ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, Vol.20, No.8, August 2009, pp.2241−2254
doi: 10.3724/SP.J.1001.2009.03318

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 一种基于拓扑势的网络社区发现方法*

淦文燕[1], 赫 南[2+], 李德毅[3], 王建民[1]

[1](清华大学 软件学院,北京 100084)

[2](北京航空航天大学 计算机科学与技术系,北京 100191)

[3](电子系统工程研究所,北京 100039)

## Community Discovery Method in Networks Based on Topological Potential

GAN Wen-Yan[1], HE Nan[2+], LI De-Yi[3], WANG Jian-Min[1]

[1](School of Software, Tsinghua University, Beijing 100084, China)

[2](Department of Computer Science, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

[3](Institute of Electronic System Engineering, Beijing 100039, China)

+ Corresponding author: E-mail: NanHe.2006@gmail.com

**Gan WY, He N, Li DY, Wang JM. Community discovery method in networks based on topological potential. *Journal of Software*, 2009,20(8):2241−2254.** http://www.jos.org.cn/1000-9825/3318.htm

**Abstract**: Inspired from the idea of data fields, a community discovery algorithm based on topological potential is proposed. The basic idea is that a topological potential function is introduced to analytically model the virtual interaction among all nodes in a network and, by regarding each community as a local high potential area, the community structure in the network can be uncovered by detecting all local high potential areas margined by low potential nodes. The experiments on some real-world networks show that the algorithm requires no input parameters and can discover the intrinsic or even overlapping community structure in networks. The time complexity of the algorithm is $O(m+n^{3/\gamma})\sim O(n^2)$, where $n$ is the number of nodes to be explored, $m$ is the number of edges, and $2<\gamma<3$ is a constant.

**Key words**: topological potential; data field; community discovery; complex network

**摘 要**: 从数据场思想出发,提出了一种基于拓扑势的社区发现算法.该方法引入拓扑势描述网络节点间的相互作用,将每个社区视为拓扑势场的局部高势区,通过寻找被低势区域所分割的连通高势区域实现网络的社区划分.理论分析与实验结果表明,该方法无须用户指定社区个数等算法参数,能够揭示网络内在的社区结构及社区间具有不确定性的重叠节点现象.算法的时间复杂度为 $O(m+n^{3/\gamma})\sim O(n^2)$,$n$ 为网络节点数,$m$ 为边数,$2<\gamma<3$ 为一个常数.

**关键词**: 拓扑势;数据场;社区发现;复杂网络

**中图法分类号**: TP393 **文献标识码**: A

结构决定功能是系统科学的基本观点.如果将系统内部的各个元素抽象为节点,元素之间的关系视为连接,

* Supported by the National Natural Science Foundation of China under Grant No.60675032 (国家自然科学基金); the National Basic Research Program of China under Grant Nos.2007CB310800, 2007CB311003 (国家重点基础研究发展计划(973))

Received 2007-07-28; Revised 2007-12-17; Accepted 2008-03-14

系统就构成一个具有复杂连接关系的网络.现实世界中复杂网络无处不在,如技术系统中的因特网、电力网、交通网,社会系统中的人际关系网、合作网、引文网,以及生物系统中的神经网、新陈代谢网、蛋白质相互作用网等等.复杂网络的研究表明[1−6],这些看似毫不相干的、形态各异的真实网络具有某些相同的拓扑性质,受制于某些基本的演化法则.

1998 年,Watts 和 Strogatz 在《Nature》上发表论文阐述实际复杂网络的小世界效应(small-world effect)[1],即节点间具有较小的平均最短路径长度,对数依赖于网络的规模.1999 年,Barabási 和 Albert 在《Science》上发表论文指出,许多真实网络的度分布遵循幂律分布,称为无标度网络(scale-free network)[2,3].而大量实证研究表明,真实网络不仅具有小世界和无标度等特性,还呈现明显的社区结构(community structure)[7−11].

所谓社区(community)[7−11],是指网络中的节点内聚子图,子图内部的节点间存在较多的连接,不同子图的节点间连接相对稀少,如图 1 所示.

社区结构是网络模块化与异质性的反映,表示真实网络可以看作是由许多不同类型节点组合形成的,如人际关系网络中的朋友圈子、引文网络中针对同一主题的相关论文、新陈代谢或蛋白质网络中的功能子团等等.深入研究网络的社区结构不仅有助于揭示错综复杂的真实网络是怎样由许多相对独立而又互相关联的社区形成的,使人们更好地理解系统不同层次的结构和功能特性,而且具有重要的实用价值.例如,社会网中的社区可用于揭示具有共同兴趣、爱好或社会背景的社会团体;蛋白质网络中的社区结构可用于发现生物系统中功能相关的结构单元;万维网中的社区结构可用于提高网络搜索的性能和准确性,实现信息过滤、热点话题跟踪和网络情报分析等.因此,社区发现(community detection or discovery)已成为复杂网络领域中的一个非常重要的研究方向.

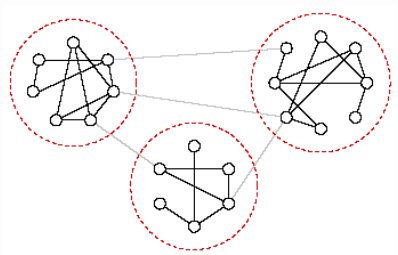


Fig.1 A simple network with community structure

图 1 一个具有社区结构的简单网络

## 1 相关工作

网络中的社区发现最早源于社会学的研究工作,相关理论涉及图论、模式识别等[12,13].传统的图分割方法总是假设网络是可分解的,待划分的子图个数由用户指定.但对大规模的真实网络进行社区划分时,社区发现方法必须解答两个方面的问题:

① 网络中是否包含社区结构.

② 如何有效发现网络内在的社区结构.

迄今为止,人们已经提出许多社区发现方法,其基本思想大都是根据某个节点内聚性度量,递归地对网络进行合并或分裂,把网络分解为嵌套的社区层次结构.典型的代表方法包括基于边介数(edge betweenness)的社区发现方法[8]和模块度优化方法(modularity optimization)[9]等.

根据社区的定义,社区间的少数连接将成为社区间通信时通信流量的必经之路,如果考虑网络中某种形式的通信并寻找到具有最高通信流量的边,则去除该边将获得网络最自然的分割.由此,Girvan 和 Newman 等人引入边介数度量网络的通信流量,提出基于边介数的社区发现算法,简称 GN 算法[8].其基本思想是,迭代计算网络中每条边的介数并去除介数最大的边,直至网络中所有边被去除,每个节点自成一个社区.

与图分割方法不同,GN 算法无须指定社区个数,可以将网络分解成任意数量的社区,但无法确定最优的社

区结构.事实上,即使明显不存在社区结构的随机网络,GN 算法也会产生强制的社区划分.此外,GN 算法的时间复杂度较高,为 $O(nm^2)$,$n$ 为网络节点数,$m$ 为边数.

针对 GN 算法的强制社区划分问题,Newman 等人引入模块度(modularity)[9]来评价社区分解的合理性,其基本思想是,一个好的社区划分内部节点连接概率应远大于具有同样度序列的随机图中内部节点的连接概率.由于模块度的定义独立于特定的社区发现算法,社区发现问题可以简化为模块度优化问题.考虑到网络可能的划分方案数与网络规模成指数关系,对大规模网络来说,穷尽搜索肯定不可行,人们引入各种启发式的模块度优化方法[14−20],如贪婪算法[14,15]、极值优化方法[16]、模拟退火[17]等.

目前,模块度优化方法已成为复杂网络社区发现的基准方法,并得到广泛的应用.然而有研究表明[21,22],模块度定义存在内在的分辨率限制,倾向于发现规模相似的社区结构.考虑到真实网络的异质性,其内在社区的大小可能很不均匀.Arenas 等人[23,24]通过对 E-mail 网络、Jazz 音乐家合作网以及科学家合作网等进行社区分析发现,这些真实网络的社区大小近似服从幂律分布,即网络中只存在极少数的大社区,绝大多数社区只包含很少的节点.显然,对于普遍具有无标度特性的真实网络来说,模块度优化方法很难保证发现真正最优意义上的社区结构.

从数据场的思想出发,本文提出一种基于数据场的社区发现算法.该方法引入拓扑势场描述网络节点间的相互作用,将每个社区视为拓扑势场的局部高势区,通过寻找被低势区域所分割的连通高势区域实现网络的社区划分.真实网络的测试结果显示,该方法无须用户指定社区个数等算法参数,能够有效揭示网络内在的社区结构及社区间具有不确定性的重叠节点现象,具有相对较好的算法性能.

本文第 2 节介绍节点拓扑势的基本思想及其形式化描述.第 3 节给出基于拓扑势的网络社区发现方法.第 4 节给出具体算法描述及性能分析.第 5 节对算法参数进行讨论,并给出参数优选方法.第 6 节给出真实网络数据的测试与比较结果.第 7 节对全文进行总结.

## 2 拓扑势的引入

场的概念最早是 1837 年由英国物理学家法拉第提出的,用于描述物质粒子间的非接触相互作用.势场是基础物理学中讨论得最多的物理场,其主要特征是,对应描述场的标量势函数,一定存在定义在空间上的矢量强度函数,二者可以通过微分算子∇相互联结.此外,空间任意一点的势值与某个代表场源强度的参量(如质点的质量或点电荷的电量等)成正比,与该点到场源间的距离成递减关系.对重力势场和静电势场来说,势值的大小与距离成反比,在距离场源很远的地方仍然存在着场力的作用,代表长程场.而对核子的中心势场来说,势值随着距离的增长急剧下降,相应力场很快衰减为 0,代表短程场.

受上述物理场思想的启发,我们将网络 $G$ 看作一个包含 $n$ 个节点及其相互作用的物理系统,每个节点周围存在一个作用场,位于场中的任何节点都将受到其他节点的联合作用.根据真实网络的模块化与抱团特性,我们认为,节点间相互作用具有局域特性,每个节点的影响能力会随网络距离的增长而快速衰减.根据数据场的相关讨论[25,26],我们倾向于采用代表短程场且具有良好数学性质的高斯势函数描述节点间的相互作用,并称相应的场为拓扑势场.

给定网络 $G=(V,E)$,其中,$V=\{v_1,\ldots,v_n\}$ 为节点的非空有限集,$E\subseteq V\times V$ 为节点偶对或边的集合,$|E|=m$.根据数据场的势函数定义[25],任意节点 $v_i\in V$ 的拓扑势可表示为

$$\varphi(v_i)=\sum_{j=1}^{n}\left(m_j\times \mathrm{e}^{-\left(\frac{d_{ij}}{\sigma}\right)^2}\right) \tag{1}$$

其中,$d_{ij}$ 表示节点 $v_i$ 与 $v_j$ 间的网络距离,本文采用最短路径长度来度量;影响因子 $\sigma$ 用于控制每个节点的影响范围;$m_i\geq 0$ 表示节点 $v_i(i=1,\ldots,n)$ 的质量,可以用来描述每个节点的固有属性.在真实网络中,节点的固有属性有丰富的物理含义,如城市交通网中城市的规模、人际关系网中个体的社会背景与活动能力、通信网络中节点的存储能力等等.本文中,我们忽略节点固有属性的差异性,假设每个节点的质量相等且满足归一化条件,由此得到简

化的拓扑势公式:

$$\varphi(v_i)=\frac{1}{n}\sum_{j=1}^{n}\mathrm{e}^{-\left(\frac{d_{ij}}{\sigma}\right)^2} \tag{2}$$

根据高斯函数的数学性质,对于给定的$\sigma$值,每个节点的影响范围近似为$\lfloor 3\sigma/\sqrt{2}\rfloor$跳的局域区域,当距离大于$\lfloor 3\sigma/\sqrt{2}\rfloor$跳时,单位势函数很快衰减为 0,指示着短程场作用.分析给定$\sigma$值的拓扑势函数,处于网络连接密集区域的节点具有较高的拓扑势,而拓扑势值最大的节点附近连接也最密集,即拓扑势可用作网络局域连接密集程度的一种估计.

参照前面所给的社区定义(即社区内部节点间连接密集,社区之间的连接相对稀疏),每个社区对应拓扑势场的一个局部高势区,靠近社区中心位置的节点由于相互连接紧密具有较大的拓扑势,可以视为社区的中心或代表节点,而位于社区边界的节点由于连接稀疏而具有相对较小的拓扑势.由此,通过寻找拓扑势场中被低势区域所分割的连通高势区域可实现网络社区的自然划分.

## 3 基于拓扑势的社区发现方法

给定网络 $G=(V,E)$及其对应某个$\sigma$值的拓扑势场,其中,$V=\{v_1,\ldots,v_n\}$为节点集,$E\subseteq V\times V$ 为边集,$|E|=m$,基于拓扑势的社区发现方法可形式化描述如下:

1) 拓扑势吸引(topology-potential attraction):已知局部极大势值节点 $v^*$,$\forall v\in V$,如果存在节点集$\{v_0,v_1,\ldots,v_k\}\subseteq V$,使得 $v_0=v$,$v_k=v^*$且 $v_i$ 位于 $v_{i-1}$ 的势值上升方向,$0<i<k$,则称 $v$ 被 $v^*$拓扑势吸引.
2) 单代表点社区(one-representative community):已知局部极大势值节点 $v^*$,如果存在子集 $C\subseteq V$,使得$\forall v\in C$ 都被 $v^*$拓扑势吸引,则称 $C$ 为以 $v^*$为代表点的社区.
3) 多代表点社区(multi-representatives community):已知局部极大势值节点集合 $A\subset V$,如果存在子集 $C\subseteq V$,使得以下条件成立,则称 $C$ 为代表点集合为 $A$ 的多代表社区:a) $\forall v\in C,\exists v^*\in A$,使得 $v$ 被 $v^*$拓扑势吸引;$\forall v^*\in A,\exists w^*\in A$ 且 $w^*\neq v^*$,使得 $v^*$与 $w^*$的距离 $d(v^*,w^*)<\lfloor 3\sigma/\sqrt{2}\rfloor$.

分析上述定义,给定网络及其拓扑势场,通过遍历每个节点的最大势值上升方向可以搜索所有的局部极大势值节点,从而确定社区划分的个数及每个社区的代表节点.其中,单代表点社区中只存在 1 个局部极大势值节点,而多代表点社区中存在多个相距很近的局部极大势值节点,形成势场分布的山脊或高原.考虑到对真实复杂网络进行社区划分时,较为宏观的社区划分更有助于人们理解网络系统的整体结构和功能特性,我们将距离小于影响范围$\lfloor 3\sigma/\sqrt{2}\rfloor$的邻近的局部极大势值节点合并为一个多代表点集合.

根据社区代表点进行网络划分时,$\forall v\in V$,如果 $v$ 不是局部极大势值节点,则 $v$ 或者唯一地被某个社区的代表点吸引,或者被多个社区的代表点同时吸引.若 $v$ 唯一地被某个社区 $C\subseteq V$ 的代表点吸引,则称 $v$ 为社区 $C$ 的内部节点或私有节点;否则,称为边界节点.

确定边界节点的所属社区本质上是迭代扩展社区内部节点的过程.具体来说,假设网络 $G$ 可以划分为 $t$ 个社区 $C_1,\ldots,C_t$,令$(a_{ij})_{n\times n}$为网络邻接矩阵,对任意边界节点 $v_i\in V$,可以引入如下的效益函数(公式(3))来评价其划分方案,并将其划分给具有最大效益的邻近社区.如果多个划分方案的效益相等,则 $v_i$ 被视为一个重叠节点,即该节点的划分具有不确定性.

$$Q_{s=1,\ldots,t}(v_i)=\sum_{v_j\in V_s}a_{ij}-\sum_{v_k\notin V_s}a_{ik} \tag{3}$$

## 4 社区发现算法描述

**算法 1**. 基于拓扑势的社区发现算法(a community discovering algorithm based on topological potential).

输入:网络 $G=(V,E)$和影响因子$\sigma$,其中,$V=\{v_1,\ldots,v_n\}$,$|E|=m$.

输出:社区 $C_1,\ldots,C_t$.

算法步骤:

(1) $[\varphi(v_1),\ldots,\varphi(v_n)]$=*Cal_TopologicalPotential*$(G,\sigma)$; //计算节点拓扑势

(2) $V_{rep}$=*Searching_*Max*PotentialNodes*$(G,\varphi(v_1),\ldots,\varphi(v_n))$; //采用最大势值上升方向指引的爬山法搜索局部极大势值节点,确定社区代表点集合

(3) $[C_1,\ldots,C_t]$=*Community_Detecting*$(G,V_{rep},\varphi(v_1),\ldots,\varphi(v_n))$; //根据社区代表点形成网络的社区划分

(4) 输出社区 $C_1,\ldots,C_t$.

分析上述算法步骤,步骤(1)根据给定$\sigma$值计算网络节点的拓扑势.如果考虑网络中所有节点的影响,则计算拓扑势的时间复杂度取决于计算所有节点对之间最短路径长度的时间复杂度,至少为 $O(nm)$,稀疏图情况下为 $O(n^2)$.根据高斯函数的数学性质,每个节点的影响范围为 $l=\lfloor 3\sigma/\sqrt{2}\rfloor$ 跳以内的邻近节点,节点拓扑势计算公式可简化为

$$\varphi(v)=\frac{1}{n}\sum_{j=1}^{l}n_j(v)\times \mathrm{e}^{-\left(\frac{j}{\sigma}\right)^2},\forall v\in V \tag{4}$$

其中,$n_j(v)$为节点 $v$ 的 $j$ 跳邻居节点数.此时,计算所有节点拓扑势的开销为 $\sum_{v\in V}\sum_{j=1}^{l}n_j(v)$ .当 $l$=1 时,其时间复杂度为 $O(m)$;当 $l$=2 时,其时间复杂度近似为 $O(m+n^{3/\gamma})$[27],$2<\gamma<3$ 为一个常数;随着 $l$ 的增大,当 $l$ 趋近于网络的平均最短路径长度时,$\sum_{j=1}^{l}n_j(v)$ 趋近于 $n$,总的拓扑势计算开销也趋近于 $O(n^2)$.

步骤(2)通过遍历每个节点的最大势值上升方向搜索所有的局部极大势值节点.爬山过程会对每个已遍历节点进行标记,当遇到某个已遍历节点时,表明此次爬山指向一个已发现的局部极大势值节点,则随机选择一个新的节点开始爬山.该步骤的时间复杂度为 $O(m)$.

步骤(3)根据代表点集合 $V_{rep}$ 实现网络的社区划分.令代表点个数为 $n_{rep}<<n$,根据势值大小自上而下遍历每一个待划分节点,确定其所属社区的时间复杂度为 $O(n_{rep}\times n)\sim O(n)$.因此,基于拓扑势的社区发现算法总的时间复杂度取决于节点拓扑势的计算开销,最坏情况下为 $O(n^2)$,最好情况下为 $O(m+n^{3/\gamma})$.

## 5 算法参数的讨论

基于拓扑势的社区发现算法只有 1 个算法参数,即影响因子$\sigma$.根据数据场中关于影响因子$\sigma$的讨论[25,26],可引入拓扑势场的势熵来衡量$\sigma$值的合理性.

给定网络 $G=(V,E)$及其对应某个$\sigma$值的拓扑势场,其中,$V=\{v_1,...,v_n\}$,$|E|=m$,令 $v_1,\ldots,v_n$ 的势值为 $\varphi(v_1),\ldots,\varphi(v_n)$,相应拓扑势场的势熵可定义为

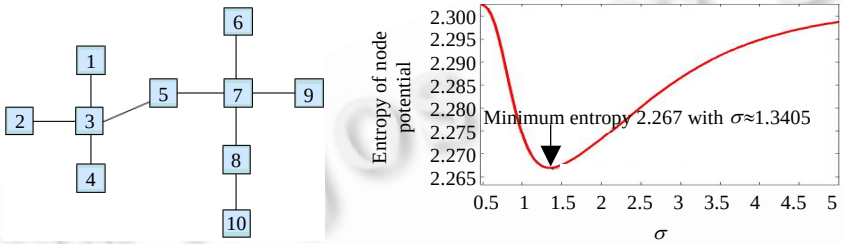
$$H=-\sum_{i=1}^{n}\frac{\varphi(v_i)}{Z}\log\left(\frac{\varphi(v_i)}{Z}\right) \tag{5}$$

其中,$Z=\sum_{i=1}^{n}\varphi(v_i)$ 为标准化因子.考虑如图 2(a)所示的具体网络,计算不同$\sigma$值所对应的拓扑势熵,可得到图 2(b)所示的关系曲线.当$\sigma$趋近于 0 时,势熵趋近最大值 log(10)≈2.3025;随着$\sigma$值的递增,势熵开始减小,在某个优化$\sigma$处($\sigma$≈1.3405)达到最小值 2.267,然后又逐渐增大,当$\sigma$值大于网络直径 5 时,再次趋于最大值.

优化$\sigma$是一个单变量非线性函数 $H(\sigma)$的最小化问题,求解此类问题有很多标准算法,如简单试探法、随机搜索法、模拟退火法等.考虑到迭代计算节点拓扑势的时间开销较大,实际求解时,可先近似估计$\sigma$的优化区间,再精确搜索其优化值.具体来说,根据高斯势函数的性质,每个节点的影响范围近似为 $\lfloor 3\sigma/\sqrt{2}\rfloor$ 跳的局域区域.当 $0<\sigma<\sqrt{2}/3$ 时,节点间没有相互作用,每个节点的势值为 1/$n$,势熵取得最大值 log($n$);随着$\sigma$的增大,当 $\sqrt{2}/3\le\sigma<2\sqrt{2}/3$ 时,每个节点只影响其邻居节点,其拓扑势与节点度(即节点的连接边数)相差一个比例常数,

即 $\varphi(v)\approx\frac{\deg(v)}{n}\times e^{-\left(\frac{1}{\sigma}\right)^2}$;当 $2\sqrt{2}/3\le\sigma<\sqrt{2}$ 时,每个节点影响 2 跳以内的邻居节点,拓扑势可近似计算为 $\varphi(v)\approx\frac{\deg(v)}{n}\times e^{-\left(\frac{1}{\sigma}\right)^2}+\frac{n_2(v)}{n}\times e^{-\left(\frac{2}{\sigma}\right)^2}$, $n_2(v)$为节点的 2 跳邻居数;依此类推,当 $\sigma\ge\sqrt{2}D/3$ 时,$D$ 为网络直径,相距最远的节点对之间都存在相互作用,计算任意一点的拓扑势要考虑所有节点的影响.

由此,令$\sigma$分别取离散值 $\sqrt{2}/3,2\sqrt{2}/3,...$, 迭代计算节点拓扑势和相应的拓扑势熵.根据前面的讨论,对应上述$\sigma$取值,拓扑势熵应先递减然后开始递增,即存在一个极小值.令极小势熵对应的$\sigma$值为 $\sqrt{2}p/3$, $p\in\mathbf{N}$(其中 **N** 为自然数集合),进一步采用以 $(\sqrt{2}(p-1)/3,\sqrt{2}(p+1)/3)$ 为初始搜索区间的简单试探法等搜索满足精度要求的最小拓扑势熵及其对应的优化$\sigma$值.




(a) A simple network
(a) 一个简单网络拓扑

(b) Plot of potential entropy $H$ vs. influence factor $\sigma$
(b) 拓扑势熵 $H$ 与影响因子$\sigma$的关系曲线

Fig.2 Optimal choice of influence factor$\sigma$
图 2 影响因子$\sigma$的优选

分析优化$\sigma$的时间复杂度,确定$\sigma$优化区间的时间复杂度取决于平均 $p$ 跳内邻居节点数,最好情况下,算法只搜索 3 跳邻居即可确定$\sigma$优化区间,时间复杂度为 $O(m+n^{3/\gamma})$;最坏情况下,算法须搜索每个节点的$\langle l\rangle$跳邻居,$\langle l\rangle$等于网络的平均最短路径长度.由于连通图中任意节点的平均 $t$=1,…,$\langle l\rangle$跳邻居数的累加和为 $n$−1,时间复杂度为 $O(n\times(n-1))\sim O(n^2)$.一旦确定$\sigma$的优化区间,搜索满足精度要求的最小势熵就只涉及已知的 $t$=1,…,$p$ 跳邻居数,时间复杂度为 $O(n\times s)$,$s$ 为迭代次数.因此,优化算法总的时间复杂度取决于确定$\sigma$优化区间的时间复杂度,最好情况下为 $O(m+n^{3/\gamma})$,最坏情况下为 $O(n^2)$.

## 6 实验结果与比较

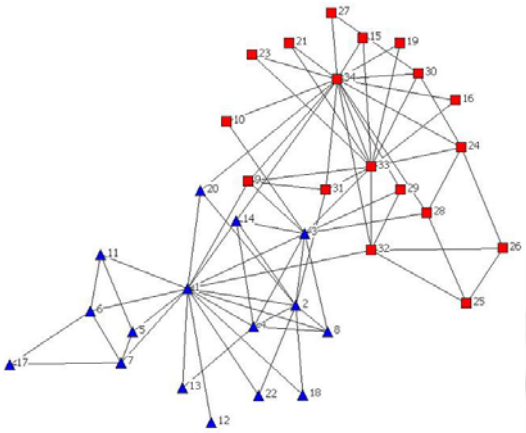
这里,我们采用 3 个真实的网络数据来检验基于拓扑势的社区发现方法的有效性.比较方法采用 GN 算法和模块度优化方法等.

### 6.1 Zachary社会关系网

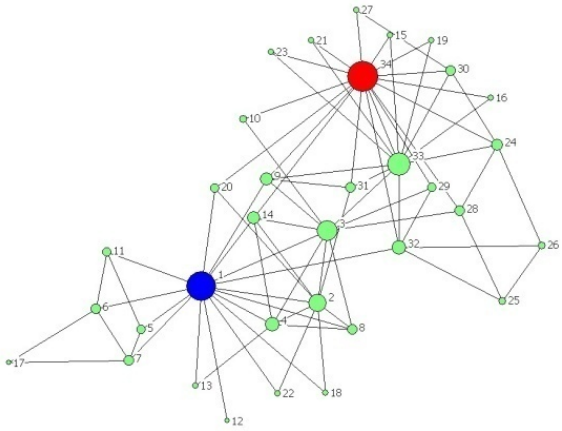
Zachary 社会关系网是复杂网络与社会网分析领域中常用的一个小型测试网络[28].20 世纪 70 年代,Wayne Zachary 用 3 年时间(1970 年~1972 年)观察美国一所大学空手道俱乐部成员间的社会关系,并构造如图 3(a)所示的俱乐部成员社会关系网(Zachary's karate club network).网络包含 34 个节点、78 条边.每个节点表示一个俱乐部成员,节点间的连接表示两个成员经常一起出现在俱乐部活动(如空手道训练、俱乐部聚会等)之外的其他场合,即在俱乐部之外他们可以被称为朋友.调查过程中,该俱乐部因为主管 John A.($v_{34}$)与教练 Mr. Hi($v_1$)之间的争执而分裂成两个各自以他们为核心的小俱乐部,图中不同形状的节点代表分裂后的小俱乐部成员.该网络作为一个真实的小型社会关系网,常常被用于测试社区发现方法的有效性[8−10,14,16,29−35].

采用基于拓扑势的社区发现算法对 Zachary 网络进行社区划分.令$\sigma$取优化值 1.020 3,相应的拓扑势场分布如图 3(b)所示(图中节点的大小与其拓扑势值成正比),显然存在两个局部极大值节点,分别对应真实社区结构的

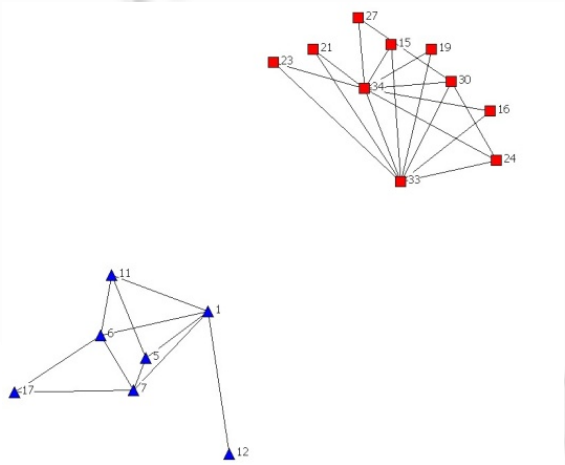
两个核心节点 $v_1$,$v_{34}$;以局部极大势值节点为社区代表节点,初始的社区划分如图 3(c)所示,迭代划分 17 个边界节点后所得的社区结构如图 3(d)所示,其中,$v_{10}$ 与两个社区的连接一样多,可以看作重叠节点.根据邻居节点的拓扑势值,我们将 $v_{10}$ 划分给最大势邻居 $v_{34}$ 所在的社区.显然,基于拓扑势的社区发现方法所得到的社区结构与 Zachary 网络内在的社区结构是相符的.



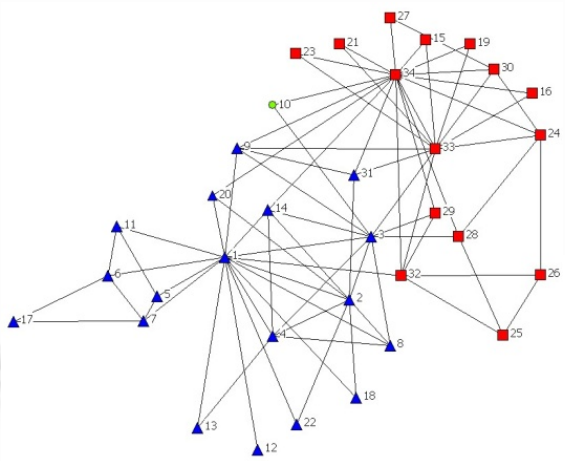
(a) Real communities of the network
(a) 真实社区结构



(b) Distribution of topological potential over nodes ($\sigma$=1.0203)
(b) 节点的拓扑势分布($\sigma$=1.0203)



(c) Initial partition of network
(c) 初始社区划分

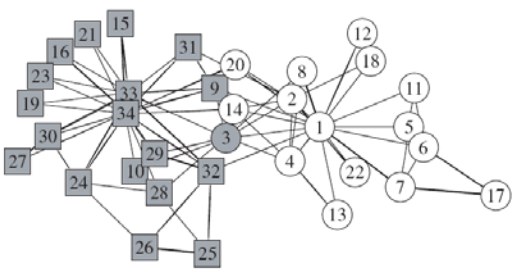


(d) Final communities discovered
(d) 最终的社区划分结果

Fig.3 Detecting community structure for the Zachary's network by the community discovery algorithm based on topological potential

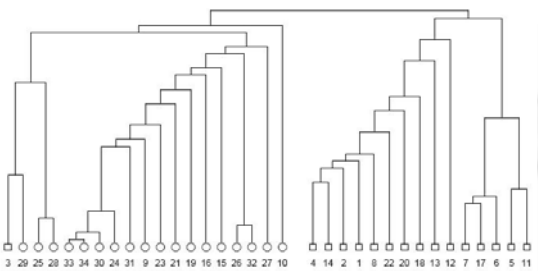
图 3 采用基于拓扑势的社区发现方法分析 Zachary 网络的社区结构

采用 GN 算法[8]、模块度优化方法[9]和谱二分法[35]分析 Zachary 网络的社区结构,划分结果如图 4 所示.其中,节点 $v_3$ 都被误分,还有一些其他方法也产生了类似的划分结果[31,33,34].根据图 4(c),模块度优化方法产生的最优划分对应 5 个小社区.与真实网络内在的社区结构相比,该方法倾向于产生粒度较小的社区结构,而且节点 $v_3$ 也被误分.与上述方法相比,基于拓扑势的社区发现方法可以有效揭示 Zachary 网络的内在社区结构且无须用户指定社区个数等算法参数.



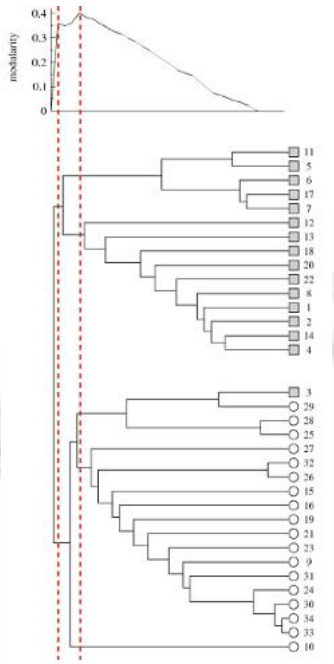
(a) Communities discovered by the spectral partitioning method

(a) 谱二分法的划分结果



(b) Communities discovered by GN algorithm

(b) GN 算法的划分结果



(c) Communities discovered by the modularity optimization method

(c) 模块度优化方法的划分结果

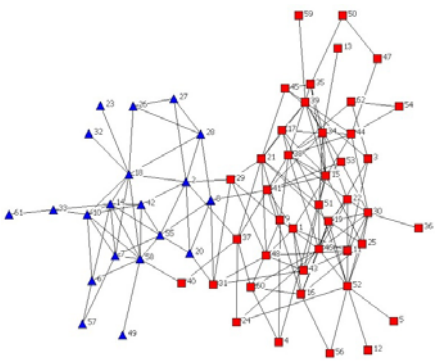
Fig.4 Detecting community structure for the Zachary's network by other algorithms

图 4 采用其他社区发现方法分析 Zachary 网络的社区结构

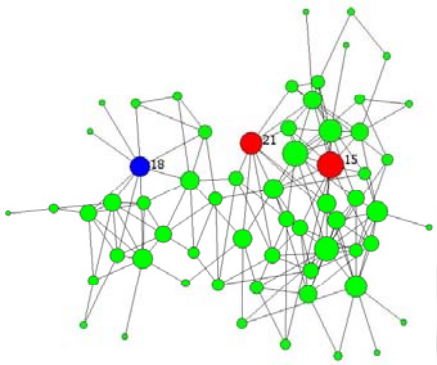
### 6.2 海豚关系网络

海豚关系网也是社会网分析中常用的一个真实网络[36,37].Lusseau 等人对栖息在新西兰 Doubtful Sound 峡湾的一个宽吻海豚群体(该群体由 2 个家族共 62 只宽吻海豚组成)进行长达 7 年的观察,并构造如图 5(a)所示的海豚关系网.图中节点代表一个海豚,边表示两个海豚之间接触频繁,共有 62 个节点、159 条边.图中不同形状的节点代表属于不同家族的海豚成员,较大的海豚家族包含 42 个成员节点,而较小的家族仅包含 20 个节点.

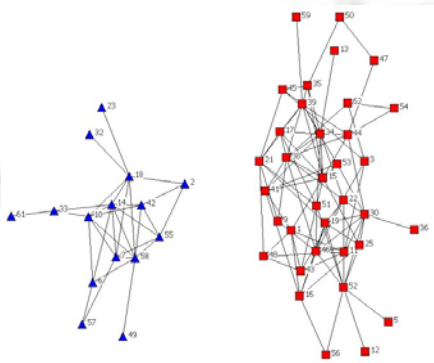
采用基于拓扑势的社区发现方法分析海豚关系网的社区结构,令$\sigma$取优化值 1.178 5,节点的拓扑势分布如图 5(b)所示,存在 3 个局部极大值节点 $v_{15}$,$v_{18}$,$v_{21}$,其中,节点 $v_{15}$,$v_{21}$ 由于相距较近被合并为一个多代表点社区.以社区代表点为中心,扩展每个社区的私有节点,可以得到图 5(c)所示的初始社区划分.迭代划分 12 个边界节点后所得的社区结构如图 5(d)所示,其中,$v_{40}$ 与两个社区的连接一样多,可以看作重叠节点.对比图 5(d)所示的划分结果与图 5(a)所示的社区结构可以发现,该划分结果能够有效反映海豚关系网的内在社区结构.采用 GN 算法[8]分析海豚关系网的社区结构,令社区个数为 2,划分结果只有 $v_{40}$ 被误分,显示了该方法的有效性,但需要用户指定社区个数.而模块度优化方法[9]得到的最优划分倾向于将真实社区结构中的大社区分解为 4 个小社区,如图 6 所示,节点 $v_{40}$ 同样被误分.



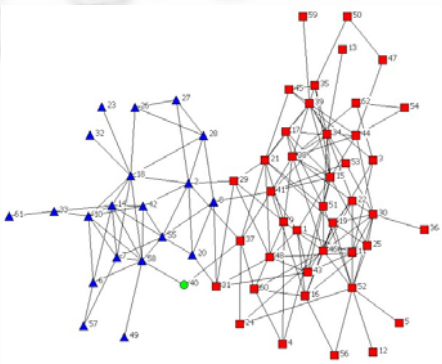
(a) Real communities of the network
(a) 真实社区结构




(b) Distribution of topological potential over nodes ($\sigma$=1.1785)
(b) 节点的拓扑势分布($\sigma$=1.1785)



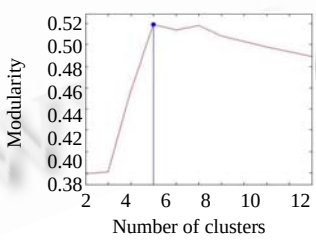
(c) Initial partition of network
(c) 初始社区划分



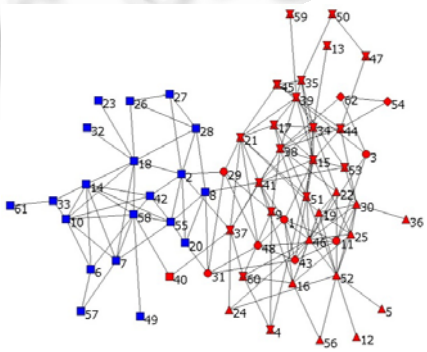
(d) Final communities discovered
(d) 最终的社区划分结果

Fig.5 Detecting community structure for the bottlenose dolphin network by the community discovery algorithm based on topological potential

图 5 采用基于拓扑势的社区发现方法分析海豚关系网络的社区结构




(a) Optimization of the modularity
(a) 模块度的优化



(b) Communities corresponding to the optimal modularity
(b) 最优模块度对应的社区划分

Fig.6 Detecting community structure for the bottlenose dolphin network by the modularity optimization method

图 6 采用模块度优化方法分析海豚关系网络的社区结构

### 6.3 中国电影演员合作网络

最后,采用基于拓扑势的社区发现方法分析 2006 年中国电影演员合作网[38]的最大连通子图的社区结构.该数据来自国内著名的网络电影社区——MTime 网站,网络中每个节点代表一个演员,边代表两个演员共同出演过同一部电影,即存在合作关系.网络共有 587 个节点、1 725 条边,如图 7 所示.

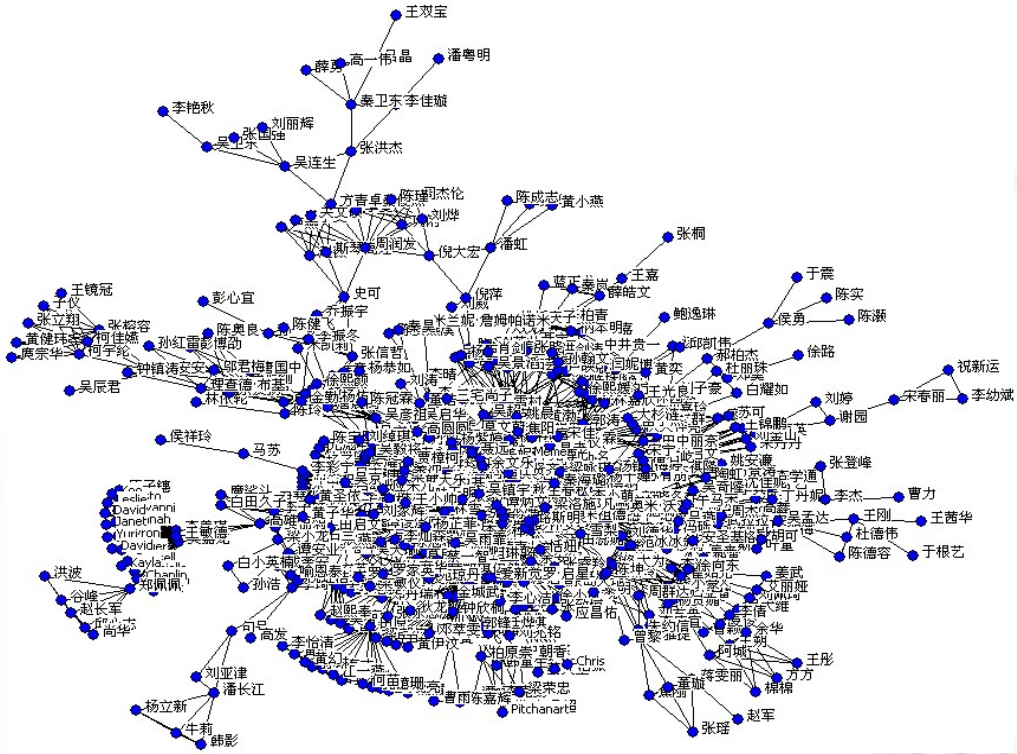



Fig.7 A Chinese movie actor collaboration network in 2006

图 7 2006 年中国电影演员合作网

采用基于拓扑势的社区发现方法分析其社区结构,令$\sigma$=1.748,所得的节点拓扑势分布如图 8 所示,存在 13 个局部极大值节点(即拓扑吸引子),分别对应电影演员曾志伟、冯德伦、夏雨、李美琪、王敏德、吴嘉龙、周润发、潘长江、柯宇纶、柯佳嬿、张榕容、秦卫东与侯勇.以 13 个拓扑吸引子为代表点对网络进行社区划分,最终形成如图 9 所示的 8 个社区,社区规模见表 1.其中,以曾志伟与冯德伦为代表的社区比其他 7 个小社区规模要大得多,这与 Arenas 等人[23,39]关于社会网社区规模极不均匀、近似服从幂律分布的观察结论是吻合的.

采用模块度优化方法分析 2006 年中国电影演员合作网的社区结构,优化所得的社区个数为 33,对应的社区划分如图 10 所示.对比图 9 与图 10 可以发现,模块度优化方法发现的 33 个社区中有 7 个小社区与拓扑势方法发现的小社区非常类似,但它倾向于将拓扑势方法所得到的包含 446 个节点的大社区分解为 26 个小社区.根据模块度优化方法在海豚关系网社区划分中的类似表现,我们有理由认为,由于受其内在的分辨率的限制,该方法倾向于将网络分解为较多的细粒度社区,不能有效体现真实网络的内在社区结构.

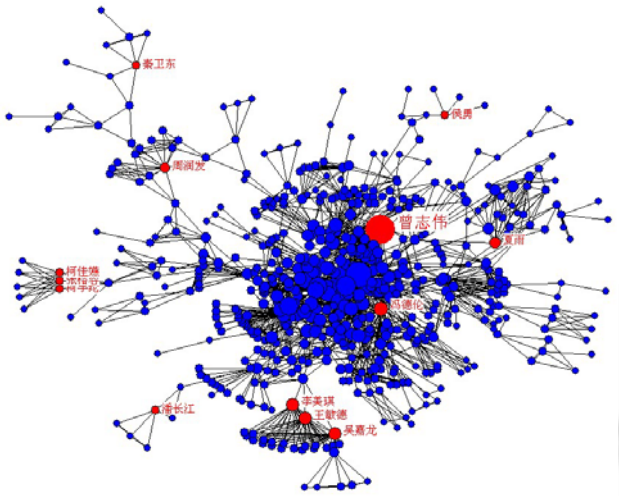



Fig.8 Distribution of topological potential over nodes in the 2006 Chinese movie actor collaboration network ($\sigma$=1.748)

图 8 2006 年中国电影演员合作网的节点拓扑势分布($\sigma$=1.748)

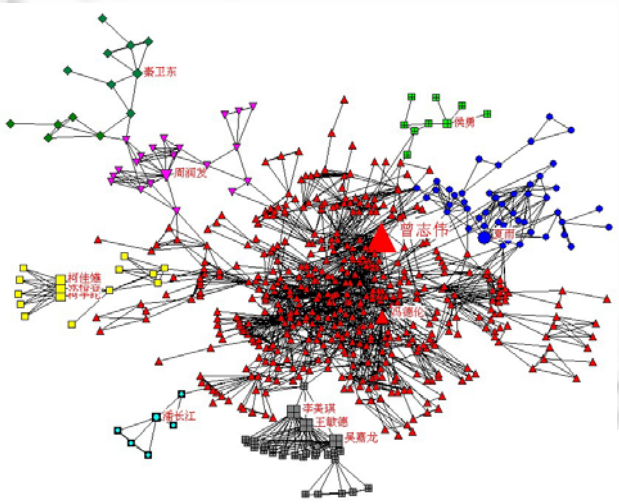



Fig.9 Detecting community structure for the 2006 Chinese movie actor collaboration network by the community discovery algorithm based on topological potential

图 9 采用基于拓扑势的社区发现方法对 2006 年中国电影演员合作网进行社区划分

**Table 1** Communities of the 2006 Chinese movie actor collaboration network

**表 1** 2006 年中国电影演员合作网的社区划分

| No. | Name of representatives | Size | Number of overlap nodes |
|---|---|---|---|
| 1 | 曾志伟,冯德伦 | 446 | 1 |
| 2 | 夏雨 | 49 | 0 |
| 3 | 李美琪,王敏德,吴嘉龙 | 28 | 0 |
| 4 | 周润发 | 21 | 1 |
| 5 | 柯宇纶,柯佳嬿,张榕容 | 15 | 5 |
| 6 | 秦卫东 | 13 | 5 |
| 7 | 侯勇 | 9 | 1 |
| 8 | 潘长江 | 6 | 0 |

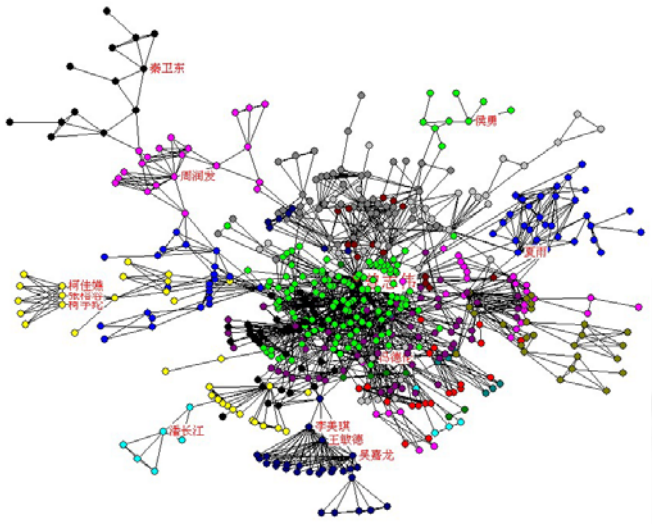



Fig.10 Detecting community structure for the 2006 Chinese movie actor collaboration network obtained by the modularity optimization method (optimal number of communities is 33)

图 10 采用模块度优化方法对 2006 年中国电影演员合作网进行社区划分(最优社区个数为 33)

## 7 结束语

社区发现是复杂网络领域中的一个重要研究方向.常用社区发现方法或者需要用户预先提供有关社区结构的先验知识,或者对社区结构存在不合理的假设,不能有效地揭示真实复杂网络的内在社区结构.本文从数据场思想出发,提出一种基于拓扑势的社区发现算法.该方法引入拓扑势场描述网络节点间的相互作用,将每个社区视为拓扑势场的局部高势区,通过寻找被低势区域所分割的连通高势区域实现网络的社区划分.采用真实网络测试本文方法的有效性,实验结果显示,该方法无须用户指定社区个数等算法参数,能够有效地揭示网络内在的社区结构及社区间具有不确定性的重叠节点现象,具有相对较好的算法性能,时间复杂度为 $O(n^2)$~$O(m+n^{3/\gamma})$, $n$ 为网络节点数,$m$ 为边数,$2<\gamma<3$ 为一个常数.

**References**:

[1] Watts DJ, Strogatz SH. Collective dynamics of small world networks. Nature, 1998,393(4):440−442.

[2] Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999,286(5439):509−512.

[3] Barabási A, Bonabeau E. Scale-Free networks. Scientific American, 2003,288(5):60−69.

[4] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. ACM SIGCOMM Computer Communication Review, 1999,29(4):251−262.

[5] Newman MEJ. The structure and function of complex networks. SLAM Review, 2003,45(2):167−256.

[6] Zhou T, Bai WJ, Wang BH, Liu ZJ, Yan G. A brief review of complex networks. Physics, 2005,34(1):31−36 (in Chinese with English abstract).

[7] Borgs C, Chayes J, Mahdian M, Saberi A. Exploring the community structure of newsgroups. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: AAAI, 2004. 783−787.

[8] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences of the United States of America, 2002,99(12):7821−7826.

[9] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2). 026113.

[10] Newman MEJ. Modularity and community structure in networks. Proc. of the National Academy of Sciences of the United States of America, 2006,103(23):8577−8582.

[11] Wang L, Dai GZ. Community finding in complex networks: theory and applications. Science & Technology Review, 2005,23(8): 62−66 (in Chinese with English abstract).

[12] Scott J. Social Network Analysis: A Handbook. 2nd ed., London: Sage Publications, 2000.

[13] West DB. Introduction to Graph Theory. Upper Saddle River: Prentice Hall, 2001.

[14] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6). 066133.

[15] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6). 066111.

[16] Duch J, Arenas A. Community identification using extremal optimization. Physical Review E, 2005,72(6). 027104.

[17] Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. Nature, 2005,433(7028):895−900.

[18] Ruan JH, Zhang WX. Identification and evaluation of weak community structures in networks. In: Proc. of the 21st National Conf. on Artificial Intelligence (AAAI 2006). The AAAI Press, 2006. 470−475.

[19] White S, Smyth P. A spectral clustering approach to finding communities in graph. In: Kargupta H, Srivastava J, Kamath C, Goodman A, eds. Proc. of the SIAM Int'l Conf. on Data Mining. Newport Beach: SIAM, 2005. 76−84.

[20] Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics, 2005, 2005(9). P09008.

[21] Muff S, Rao F, Caflisch A. Local modularity measure for network clusterizations. Physical Reviews E, 2005,72(5). 056107.

[22] Fortunato S, Barthelemy M. Resolution limit in community detection. Proc. of the National Academy of Sciences of the United States of America, 2007,104(1):36−41.

[23] Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerà R. Community analysis in social networks. European Physics Journal B, 2004,38(2):373−380.

[24] Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-Similar community structure in organizations. Preprint cond-mat/0211498, 2002. http://arxiv.org/abs/cond-mat/0211498

[25] Gan WY. Study on the clustering problem in data mining [Ph.D. Thesis]. Nanjing: University of Science and Technology, 2003 (in Chinese with English abstract).

[26] Gan WY, Li DY, Wang JM. An hierarchical clustering method based on data fields. Chinese Journal of Electronics, 2006,34(2): 258−262 (in Chinese with English abstract).

[27] Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. Physical Review E, 2001,64(2). 026118.

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

[29] Newman MEJ. Detecting comunity structure in networks. European Physics Journal B, 2004,38(2):321−330.

[30] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality. Physical Review E, 2004,70(5). 056104.

[31] Gfeller D, Chappelier JC, Rios PDL. Finding instabilities in the community structure of complex networks. Physical Review E, 2005,72(5). 056135.

[32] Danon L, Díaz-Guilera A, Arenas A. Effect of size heterogeneity on community identification in complex networks. Journal of Statistical Mechanics: Theory and Experiment, 2006,2006(11). P11010.

[33] Newman MEJ, Girvan M. Mixing patterns and community structure in networks. Preprint cond-mat/0210146, 2002. http://arxiv.org/abs/cond-mat/0210146

[34] Bagrow JP, Bollt EM. A local method for detecting communities. Physical Review E, 2005,72(4). 046108.

[35] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. Physics Reports, 2006, 424(4-5):175−308.

[36] Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations—Can geographic isolation explain this unique trait? Behavioral Ecology and Sociobiology, 2003,54(4):396−405.

[37] Lusseau D, Newman MEJ. Identifying the role that animals play in their social networks. Proc. of the Royal Society B: Biological Sciences, 2004,271(Suppl_6):S477−S481.

[38] He N, Gan WY, Li DY, Kang JC. The topological analysis of a small actor collaboration network. Complex Systems and Complexity Science, 2006,3(4):1−10 (in Chinese with English abstract).

[39] Guimera R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-Similar community structure in organizations. Preprint cond-mat/0211498, 2002. http://arxiv.org/abs/cond-mat/0211498

**附中文参考文献:**

[6] 周涛,柏文洁,汪秉宏,刘之景,严钢.复杂网络研究概述.物理,2005,34(1):31−36.

[11] 王林,戴冠中.复杂网络中的社区发现——理论与应用.科技导报,2005,23(8):62−66.

[25] 淦文燕.聚类——数据挖掘中的基础理论问题研究[博士学位论文].南京:解放军理工大学,2003.

[26] 淦文燕,李德毅,王建民.一种基于数据场的层次聚类方法.电子学报,2006,34(2):258−262.

[38] 赫南,淦文燕,李德毅,康建初.一个小型演员合作网的拓扑性质分析.复杂系统与复杂性科学,2006,3(4):1−10.

**淦文燕**(1971－),女,江西九江人,博士,副教授,主要研究领域为数据挖掘,人工智能,复杂网络.

**李德毅**(1944－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为不确定性人工智能.

**赫南**(1983－),男,硕士,主要研究领域为复杂网络,不确定性人工智能.

**王建民**(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据管理与信息系统,工作流技术,软件度量与测试,数据网格与数据挖掘.