PROGRESS IN COMPLEXITY SCIENCE

复杂性科学进展

主　编　周　涛　蒋　晓
副主编　汪小帆　史定华　汪秉宏　吕琳媛　荣智海

电子科技大学出版社

PROGRESS IN COMPLEXITY SCIENCE

复杂性科学进展

主 编 周 涛 蒋 晓
副主编 汪小帆 史定华 汪秉宏 吕琳媛 荣智海

电子科技大学出版社

图书在版编目（CIP）数据

复杂性科学进展 / 周涛，蒋晓主编. —成都：
电子科技大学出版社，2015.3
ISBN 978-7-5647-2884-7

Ⅰ. ①复… Ⅱ. ①周… ②蒋… Ⅲ. ①复杂性理论—
文集 Ⅳ. ①TP301.5-53

中国版本图书馆 CIP 数据核字（2015）第 059243 号

复杂性科学进展

主　编　周　涛　蒋　晓
副主编　汪小帆　史定华
　　汪秉宏　吕琳媛　荣智海

出　　版：电子科技大学出版社（成都市一环路东一段 159 号电子信息产业大厦　邮编：610051）
策划编辑：张　琴　汤云辉
责任编辑：汤云辉
主　　页：www.uestcp.com.cn
电子邮箱：uestcp@uestcp.com.cn
发　　行：新华书店经销
印　　刷：成都蜀通印务有限责任公司
成品尺寸：185mm×260mm　　印张 24　　字数 642 千字
版　　次：2015 年 8 月第一版
印　　次：2015 年 8 月第一次印刷
书　　号：ISBN 978-7-5647-2884-7
定　　价：78.00 元

复杂性科学：二十一世纪的中国贡献？

各位读者拿到手中这本沉甸甸的文集，是电子科技大学学报“复杂性科学专栏”五年来所发表的论文的精粹。参与建设学报的“复杂性科学专栏”，是我来到电子科技大学之后的第一件具有相当持续性的学术活动——这个文集可以看成是一个阶段性的小结。追溯到专栏的构思和运营，有两个人是我特别要感谢的：一是上一届学报编辑部的主任周小佳先生，他创造性地提出了建设特色专栏的想法，而且在电子科技大学这样一个以电子信息为主要方向的工科主导型的大学，首先推出了“复杂性科学”这个和我们主流研究似乎风马牛不相及的方向；二是现在电子科技大学学报的主编蒋晓女士，她一直管理和运营“复杂性科学专栏”，保证了这个专栏每一个细节的高质量——如同她本人一般地美丽精致。

“复杂性科学专栏”已经获得了不俗的战绩。专栏发表文章的数量不到《电子科技大学学报》的8%，但是却贡献了超过1/3的引用，单篇引用率是非专栏文章的5倍以上。专栏论文平均每篇被引用约7次，已经达到甚至超过了中文顶尖期刊的平均引用情况。专栏还吸引了相关领域大量国内著名的教授，包括汪小帆、史定华、汪秉宏、陆君安、陈关荣、李翔、狄增如、樊瑛、方锦清、刘宗华、吕金虎等等；以及相关领域一大批国内非常活跃的青年科学家，包括吕琳媛、闫小勇、荣智海、唐明、王文旭、吴枝喜、李大庆、韩筱璞、闫强、许小可、王璞、张子柯、严钢、刘建国、刘润然、章忠志等等。

专栏之所以能够取得一定的成绩，我觉得有三个方面的原因。第一是服务到位，我们免除了版面费，提高了作者稿酬，缩短了发表时间，允许任何IP地址免费下载，根据图片质量和需求免费彩色印刷，对论文篇幅不做限制，等等。我自己2013年就在这个专栏上发表过一篇介绍人类动力学的研究进展的综述论文，长度为60页，是学报论文平均长度的十多倍。第二是宣传到位，我们建立了包含近千人的国内复杂性科学研究人员通讯录，将每期专栏的电子文档发给相关学者，经常组队参加国内复杂性科学领域的会议，并在会上将专栏的文章及宣传资料向与会专家赠阅，扩大专栏的影响力。这次出版文集，也是一个宣传和扩大影响力的尝试。第三是专栏文章的形式新颖，我们不仅有原创研究论文，还接受发表各种长度的综述和评述，包括对国内研究成果的回顾和未来重要研究方向的展望。

这次入选的32篇论文，有30篇论文是我们截止到2014年年底专栏文章中精选出来的，综合考虑了论文的相关性以及文章内容的长期影响力，因此倾向于选择综述和评述类的文章。还有两篇论文，是在专栏成立前发表的，包括汪小帆等人关于社团结构算法的综述和赵飞等人关于维基百科的综述。选中这两篇是因为从影响力和内容关联程度而言，都非常合适，而且发表时间和专栏的成立很接近。遴选专栏中优秀的论文并结集出版，这是第一次尝试，但肯定不是最后一次，所以非常希望各位读者、作者一如既往地支持我们。

回到复杂性科学研究本身。二十世纪复杂性科学的研究，出现了很多具有国际重大影响力的流派和理论，例如普利高津的耗散结构理论、托姆的突变论、哈肯的协同学、巴克的自组织临界理论等等，复杂系统的一些突出特征，例如非线性、不确定性、自组织性、涌现性等等，也获得了较广泛的认可。但是，由于复杂性科学主要处理的对象，是社会、

经济、生命等复杂异质系统，难以如理论物理一样发现简洁而又可以通过实验验证的“基本原理”，更不可能形成类似基础数学中的公理和定理。因此，尽管有一些有影响力的理论，但是这些理论在解释现实上仍然存在不足，或者是解释能力的边界依然不清楚，远远没有达到一统江湖的地位。事实上，如果说二十世纪后半叶是复杂性科学的战国时代，还有几个大流派如“战国七雄”，那么二十一世纪的复杂性科学倒退回了春秋早期，小邦弱国遍野，一时间看不到什么明确的理论或者学派可以给出复杂性科学的宏观叙事——复杂性科学领域的代表性学者都已经进入到疾病传播、城市交通、生物代谢网络、社交网络分析等等具体的分支中。从好的一面说，复杂性科学的理念和方法在一些具体的分支产生了相比二十世纪更明确的贡献；从坏的一面说，好像没有重要的科学力量在思考复杂性科学整体性的问题——或者说大家对于这方面的思考开始持有一种悲观的态度。

二十一世纪已经过去15%了，仅从已经产生的成果看，丝毫没有看出复杂性科学要产生什么惊天动地的结果，只是润物无声地在各个科学分支发挥具体的作用。但是，“数据化”浪潮的到来，使得很多点滴行为的数据都被忠实地记录了下来，而这些“中间过程”的数据往往是我们以前没有办法获取的——例如，我们很早以前通过问卷调查可以建立小规模的社交网络，这是一个结果性的数据；后来我们通过Facebook这样的系统，一下子得到很大的网络，但是往往无法得到每一条链接建立的时间；但是现在，我们不仅能够获得每一条链接建立的时间，还知道建立这个链接的时候，研究对象在看什么、在点击什么以及这个对象多年的购买记录和移动轨迹。于人如此，于人造的系统也如此，通过传感器网络和城市信息系统，我们可以知道一个城市某时某刻各个地方发生的细节。于生命系统也如此，可穿戴的传感设备可以实时监测到生命体的很多特征。观察从起点到终点的一切过程，这种“大数据”的能力，使得我们似乎被提升了一个维度，可以把原因和结果之间的所有细节铺呈开来。“大数据”或许是复杂性科学抱负得以实现的一把钥匙，能够把针对复杂性科学进行整体性研究的能力提升到一个新的阶段，从而有希望在二十一世纪的下面两个15%中产生一些可以持续10个20个15%的基石性的成果。

尽管二十世纪复杂性科学的主流理论已经式微，但是我们要牢牢记住，这些在科学历史舞台上曾经辉煌而又黯淡落幕的理论，并没有一个是源自中国的贡献。而我们中国的学者，不过是在舞台下看一幕幕的剧，甚至说，隔舞台都很远，不过是一边啃鸭掌，一边看着电视的录播。上面的那些感怀与愿景，也不过是鸭掌水平的感怀，因为在复杂性科学历史和现在的舞台上，还没有我们中国学者站立的地方。通过半个世纪的努力，或许我们中的一小部分人，已经拿到门票可以去看现场剧了，但是舞台下面“请保持安静”的标语，依然让作为观众的我们憋屈。或许，是我们要准备登上舞台的时候了！因为，中国的科技经济能力，已经可以支撑我们去思考重大的科学问题。我们要逐步清除以跟踪性研究为主要研究方法，以为主流学者抬轿子为主要学术贡献的思路，努力探索原创性的理论与方法。

霍金说“二十一世纪是复杂性的世纪”，汤恩比说“二十一世纪是中国人的世纪”。不知道“中国人”和“复杂性”能否在这个世纪剩下的时间中交相辉映。

以为序！

周　涛

目　录

复杂网络中的社团结构算法综述

汪小帆，刘亚冰

上海交通大学电子信息与电气工程学院 上海 闵行区 200240

【摘要】社团结构是复杂网络的一个极其重要的特性，网络社团结构挖掘在生物学、计算机科学和社会学等多个领域都具有很重要的意义。近年来，针对不同类型的大规模复杂网络，人们提出了很多寻找社团结构的算法。该文综述了该领域最新的比较有代表性的一些算法，重点分析了基于模块度指标的改进算法，能够体现社团层次性和重叠性的新算法，衡量社团划分算法好坏的基准图。最后展望了该领域的未来研究方向。

关 键 词 复杂网络；社团结构；层次性；模块度函数；重叠性

中图分类号 N941；TP311.13 **文献标识码** A

doi:10.3969/j.issn.1001-0548.2009.03.007

Overview of Algorithms for Detecting Community Structure in Complex Networks

WANG Xiao-fan and LIU Ya-bing

School of Electronics, Information & Electrical Engineering, Shanghai Jiao Tong University Minhang Shanghai 200240

Abstract Community structure is a very important property of complex networks. Detecting communities in networks is of great importance in biology, computer science, sociology and so on. In recent years, a lot of community discovery algorithms have been proposed aiming at different kinds of large scale complex networks. In this paper, we review some latest representative algorithms, focusing on the improved methods based on the modularity function, the algorithms which can detect overlapping and hierarchical community structure in networks, and the benchmark in detecting communities. Finally, some future directions are pointed out.

Key words complex network; community structure; hierarchical structure; modularity function; overlapping communities

网络中的社团结构[1]是指一组相互之间有着比较大的相似性而与网络中的其他部分有着很大不同的节点的群。也就是说，在社团内部，节点之间的联系非常紧密，而社团之间的联系相对而言比较稀疏。寻找社团结构并对其进行分析是了解现实生活中各种网络组织结构的一种很重要的方法，并在生物学、计算机科学以及社会学等领域都有着广泛的应用[2]。如社会网络中的社团结构使得人们能够清晰地了解他们区别于其他社团的一些特质或者信仰等；在生物分子反应网络中，聚合到一起形成功能性模块的节点往往担当特定的角色或具有特定的功能。寻找网络社团结构的算法有很多，一些经典算法，如图形分割经典问题、社会学中的聚类分析等可参考[1]。

基金项目：国家自然科学基金(60674045，60731160629)；上海市优秀学科带头人计划(07XD14017).

作者简介：汪小帆(1967－)，男，博士，长江学者特聘教授，主要从事复杂网络分析与控制研究方面的研究.

1 基于模块度指标的改进算法

社团模块度指标Q是用于刻画社团特性强弱的参数，定义如下[3]：

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j) \qquad (1)$$

式中，k_i和k_j是节点的度值；C_i是节点i所属社团；m是网络总边数。当$C_i=C_j$时$\delta(C_i,C_j)=1$，否则为0。Q值在0～1之间，一般以$Q=0.3$作为网络具有明显社团结构的下界。

模块度是应用最为广泛的评判社团特性强弱的指标。对其进行优化是个NP难题，且具体的算法时间复杂度比较高，不能应用于大规模的网络。近年来提出了很多基于模块度的改进算法，这些算法或将原算法拓展到各种有向、加权网络中，或能够体现社团的层次性和重叠性，或具有较小的时间复杂度，可用于大规模网络，或解决模块度优化中存在的分辨率问题。

1.1 分辨率问题

文献[4]对模块度优化问题进行了分析，指出了该方法存在的内在缺陷，即基于模块度的优化方法不能检测出小于一定规模的小社团，而这个规模的阈值取决于整个网络的规模以及模块内部的连接度值。对现实中大量的社会、生物网络等的测试结果表明，模块度优化算法不能探测到很多实际存在的社团，会遗失网络的一些小社团。这种分辨率问题并不取决于特定的网络结构，而是由于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问题。

1.2 模块度指标的拓展

文献[5]将式(1)定义拓展到了加权网络情形：式(1)中k_i和k_j分别用节点所有连接边权重S_i和S_j取代，m用所有边的总权重和W取代。

如果网络是有向的，那么其两种可能的方向的概率就取决于节点的入度和出度，故有向网络中的模块度可定义如下[6-7]：

$$Q_{\mathrm{d}}=\frac{1}{m}\sum_{ij}\left(A_{ij}-\frac{k_i^{\mathrm{out}}k_j^{\mathrm{in}}}{m}\right)\delta(C_i,C_j) \qquad (2)$$

式中，k_i^{in}和k_j^{out}分别表示节点i的入度和节点j的出度。对于一般的有向加权网络，文献[6]也提出了相应的模块度定义：

$$Q_{\mathrm{gen}}=\frac{1}{W}\sum_{ij}\left(W_{ij}-\frac{s_i^{\mathrm{out}}s_j^{\mathrm{in}}}{W}\right)\delta(C_i,C_j) \qquad (3)$$

对于具有重叠现象的网络，文献[8]最近提出了一种较简单的模块度定义：

$$Q=\frac{1}{2m}\sum_{ij}\frac{1}{O_iO_j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j) \qquad (4)$$

式中，O_i表示节点i所存在的社团的个数，一条边在社团中的贡献越小，包含这条边的两个端点的社团数目就越多。文献[9]对具有重叠性网络的模块度给出了更一般的定义。该定义考虑了网络的方向性，可以同时处理有向网络的数据，其表达式为：

$$Q_{\mathrm{ov}}=\frac{1}{m}\sum_{c=1}^{n_c}\sum_{i,j}\left[r_{ijc}A_{ij}-s_{ijc}\frac{k_i^{\mathrm{out}}k_j^{\mathrm{in}}}{m}\right] \qquad (5)$$

式中，r_{ijc}和s_{ijc}分别表示边i、j在网络中贡献的总和与相应空模型中边的贡献。若社团间没有重叠现象，则$r_{ijc}=s_{ijc}=\delta_{c_ic_jc}$，其中$c_i$和$c_j$分别表示节点$i$和$j$所在社团的编号。

上述模块度定义并不适用于具有负权重边的网络。文献[10]在社团的定义中引入了负相关性或者不相容性，当节点之间的连接是基于不相似性就称之为节点是负相关的；将正相关性和负相关性结合起来形成社团的定义，称之为双重相关性。文献[10]还提出了一种双重相关模型性检测方法(DAMM)，对正负权重的边都进行考虑，从而拓展了社团算法所能分析的网络范围，并引入一种称为公共重叠的方法，可以用来测

量社团划分结果的精度。文献[11-12]也给出了类似的模块度定义。

2 基于层次性和重叠性的算法

2.1 层次性

网络中的节点可能具有不同层次的组织结构，如大社团内部可能含有较小规模的社团，以此类推。在这种情况下，就称该网络具有层次性社团结构。

文献[13]较早提出了层次性聚类算法，这种聚类技术可以体现出图的多层次结构，在社会网络分析、生物网络等领域都有应用。文献[13]指出，层次聚类方法在一个网络不具有层次性时，也有可能会得到一种层次化的结果。此外，社团中的节点有可能并没有被正确地划分，而且某些在模块中具有关键性作用的节点或边也可能会丢失掉，这种情况在处理大型网络数据时较为明显。为此，文献[14]提出了一种自上而下的分裂算法，它可以区分出不具有社团特性的网络，或具有社团特性但不具有层次性的网络，或具有层次性社团的网络。但由于其计算量大，该算法在处理大型数据时并不适用。

文献[15]提出了一种能够探测到层次化社团结构的凝聚算法BGLL，该算法可分为两个阶段。

第一阶段，首先进行社团初始化，网络中的每个节点都分配一个社团编号，这样每个节点都被看作是一个社团。然后，对于任意节点 i、j，根据式(2)计算当节点 i 加入到它的每个邻居节点 j 所在的社团时，对应社团模块度的增量 ΔQ：

$$\Delta Q=\left[\frac{\sum_{\text{in}}+k_{i,\text{in}}}{2m}-\left(\frac{\sum_{\text{tot}}+k_i}{2m}\right)^2\right]-\left[\frac{\sum_{\text{in}}}{2m}-\left(\frac{\sum_{\text{tot}}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right] \tag{6}$$

式中，$\sum_{\text{in}}$ 是社团内部所有边的权重和；$\sum_{\text{tot}}$ 是所有与社团内部节点相关联的边的权重和；k_i 是所有与点 i 相关联的边的权重和；$k_{i,\text{in}}$ 是节点 i 与社团 C 相连接的所有边的权重和。

当 ΔQ 为正值时，选出对应最大值的那个邻居节点，把点 i 加入到该邻居节点所在的社团中；若所有 ΔQ 都为负值，则节点 i 留在初始社团中。这种社团的合并过程重复进行，直到整个网络不再出现合并现象时，划分出了第一层的社团。

第二阶段，首先构造一个新网络，该新网络的节点是第一阶段探测出的各个社团，节点之间连边的权重是两个社团之间所有连边的权重和。然后，用第一阶段中的算法再次对该新网络进行社团划分，得到第二层的社团结构。以此类推，直到不能再划分出更高一层的社团结构为止。BGLL算法本身就能够生成一种层次性的社团结构。

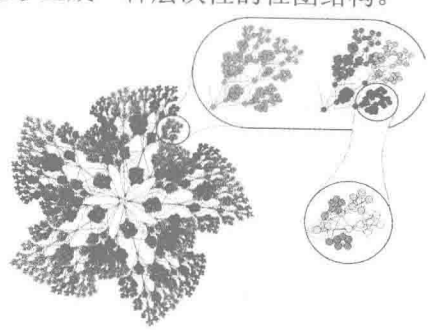


图1 具有约2万节点的网络的层次性社团

图1是应用BGLL算法对一个包含大约2万个节点的网络进行层次性社团结构划分得到的结果。BGLL算法有以下优点：(1) 计算速度很快，可用于大规模网络。(2) 是一种自下而上的凝聚过程，不会出现对小规模社团的探测遗漏现象，即解决了分辨率问题。(3) 可应用于大规模的加权网络。

该算法是基于模块度的改进算法，本文在其基础上又进行了改进，使其可以同时实

现社团的层次性并寻找到最具有重叠性倾向的节点。将一个节点 k 从它现在所属的社团 A 移到另一个社团 B 所引起的 Q 值变化越小，则把 k 划在 A 或划在 B 对整体社团划分的优劣影响越小，因此可以说节点 k 同时属于社团 A 和 B，是具有重叠性的节点。这就是在BGLL层次性算法基础上再寻找社团间重叠节点的主要思想。

本文算法的主要流程如下：首先，选中网络中的任一节点 i，然后将它移出目前所在的社团。然后，对于每个和这个节点有连边的社团 C，都计算出将这个节点加入社团 C 的模块度增益：

$$d_{\mathrm{q}}=\frac{1}{2m}\left(n_{i,C}-\frac{k_i k_C}{2m}\right) \tag{7}$$

式中，$n_{i,C}$ 是节点 i 和社团 C 之间的所有连边的权值和。用式(7)可算出将节点加入其他社团的模块度变化，然后将这个变化和将节点保留在原社团的模块度变化进行比较，如果差值小于某个限度，就认为这个节点具有重叠性。至于这个限度如何选取，一般认为节点加入两个不同社团的 d_{q} 的差值不超过 $1/2m$，就可认为该节点具有重叠性。当然还可以对该限度进行调整从而得到不同精确度的结果。

2.2 重叠性

社团结构的一个重要特征就是具有重叠性[16]，它是指网络中存在一些“骑墙节点”，它们同时被多个社团包含，属于这些社团的交叉部分。在真实网络中，社团结构的重叠性十分明显。已有的大部分算法都只能得到标准化划分，即一个节点只属于一个社团。然而节点通常都是属于多个社团的，现实网络也是由许多彼此重叠互相关联的社团构成。因此重叠性社团成为近年来研究热点，下面探讨寻找重叠性社团的算法。

2.2.1 基于派系过滤算法的改进算法

文献[16]提出的派系过滤算法(CMP)可以用来分析具有重叠性的社团结构。在此基础上文献[17-18]又对CMP进行拓展，提出了CMPw和CMPd两种算法，将模块度指标分别拓展到加权和有向网络中。

文献[17]针对加权网络提出了一种带有权重的派系过滤算法，即CMPw，该算法主要是基于拥有较高子图强度的 k-派系过滤的概念。首先给出 k 派系子图 C 的强度 $I(C)$ 定义，它是子图 C 中所有边的连接强度的几何平均数。k 派系子图 C 在其两个节点 i、j 之间有 $k(k-1)/2$ 条边，并且其强度定义如下：

$$I(C)=\left(\prod_{\substack{i<j\\ i,j\in C}} w_{ij}\right)^{2/k(k-1)} \tag{8}$$

在加权网络中，CMP是通过去除小于一个固定的权重临界值 W 的弱连接并将其他保留的连接看作无权重的网络来寻找社团，因此社团中的所有连接权值都高于 W。CPMw 则是当 k-派系的强度大于固定临界值 I 时定义成为一个社团，但社团中经常会包含连接强度小于 I 的连接边。

当加权网络中具有强连接的边倾向与强连接的边相连时，这两种算法的结果相似。但如果在CPM 中所去除的弱连接的边是由系统本身的高噪音所导致的，这时设置 $I=W$，CPMw 方法就能够得到比原来方法更好的结果。

文献[18]通过拓展CPM方法又提出了寻找有向网络中社团的算法。为了比较社团中的各个节点，引进了相对出度(相对入度)的定义，也就是一个节点指向(被指向)社团中的其他所有节点的连接的相对权重。社团 α 中节点 i 的相对入度与相对出度的定义如下：

$$D^{\alpha}_{i,\mathrm{in}}=\frac{d^{\alpha}_{i,\mathrm{in}}}{d^{\alpha}_{i,\mathrm{in}}+d^{\alpha}_{i,\mathrm{out}}},\quad D^{\alpha}_{i,\mathrm{out}}=\frac{d^{\alpha}_{i,\mathrm{out}}}{d^{\alpha}_{i,\mathrm{in}}+d^{\alpha}_{i,\mathrm{out}}} \tag{9}$$

式中，$d^{\alpha}_{i,\mathrm{in}}$ 与 $d^{\alpha}_{i,\mathrm{out}}$ 分别表示指向该节点的邻

居数目和该节点所指向的邻居数目。在加权网络中，只需将相应的度值改为边权值之和即可。

文献[16]同时又给出了有向 k 派系的定义，该定义具有很强的限制性。该算法具有很大的改善空间，通过给出更合适的定义对于寻找有向网络中的社团有可能得到更好的结果。

2.2.2 基于GN算法的改进算法

文献[19]同样提出了一种用于探测社团结构重叠性的算法CONGA。CONGA算法是在传统的GN算法[20]的基础上改进实现的。GN算法通过不断地移除网络中的边来实现社团的凝聚；而CONGA算法在此基础上增加了一个节点分裂过程，即假定节点 i 同时属于 n 个社团，将节点 i 复制 n 个，分别放入这 n 个社团中，使得节点在社团形成过程中能够同时被多个社团包含，从而实现了重叠性节点的探测任务。

2.3 层次性和重叠性同时体现的算法

2.3.1 一种基于适应度函数局部最优化的方法

下面介绍一种既可以找到重叠性社团又可以发现层次性结构的LFM算法[21]。这种基于对一个适应度函数局部最优化的方法速度很快，可分析多达数百万节点的大规模网络。该算法是基于极值优化的思想进行社团结构划分的代表算法之一。在LFM算法中，社团的划分通过对如下定义的适应度函数 f_G 进行优化取最大值来实现：

$$f_G = \frac{k_{\text{in}}^G}{(k_{\text{in}}^G + k_{\text{out}}^G)^\alpha} \tag{10}$$

式中，k_{in}^G（k_{out}^G）是社团 G 内(外)部包含的边的权重之和；α 是一个控制社团规模的参数，也称为解参数。对于网络中的任意一个节点，定义点 i 对社团 G 的适应度函数 f_G^i 为：

$$f_G^i = f_{G+i} - f_{G-i} \tag{11}$$

式中，$f_{G+i}(f_{G-i})$ 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。当 $f_G^i > 0$，说明点 i 加入社团 G 能使其适应度函数增大，因而应当被包含在社团 G 中；当 $f_G^i < 0$，则点 i 应从社团 G 中移除。

社团划分的具体过程可以通过以下步骤完成：(1) 初始化，随机选择一个孤立节点作为社团 G 的初始成员，$k_{in}^G = 0$。(2) 根据式(10)计算社团 G 的所有邻居节点对 G 的适应度贡献。(3) 选出适应度函数值最大的邻居节点，把它加到社团 G 中，得到新的社团 G'。(4) 根据式(11)重新计算社团 G' 中所有节点对社团 G' 的适应度贡献。(5) 选出适应度函数值为负的节点，将其从社团 G' 中删除，得到新的社团 G''。(6) 如果步骤(5)发生，则返回步骤(4)。如果所有节点的适应度贡献值都为正，则返回步骤(2)。上述过程循环执行，直到社团 G 的所有邻居节点对 G 的适应度贡献值都为负值时停止。此时社团 G 的适应度函数达到了最大值，完成了第一个社团的探测。然后继续选取孤立节点重复以上过程，直到网络中所有节点都已经被划分到至少一个社团为止。在这个过程中，有部分节点会被划分到不止一个社团中去，这就是所谓的“骑墙节点”，体现了社团结构的重叠性。

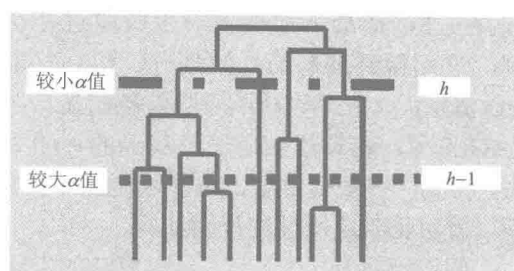


图2 解参数对应网络的层次性

图2是网络的树状图，较小 α 值能把网络划分为一些规模较大的社团，较大 α 则对应数量较多的小社团。通过选择一系列 α 值对网络进行社团划分就能够得到社团结构

的层次性。

该算法的时间复杂度主要取决于社团的大小和其重叠的程度。最坏情况下，即当社团规模和节点数 n 在同一数量级时，时间复杂度为 $O(n^2)$，但大部分情况算法会更快。该方法可拓展到加权网络，此时适应度公式中将边数之和替换成边权重和即可。

2.3.2 一种基于凝聚算法的新算法EAGLE

EAGLE也是一种能够同时探测出社团层次性和重叠性的算法[22]。它通过凝聚的方法来划分社团，但它与传统的凝聚算法在研究对象上有着很大的区别。传统凝聚算法的研究对象是网络中的节点，通过节点之间的不断凝聚来实现社团划分，而EAGLE的研究对象则是网络中的极大团，通过极大团之间的不断凝聚来实现社团划分。极大团是指网络中不能再被分割为子团体的最大节点集，EAGLE选择了Bron-Kerbosch算法[23]来寻找网络的极大团。

EAGLE算法分为两个步骤：(1) 生成网络的树状图；(2) 在生成树上选择合适位置断开，得到相应的社团划分。为了评判划分结果的优劣，该算法提出了一种新的模块度指标：

$$\mathrm{EQ}=\frac{1}{2m}\sum_{i}\sum_{v\in C,w\in C}\frac{1}{O_vO_w}\left[A_{vw}-\frac{k_vk_w}{2m}\right] \quad (12)$$

式中，Q_v 是节点 v 所属的社团的数目。通常选择在EQ值最大的位置对生成树进行切割，进而得到理想的社团划分。仿真发现EQ值越大，社团结构的重叠性越明显。在EAGLE中，将算法应用于已找到的社团中去，得到一些规模更小、联系更紧密的子社团，就可得到层次性的社团结构。

假设网络有 n 个节点，经过算法的步骤(1)，可以划分出 s 个极大团，h 是相邻的极大团(即有连边的两个极大团)的对数。算法步骤(1)的时间复杂度为 $O(n^2+(h+s)s)$，步骤(2)的时间复杂度为 $O(n^2s)$。在算法的步骤(1)中，寻找网络的极大团本身就是一个NP问题，但由于实际网络的稀疏性，这个过程还是很快的。

2.3.3 一种从边的角度考虑同时体现重叠性和层次性的算法

传统的社团结构划分都是从节点的角度出发，把网络中的节点看作是研究对象，根据节点之间的相似度，把它们划分成一个个的社团，这样的处理方法很难解决“骑墙节点”的归属问题。文献[24]针对传统观点的不足，提出了从边的角度出发，按照边之间的相似度对网络进行社团划分，避免了“骑墙节点”对结果的影响。这是因为网络中边的社团归属性是唯一确定的，只能被一个社团所包含。该算法基本思想是根据边的凝聚过程得到网络的层次树结构，在合适位置对生成树进行切割，从而得到网络的社团结构。为了实现边的凝聚过程，首先定义相邻边的相似度 S。

对于相邻边 e_{ik} 和 e_{jk}，若它们同时连接到公共节点 k 上，则称节点 k 为基节点，节点 i、j 称为关联节点。对于任意节点 i，定义节点 i 的广泛邻居节点为 $n_+(x)$，则：

$$n_+(i)\equiv\{x\,|\,d(i,x)\leqslant 1\} \quad (13)$$

式中，$d(i,x)$ 为节点 i 和 x 之间的最短路径长度。边的相似度 S 的定义如下：

$$S(e_{ik},e_{jk})=\frac{\left|n_+(i)\cap n_+(j)\right|}{\left|n_+(i)\cup n_+(j)\right|} \quad (14)$$

然后就可以进行边的凝聚过程。首先，进行社团的初始化，将每一条边看作是一个社团。然后，计算相邻社团之间的边相似度，选出相似度最大的两个社团，将它们合并。凝聚过程反复执行，直到整个网络凝聚成一个社团为止。经过上述步骤后就可以得到网络的层次树状图。

文献[24]还定义了一个新的评判标准——分区密度 D，在层次树状图中选择分区密度达到最大值的位置，对其进行切割，得到社团划分。假定 $P=\{P_1,P_2,\cdots,P_C\}$ 是对包含有 E 条边的网络进行社团划分的结果，网

络被分为 C 个社团。第 c^{th} 社团包含的边的个数为 m_c，$m_c=|P_c|$。这些边所覆盖的节点为 n_c，则社团 c 的分区密度定义如下：

$$D=\frac{2}{M}\sum_c m_c\frac{m_c-(n_c-1)}{(n_c-1)(n_c-2)} \tag{15}$$

如图3所示，其中虚线标注的最佳划分对应着分区密度 D 最大值。根据这样的方法得到的社团结构，与从节点角度出发进行社团划分相比，更容易发现社团的重叠部分。社团结构重叠性的探测过程转化为一个极值优化问题。

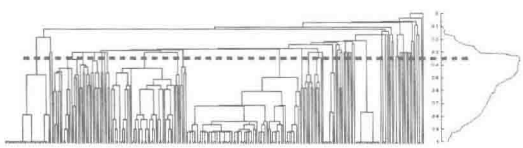


图3 边的层次性划分结果

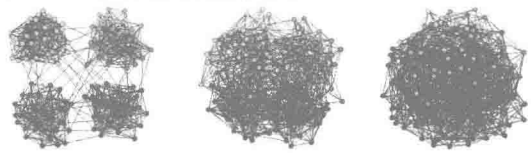
文献[25]也同样从边的角度提出了划分社团的思想，引入了一系列类似于模块度指标的质量函数并采用了文献[15]提出的寻找层次性技术以及文献[26]提出的多层次性算法，但同时改进了所采用的模块度指标使之能够包含更大步长的随机游走，通过调节解参数即随机游走的步长可以得到更精确的划分结果。

正如上文所阐述的，在寻找社团过程中重叠性与层次性存在着“冲突”，到目前为止，能够同时解决社团这两大特性的算法并不多，这一问题也值得继续研究。

3 评判基准图

近年来涌现了大量的寻找社团的算法，这就带来一个问题：这些算法好不好，其结果能否反映网络中真实的社团结构？当要处理某一个网络数据时，选择哪一个算法更好？为此，需要有一些已知其社团特性的基准图(benchmark graphs)用于检验和比较算法的性能。

文献[27]在人工 l-分割模型基础上提出了一种特例[28]，规定 $l=4$，$g=32$，则 $n=128$，同时规定 $\langle k\rangle=16$。这一类网络模型也已经成为最为广泛使用的标准化基准图。当 $z_{\text{out}}<8$ 时构造的网络具有较明显的社团特性，故 $z_{\text{in}}=z_{\text{out}}=8$ 可看作该基准图是否具有明显社团特征的阀值。图4是三种不同参数下的基准图网络，其中图4c中的4个社团已经不容易看出来了。



a. $z_{\text{in}}=15$　　b. $z_{\text{in}}=11$　　c. $z_{\text{in}}=8$

图4 3种不同参数下的基准图网络

很多算法在测试 z_{out} 较小的基准图时能够给出很好的结果，但是当基准图中的 z_{out} 超过8时就不能准确地检测出其中的社团。文献[29]又将文献[28]的基准图拓展到了加权网络，给社团内部和社团之间的边分配不同的权重。

人工 l-分割模型存在局限性，即所有节点必须具有相同度值，社团也都是相同规模，而这与现实网络的特性是不符合的，因此如何在模型中考虑到节点度和社团规模的异质性是一个很重要的挑战。文献[30]提出了一种新的模型，其构造基准图的步骤如下：

(1) 每一个节点给定一个度值，保证节点度分布服从指数为 γ ($2\leqslant\gamma\leqslant3$)的幂律分布，选定分布的两个极值 $k_{\min}$ 和 $k_{\max}$ 从而保证整个网络的节点平均度为 $\langle k\rangle$。(2) 每一个节点以概率 μ 与所在社团中其他节点相连，以概率 $1-\mu$ 与其他社团中的节点相连($0\leqslant\mu\leqslant1$被称为混合参数)。(3) 社团规模服从指数为 β ($1\leqslant\beta\leqslant2$)的幂律分布，且所有社团大小的总和等于图中节点数 N，选定社团的最大和最小规模以满足在社团定义中所加的约束：$s_{\min}>k_{\min}$ 和 $s_{\max}>k_{\max}$，从而保证任何一种度值的节点都可被至少一个社团包含。(4) 开始时，节点都是孤立

的，不属于任何社团。首先，一个节点加入一个随机选择的社团中，若该社团大小超过了节点的度，即在该社团内部其邻居节点的数目，则该节点确定加入该社团，否则就不加入。依此类推反复迭代，不断将孤立节点放入一个随机选择的社团中，直到没有孤立节点为止。(5) 为了加强混合参数μ对内部邻居节点的影响，进行随机重连，即保证所有节点的度不变，仅改变社团内部和外部的度值。

图5为具有500个节点的基准图实现结果。

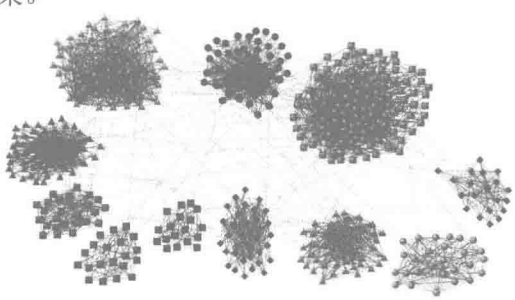


图5 具有500个节点的基准图实现结果

文献[31]将该新的基准图拓展成为有向加权网络，并同时能够包含重叠性节点，提出了一种更一般的模型，该模型同样考虑了社团间的重叠现象。

一些常用的已知社团结构的现实网络包括空手道俱乐部成员关系网络(34个节点，78条边)，海豚家族关系网络(62个节点，159条边)，美国大学生足球俱乐部网络(115个节点，616条边)等。

4 未来研究方向及展望

尽管近年来涌现了很多寻找不同网络类型的社团结构的算法，但是还存在许多深入的问题。首先，到目前为止对社团结构仍然缺乏一个明确并达成共识的定义；其次，需要定义一组可靠的基准网络，用来测试并比较各种算法以及它们所得到的社团划分结果的质量。现在大多数算法还是通过模块度这一指标进行比较，但是本文也提到了模块度指标的缺陷，因此能否提出一种更优的评判指标也是未来一个很重要的研究内容。

尽管已有很多划分社团的算法，可当面对诸多算法和一个实际要处理的网络数据时，该如何选择算法，所以对于算法的选择问题也是未来要考虑的一个重点。如基于模块度优化的方法可能是现在最常用的一种算法，但该算法并不是最优的解决方案。

动态网络的社团划分应该给予更多的关注。现在已经可以得到一些具有时间节点特性的动态网络数据，如何分析网络的演化并给出社团产生和随时间相互作用的机理也是一个很具有挑战性的课题。

从实际应用的角度看，对网络社团划分的结果的分析是很关键的[32]。从所得的社团划分结果中到底能得到哪些信息，希望能够分析节点间所隐藏的关系，也就是在划分社团之前所看不出来的特性，这些结果能够告诉哪些节点间的关系，能够展示网络的哪些特性等，这些问题都需要深入研究。

参 考 文 献

[1] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社, 2006.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Theory of complex networks and its application[M]. Beijing: Tsinghua University Press, 2006.

[2] FORTUNATO S, CASTELLANO C. Community structure in graphs[J/OL]. Eprint arXiv, 2007, 0712: 2716. [2009-03-10]. http://www.arXiv.org.

[3] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69 (2): 026113.

[4] FORTUNATO S, BARTHELEMY M. Resolution limit in community detection[J]. PPNAS, 2007, 104(1): 36-41.

[5] NEWMAN M E J. Analysis of weighted networks[J]. Phys Rev E, 2004, 70: 056131.

[6] ARENAS A, DUCH J, FERNANDEZ A, et al. Community structure in directed networks[J]. New J Phys, 2007, 9: 176.

[7] NEWMAN M E J, LEICHT E A. Community structure in directed networks[J]. Proc Natl Acad Sci USA, 2007, 104: 9564.

[8] SHEN H, CHENG X, CAI K, et al. Detect

overlapping and hierarchical community structure in networks[J]. Physica A, 2009, 388: 1706-1712.
[9] NICOSIA V, MANGIONI G, CARCHIOLO V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. J Stat Mech, 2009, 3: 03024.
[10] KAPLAN T D, FORREST S. A dual assortative measure of community structure[J]. Eprint arXiv, 2008, 0801: 3290. [2009-03-10]. http://www.arXiv.org.
[11] GOMEZ S, JENSEN P, ARENAS A. Analysis of community structure in networks of correlated data[J]. Eprint arXiv, 2008, 0812: 3030. [2009-03-10]. http://www. arXiv.org.
[12] TRAAG V A, BRUGGEMAN J. Community detection in networks with positive and negative links[J]. Eprint arXiv, 2008, 0811: 2329. [2009-03-10]. http://www.arXiv.org.
[13] NEWMAN M E J. Detecting community structure in networks[J]. Eur. Phys. J. B, 2004, 38(2): 321.
[14] SALES-PARDO M R, GUIMERA A, MOREIRA A, et al. Extracting the hierarchical organization of complex systems [J]. Proc. Natl. Acad. Sci. USA, 2007, 104: 15224.
[15] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of community hierarchies in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10: 10008.
[16] PALLA G., DERENYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435: 814-818.
[17] FARKAS I, ÁBEL D, PALLA G, et al. Weighted network modules[J]. New J Phys, 2007, 9(6): 180.
[18] FARKAS I, ÁBEL D, PALLA G, et al. Directed network modules[J]. New J Phys, 2007, 9(6): 186.
[19] GREGORY S. A fast algorithm to find overlapping communities in networks[J]. Machine Learning and Knowledge Discovery in Databases, 2008, 408-423.
[20] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci. USA, 2002, 99(6): 7821-7826.
[21] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New J. Phys. 2009, 11: 033015.
[22] SHEN Hua-wei, CHENG Xue-qi, CAI Kai, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A, 2009, 388(8): 1706-1712.
[23] BRON C, KERBOSCH J. Algorithm 457: finding all cliques of an undirected graph[J]. Commun ACM, 1973, 16(9): 575-577.
[24] AHN Y-Y, BAGROW J P, LEHMANN S. Communities and Hierarchical Organization of Links in Complex Networks[J/OL]. Eprint arXiv, 2009, 0903: 3178. [2009-03-12]. http://www.arXiv.org.
[25] EVANS T S, LAMBIOTTE R. Line graphs. Link partitions and overlapping communities[J/OL]. Eprint arXiv, 2009, 0903: 2181. [2009-04-01]. http://www.arXiv.org.
[26] NOACK A, ROTTA R. Multi-level algorithms for modularity clustering[J]. Eprint arXiv, 2008, 0812: 4073. [2009-04-16]. http://www.arXiv.org.
[27] CONDON A, KARP R M. Algorithms for graph partitioning on the planted partition model[J]. Random Struct. Algor, 2001, 18(2): 116-140.
[28] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA, 2002, 99, 8271-8276.
[29] FAN Y, LI M, ZHANG P, et al. Accuracy and precision of methods for community identification in weighted networks[J]. Physica A, 2007, 377: 363-372.
[30] LANCICHINETTI A, FORTUNATO S, RADICCHI F. New benchmark in community detection[J]. Eprint arXiv, 2008, 0805: 4770v2. [2009-04-16]. http://www.arXiv.org.
[31] SAWARDECKER E N, SALES-PARDO M, AMARAL L A N. Detection of node group membership in networks with group overlap[J]. Eur Phys J B, 2009, 67: 277.
[32] FORTUNATO S. Community detection in graphs[J/OL]. Eprint arXiv, 2009, 0906: 0612v1. [2009-04-20]. http://www.arXiv.org.

维基百科研究综述

赵 飞[1,2]，周 涛[1,3,4]，张 良[1,2]，马鸣卉[1,5]，
刘金虎[1,6]，余 飞[1,5]，查一龙[1,7]，李睿琪[1,7]

1. 电子科技大学互联网科学中心 成都 610054; 2. 电子科技大学经济与管理学院 成都 610054;
3. 中国科学技术大学近代物理系 合肥 230026; 4. 瑞士弗里堡大学物理系 弗里堡 1700;
5. 电子科技大学示范性软件学院 成都 610054; 6. 电子科技大学应用数学学院 成都 610054;
7. 电子科技大学国际化软件人才实验班 成都 610054

【摘要】对维基百科的相关研究进行了综述。介绍了维基百科的历史发展与特点，宏观统计性质以及演化规律等方面，特别强调了复杂网络的思想和方法在分析维基百科中的应用。讨论了维基百科在社会、经济、文化、教育方面的意义和价值。最后提出了相关研究可能的发展方向，特别突出了复杂网络与人类动力学和维基百科研究可能的深入结合。

关 键 词 复杂网络; 演化规律; 统计性质; 维基百科
中图分类号 N941 文献标识码 A
doi:10.3969/j.issn.1001-0548.2010.03.001

Research Progress on Wikipedia

ZHAO Fei[1,2], ZHOU Tao[1,3,4], ZHANG Liang[1,2], MA Ming-hui[1,5],
LIU Jin-hu[1,6], YU Fei[1,5], ZHA Yi-long[1,7], and LI Rui-qi[1,7]

1. Web Sciences Center, University of Electronic Science and Technology of China Chengdu 610054;
2. School of Economy and Management, University of Electronic Science and Technology of China Chengdu 610054;
3. Department of Modern Physics, University of Science and Technology of China Hefei 230026;
4. Department of Physics, University of Fribourg Fribourg Switzerland 1700;
5. School of Software Engineering, University of Electronic Science and Technology of China Chengdu 610054;
6. School of Applied Mathematics, University of Electronic Science and Technology of China Chengdu 610054;
7. Experimental Class of International Software Professionals, University of Electronic Science and Technology of China Chengdu 610054

Abstract The rapid development of web technology has promoted the emergence and organization of the collaborative Wiki systems. This paper introduces the Wikipedia's history, macro-level statistical properties, evolution regularities, and so on. Especially the application of the motivation and methods of complex network study in analyzing the Wikipedia is emphasized. Wikipedia's significance and impacts on society, economy, culture and education are also discussed. Finally, some open questions are outlined for future research; especially the connection between Wikipedia and the new development in complexity sciences, such as the studies of

基金项目：国家973计划(2006CB705500)；国家863计划(2007AA01Z440)；自然科学基金重大研究计划(90924011)；国家自然科学基金重点项目(10635040)；国家自然科学基金面上项目(60973069, 60973120, 60903073).
作者简介：赵飞(1985－)，男，硕士，主要从事复杂网络方面的研究.