基于拓扑势的重叠社区及社区间结构洞识别
——兼论结构洞理论视角下网络的脆弱性

李泓波[1],张健沛[2],杨 静[2],白劲波[3,4],初 妍[2]
(1.肇庆学院计算机学院,广东肇庆 526061;2.哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001;
3.哈尔滨工程大学经济管理学院,黑龙江哈尔滨 150001;4.黑龙江工程学院计算机科学与技术学院,黑龙江哈尔滨 150050)

摘 要: 社会网络和复杂网络上的社区识别已经成为当前研究的热点和前沿课题.针对目前社区识别方法不能兼具较低时间复杂度、无须专家知识或先验知识和允许存在重叠节点的不足,提出了基于拓扑势理论的重叠社区识别方法.通过提出的重叠节点社区归属不确定性测度,该方法同时实现了社区间结构洞的识别.实验验证了该方法的有效性.另外,文章在理论证明的基础上提出了影响因子优化算法;论证了结构洞理论视角下网络的脆弱性.

关键词: 网络; 重叠社区; 结构洞; 识别; 拓扑势; 影响因子; 不确定性测度; 脆弱性
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2014)01-0062-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.01.010

Identification of Overlapping Communities and Structural Holes Between Communities Based on Topological Potential
— Also on the Fragility of Network from the Perspective of the Structural Hole Theory

LI Hong-bo[1], ZHANG Jian-pei[2], YANG Jing[2], BAI Jin-bo[3,4], CHU Yan[2]
(1. *School of Computer Science, Zhaoqing University, Zhaoqing, Guangdong* 526061, *China*; 2. *College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*; 3. *School of Economics and Management, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*; 4. *School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, Heilongjiang* 150050, *China*)

Abstract: Community identification has been a hot spot and a cutting-edge topic among researchers. Since none of the present community identification methods simultaneously meets the requirements, such as lower time complexity, independence of expertise or experiences, allowance for overlapping nodes and so on, an overlapping community identification method is proposed based on topological potential theory. This method can also identify the structural holes in communities at the same time by the presented uncertainty measure of the community identity of the overlapping nodes, and its effectiveness is verified by experiments. In addition, an influence factor optimization algorithm is proposed and network fragility is discussed and prooved from the perspective of structural hole theory.

Key words: network; overlapping community; structural holes; identification; topological potential; influence factor; uncertainty measure; fragility

1 引言

研究表明,诸多看起来形态各异的网络普遍存在着聚簇效应[1~5],对其进行聚簇发现即为社区识别.由于21世纪第二个十年将进入社会网络和物联网时代[6],而且许多国家都将面向网络化社会的研究和社会计算研究提升为国家战略[7],因此作为社会计算重要研究内容的社区识别已经成为学术界研究的热点和前沿课题.本文将基于源自核子物理学中短程场的TP(Topological Potential,拓扑势)理论,提出影响因子优化算法和网络重叠社区及社区间结构洞识别方法.

2 相关背景

与经典的网络社区识别方法相比[8~12],基于TP理论的社区识别方法[13,14]具有综合比较优势,如无需专家知识、可以进行重叠社区识别、具有较小的时间开销

收稿日期:2012-01-29;修回日期:2013-08-28;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61073041,No.61073043);黑龙江省自然科学基金(No.F200901,No.F200917);黑龙江省教育厅科学技术研究基金(No.12531529);哈尔滨市优秀学科带头人基金(No.2010RFXXG002,No.2011RFXXG015);高等学校博士学科点专项科研基金(No.20112304110011)

等,是一种颇具活力的新方法.但是,该种方法也存在着不足:

(1)重叠节点过于稀疏,缺乏现实合理性.对此,下文的对比实验中将有详尽的体现.

(2)优化影响因子的确定算法或者效率较低,或者不能找到真正的优化值.

(3)与其他方法一样,亦不能实现边界节点社区归属亦此亦彼的不确定性程度的定量刻画.

此外,与其他方法一样,该种方法也不能识别出社区间的结构洞.

结构洞(Structural holes)理论是美国社会学家和管理学家 Burt 提出的社会网络分析理论,融合了社会网络的交换论、弱关系强度理论等多种理论[15].Burt 认为"非冗余的联系人由结构洞连接,结构洞是两个行动者之间的非冗余的联系".本质上,结构洞表示的是一种三方之间的非冗余联系.例如,在图 1 中,节点 E 与节点对 AB、AC 构成的关系结构都是结构洞,而节点 E 与节点对 BC 构成的关系结构不是结构洞(因为节点 B 和 C 还有直接联系,即冗余联系).

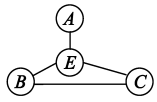


图1　节点间结构洞示意图

在图 1 中,节点 E 被 Burt 称为结构洞的占据者.如果将图 1 中的节点 A、B、C 看成网络社区,而将节点 E 看作可同时归入多个社区的重叠节点集合,那么 E 即为社区间的结构洞占据者.因为一个结构洞占据者往往对应着一个或多个结构洞,找到社区间的结构洞占据者就相当于找到了所有社区间的结构洞,所以下文将结构洞占据者简称为结构洞,并称对其的识别为结构洞识别.

既鲁棒又脆弱是复杂系统最重要和最基本的特征之一[16].文献[17]中阐释了无标度(Scale-free)网络的鲁棒性(Robustness)和脆弱性(Fragility).这里的脆弱性是指选择性地去除网络中的一些节点(如度数最高的节点)从而使得大多数节点不再连通而导致的网络信息运载(Information-carrying)失效.

本文将从结构洞理论视角来认识和理解网络的脆弱性.结构洞理论认为,同一社区内的各个节点所掌握的信息和资源是相似的,如果想获得异质信息和资源,只能以结构洞为中介从其他社区获得.可见,结构洞在社区间起着信息阀门和中转站的作用.如果去除社区间的结构洞,则会导致信息只能在社区内流动而不能在社区间流动.这种社区间的信息流通失效,当然也是网络脆弱性的一种表现.而且,在此种策略下网络呈现出的这种脆弱性并不要求损毁具有高度数的节点.在实际网络中,高度数节点往往具有更安全、更严密的保护措施,瘫痪这样的节点一般代价较高,且不易成功.目前来看,结构洞识别算法还十分鲜见.

限于篇幅,下文中涉及到的基于 *TP* 理论的社区识别方法的一些概念和公式请参考文献[13,18,19].

3　若干相关定理

在进行算法设计之前,先引入文献[19]中的几个定理.

对于一个具有 n 个节点的网络 G 来说:

定理 1　如果网络 G 可由完全图表示,那么其 TPE(Topological Potential Entropy,拓扑势熵)$H(\sigma)=\ln n$,$\sigma\in(0,+\infty)$.

定理 2　$\lim\limits_{\sigma\to+\infty}H(\sigma)=\ln n$.

定理 3　$\lim\limits_{\sigma\to0^{+}}H(\sigma)=\ln n$.

定理 4　$H\left(\dfrac{\phi_{v_1}(\sigma)}{Z(\sigma)},\dfrac{\phi_{v_2}(\sigma)}{Z(\sigma)},\cdots,\dfrac{\phi_{v_n}(\sigma)}{Z(\sigma)}\right)\leq\ln n$.

4　影响因子优化算法 IFO

目前的基于 TP 理论的网络社区识别方法,在确定优化的节点影响范围 h 时,首先要求出 TPE 最小值所对应的优化影响因子 σ_{opt},然后才能计算出 h.文献[20]中提供了一个求取 σ_{opt}的高效算法,该算法实质上是一种单因素单峰目标函数优选法——0.618 法.0.618 法采用迭代方式逐步缩小优选区间,当前迭代选优区间至少占前一迭代优选区间的 0.618;除在初始区间上需要进行两次实验外,以后的各优选区间上只需进行一次实验.由于 0.618^{n-1}可以小于任何预先给定的正数,所以 0.618 法可使目标函数值位于任何给定的精度范围内.根据指数函数的性质,易知 0.618 法达到精度要求非常迅速.例如,要求目标函数的优化值精度不超过 0.0001,只需计算目标函数值 $\left\lfloor 1+\dfrac{\lg 0.0001}{\lg 0.618}\right\rfloor=20$ 次.

由于目前尚不能从理论和实证研究层面上给出 σ_{opt}的存在范围,又考虑到 0.618 法的收敛速度较快,所以给定一个尽量大的初始选优区间是一个非常自然的做法.然而,如此做法却会导致错误的 σ_{opt}值:由定理 2 和定理 3 知,当 σ 趋向 0 或 $+\infty$ 时 $H(\sigma)$趋向最大值 $\ln n$(定理 4),又由于数值在计算机内部的离散表示机制,远离优化值而趋向 0 的两点或趋向 $+\infty$ 的两点极可能被认为是函数等值的,依照 0.618 法这样的点对会被确定为新的优选区间.例如,在图 2 中优选区间可能被确定为远离优化值的区间 I_1 或 I_2,而这两个区间中显然不含 σ_{opt}.为了克服这一问题,本文将提供一个相对高效的优化算法 IFO(Influence Factor Optimization).

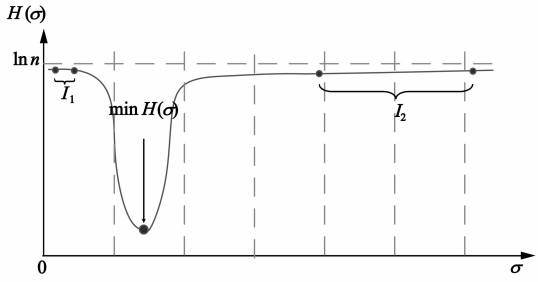


图2 可能出现的优选区间

由 TP 理论易知,如果优化的影响因子 σ_{opt}在区间$[\sqrt{2}h/3,\sqrt{2}(h+1)/3)$内取得,那么优化的节点影响范围就等于 h.受此启发,算法 IFO 总的指导思想是先确定 σ_{opt}所在的区间$[\sqrt{2}h/3,\sqrt{2}(h+1)/3)$后再求其具体值.具体地说,IFO 算法首先预设一个足够小的正数 δ(如令 $\delta=\sqrt{2}/300$),然后以 $H(\sqrt{2}(h+1)-\delta)$是否小于 $H(\sqrt{2}(h+1))$为判断条件确定 σ_{opt}所在的区间,最后在确定的区间里调用 0.618 法确定满足一定精度要求的 σ_{opt}存在范围. IFO 确定 σ_{opt}所在区间原理如图 3 所示,算法详细描述如算法 1 所示.

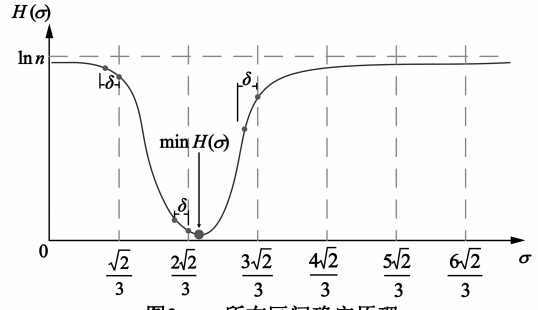


图3 σ_{opt}所在区间确定原理

算法 1 影响因子优化算法 IFO

```
Input: the original value of sigma and the value of delta
Output: optimized influence factor sigma
1  step = Sqrt(2)/3;
2  l = sigma;   r = step;   h = 1;
3  if (H(l) = = H(r))
4  {
5     print "the network is a complete graph";
6     exit;
7  }
8  while (H(r - delta) ≥ H(r))
9  {
10    l = h * Sqrt(2)/3; r = l + step; h + + ;
11 }
12 sigma = Func(√2(h - 1)/3, √2h/3);
13 print sigma;
```

由于 σ 在数值 0 处没有定义,所以在 IFO 算法中首先要求输入其初始值.该值应为一个尽量靠近 0 的极小正数,当然算法实现时也可在程序中指定一个默认值(如 0.0001)而避免用户输入.根据定理 1,可知完全图表示的网络的 $H(\sigma)$图像为一直线,算法 1 的 3 - 7 行用于防止多余的运算及避免确定出错误的 σ_{opt};第 12 行中的 Func 函数则用于调用文献[20]中影响因子优化算法.

若设 σ_{opt}存在于区间$[\sqrt{2}h/3,\sqrt{2}(h+1)/3)$内,与文献[20]中方法相比,虽然 IFO 算法比其多计算了 $2h$ 次 TPE,但是可以避免出现错误的 σ_{opt}值.根据网络的小世界效应(六度分离理论),一般网络,尤其是社会关系网络,直径平均为 6;又根据 σ_{opt}的计算机制和网络的聚簇效应,一般网络优化的节点影响范围 h 的值平均为 3,所以在一般情况下,IFO 算法只比文献[20]中方法平均多计算 6 次 TPE.因而, IFO 算法的时间复杂度不超过 $O(n^2)$.

为了考察 IFO 算法的有效性,现将其与应用遗传算法 GA 的影响因子优化方法进行对比,如表 1 所示.分析表 1 可知,IFO 算法与 GA 求得的 σ_{opt}是基本吻合的,而且二者求得的优化节点影响范围 h 是完全一致的.因此,IFO 算法是有效和准确的.

表 1 IFO 算法与遗传算法的结果对比

Network names	GA		IFO	
	σ_{opt}	h	σ_{opt}	h
Word adjacencies[21]	1.0045	2	1.0043	2
Neural[21]	0.8416	2	0.8415	2
Les miserables[21]	1.0435	2	1.0435	2
Books about US politics[21]	0.9803	2	0.9802	2
Power grid[21]	3.9050	8	3.9048	8
American college football[21]	1.6248	3	1.6250	3
Zachary's karate club[21]	1.0205	2	1.0204	2
Dolphin social network[21]	1.1787	2	1.1782	2

5 网络重叠社区及社区间结构洞识别

5.1 重叠节点社区归属不确定性测度

从研究历程来看,社区识别研究经历了非重叠方法和重叠方法两个阶段.为了满足数据挖掘、知识发现和智能决策等研究的需要,我们曾提出了刻画和表征重叠节点归属不同社区的测度公式[18]

$$p_k(v_i)=\frac{attC_k(v_i)}{\sum_{k\in I}attC_k(v_i)} \tag{1}$$

其中 I 为网络 G 的社区代表节点集合(Representative node Set of network G,记为 RepSet)的指标集(Index Set)

(本文用代表节点编号表示社区编号,如某代表节点编号为88,则以该代表节点扩展而成的社区编号也为88),$attC_k(v_i)$ ($k=1,\cdots,t$)则由下式确定

$$attC_k(v_i)=\frac{1}{n}\sum_{j=1}^{h} m_j(v_i)\times e^{-\left(\frac{j}{\sigma_{opt}}\right)^2} \tag{2}$$

其中的 $m_j(v_i)$为满足以下条件的节点的个数之和:(1)为 v_i 的第 j 跳邻居节点;(2)为可归入社区 C_k 的重叠节点;(3)为唯一属于社区 C_k 的节点.

由于式(2)在计算重叠节点对 v_i 的作用力时比较粗略(例如,一个可归入 k 个社区的重叠节点,在计算其对作用范围内的某一节点的作用力时,应该对该作用力乘以系数 $1/k$,而不应像式(2)那样将重叠节点和非重叠节点的作用力进行同样处理),故将式(2)改为

$$attC_k(v_i)=\frac{1}{n}\sum_{j=1}^{h}\left(n_j(v_i)+\sum_{k=1}^{m_j(v_i)}\frac{1}{t(v_k)}\right)e^{-\left(\frac{j}{\sigma_{opt}}\right)^2} \tag{3}$$

其中,$n_j(v_i)$为节点 v_i 的第 j 跳且属于社区 C_k 的非重叠节点数目,$m_j(v_i)$为节点 v_i 的第 j 跳的重叠节点数目,$t(v_k)$($k=1,2,\cdots,m_j(v_i)$)为 v_i 的第 j 跳中第 k 个重叠节点可归入的社区数目.

5.2　识别算法 CASHI

本部分利用重叠节点的社区归属不确定性测度和影响因子优化算法 IFO,给出重叠社区及社区间结构洞识别算法 CASHI(Community and Structural Hole Identification).CASHI 算法最复杂的部分为社区生成部分,在最坏的情况下其时间复杂度也不超过 $O(n^2)$.

算法 2　CASHI 算法

输入:网络 G.

输出:识别得到的社区及社区间的结构洞.

算法步骤:

步骤 1　计算 G 中每个节点的 TP;

步骤 2　计算网络 G 的 TPE;

步骤 3　利用 IFO 算法确定最优的影响因子,进而确定优化的节点影响范围 h;

步骤 4　根据计算出的 h 重新计算每个节点的 TP 后,采用盲人登山法思想确定社区代表点集合 RepSet;

步骤 5　在跳数 s 不大于 h 的条件下,执行如下操作:从 RepSet 中的每个代表节点出发,沿拓扑势吸引链下降方向以广度优先方式逐跳交替搜索被其吸引的节点,以实现对各个社区的拓展,并同时确定其内部节点、结构洞(重叠节点),以及计算结构洞(重叠节点)的社区归属不确定性;

步骤 6　输出识别出的社区以及社区间的结构洞.

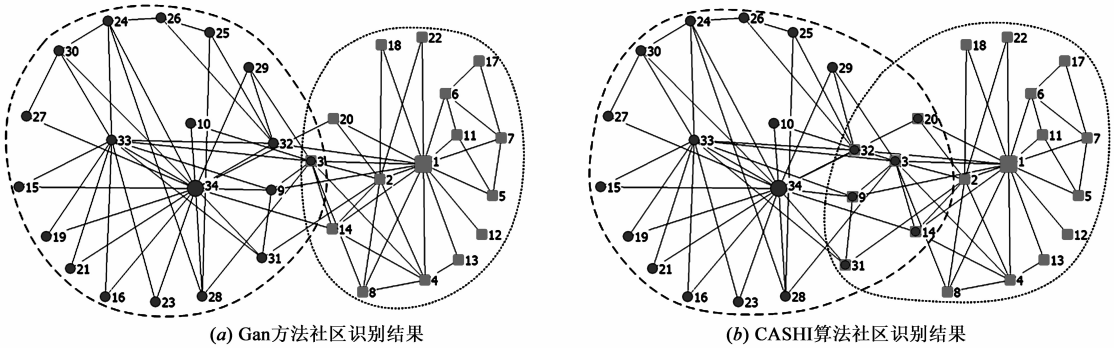
按照 Newman 等学者的定义,社区是存在于网络中的其内部联系相对紧密而相互间联系相对稀疏的结构.又由文献[13]等文献中关于优化影响因子和节点优化影响范围的定义,易知 CASHI 算法中的重叠节点即为社区间的结构洞.

5.3　实验及实验结果分析

5.3.1　对比实验

目前基于 TP 理论的社区识别方法在具体实现时虽然存在差异,但社区识别结果基本一致.为进行比较,本小节将以经充分验证并具有代表性的文献[13]中方法(以下简称其为 Gan 方法)为对象进行对比实验.

首先将在 Zachary 空手道俱乐部网络[21]和 Books about US Politics 网络[21]两个标准数据集上进行对比实验,然后再在一个人工网络(Artificial Ntwork)上进行对比实验.由于人工网络具有较大的随机性,有利于验证方法的有效性,所以近来比较流行.在该人工网络上共预设了 3 个社区,每个社区 15 个节点,社区内部节点的连接概率为 $p_{in}=0.4$,社区间的连接概率为 $p_{out}=0.005$.这 3 个网络上的对比结果如图 4 ~ 6 所示,其中用点线圈起的部分为识别出的社区,每个社区中的大图标节点为该社区的代表点.



(*a*) Gan方法社区识别结果　　(*b*) CASHI算法社区识别结果

图4　Zachary空手道俱乐部网络上的实验对比

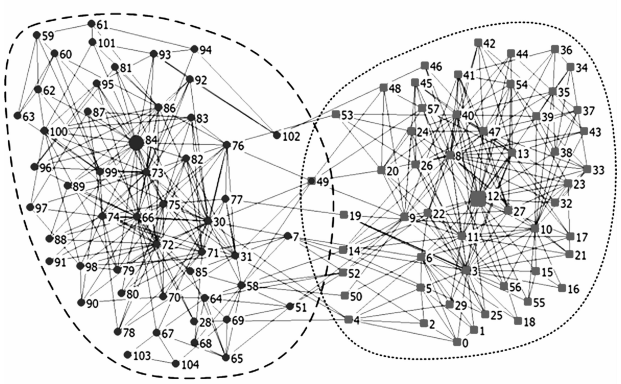
5.3.2　实验分析

(1)不确定性测度和 CASHI 算法有效性分析

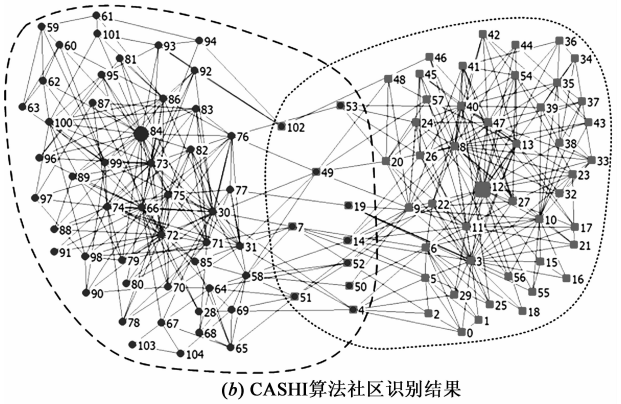
本部分将分为两个层面:首先,阐释和分析不确定性测度与强社区结构的关系,以说明该测度的合理性;其次,分析说明不确定性测度和 CASHI 算法的有效性.

①在进行分析之前,有必要回顾一下 Gan 方法中

提出的用于判断重叠节点社区归属的效益函数和 Raddichi 等人定义的强社区结构[11]. Gan 方法中用于判断重叠节点 v_i社区归属的效益函数定义为



(*a*) Gan方法社区识别结果



(*b*) CASHI算法社区识别结果

图5 Books about US politics网络上的实验对比

$$Q_{k=1,2,\cdots,t}(v_i)=\sum_{v_j\in C_k}a_{ij}-\sum_{v_m\notin C_k}a_{im} \tag{4}$$

其中 t 为识别出的网络社区个数，C_k 为识别出的社区，a_{ij}和 a_{im}为网络邻接矩阵中的元素. 只有当 v_i与所有相邻社区的效益函数值都相等时，才能判定 v_i为重叠节点. 该条件比较严苛，也是造成 Gan 方法产生的重叠节点比较稀疏的原因. 事实上，该函数来源于 Raddichi 等人定义的强社区结构. 强社区结构要求社区 C 内的任一节点 v 满足

$$k_v^{\text{in}}(C)>k_v^{\text{out}}(C) \tag{5}$$

其中 $k_v^{\text{in}}(C)$代表 v 与 C 内部的连线数目，$k_v^{\text{out}}(C)$代表 v 与 C 外部的连线数目. 对于强社区结构来说，只有当 $k_v^{\text{in}}(C)=k_v^{\text{out}}(C)$时，$v$ 才具备成为重叠节点的可能性. 由此可见，二者在决定重叠节点的社区归属时采用的判决标准在本质上具备极大的一致性.

为便于比较和分析，对式(4)进行标准化处理

$$\Theta_k(v_i)=\frac{Q_k(v_i)}{d(v_i)} \tag{6}$$

其中 $d(v_i)$为重叠节点 v_i的度. 很明显，$\Theta_k(v_i)$与 $Q_k(v_i)$是完全相关的，二者之间只差一个系数 $1/d(v_i)$. 分析表 2～4，又知本文提出的测度与标准化后的 Gan 方法中的效益函数 $\Theta_k(v_i)$在数值上具有极高的相似性. 因此，如果将具有高测度值作为重叠节点社区归属的判断依据，则会使重叠节点的社区归属满足强社区结构要求.

表 2 空手道网络中重叠节点及其 p_k和 Θ_k值

No.	p_{34}	Θ_{34}	p_1	Θ_1
3	0.5012	0.5000	0.4988	0.5000
9	0.5785	0.6000	0.4215	0.4000
14	0.3558	0.3000	0.6442	0.7000
20	0.3968	0.3333	0.6032	0.6667
31	0.6338	0.6250	0.3662	0.3750
32	0.7686	0.8333	0.2314	0.1667

表 3 Books about US politics **网络中重叠节点及其 p_k和 Θ_k值**

No.	p_{84}	Θ_{84}	p_{12}	Θ_{12}
4	0.4191	0.4375	0.5809	0.5625
7	0.5802	0.5714	0.4198	0.4286
14	0.1731	0.1667	0.8269	0.3333
19	0.1940	0.2000	0.8060	0.8000
49	0.5345	0.5000	0.4655	0.5000
50	0.3300	0.3333	0.6700	0.6667
51	0.7033	0.7500	0.2967	0.2500
52	0.4843	0.5000	0.5157	0.5000
53	0.2351	0.2000	0.7649	0.8000
102	0.7500	0.7500	0.2500	0.2500

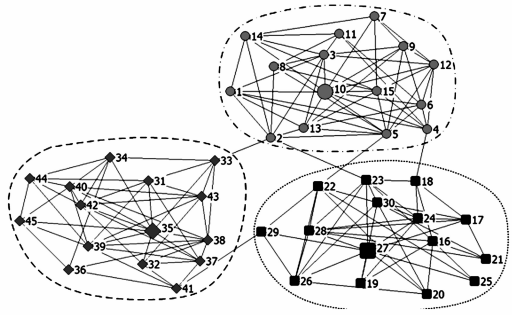
表 4 人工网络中重叠节点及其 p_k和 Θ_k值

No.	p_{35}	Θ_{35}	p_{27}	Θ_{27}	p_{10}	Θ_{10}
18	—	—	0.6655	0.7000	0.3345	0.3000
22	—	—	0.7336	0.7857	0.2664	0.2143
23	—	—	0.7162	0.7500	0.2838	0.2500
29	0.2983	0.2858	0.7017	0.7143	—	—
30	—	—	0.8628	0.8750	0.1372	0.1250
33	0.7928	0.8334	—	—	0.2072	0.1667

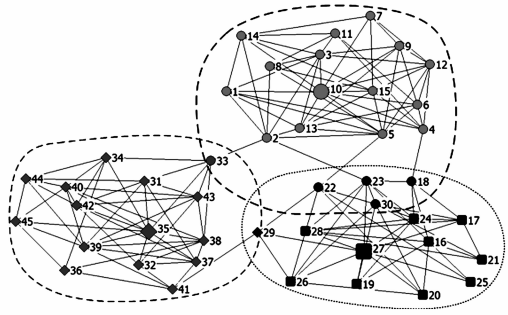
②由图 4～6 知，与 Gan 方法相比，CASHI 算法识别出的重叠节点明显增多. 手段服务于目的，增加重叠节点数量首先是为了避免人为割裂边界节点与其他社区的联系而造成的信息损失，其次是为了识别出社区间的所有结构洞，最后是为了克服 Gan 方法产生的重叠节点过于稀少而造成的缺乏现实合理性的不足.

可以设想，如果本文提出的不确定性测度比较科学、合理，则当以高测度值为依据来判定重叠节点的社区归属时，虽然 CASHI 算法与 Gan 方法处理重叠节点

的策略存在差异,但二者识别的总体结果应基本吻合.实验证明确实如此:以高测度值为依据来判定重叠节点的社区归属,则前述3个网络上的社区识别结果如图7、8和图6(*a*)所示.图7与图4(*a*)相比,只是在图7的社区 C_1 中少了节点3;图8与5(*a*)相比,只是在社区 C_{12} 中少了结点49.这充分说明了不确定性测度的有效性,同时也说明了CASHI算法的有效性.



(*a*) Gan方法社区识别结果



(*b*) CASHI算法社区识别结果

图6　人工网络上的实验对比

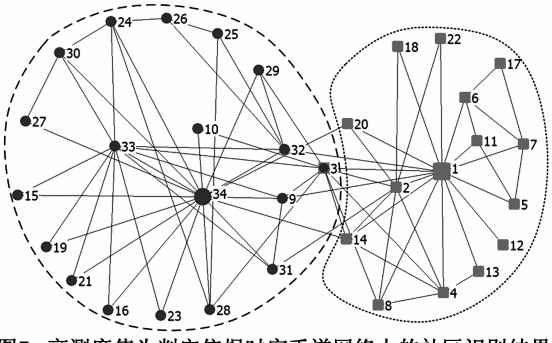


图7　高测度值为判定依据时空手道网络上的社区识别结果

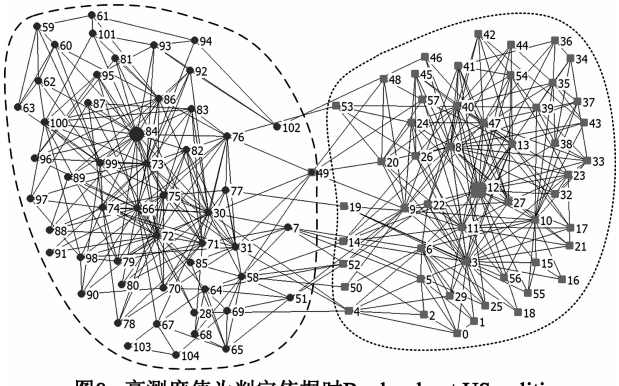


图8　高测度值为判定依据时Books about US politics网络上的社区识别结果

需要指出的是,网络中的某些节点的社区归属历来存在争议.例如,空手道网络中的节点3,既有算法将其归入左面的社区,也有算法将其归入右面的社区.对于这些有争议的节点,如空手道网络中的节点3或Books about US politics网络中的节点49,将其归入任意一个相连社区都有一定的合理性.由表2和表3易知,这两个节点归属相连社区的不确定值都非常接近0.5,这进一步验证了重叠节点社区归属不确定测度的合理性.

(2)结构洞理论下网络脆弱性分析

根据结构洞理论,可以判定图4(*b*)、图5(*b*)和图6(*b*)中社区交叠部分由社区间的结构洞构成.对前述3个网络进行选择性攻击——瘫痪社区间的结构洞,结果如图9~11所示.可见,正如前文分析的那样,对网络社区的结构洞进行攻击确可使得原本连通的网络分裂为互不连通的社区.

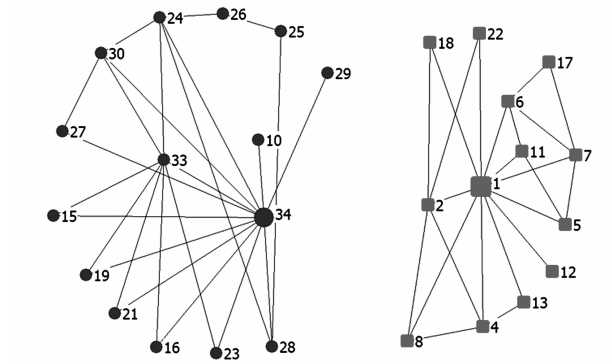


图9　除去结构洞后的空手道网络

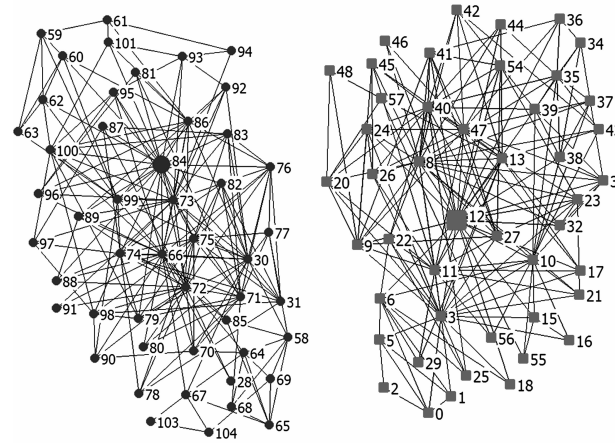


图10　除去结构洞后的Books about US politics网络

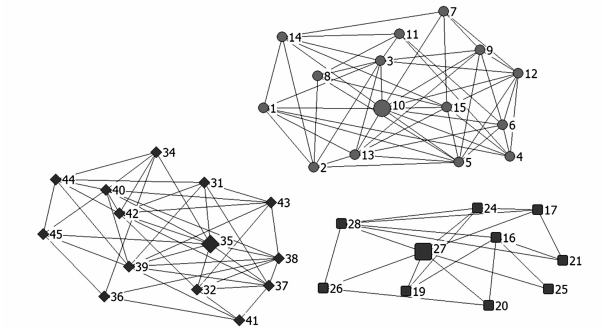


图11 除去结构洞后的人工网络

表5从网络节点的平均度数(NNMD)、度数标准差(SDV)、结构洞节点的平均度数(SHMD)几个方面给出了3个网络的统计信息.分析表5,易知引起网络脆弱性的结构洞节点既非拥有最大度数的明星节点也非拥有最少度数的边缘节点,而只是一些拥有平均度数的普通节点.这种由破坏普通节点而引发的网络脆弱性明显区别于破坏最大度数节点而引发的网络脆弱性.

表5 3个网络的若干统计信息

Network Names	NNMD	SDV	SHMD
Zachary's karate club	4.588	3.820	5.500
Books about US politics	8.400	5.449	6.100
Artificial Network	6.800	1.759	6.333

6 结束语

基于TP理论,本文一方面在理论证明的基础上提出了影响因子优化算法,克服了以往算法由于优化区间选择不当而导致的不能找到影响因子真正的优化值、效率较低等不足;另一方面,通过引入重叠节点社区归属不确定性测度,提出了网络社区和社区间结构洞识别算法.以上两个算法的有效性都在实验中得到了验证.另外,文章从结构洞理论视角出发,探讨了蓄意攻击社区间的结构洞节点而引发的网络脆弱性.

参考文献

[1] 潘磊,金杰,王崇骏,等.社会网络中基于局部信息的边社区挖掘[J].电子学报,2012,40(11):2255-2262.
Pan L,Jin J,Wang C J,et al.Detecting link communities based on local information in social networks[J].Acta Electronica Sinica,2012,40(11):2255-2262.(in Chinese)

[2] Gao J,Liang F,et al.On community outliers and their efficient detection in information networks[A].Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C].Washington,DC,USA:ACM,2010.813-822.

[3] Xu K Q,Li J X,Liao SS.Sentiment community detection in social networks[A].Proceedings of iConference[C].Seattle,WA,USA:ACM,2011.804-805.

[4] 张伟哲,王佰玲,等.基于异质网络的意见领袖社区发现[J].电子学报,2012,40(10):1927-1932.
Zhang W Z,Wang B L,et al.Public opinion leader community mining based on the heterogeneous network[J].Acta Electronica Sinica,2012,40(10):1927-1932.(in Chinese)

[5] Kamath K Y,Caverlee J.Identifying hotspots on the real-time web[A].Proceedings of the 19th ACM International Conference on Information and Knowledge Management[C].Toronto,Ontario,Canada:ACM,2010.1837-1840.

[6] 刘挺.社会计算[J].中国计算机学会通讯,2011,7(12):6-7.
Liu T.Social computing[J].Communication of the CCF,2011,7(12):6-7.(in Chinese)

[7] 毛文吉,曾大军,柯冠岩,等.社会计算的研究现状与未来[J].中国计算机学会通讯,2011,7(12):8-11.
Mao W J,Zeng D J,Ke G Y,et al.The current statusand future of social computing research[J].Communication of the CCF,2011,7(12):8-11.(in Chinese)

[8] Girvan M,Newman M E J.Community structure in social and biological networks[A].Proceedings of the National Academy of Sciences of the United States of America[C].USA:PMC,2002.7821-7826.

[9] Fortunato S,Latora V,Marchiori M.Method to find community structures based on information centrality[J].Physical Review E:Statistical,Nonlinear,and Soft Matter Physics,2004,70(5):056104.

[10] Newman M E J.Fast algorithm for detecting community structure in networks[J].Physical Review E:Statistical,Nonlinear,and Soft Matter Physics,2004,69(6):66-133.

[11] Radicchi F,Castellano C,Cecconi F,et al.Defining and identifying communities in networks[A].Proceedings of the National Academy of Sciences of the United States of America[C].USA:PMC,2004.2658-2663.

[12] Newman M E J,et al.Finding and evaluating community structure in networks[J].Physical Review E:Statistical,Nonlinear,and Soft Matter Physics,2004,69(2):026113.

[13] 淦文燕,赫南,李德毅,等.一种基于拓扑势的网络社区发现方法[J].软件学报,2009,20(8):2241-2254.
Gan W Y,He N,Li DY,et al.Community discovery method in networks based on topological potential[J].Journal of Software,2009,20(8):2241-2254.(in Chinese)

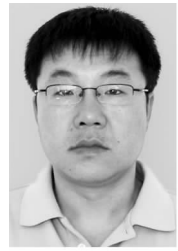
[14] Han Y N,Li D Y,Wang T.Identifying different community members in complex networks based on topology potential[J].Frontiers of Computer Science,2011,5(1):87-99.

[15] Burt R S.Structural holes:the social structure of competition Cambridge[M].MA:Harvard University Press,1992.

[16] 汪小帆,等.复杂网络理论及其应用[M].北京:清华大

学出版社,2006.
Wang X F, et al. Theory and application of complex network [M]. Beijing: Tsinghua University Press, 2006. (in Chinese)
[17] Albert R, Jeong H, Barabasi A L. Error and attack tolerance of complex networks[J]. Nature, 2000, 406: 378 – 382.
[18] Zhang J P, Li H B, Yang J, Bai J B, et al. Community discovery method with uncertainty measure of overlapping nodes based on topological potential[J]. Journal of Harbin Institute of Technology (New Series), 2012, 19(2): 16 – 22.
[19] 张健沛,李泓波,杨静,等.基于归属不确定性的变规模网络社区识别[J].电子学报,2012,40(12):2512 – 2518.
Zhang J P, Li H B, et al. Variable scale network overlapping community identification based on identity uncertainty[J]. Acta Electronica Sinica, 2012, 40(12): 2512 – 2518. (in Chinese)
[20] Han Y N, Li D Y. A novel measurement of structure properties in complex networks[J]. Lecture Notes of the Institute for Computer Sciences, Social Information and Telecommunications Engineering, 2009, (5): 1292 – 1297.
[21] Newman M E J. Network Data[OL]. http://www – personal.umich.edu/ ~ mejn/netdata/, 2013 – 04 – 19/2013 – 07 – 20.
[22] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology Sociobiology, 2003, 54(4): 396 – 405.

作者简介



李泓波　男,1971 年生,黑龙江人,工学博士.主要研究方向为社会网络、复杂网络上的社区发现、挖掘和分析,以及群体智能和数据挖掘等.
E-mail: islhb@126.com

张健沛(通讯作者)　男,1956 年生,黑龙江人,哈尔滨工程大学教授、博士生导师,主要从事数据库与知识库、数据挖掘、软件理论、社会计算等领域的研究.
E-mail: zhangjianpei@hrbeu.edu.cn