

由于空手道俱乐部网络具有已知的社区结构而广泛用于对各种社区检测算法的测试。这个网络包括 34 个节点、78 条边,描述的是美国某大学的一个空手道俱乐部会员之间的社交关系。一次偶然的机会,俱乐部的管理员和教练员产生了冲突,最终俱乐部分裂成以管理员和教练员为核心的两派。采用本文算法,将网络划分成了两个具有重叠点(节点 10)的社区,社区划分的模块度 Q_0 为 0.4214,如图 3 所示。此网络在文献[12,14,19]也进行了讨论。在文献[12]中,检测出了 4 个社区,在文献[14,19]中划分出了两个具有一个重叠点的社区。本文算法划分得到的两个重叠社区和实际的社区结构的情况对应较好,节点 10 和每一个社区都只有一条连边,把它作为两个社区的重叠点是合乎情理的。

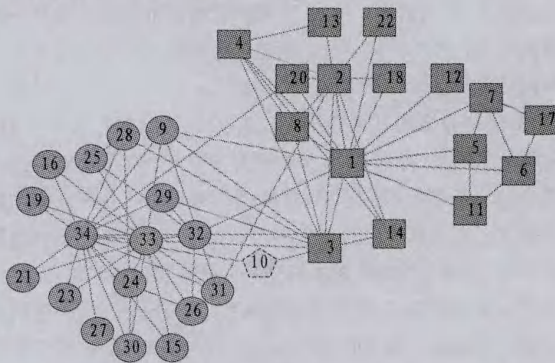


图3 空手道俱乐部网络的两个重叠社区

4.2 科学家合作网络

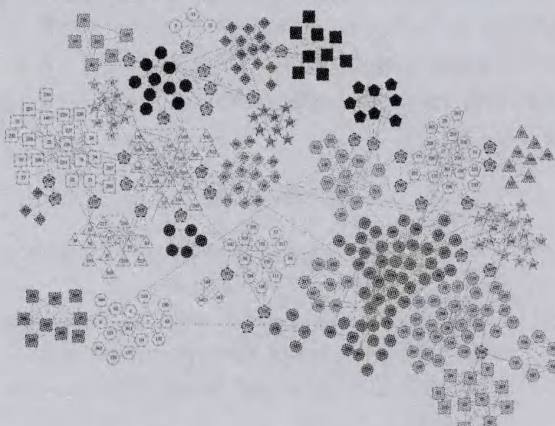


图4 科学家合作网络的社区结构

这个测试实例是复杂网络领域的科学家合作发表论文形成的网络,如果两个科学家合作发表了一篇论文,那么他们之间就有一条连边。网络总共包含 1589 名科学家,本文只取其中的一个最大连通分支(包含 379 个节点)作为测试算例。本文算法将网络划分成了 28 个紧密的重叠社区,其中包括一些

较小的社区,社区划分的模块度 Q 为 0.8919,其社区结构如图 4 所示。在图 4 中,虚框正五边形所示节点为社区间的重叠点,分析实际网络发现,其中大多数重叠点都是一些交叉领域的科学家,如节点 106(Kleinberg, J),节点 131(Crucitti, P)和节点 214(Schnitzler, A)等。此网络在文献[14, 19]中也进行了讨论,其社区划分的模块度分别为 0.8531 和 0.8156。

4.3 互联网络

这个网络是 Newman 根据 2006 年 7 月 22 日互联网上的一个自治系统的快照建立的。数据可以从 <http://www-personal.umich.edu/~mejn/netdata> 下载。网络包含了 22963 个节点、48436 条边。采用本文算法将社区划分成了 264 个重叠社区,其模块度为 0.58。

从 3 个实际的测试实例可以看出,本文算法可以有效地检测网络中的社区结构,而且可用于大规模网络的重叠社区结构检测,从而有助于更加深入地理解网络的结构和功能。

结束语 挖掘网络的社区结构对于深入理解网络所具有的功能、基于社区结构的推荐、网络核心节点的挖掘等都有重要作用。本文基于两阶段策略,提出了一种重叠社区的划分算法,算法的第一阶段,社区粗划分,得到内部联系非常紧密的小社区;算法的第二阶段,对联系紧密的社区进一步合并而得到更为合理的社区划分。用包括大规模网络在内的 3 个实际网络对算法的有效性进行了验证。对于空手道俱乐部网络,算法划分出的社区结构和实际的社区结构吻合非常好。对科学家网络,得到了 28 个联系紧密的社区,并且挖掘出了其中的交叉领域的科学家,对网络的深入研究具有重要的指导意义。本文算法在具有 22963 个节点的大规模上也有很好的表现。然而,社区挖掘本身是一个极其复杂的问题,尤其针对超大规模网络,如何设计高效的社区划分算法,是一个极具挑战性的问题。对有向网络的社区划分研究,目前研究也非常少,而现实中存在相当一部分复杂系统,如在线社交网络,用有向网络更能准确描述它们的一些内在结构特征,如何设计快速有效的算法检测有向网络中的社区也是一个极有价值的研究问题。随着人类动力学研究的推动,含时网络^[23,24]的研究也逐渐深入,挖掘含时网络中的社区结构对信息传播与控制、知识发现等都具有重要的作用。下一步将对这些问题展开深入研究。

参 考 文 献

- [1] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45: 167-256
- [2] Albert R, Barabási A-L. Statistical mechanics of complex networks[J]. Reviews of Modern Physics, 2002, 74: 47-79
- [3] Strogatz S H. Exploring complex networks[J]. Nature, 2001, 410: 268-276
- [4] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69: 026113
- [5] Duch J, Arenas A. Community detection in complex networks using extremal optimization[J]. Physical Review E, 2005, 72: 027104
- [6] Clauset A. Finding local community structure in networks[J]. Physical Review E, 2005, 72: 026132
- [7] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74: 036104
- [8] Chen D B, Fu Y, Shang M S. A fast and efficient heuristic algorithm for detecting community structures in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388: 2741-2749
- [9] Li X L, Tan A, Yu P S, et al. Ecode: Event-based community detection from social networks[C]// Proceedings of 16th International Conference on Database Systems for Advanced Applications, Part I, Lecture Notes in Computer Science, Hong Kong, China, 2011, 6587: 22-37
- [10] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structures in complex networks[J]. New Journal of Physics, 2009, 11: 033015
- [11] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435: 814-818
- [12] Shen H W, Cheng X Q, Guo J F. Quantifying and identifying the overlapping community structure in networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, P07042
- [13] Shen H W, Cheng X Q, Cai K, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388: 1706-1712
- [14] Gregory S. Finding overlapping communities using disjoint community detection algorithms[J]. Studies in Computational Intelligence, 2009, 207: 47-61
- [15] Chen D B, Shang M S, Lv Z, et al. Detecting overlapping communities of weighted networks via a local algorithm[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389: 4177-4187
- [16] Gregory S. Fuzzy overlapping communities in networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2011, P02017
- [17] Lázár A, ábel D, Vicsek T. Modularity measure of networks with overlapping communities[J]. Europhysics Letters, 2010, 90: 18001
- [18] Nepusz T, Petróczy A, Négyessy L, et al. Fuzzy communities and the concept of bridgeness in complex networks[J]. Physical Review E, 2008, 77: 016107
- [19] Shang M S, Chen D B, Zhou T. Detecting overlapping communities based on community cores in complex networks[J]. Chinese Physics Letters, 2010, 27: 058901
- [20] Hu Y, Chen H, Zhang P, et al. Comparative definition of community and corresponding identifying algorithm[J]. Physical Review E, 2008, 78: 026121
- [21] Nandini R U, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76: 036106
- [22] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33: 452-473
- [23] Pan R K, Saramäki J. Path lengths, correlations, and centrality in temporal networks[J]. Physical Review E, 2011, 84: 016105
- [24] Holme P, Saramäki J. Temporal networks[R]. Physics Reports, in press, arXiv: 1108. 1780, 2011