一种重叠社区发现的启发式算法

万雪飞,陈端兵,傅 彦
WAN Xue-fei,CHEN Duan-bing,FU Yan

电子科技大学 计算机科学与工程学院,成都 610054
School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China
E-mail:wxf2005@tom.com

WAN Xue-fei,CHEN Duan-bing,FU Yan.Heuristic algorithm for detecting overlapping communities.Computer Engineering and Applications,2010,46(3):36-38.

Abstract:A heuristic algorithm is proposed to detect overlapping communities in this paper.The algorithm is based on the local modularity and the vertex with greatest degree is considered as the initial community.Then expanding the community by putting the adjacent vertex which has maximal contribution into it.Furthermore,the algorithm takes into account the overlapping property of community.If there are border vertices which contribute greatly to more than one community,adding them to these communities.Finally,In order to make the result more reasonable,the detected communities is adjusted according to overlapping coefficient. Two typical social networks,Zachary and American College Football,are applied to the algorithm.The results show that it can detect overlapping communities rapidly and correctly and also mine the border vertices.

Key words:overlapping communities;community structure;social network;border vertex;heuristic algorithm

摘 要:提出了一种重叠社区发现的启发式算法。该算法基于局部贡献度的思想,以度最大的节点作为初始社区,逐步把对社区贡献最大的邻节点加入社区;同时考虑了社区的重叠性,若存在对多个社区贡献都很大的边界节点,则把边界节点同时加入到这些社区中。最后利用重叠系数对所划分的社区进行调整,使社区结构更加合理。对两个经典的社会网络 Zachary 和 American College Football 进行了实验测试,实验结果表明:该算法能快速准确地划分出社区,并能挖掘出社区间的边界节点。

关键词:重叠社区;社区结构;社会网络;边界节点;启发式算法
DOI:10.3778/j.issn.1002-8331.2010.03.011 **文章编号**:1002-8331(2010)03-0036-03 **文献标识码**:A **中图分类号**:TP301.6

1 引言

现实中的很多系统都可以用复杂网络来描述。复杂网络中的节点可表示为复杂系统中的个体,节点之间的边则是系统中个体之间按照某种规则而自然形成的一种关系。现实世界中包含着各种类型的复杂网络,如社会网络[1-4](朋友关系网络及合作网络等)、技术网络[2-3,5-7](Internet、万维网及电力网等)、生物网络[2,8-10](神经网络、食物链网络以及新陈代谢网络等)。这些网络都具有一种普遍的特性——社区结构(community structure)。大量实证研究表明,许多网络是异构的,即复杂网络不是一大批性质完全相同的节点随机地连接在一起,而是许多类型的节点的组合。相同类型的节点之间连接紧密,不同类型的节点之间的连接稀疏(如图 1 所示)。把同一类型的节点以及这些节点之间的边所构成的子图称为网络社区(community)。

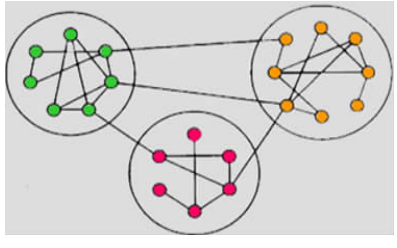

图 1 具有社区结构的网络图

在复杂网络中搜索或发现社区,有助于人们理解和开发网络,具有重要社会价值,由此出现了许多社区发现算法。较早社区发现算法包括 Kernighan-Lin 算法[11]和基于 Laplace 图特征值的谱平分法[12]。Newman 和 Girvan 提出的 GN 算法[13]是一种基于边界数(Edge Betweenness)的分割方法,将网络划分成越来越小的部分。该算法计算量大,并且划分出的社区没有一个定量描述。因此 Newman 和 Girvan 之后又提出了度量网络社区

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z414,No.2007AA01Z440);国家 242 信息安全计划项目(the Chinese Information Security Research Plan under Grant No.(242)2007B27);四川省应用技术研究与开发项目支撑计划(the Research and Development Project on Application Technology of Sichuan under Grant No.2008GZ0009);中国博士后科学基金资助项目(China Postdoctoral Science Foundation under Grant No.20080431273)。

作者简介:万雪飞(1986-),男,硕士研究生,主要研究方向为数据挖掘;陈端兵(1971-),男,博士,讲师,主要研究方向为数据挖掘,社会计算,NP 难问题高效求解等;傅彦(1962-),女,教授,博士生导师,主要研究方向为数据挖掘,信息安全,模式识别等。

收稿日期:2009-02-17 修回日期:2009-04-07

质量的标准——模块度 $Q=\sum_r (e_{rr}-a_r^2)$[14]，其中 e_{rr} 表示社区 r 内部连边占网络边数的比例；a_r 表示社区 r 内部连边及与外部连边之和占网络总边数的比例，成为社区发现研究中的重要参考模型。Radicchi 等人在 GN 算法基础上作了进一步改进，提出了快速分裂算法[15]。

目前的大多算法将一个节点仅归属于一个社区。然而在现实自然界中，事物具有多样性的特点，一种事物往往可归属到不同的类别中，社区间必定存在重叠的现象，即一个节点可属于多个社区。例如：某个体有多种喜好，根据不同的喜好归可属于不同的群体(社区)中。近年来，一些学者在重叠社区挖掘方面作了一些初步的研究，如 Gergely Palla 提出的 k-clique[16]及 Shen Huawei 提出的 EAGLE[17]算法等。

考虑网络中的节点具有多样性的特点，可归属于不同的社区中，而提出了一种重叠社区发现的启发式算法。实验证明该算法可挖掘出同类性质的节点，将网络划分成若干社区；还可以可挖掘出具有多种特性的节点，它在社区间起到了“桥梁”的作用。

2　重叠社区发现算法

2.1　基本思想

社区具有内部节点连接较多而与外部节点连接较少的特点。在实际的网络中，往往出现一些与其他节点联系很多的节点(即网络中度很大的节点)，称为“中心节点”，其他节点常以“中心节点”为中心形成社区。提出了一种社区发现的启发式算法，以“中心节点”为初始社区，然后把对社区贡献最大的邻节点依次加入到社区，当全局贡献度达到极值时，便可形成一个社区；同时，若存在对多个社区贡献度都较大边界节点，则将这些边界节点加入到多个社区中。提取出社区后，社区的节点和边并不从网络中删除，方便挖掘社区间的边界节点。

2.2　基本概念

为了清晰说明提出的算法，首先定义几个相关概念。

(1) 局部贡献度 q[18]

$$q=\frac{L_{in}}{L_{in}+L_{out}} \tag{1}$$

其中，L_{in} 为与社区内部的连边数；L_{out} 为与社区外部的连边数；q 为节点对社区的贡献度。q 越大，和社区内部的连边数越多，加入的点对社区的贡献越大；反之越小。

(2) 全局贡献度 Q

采用全局贡献度 Q 来表示在社区挖掘过程中，当前的最大贡献度，利用此指标来判断社区结构是否达到最佳状态。

(3) 重叠系数

社区 C_i 和 C_j 的重叠系数 S 定义为：

$$S=\frac{|C_i\cap C_j|}{|C_i\cup C_j|} \tag{2}$$

其中，$|C_i\cap C_j|$为社区 C_i 和 C_j 的公共节点数目，$|C_i\cup C_j|$为社区 C_i 和 C_j 的所有节点数目。

2.3　社区发现算法

算法分成两个阶段，第一阶段为挖掘网络中的社区，第二阶段为对挖掘出的社区进行调整。

2.3.1　社区挖掘

步骤 1　计算网络中各节点的度，选择度最大的节点 i 作为初始社区 C_i，并将节点 i 做上标记；初始化全局贡献度 $Q=0$；

步骤 2　找出网络中所有与社区 C_i 相连的节点，并放入邻节点集 U 中；

步骤 3　对 U 中每个节点 j，根据公式(1)计算节点 j 对社区的贡献度 q_j。若贡献度最大的节点 j 的贡献度 $q_{max}=\max_{j\in U}\{q_j\}\geqslant Q$，则将节点 j 加入社区 C_i，将节点 j 做上标记，同时更新全局贡献度：$Q=q_{max}$，返回步骤 2 继续执行；否则，转步骤 4；

步骤 4　全局贡献度 Q 已达到极大值，得到社区 C_i。

步骤 5　若网络中已没有未作标记的节点，则网络中的所有社区已检测出来，过程结束；否则从未标记的节点中选择度最大的节点作为新的初始社区 C_i，返回步骤 2 继续执行。

2.3.2　社区调整

由于存在一些边界节点属于多个社区，当社区 C_i 和 C_j 的重叠系数达到阈值 T(取为 0.7)时，可以认为 C_i 和 C_j 联系很紧密，因此将社区 C_i 和 C_j 合并。社区调整具体流程如下：

步骤 1　计算任意两个社区 C_i 和 C_j 间的重叠系数 S；

步骤 2　当 S 大于阈值 T 时，将 C_i 和 C_j 合并成一个社区；

步骤 3　若任意两个社区的重叠系数都小于阈值 T，调整结束，否则，返回步骤 1 继续调整。

经过调整后，将网络最终划分成若干个社区，节点也可从属于多个社区。这样，更加符合自然界“物以类聚”以及“事物的多样性”的规律。整个社区挖掘算法流程如图 2 所示。

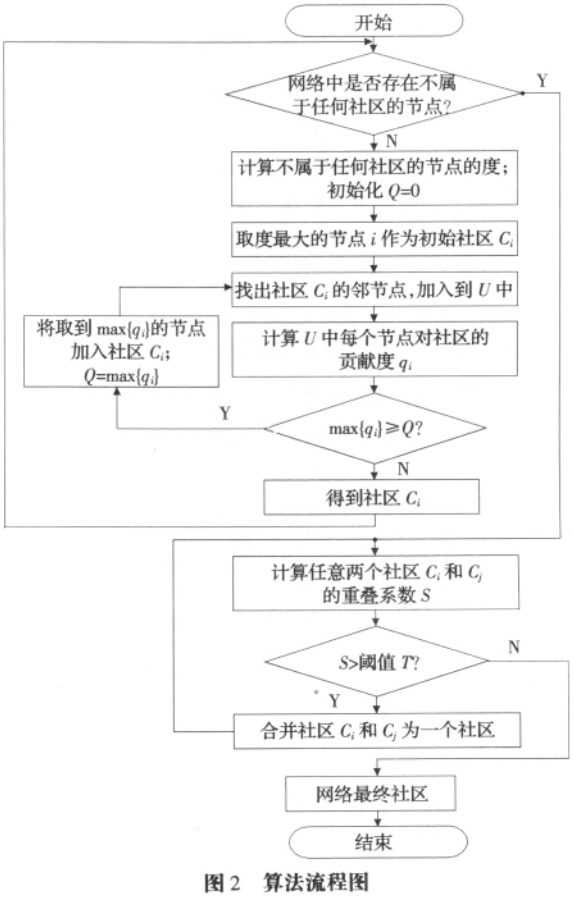

图 2　算法流程图

2.4　算法复杂性分析

在执行第一阶段时，步骤 2 和步骤 3 是为计算量最大，时间复杂度为 $O(n^2)$，n 为网络中的节点数；执行第二阶段时，最坏情况下所需的时间复杂度为 $O(n)$。由此，算法的时间复杂度为 $O(n^2)$。

3 实验结果及分析

利用两个经典的数据集：(1) 34 个节点 Zachary's Karate Club Network [19]；(2) 115 个节点 American College Football Network[20]，对该算法进行测试，结果表明，所划分出的社区具有较好的效果。

3.1 实验 1

经典数据集 Zachary's Karate Club Network[19]包含 34 个节点，具体网络如图 3 所示，不同圈标识不同的原始社区，该网络包括两个社区。

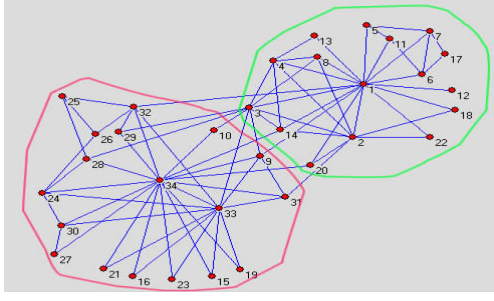

图 3 Zachary's Karate Club Network
(节点表示俱乐部的成员，连边表示成员间的关系)

利用提出的算法所划分出的社区如图 4 所示。从图 4 可以看出，所提出的算法将初始网络划分成 4 个社区，其中节点 9、10、31 同时归属于两个社区。算法所划分的社区和图 3 没有大的区别，只是分别对图 3 中的两个社区再细分一次，将节点集 (5,6,7,11,17) 和 (25,26,32,29) 抽取出来作为单独的社区，形成了图 4 所示的 4 个社区。事实上，节点集 (5,6,7,11,17) 内部联系较紧密，可抽出来作为单独的社区；节点集 (25,26,29,32) 在原始的社区中和其他节点联系不多，也可作为一个社区。并且节点 9、10 及 31 为两个社区的边界节点，对两个社区的贡献度相当，可归属于两个不同的社区。

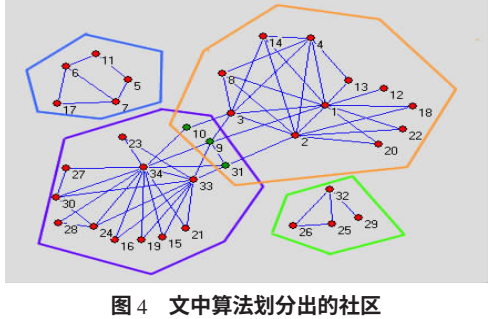

图 4 文中算法划分出的社区

在进行社区调整时，由于任意两社区的重叠系数小于阈值 T，故不需调整。

3.2 实验 2

经典数据集 American College Football[20]有 115 个节点，具体网络如图 5 所示。

利用提出的算法挖掘 American College Football[20]的社区，共挖掘出 10 个社区，结果如图 6 所示。

算法所划分的 10 个社区与实际 American College Football 网络的划分[20]绝大数都吻合，准确率达到 85%以上。实际的网络中，有一个社区内部连边数要比与社区间的连边数少，故在使用该算法时，会将该社区分散到其他社区中。

此外，还挖掘出了社区间的边界节点，图 6 中的边界节点所属社区如表 1 所示。

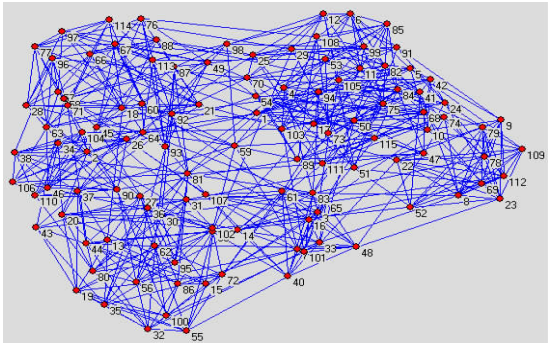

图 5 American College Football Network
(节点代表足球队，节点之间的连边表示两个球队间的常规赛)

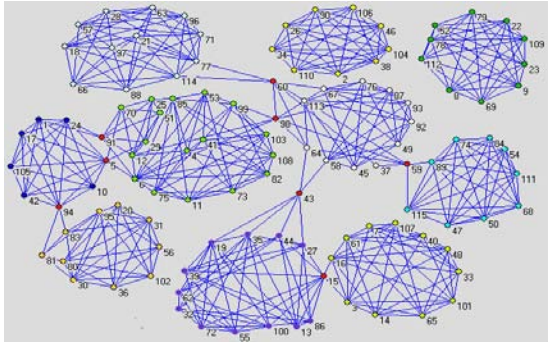

图 6 文中算法对 American College Football 网络所划分出的 10 个社区

表 1 边界节点及所属社区

边界节点	所属社区	
	社区 A	社区 B
5,91	节点 (1,5,10,17,24,42,91,94,105) 构成的社区	节点 (4,5,6,11,12,25,29,41,51,53,70 ,73,75,82,85 ,88,91,98,99,103,108) 构成的社区
94	节点 (1,5,10,17,24,42,91,94,105) 构成的社区	节点 (20,30,31,36,56,80,81,83,94,95,102) 构成的社区
60	节点 (18,21,28,57,60,63,66,71,77,88,96,97,114) 构成的社区	节点 (37,43,45,49,58,59,60,64,67,76,87,92,93,98,113) 构成的社区
98	节点 (37,43,45,49,58,59,60,64,67,76,87,92,93,98,113) 构成的社区	节点 (4,5,6,11,12,25,29,41,51,53,70 ,73,75,82,85,88,91,98 ,99,103,108) 构成的社区
43	节点 (37,43,45,49,58,59,60,64,67,76,87,92,93,98,113) 构成的社区	节点 (13,15,19,27,32,35,39,43,44,55,62,72,86,100) 构成的社区
15	节点 (13,15,19,27,32,35,39,43,44,55,62,72,86,100) 构成的社区	节点 (3,7,14,15,16,33,40,48,61,65,101,107) 构成的社区
59	节点 (37,43,45,49,58,59,60,64,67,76,87,92,93,98,113) 构成的社区	节点 (47,50,54,59,68,74,84,89,111,115) 构成的社区

4 结论

提出了一种基于局部信息的重叠社区发现算法，突破了传统算法中网络节点仅属于一个社区的思想。利用两个经典数据集对算法的有效性进行了验证，实验结果表明：该算法能较准确地找出网络中存在的社区，并且挖掘出了属于多个社区的边界节点，更加符合现实自然的规律。未来工作将致力于算法的边界节点的研究以及算法效率的进一步提高。

(下转 41 页)

$t_S(y_0,z)\geqslant\alpha_t, 1-f_S(y_0,z)\geqslant\alpha_f$,所以

$$t_{R\circ S}(x,z)=\max_{y\in Y}\{\min\{t_R(x,y),t_S(y_0,z)\}\}\geqslant\alpha_t$$

$$1-f_{R\circ S}(x,z)=1-\min_{y\in Y}\{\max\{f_R(x,y),f_S(y_0,z)\}\}\geqslant\alpha_f$$

即 $R\circ S$ 满足条件(M)。

设 R,S 满足条件(M),若有 $x\in X, z_1,z_2\in Z$ 且满足 $t_{R\circ S}(x,z)\geqslant\alpha_t, 1-f_{R\circ S}(x,z)\geqslant\alpha_f, t_{R\circ S}(x,z)\geqslant\alpha_t, 1-t_{R\circ S}(x,z)\geqslant\alpha_f$ 则由

$$t_{R\circ S}(x,z)=\max_{y\in Y}\{\min\{t_R(x,y),t_S(y,z)\}\}\geqslant\alpha_t$$

$$1-f_{R\circ S}(x,z)=1-\min_{y\in Y}\{\max\{f_R(x,y),f_S(y_0,z)\}\}\geqslant\alpha_f$$

$\exists y_1\in Y$,使 $t_R(x,y_1), t_S(y_1,z_1)\geqslant\alpha_t, 1-f_R(x,y_1), 1-f_S(y_1,z)\geqslant\alpha_f$,同理,$\exists y_2\in Y$,使 $t_R(x,y_2), t_S(y_2,z_2)\geqslant\alpha_t, 1-f_R(x,y_2), 1-f_S(y_2,z_2)\geqslant\alpha_f$,即

$$t_R(x,y_1)\geqslant\alpha_t, 1-f_R(x,y_1)\geqslant\alpha_f, t_R(x,y_2)\geqslant\alpha_t, 1-f_R(x,y_2)\geqslant\alpha_f$$

$$t_S(y_1,z_1)\geqslant\alpha_t, 1-f_S(y_1,z_1)\geqslant\alpha_f, t_S(y_2,z_2)\geqslant\alpha_t, 1-f_S(y_2,z_2)\geqslant\alpha_f$$

由于 R 满足条件(M),所以 $y_1=y_2$,从而 $t_S(y_1,z_1)\geqslant\alpha_t, 1-f_S(y_1,z_1)\geqslant\alpha_f$,再由 S 满足条件(M),即得 $z_1=z_2$。这说明 $R\circ S$ 也满足条件(M)。

推论 1 设 $R:X\to Y, S:Y\to Z$ 是 Vague 映射,则 $R\circ S:X\to Z$ 也是 Vague 映射,且

(1) 当 R,S 是 Vague 满射时,$R\circ S$ 也是 Vague 满射;

(2) 当 R,S 是 Vague 单射时,$R\circ S$ 也是 Vague 单射;

(3) 当 R,S 是 Vague 双射时,$R\circ S$ 也是 Vague 双射。

定理 3 设 R 是从 X 到 Y 的 Vague 关系,R^{-1} 是 R 的逆 Vague 关系,则

(1) R 满足条件(M$_1$)$\Leftrightarrow R^{-1}$ 满足条件(M$_1$);

(2) R 满足条件(M$_2$)$\Leftrightarrow R^{-1}$ 满足条件(M$_2$)。

证 仅证(1)。若 R 满足条件(M),则对任意的 $x\in X$,存在 $y\in Y$,使 $t_R(x,y)\geqslant\alpha_t, 1-f_R(x,y)\geqslant\alpha_f$,于是 $t_{R^{-1}}(y,x)=t_R(x,y)\geqslant\alpha_t, 1-f_{R^{-1}}(y,x)=1-f_R(x,y)\geqslant\alpha_f$,即 R^{-1} 满足条件(M$_1$)。

反过来,若 R^{-1} 满足条件(M$_1$),则对任意的 $x\in X$,存在 $y\in Y$,使 $t_{R^{-1}}(y,x)=t_R(x,y)\geqslant\alpha_t, 1-f_{R^{-1}}(y,x)=1-f_R(x,y)\geqslant\alpha_f$,从而 $t_R(x,y)=t_{R^{-1}}(y,x)\geqslant\alpha_t, 1-f_R(x,y)=1-f_{R^{-1}}(y,x)\geqslant\alpha_f$,即 R 也满足条件(M)。

4 结论

Zadeh 模糊集理论及应用,特别是在知识处理中的应用,虽然也在进一步发展,但已渐趋成熟,而 Vague 集理论在用作知识处理时,尚有许多方面处于探讨阶段。由于这一理论正在发展之中,且其数学描述较之 Zadeh 模糊集理论更加符合客观世界模糊对象的本质,是求解不确定性问题、处理不完全信息、进行不精确推理的更为有效的数学方法,因而形成新的研究热点。在已有的 Vague 关系及其合成运算定义的基础上,进一步研究了它们的一些特殊性质。最后定义了 Vague 映射,并给出了相应的定理,是对 Vague 集理论的有益补充,希望能成为 Vague 理论的发展的一个工具。

参考文献:

[1] Pawlak Z.Rough Sets[J].Bull Polish Acacl Sci Tech,1982:591-596.

[2] 何新贵.模糊知识处理的理论与技术[M].2 版.北京:国防工业出版社,1999.

[3] Gau Wen-Lung,Daniel J.Vague sets[J].IEEE Transactions on Systems,Man,and Cybemetics,1993,23(2):610-614.

[4] 李凡,卢安,饶勇.Vague 集的运算规则[J].计算机科学,2000,27(9):15-17.

[5] 梁家荣.Vague 关系[J].计算机工程与应用,2005,41(30):10-12.

(上接 38 页)

参考文献:

[1] Wasserman S,Faust K.Social network analysis [M].London:Cambridge University Press,1994.

[2] Watts D J,Strogatz S H.Collective dynamics of 'small-world' networks[J].Nature,1998,393:440-442.

[3] Amaral L A N,Scala A,Barthélémy M,et al.Classes of small-world networks[C]//Proceedings of the National Academy of Sciences,2000,97:11149-11152.

[4] Newman M E J.The structure of scientic collaboration networks[C]//Proceedings of the National Academy of Sciences,2001,98(2):404-409.

[5] Faloutsos M,Faloutsos P,Faloutsos C.On power-law relationships of the internet topology[J].Computer Communications Review,1999,29:251-262.

[6] Albert R,Jeong H,Barabasi A L.Diameter of the world-wide web[J].Nature,1999,401(6749):130-131.

[7] Broder A,Kumar R,Maghoul F,et al.Graph structure in the web[J].Computer Networks,2000,33:309-320.

[8] Williams R J,Martinez N D.Simple rules yield complex food webs [J].Nature,2000,404(6774):180-183.

[9] Jeong H,Tombor B,Albert R,et al.The large-scale organization of metabolic networks[J].Nature,2000,407(6804):651-654.

[10] Fell D A,Wagner A.The small world of metabolism [J].Nature Biotechnology,2000,18:1121-1122.

[11] Kernighan B W,Lin S,An efficient heuristic procedure for partitioning graphs[J].Bell System Technical Journal,1970,49:291-307.

[12] Fiedler M.Algebraic connectivity of graphs[J].Czech Math J,1973,23:298-305.

[13] Girvan M,Newman M E J.Community structure in social and biological networks[C]//Proceedings of the National Academy of Sciences,2002,99:7821-7826.

[14] Newman M E J,Girvan M.Finding and evaluating community structure in networks[J].Physical Review E,2004,69.

[15] Radicchi F,Castellano C,Cecconi F,et al.Defining and identifying communities in networks[C]//Proceedings of the National Academy of Sciences,2004,101(9):2658-2663.

[16] 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.

[17] Shen Hua-Wei,Cheng Xue-Qi,Cai-Kai,et al.Detect overlapping and hierarchical community structure in networks[Z]//arXiv,2008.

[18] Clauset A.Finding local community structure in networks[J].Physical Review E,2005,72.

[19] Zachary W W.An information flow model for conflict and fission in small groups[J].Journal of Anthropological Research,1977,33(4):452-473.

[20] Girvan M,Newman M E J.Community structure in social and biological networks[C]//Proceedings of the National Academy of Sciences,2002,99(12):7821-7826.