第38卷 第17期
Vol.38 No.17

计 算 机 工 程
Computer Engineering

2012年9月
September 2012

·软件技术与数据库·

文章编号：1000—3428(2012)17—0056—03 文献标识码：A 中图分类号：TP301.6

基于组合模型的局部搜索弱社团结构发现算法

叶 慧，李 旻

(华南师范大学计算机科学系，广州 510631)

摘 要：针对复杂网络社团结构发现算法中全局模块度存在的分辨率缺陷问题，即不能发现很多实际存在的小社团，甚至发现的社团不满足普通意义上的社团定义，给出一种新型的多目标整数规划模型。结合弱社团定义、局部适应度和全局模块度标准，提出一种高效的启发式算法，发现网络的层次重叠社团。实验结果表明，该算法克服全局模块度的缺陷，能充分挖掘出小社团，具有较高的效率。

关键词：复杂网络；弱社团结构；全局模块度；局部适应度；多目标整数规划

Local Searching Weak Community Structure Discovery Algorithm Based on Combinatorial Model

YE Hui, LI Min

(Dept. of Computer Science, South China Normal University, Guangzhou 510631, China)

【Abstract】 Existing complex network community algorithms mostly take the global modularity as a criterion of searching the best community structure. However, it is revealed to suffer a resolution limit that may fail to discover small known qualified communities and discover some unqualified communities. Aiming at this problem, combining weak community definition, local fitness and global modularity, this paper presents a new multi-objective integer programming model and an efficient heuristic algorithm. It successfully discovers networks' hierarchical and overlapping community structure. Experimental results show that the algorithm overcomes the disadvantages and fully discovers small communities with high efficiency.

【Key words】 complex network; weak community structure; global modularity; local fitness; multi-objective integer programming

DOI: 10.3969/j.issn.1000-3428.2012.17.016

1 概述

随着信息技术的发展，人们对数据研究从着眼于孤立个体逐步发展到着眼于群体，乃至对网络的研究。复杂网络[1]作为真实复杂系统的一种拓扑抽象，反映了真实系统的内在结构、演化方向和行为规律，是研究真实复杂系统的重要工具。为了研究复杂网络，学者们通过大量实验发现了复杂网络具有小世界和无标度2个重要的统计特性。2002年，Girvan和Newman提出复杂网络除了小世界和无标度特性外，还具有社区结构的特征[2]，引起了学术界的广泛关注。社区是由网络中带有相似属性的一组节点的集合，它具有社区内节点关系紧密和社区间节点关系松散的特点。

2004年，Newman和Girvan提出了度量网络社区质量的标准——全局模块度[3]，基于GN算法的改进[4]或全局模块度的变体[5](扩展到有向或加权网络)产生了很多新的算法，全局模块度也因此成为被普遍认同的衡量社区结构的标准。文献[6]提出全局模块度存在分辨率问题，即较小规模的社团不能够被检测出来，而且探测出的部分社团不满足社团的定义。这种分辨率问题并不取决于特定的网络结构，而是由在模块度的定义中将相互连接的社团间的连接边数和整个网络的总边数进行比较造成的。同样，表达式类似于模块度指标的其他评判指标本质上也可能具有这种分辨率的问题。局部适应度[7]基于社团是一个局部结构的概念，一个节点不知道整个网络的规模，节点形成社团是基于网络的局部信息。局部适应度对层次重叠社团研究发挥了巨大的作用，且大大提高了算法速度，但目前还没有全局模块度那么被广泛接受。

针对以上2种社团衡量标准的优缺点，本文提出了一种多目标整数规划模型，探测尽可能多的小社团和获得尽可能大的全局模块度。根据已有社团探测算法存在的一些问题，同时考虑全局模块度、局部适应度和弱社团定义[8]，并针对新模型提出了一种启发式算法实现层次重叠社团的探测，弥补了以上方法的各种缺陷。

2 多目标整数规划模型

无向网络 $G=(V,E)$，邻接矩阵 $\boldsymbol{A}=[a_{ij}]$，V_S 是 V 的一个子集。n 是节点的个数，同时也是社团的最大个数，L

作者简介：叶 慧(1983－)，女，硕士研究生，主研方向：复杂网络，数据挖掘；李 旻，工程师、硕士研究生

收稿日期：2011-10-24 **修回日期：**2011-12-12 **E-mail：**yehu124@qq.com

是边的数量。

定义 1 全局模块度

$$Q=\sum_j(A_j-k_ik_j/2L)\delta(C_i,C_j)/2L$$

其中，k_i和k_j是节点的度值；C_i是节点 i 所属社团。当$C_i=C_j$时$\delta(C_i,C_j)=1$，否则为 0。

z_{lk}表示边e_l是否属于社团C_k，$l=1,2,\cdots,L$，$k=1,2,\cdots,n$，$e_l=(v_i,v_j)$表示连接点v_i和v_j的边，x_{ik}是 0、1 变量，表示点v_i是否属于社团C_k。社团应满足以下约束条件：

$$z_{lk}\leqslant x_{ik},\ z_{lk}\leqslant x_{jk},\ x_{ik}+x_{jk}-1\leqslant z_{lk}$$

y_k是 0、1 变量，表示社团C_k是否为空。社团应满足以下约束条件：

$$y_k\leqslant\sum_{i=1}^{n}x_{ik}\leqslant ny_k$$

定义 2 弱社团结构

如果子网络V_S满足$\sum_{i\in V_S}k_i^{in}(V_S)>\sum_{i\in V_S}k_i^{\text{out}}(V_S)$，即社团内部节点间的相互连接比这些节点与社团外部的节点的联系更加紧密，则称V_S为该网络的弱社团结构。社团应满足以下约束条件：

$$2\sum_{l=1}^{L}z_{lk}\geqslant\sum_{j=1}^{n}\sum_{i=1}^{n}x_{ik}a_{ij}-2\sum_{l=1}^{L}z_{lk}+y_k$$

社团探测的多目标整数规划模型如下：

$$\max\sum_{k=1}^{n}y_k$$

$$\max Q$$

$$\text{s.t.}\ \sum_{k=1}^{n}x_{ik}=1$$

$$z_{lk}\leqslant x_{ik}$$

$$z_{lk}\leqslant x_{jk}$$

$$x_{ik}+x_{jk}-1\leqslant z_{lk}$$

$$y_k\leqslant\sum_{i=1}^{n}x_{ik}\leqslant ny_k$$

$$2\sum_{l=1}^{L}z_{lk}\geqslant\sum_{j=1}^{n}\sum_{i=1}^{n}x_{ik}a_{ij}-2\sum_{l=1}^{L}z_{lk}+y_k$$

$$x_{ik}\in\{0,1\},\ y_k\in\{0,1\},\ z_{lk}\in\{0,1\}$$

$$i=1,2,\cdots,n;\ k=1,2,\cdots,n;\ l=1,2,\cdots,L$$

3 局部搜索弱社团结构算法

多目标整数规划已经被证明是 NP 难问题，精确算法的时间复杂度是指数时间，不适用大型网络。由于全局模块度有分辨率缺陷，不能检测出小社团，单独应用全局模块度或局部适应度的算法得到的社团都比较大。弱社团结构是一种相对较小的社团的定义，也就是说网络中满足弱社团结构的社团个数比较多，同时弱社团结构也是一种有物理意义且被广泛认可的社团定义。本文针对全局模块度存在的分辨率问题，结合弱社团结构定义，采用局部适应度函数标准探测每一个弱社团结构，利用问题的特殊性对这个多目标整数规划模型设计了一个启发式算法，可以成功发现复杂网络的层次重叠社团。

定义 3 社团 g 的局部适应度

$$f_g=k_{in}^g/(k_{in}^g+k_{\text{out}}^g)$$

其中，k_{in}^g、k_{out}^g是社团内部、外部适应度之和。

定义 4 节点 i 的适应度

$$f_g^i=f_{\{g+i\}}-f_{\{g-i\}}$$

其中，$f_{\{g+i\}}$、$f_{\{g-i\}}$是社团$\{g+i\}$($\{g-i\}$)的适应度函数值，反映了节点 i 加入(或被移除)社团 g 后该社团的适应度。

定义 5 社团间的归一化边权

$$c_{ij}=k_{ij}/\min(k^i,k^j)$$

其中，k_{ij}为社团C_i和社团C_j之间的连接边权和；k^i和k^j分别为社团C_i和社团C_j节点度和。

启发式算法设计的准则是获得尽可能多的小社团和更大的模块度。利用贪心算法的思想，不妨假设社团通常以度大的节点为中心节点，蔓延开来形成社团。从度大的节点出发，利用局部适应度标准寻找新节点加入社团，直到新社团满足弱社团结构定义停止，在余下的节点中重复以上过程。划分好的社团作为新网络的节点，社团间的归一化边权作为节点间的连接。在新网络上重复探测社团，直到最后生成一个社团。

通常，满足弱社团结构的社团都比较小，这样可以获得尽可能多的小社团。不同层次的社团划分会得到不同的模块度，最顶层的社团划分(即网络划分为 1 个社团)，模块度达到最大值；最底层的社团划分(即社团数量最多)对应模块度最小。根据实际网络的特性，选择一个合适层次的划分作为最后的结果。

QLF 算法具体描述如下：

输入 无向网络$G=(V,E)$

输出 层次重叠社团

(1)初始化：已经找到社团的节点集合$C_{\text{done}}=\Phi$，还没有找到社团的节点集合$V_s=V$，第 i 个社团$C_i=\Phi$，$i=1$。

(2)找具有最大度的节点v，$v\in V_s$，加入节点v到社团C_i，$C_i=C_i\cup\{v\}$，$V_s=V_s-\{v\}$，$C_{\text{done}}=C_{\text{done}}\cup\{v\}$。

(3)找C_i的邻居节点v'，v'有最大的局部适应度，$v'\in V$，加入节点v'到社团C_i，$C_i=C_i\cup\{v'\}$，$C_{\text{done}}=C_{\text{done}}\cup\{v'\}$，$V_s=V_s-\{v'\}$。

(4)如C_i不满足弱社团定义，转步骤(3)；否则$i=i+1$，$C_i=\Phi$，转步骤(2)开始下一个社团的探测。如果$V_s=\Phi$，即所有节点都找到了社团，转步骤(5)。

(5)如果$i>1$，即探测到的社团不止一个，划分好的社团作为新网络的节点，$V=\{1,2,\cdots,i\}$，新网络的连边按照归一化边权定义计算，转步骤(1)。如果$i=1$，即网络所有节点都划分为一个社团(最高层次)，算法结束。

通过上述分析，寻找度最大节点所用的时间复杂度为$O(n)$，n为网络中节点的个数；寻找邻居节点的适应度所用的时间复杂度为$O(mn)$，m为与已知点相连接的节点个数。所以，最底层的社团划分需要的时间复杂度为$O(n+mn)$。假设网络划分了 k 层，整个算法的时间复杂度为 $O(k(n+mn))$，$k<<n$，对于稀疏网络，$m<<n$，所以，算法的实际时间复杂度接近于线性时间。

4 实验与分析

4.1 19 个节点组成的三社团网络

以 19 个节点构成的三社团网络为例，该网络包含 19 个节点和 37 条边，有非常明显的社团结构，可以分成 3 个社团。根据本文提出的算法，社团最底层划分后的结果如图 1 所示，社团所有层次划分如图 2 所示。

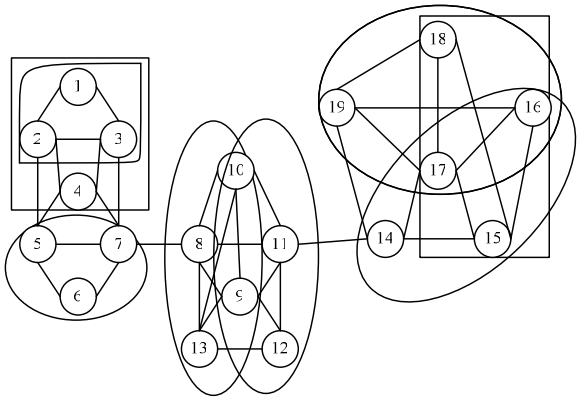

图 1 社团最底层划分结果

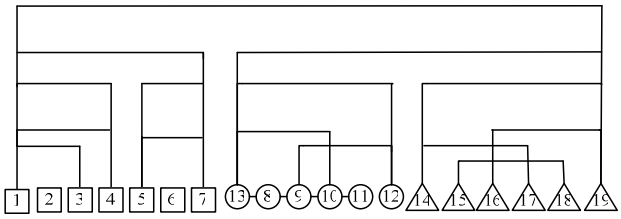

图 2 社团划分的层次树状图

如图 1 所示，社团第 1 层次划分出了 8 个小社团，通常算法都是直接划分出 3 个社团，QLF 算法成功探测出小社团，避免了模块度 Q 的分辨率问题。如图 2 所示，19 个节点的三社团划分出了 4 个层次(不同线宽代表不同的层次)，第 2 层划分为 3 个社团，对应正确的社团结构。

4.2 Zachary 空手道俱乐部网络

20 世纪 70 年代初期，Wayne Zachary 构造了美国一所大学空手道俱乐部成员间的社会关系网。整个网络由 34 个节点和 78 条边组成，节点代表俱乐部的成员，边代表成员之间的关系，如图 3 所示。

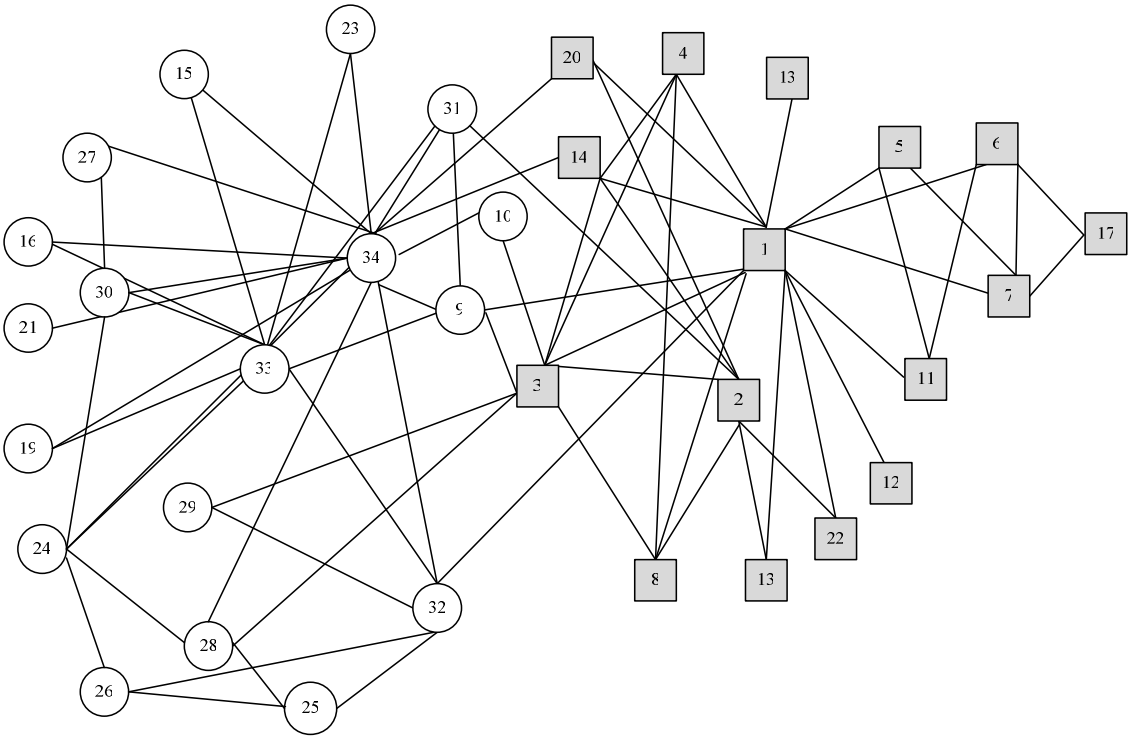

图 3 空手道俱乐部内部成员的关系网络

在调查过程中，该俱乐部的主管与校长之间因是否抬高俱乐部收费的问题产生了争执。结果该俱乐部分裂成了 2 个分别以主管和校长为核心的小俱乐部。图 3 中节点 1 和节点 33 分别代表了俱乐部主管和校长，而方形和圆形的节点也分别代表了分裂后的小俱乐部中的各个成员。

利用 QLF 算法得到的最底层划分结果如表 1 所示。如图 4 所示呈现的是 zachary 网络的层次树状图。第 3 层次把网络划分为 2 个社团，有 2 个重叠节点{10,29}，与图 3 正确结果相比，社团划分的正确率是 100%。第 2 层的社团划分全局模块度为 0.32，社团的平均局部适应度为 0.81。

表 1 zachary 网络最底层划分结果

社团编号	包含的节点
1	{10,15,16,19,21,33,34}
2	{1,2,4,8,12,13,14}
3	{2,3,4,8,10,14,29}
4	{25，26,29,32}
5	{9,15,16,19,31,33,34}
6	{24,25,26,28}
7	{6,7,17}
8	{24,25,26,27,28,30}
9	{5,6,7,11}
10	{1,2,4,8,18,20,22}
11	{15,16,19,21,23,33,34}

(下转第 62 页)