ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.1, January 2009, pp.54−66
doi: 10.3724/SP.J.1001.2009.03464

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

复杂网络聚类方法*

杨 博[1,2], 刘大有[1,2+], LIU Jiming[3], 金 弟[1,2], 马海宾[1,2]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](吉林大学 符号计算与知识工程教育部重点实验室,吉林 长春 130012)

[3](香港浸会大学 计算机科学系, 香港)

Complex Network Clustering Algorithms

YANG Bo[1,2], LIU Da-You[1,2+], LIU Jiming[3], JIN Di[1,2], MA Hai-Bin[1,2]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering for the Ministry of Education, Jilin University, Changchun 130012, China)

[3](Department of Computer Science, Hong Kong Baptist University, Hong Kong, China)

+ Corresponding author: E-mail: dyliu@jlu.edu.cn

Yang B, Liu DY, Liu J, Jin D, Ma HB. Complex network clustering algorithms. *Journal of Software*, 2009, 20(1):54−66. http://www.jos.org.cn/1000-9825/3464.htm

Abstract: Network community structure is one of the most fundamental and important topological properties of complex networks, within which the links between nodes are very dense, but between which they are quite sparse. Network clustering algorithms which aim to discover all natural network communities from given complex networks are fundamentally important for both theoretical researches and practical applications, and can be used to analyze the topological structures, understand the functions, recognize the hidden patterns, and predict the behaviors of complex networks including social networks, biological networks, World Wide Webs and so on. This paper reviews the background, the motivation, the state of arts as well as the main issues of existing works related to discovering network communities, and tries to draw a comprehensive and clear outline for this new and active research area. This work is hopefully beneficial to the researchers from the communities of complex network analysis, data mining, intelligent Web and bioinformatics.

Key words: complex network; network clustering; network community structure

摘 要: 网络簇结构是复杂网络最普遍和最重要的拓扑属性之一,具有同簇节点相互连接密集、异簇节点相互连接稀疏的特点.揭示网络簇结构的复杂网络聚类方法对分析复杂网络拓扑结构、理解其功能、发现其隐含模式、预测其行为都具有十分重要的理论意义,在社会网、生物网和万维网中具有广泛应用.综述了复杂网络聚类方法的研究背景、研究意义、国内外研究现状以及目前所面临的主要问题,试图为这个新兴的研究方向勾画出一个较为全面

* Supported by the National Natural Science Foundation of China under Grant Nos.60496321, 60503016, 60573073, 60873149 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA10Z245 (国家高技术研究发展计划(863))

Received 2008-06-17; Accepted 2008-08-28

和清晰的概貌,为复杂网络分析、数据挖掘、智能 Web、生物信息学等相关领域的研究者提供有益的参考.

关键词: 复杂网络;网络聚类;网络簇结构

中图法分类号: TP311 **文献标识码**: A

现实世界中的诸多系统都以网络形式存在,如社会系统中的人际关系网、科学家协作网和流行病传播网,生态系统中的神经元网、基因调控网和蛋白质交互网,科技系统中的电话网、因特网和万维网等.由于这些网络具有很高的复杂性,因此被称为“复杂网络(complex network)”.复杂网络已成为当前最重要的多学科交叉研究领域之一[1−3].与小世界性[1]、无标度性[2−4]等基本统计特性相并列,网络簇结构(network cluster structure 或 network community structure)是复杂网络最普遍和最重要的拓扑结构属性之一,具有同簇节点相互连接密集、异簇节点相互连接稀疏的特点[5−9].复杂网络聚类方法旨在揭示出复杂网络中真实存在的网络簇结构.

复杂网络聚类方法的研究对分析复杂网络的拓扑结构、理解复杂网络的功能、发现复杂网络中的隐藏规律以及预测复杂网络的行为不仅具有十分重要的理论意义,而且具有广泛的应用前景,目前已被应用于恐怖组织识别、组织结构管理等社会网络分析[5,7,10−12]、新陈代谢网络分析[7,13]、蛋白质交互网络分析和未知蛋白质功能预测[14−16]、基因调控网络分析和主控基因识别[8]等各种生物网络分析以及 Web 社区挖掘和基于主题词的 Web 文档聚类[17−19]和搜索引擎[20−22]等众多领域.

由于复杂网络聚类研究具有重要的理论意义和应用价值,它不仅成为计算机领域中最具挑战性的基础性研究课题之一,也吸引了来自物理、数学、生物、社会学和复杂性科学等众多领域的研究者,掀起了一股研究热潮.从 2002 年至今,新的方法层出不穷,新的应用领域不断被拓展,《Nature》[6,7,10],《Science》[13],《Proc. of National Academy of Sciences (PNAS)》[5,8,9,23−25],《Physics Review Letter》[26,27],《IEEE Trans. on Knowledge and Data Engineering(TKDE)》[28,29],《PLOS Computational Biology》[14,30]等不同领域的权威国际杂志和多个重要的国际学术会议(如数据挖掘领域权威国际会议 ACM SIGKDD[31]和 IEEE ICDM[32],万维网领域权威国际会议 WWW[19,21]等)多次报道这方面的研究工作.复杂网络聚类方法已成为图论、复杂网络、数据挖掘等基础理论的重要组成部分和相关课程的核心内容,如康奈尔大学计算机系开设的“The Structure of Information Networks”课程和麻省理工大学电子工程和计算机系开设的“Networks and Dynamics”课程.

在以上研究背景下,本文综述了复杂网络聚类方法的研究现状以及目前面临的主要问题,试图为该研究方向勾画出一个较为全面和清晰的概貌,为复杂网络分析、数据挖掘等相关领域的研究者提供有益的参考.本文第 1 节分析复杂网络聚类问题的研究现状,重点分析 15 个具有代表性的复杂网络聚类算法.第 2 节通过实验定量分析和比较 7 种典型算法的性能.第 3 节总结全文并给出本文的一些结论.

1 复杂网络聚类方法分类与分析

复杂网络可以建模为一个图 $G=(V,E)$,V 表示网络的节点集合,E 表示连接的结合.复杂网络可以是无向图、有向图、加权图或者超图.网络簇定义为网络的稠密连通分支,具有簇内连接稠密、簇间连接相对稀疏的特点.例如,可以把人类社会抽象成一个称为“社会网络”的加权有向图.图中节点表示人,有向边表示人与人之间的社会关系,权值表示关系的强弱,路径表示由社会关系组成的“关系链”,网络簇表示由多个具有共同属性的人组成的“社团”.除社会网络之外,常见的复杂网络还有生物网络和科技网络.研究发现,尽管客观世界中的复杂系统功能各样,但它们对应的复杂网络在结构上却具有十分惊人的相似性.根据网络结构的特点,科学家把绝大多数的复杂网络归纳为 3 类:随机网络、小世界网络和无标度网络.复杂网络的核心研究内容是揭示复杂网络功能和结构之间的内在联系.目前,用于刻画复杂网络结构的重要属性是平均路长、聚类系数、度分布、网络 Motif 和网络簇结构.借助复杂网络簇结构分析方法,科学家取得了一些有关网络功能和结构的初步研究结果,如:揭示出蛋白质功能和交互关系的内在联系、网页主题和超连接的内在联系、社会组织如何随时间演化等.然而,已有的研究结果还远未揭示复杂网络功能与结构的内在联系,在理论和应用上都还存在许多亟待解决的问题.本文主要从计算方法的角度,围绕复杂网络簇结构的发现算法进行讨论,分析和比较现有算法的基本原理、特

点、不足和需要解决的问题.

目前已存在多种复杂网络聚类算法,按照所采用的基本求解策略,本文将它们中的大多数归纳为两大类:基于优化的方法(optimization based method)和启发式方法(heuristic method).前者将复杂网络聚类问题转化为优化问题,通过最优化预定义的目标函数来计算复杂网络的簇结构.例如,谱方法(spectral method)将网络聚类问题转化为二次型优化问题,通过计算特殊矩阵的特征向量来优化预定义的"截(cut)"函数.后者将复杂网络聚类问题转化为预定义启发式规则的设计问题.例如,被广泛引用的 Girvan-Newman 算法[5]的启发式规则是:簇间连接的边介数(edge betweenness)应大于簇内连接的边介数.除了以上两类方法之外,还存在其他类型的复杂网络聚类方法.按照本文的分类方法,现有复杂网络聚类方法的分类如图 1 所示,本节将具体分析各类方法的典型代表.

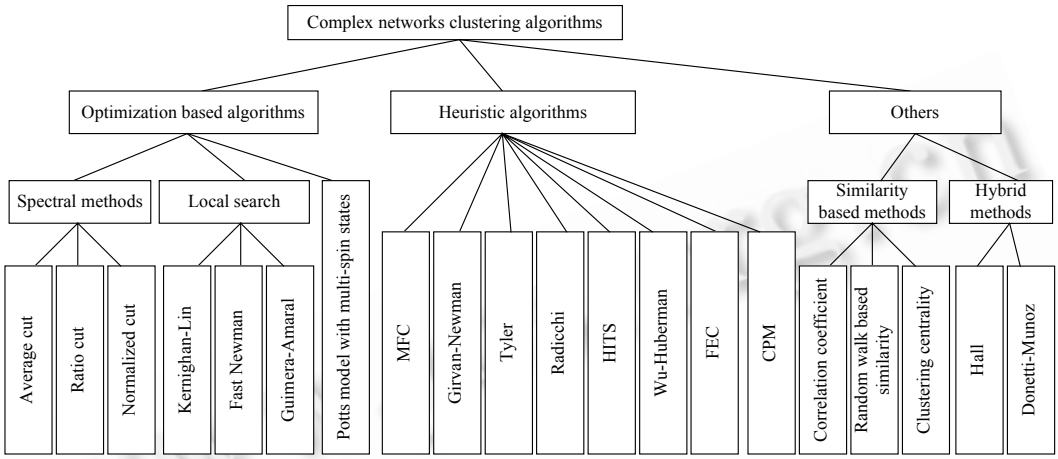


Fig.1 Classification chart of complex network clustering algorithms

图 1 复杂网络聚类算法分类图

1.1 基于优化的复杂网络聚类方法

谱方法和局部搜索方法是两类主要的基于优化的复杂聚类方法.

1.1.1 谱方法

谱方法最早用于解决图分割(graph partition)问题,近年来被应用到复杂网络聚类[23,31,33,34].谱方法采用二次型优化技术最小化预定义的"截"函数.当一个网络被划分为两个子网络时,"截"即指子网间的连接密度.具有最小"截"的划分被认为是最优的网络划分.针对不同问题,提出了不同的"截"函数.例如,针对分布式系统负载平衡提出的"平均截(average cut)"[35,36]、针对大规模集成电路(VLSI)设计提出的"比率截(ratio cut)"[37,38]以及针对图像分割提出的"规范截(normalized cut)"[39]等.已经证明,最小化以上的"截"函数是 NP 完全问题[39,40].采用矩阵分析技术,谱方法将求解最小"截"问题转化为求解带约束的二次型优化问题:$\min\{(X^TMX)/(X^TX)\}$,其中,向量 X 表示网络划分,M 表示对称半正定矩阵.对于"平均截",$M=D-A$ 表示网络的拉普拉斯矩阵(Laplacian matrix),其中 D 表示由节点度构成的对角矩阵,A 为网络的邻接矩阵;对于"规范截",$M=D^{-1/2}(D-A)D^{-1/2}$ 表示网络的规范化拉普拉斯矩阵;对于其他截函数,M 是拉普拉斯矩阵的不同变体.由拉格朗日方法,以上约束二次型的近似最优解(即网络的近似最优划分)可以通过计算 M 的第 2 小特征向量求得.一般地,n 维矩阵特征向量的计算时间为 $O(n^3)$.对于稀疏网络,采用 Lanczos 算法[41],M 的第 2 小特征向量的计算时间为 $O(m/(\lambda_3-\lambda_2))$,其中 m 表示网络连接数目,λ_2 和 λ_3 分别表示 M 的第二、第三小特征值.谱方法本质上是一种二分法,在每次二分过程中,网络被分割成两个近似平衡的子网络.当网络中含有多个簇时,谱方法递归地分割现存的子网络,直到满足预先定义的停止条件为止.

谱方法具有严密的数学理论,已发展成数据聚类的一种重要方法(称为谱聚类法),被广泛应用于图分割和

空间点聚类等领域.但是,针对复杂网络聚类,谱方法的主要不足是:1) 需要借助先验知识定义递归终止条件,即谱方法不具备自动识别网络簇总数的能力;2) 现实世界中的复杂网络往往包含多个网络簇,而谱方法的递归二分策略不能保证得到的网络划分是最优的多网络簇结构.

1.1.2 基于局部搜索的复杂网络聚类方法

Kernighan-Lin 算法(简称 KL 算法)[42]、快速 Newman 算法(简称 FN 算法)[43]和 Guimera-Amaral 算法(简称 GA 算法)[6]是 3 种典型的基于局部搜索优化技术的复杂网络聚类算法.这类算法包含 3 个基本部分:目标函数、候选解的搜索策略和最优解的搜索策略.以上 3 种算法采用了几乎相同的候选解搜索策略,但其所采用的目标函数和最优解搜索策略却不尽相同.

针对图分割问题,Kernighan 和 Lin 在 1970 年提出 KL 算法[42],该方法也可用于复杂网络聚类.KL 算法的优化目标是极小化簇间连接数目与簇内连接数目之差;其候选解搜索策略是:将节点移动到其他簇或交换不同簇的节点.从初始解开始,KL 算法在每次迭代过程中产生、评价、选择候选解,直到从当前解出发找不到更好的候选解为止.在整个搜索过程中,KL 算法只接受更好的候选解,而拒绝所有较差的候选解,因此它找到的解往往是局部最优而不是全局最优解.KL 算法最大的局限性在于它需要先验知识(如簇的个数或簇的平均规模)来产生一个较好的初始簇结构,因为该算法对初始解非常敏感,不好的初始解往往导致缓慢的收敛速度和较差的最终解.KL 算法的时间复杂性是 $O(tn^2)$,其中,n 表示网络节点个数,t 表示算法停止时的迭代次数.

2004 年,Newman 提出了基于局部搜索的快速复杂网络聚类算法 FN[43].其优化目标是极大化 Newman 和 Girvan 在同年提出的网络模块性(modularity)评价函数(他们称为 Q 函数)[44].Q 函数定义为簇内实际连接数目与随机连接情况下簇内期望连接数目之差,用来定量地刻画网络簇结构的优劣,一种计算形式如下:

$$Q=\sum_{s=1}^{K}\left[\frac{m_s}{m}-\left(\frac{d_s}{2m}\right)^2\right],$$

其中,K 表示网络簇个数,m 表示网络连接总数,m_s 表示网络簇 s 中的连接总数,d_s 表示网络簇 s 中节点度之和.

一般地,好的网络簇结构对应较大的 Q 值.候选解的局部搜索策略为:选择且合并两个现有的网络簇.从初始解开始(每个网络簇仅包含一个节点),在每次迭代中,FN 算法执行使 ΔQ 值最大化的合并操作,直到网络中只剩下一个网络簇.通过这种自低向上的层次聚类过程,FN 算法输出一棵刻画网络簇层次关系的树结构(dendrogram).FN 算法的时间复杂性是 $O(mn)$,m 和 n 分别表示网络的连接数和节点数.

采用与 FN 算法相同的优化目标,Guimera 和 Amaral 在 2005 年提出了基于模拟退火算法(simulated annealing,简称 SA)的复杂网络聚类算法 GA,并应用到新陈代谢网络分析中.2005 年 2 月刊的《Nature》报道了该工作[6].类似于 KL 算法,从初始解开始,在每次迭代中,GA 算法产生、评价、接受或拒绝由当前解产生的候选解.GA 算法产生候选解的策略是:将节点移动到其他簇、交换不同簇的节点、分解网络簇或合并网络簇.GA 算法通过计算候选解对应的 Q 值来评价其优劣,并采用模拟退火策略的 Metropolis 准则决定是否接受它,允许以一定的概率接受较差的候选解而放弃较好的候选解.因此,GA 算法具有跳过局部最优解、找到全局最优解的能力,从而具有很好的聚类精度. GA 采用的 Metropolis 准则定义如下:

$$p=\begin{cases}1, & \text{if } C_{t+1}\le C_t\\ \exp\left(-\dfrac{C_{t+1}-C_t}{T}\right), & \text{if } C_{t+1}>C_t\end{cases},$$

其中,$C_t=-Q_t$,p 表示接受 $t+1$ 时刻候选解的概率,T 表示 $t+1$ 时刻的系统温度.

GA 算法的效率完全取决于 SA 算法的效率,而后者的收敛速度通常很缓慢.据报道,在普通配置的计算机上采用 GA 算法聚类仅包含 3 885 个节点、7 260 条边的酵母菌蛋白质交互网络需要 3 天时间[14].此外,GA 算法对输入参数(如初始解、候选解搜索策略、降温(cooling)策略等)非常敏感,不同的参数设置往往导致具有较大差别的聚类结果和运行时间.

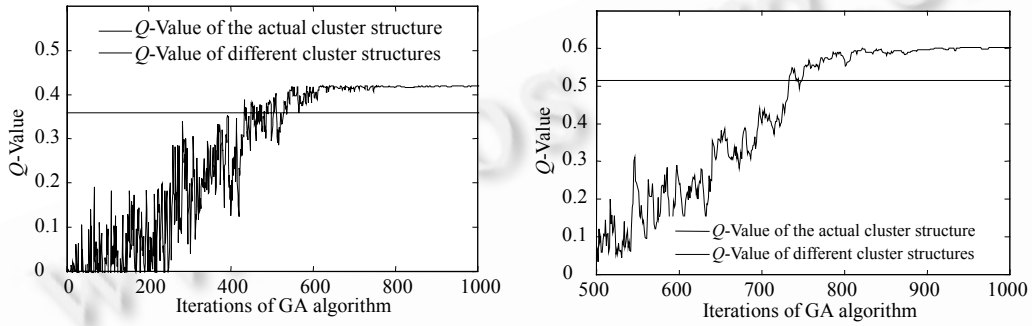
1.1.3 其他基于优化的复杂网络聚类方法

除以上两种主要方法外,还存在其他基于优化方法的复杂网络聚类方法.例如,Reichardt 和 Bornholdt 在

2004 年提出的基于多自旋状态 Potts 模型的网络聚类算法[26].在该模型中,每个网络节点被看作是一个具有多自旋状态的旋转子(spin),并且同簇内节点具有相同的自旋状态.他们认为,最优的网络簇结构应该对应最稳定的系统状态,即能量最低的状态.因此,网络聚类问题就转化为求最小化系统能量的自旋状态分布问题.他们定义了系统能量函数,并基于蒙特卡罗方法和模拟退火算法给出了相应的优化算法.

1.1.4 基于优化聚类方法的分析

采用优化方法识别出的网络簇结构完全取决于优化目标,因此“有偏”的目标函数会导致“有偏”的解(即得到的网络簇结构和真实存在的网络簇结构不符).值得注意的是,除了以上提到的 FN 算法和 GA 算法外,很多基于优化的复杂网络聚类方法都以最大化 Q 函数作为优化目标[14,23,45,46].然而,研究发现,Q 函数是有偏的,并不能完全准确地刻画最优的(或者说是真实的)网络簇结构.对于某些网络而言,其真实的网络簇结构对应的 Q 值是局部极大值,而非全局最大值.图 2 给出了 GA 算法计算两个基准社会网络(Karate 网络[47]和 Football 网络[5])的局部搜索过程.如图 2(a)所示,对于 Karate 网络而言,其真实的 2-网络簇结构对应一个局部极大值 0.37,而 GA 计算出的全局最优值 0.42 对应一个 4-网络簇结构.如图 2(b)所示,对于 Football 网络而言,其真实的 12-网络簇结构对应一个局部极大值 0.51,而 GA 计算出的全局最优值 0.60 对应一个 10-网络簇结构.



(a) (b)

Fig.2 Local search processes of the GA algorithm

图 2 GA 算法的局部搜索过程

2004 年,Guimera 等人进一步研究发现,对于某些随机网络,由于受到扰动的影响,明显不好的网络簇结构却对应相对较高的 Q 值[48].2007 年,Fortunato 和 Barthelemy 系统地研究了 Q 函数对聚类精度的影响,他们在《PNAS》上发表论文指出:对于大规模复杂网络,基于优化 Q 函数的复杂网络聚类算法倾向于找到粗糙的而不是精细的网络簇结构[24].这意味着,该类算法未必能够找到这些网络中真实存在的全部网络簇.

1.2 启发式复杂网络聚类方法

MFC(maximum flow community)算法[17]、HITS(hyperlink induced topic search)算法[20]、Girvan-Newman (GN)算法[5]及其改进[9,12]、Wu-Huberman(WH)算法[49]和 CPM(clique percolation method)算法[7]和 FEC(finding and extracting communities)算法[28]是典型的启发式复杂网络聚类算法.这类算法的共同特点是:基于某些直观的假设来设计启发式算法,对于大部分网络,它们能够快速地找到最优解或者近似最优解,但无法从理论上严格保证它们对任何输入网络都能找到令人满意的解.

2002 年,Flake 等人基于图论的最大流-最小截定理提出了复杂网络聚类算法 MFC[17].该算法的基本假设是:网络中的最大流量由网络“瓶颈”的容量决定,而在具有簇结构的网络中,网络“瓶颈”由簇间连接构成.由最大流-最小截定理可知:网络中的最大流等于最小截集的容量.因此,通过计算最小截集可以识别簇间连接.经过反复识别并删除簇间连接,网络簇能够被逐渐分离开来.Flake 等人将 MFC 应用到基于链接的 Web 网页聚类,并通过实验验证了一个非常有用的假说:通过自组织方式形成的 Web 簇是高度主题相关的.这个发现为基于主题词的

Web 网页/文本聚类提供了一个新思路,因为基于连接分析的聚类算法所需要的开销要远远低于基于内容分析的聚类算法.MFC 算法的效率由计算最小截集的时间决定,目前最快的最小截集计算方法需要 $O(mn\log(n^2/m))$ 时间[50].

由 Girvan 和 Newman 在 2002 年提出的 GN 算法也采用反复识别和删除簇间连接的策略聚类复杂网络[5].但 GN 算法采用了与 MFC 算法完全不同的启发式规则:簇间连接的边介数(edge betweenness)应大于簇内连接的边介数.连接的边介数定义为网络中经过该连接的任意两点间最短路径的条数.GN 算法通过反复计算边介数、识别簇间连接、删除簇间连接,以自顶向下的方式建立一棵层次聚类树(dendrogram).GN 算法的最大缺点是计算速度慢,由于边介数的计算开销过大($O(mn)$),GN 算法具有很高的时间复杂性($O(m^2n)$),只适合处理中小规模的网络(包含几百个节点的网络).尽管如此,GN 算法在复杂网络聚类研究中仍占有十分重要的地位,Girvan 和 Newman 工作的重要意义在于:他们首次发现了复杂网络中普遍存在的网络簇结构,启发了其他研究者对这个问题的深入研究,掀起了复杂网络聚类的研究热潮.针对 GN 算法计算速度慢的缺点,研究者提出了多种改进方法.

2003 年,Tyler 等人将统计方法引入基本的 GN 算法,提出一种近似 GN 算法[12].他们的策略是:采用蒙特卡洛方法估算出部分连接的近似边介数,而不是计算出全部连接的精确边介数.显然,这种方法计算速度的提高是以牺牲聚类精度为代价的.

考虑到 GN 算法效率低是因为边介数计算开销过大,2004 年,Radicchi 等人提出了连接聚类系数(link clustering coefficient)取代 GN 算法的边介数[9].他们认为:簇间连接应该很少出现在短回路(如三角形或四边形)中,否则,短回路中的其他多数连接也会成为簇间连接,从而显著增加簇间的连接密度.基于该出发点,他们把连接聚类系数定义为包含该连接的短回路数目,并采用如下启发式规则:簇间连接的连接聚类系数应小于簇内连接的连接聚类系数.在算法的每次迭代中,具有最小连接聚类系数的边被删除.连接聚类系数的平均计算时间是 $O(m^3/n^2)$,这一算法的时间复杂性为 $O(m^4/n^2)$.对于稀疏网络,他们的算法($O(n^2)$)要快于 GN 算法($O(n^3)$).该算法的最大局限性是:不适合处理短回路很少甚至没有的复杂网络.

针对基于连接的 WWW 聚类问题,Kleinberg 等人在 1999 年提出了著名的 HITS 算法[20].该算法本质上是一种启发式算法,所基于的基本假设是:根据连接关系,WWW 中存在权威(authority)和中心(hub)两种基本类型的页面,权威页面倾向于被多个中心页面引用,而中心页面倾向于引用多个权威页面.基于权威-中心页面间相互指向的连接关系,HITS 算法通过计算 WWW 对应的某些特殊矩阵(AA^T 和 A^TA,A 表示 Web 图的邻接矩阵)的主特征向量来发现隐藏在 WWW 中的全部由权威-中心页面构成的网络簇结构.该算法被广泛地应用于包括 Altavista 在内的多个搜索引擎中.

2004 年,Wu 和 Huberma 提出了快速启发式算法 WH[49].该算法将复杂网络建模为电路系统,网络连接看作是具有电阻的线路,不同位置的网络节点具有不同的电位势.WH 算法的启发式规则是:当在不同的簇中分别选取两个节点作为正负极后,由于簇间的电阻远远大于簇内电阻,因此,同簇节点位势应近似相同,而异簇节点位势应具有显著差异.WH 算法首先基于 Kirchhoff 方程计算出每个节点的位势,然后采用寻找最大位势差的方法区分出不同的网络簇.WH 算法是目前报道过的最快的复杂网络聚类算法,具有近似线性的时间复杂性 $O(t(n+m))$,其中,t 为计算出全部位势所需要的迭代次数.但 WH 算法需要过多的先验知识,并且通常难以获取.例如,WH 算法需要从两个不同的簇中选择正负极节点;为挖掘出多个网络簇,WH 算法需要知道网络簇的总数和每个簇的近似规模.

目前,绝大多数算法不考虑重叠网络簇结构.但在多数应用中,重叠网络簇结构更具有实际意义.例如,在语义网中,多义词允许同时出现在多个表示不同词义的网络簇中.2005 年,Palla 及其同事在《Nature》上发表文章,提出了能够识别重叠网络簇结构的 CPM 算法[7].该算法的基本假设是:网络簇由多个相邻的 k-团(k-clique)组成,相邻的两个 k-团至少共享 k–1 个节点,每个 k-团唯一地属于某个网络簇,但属于不同网络簇的 k-团可能会共享某些节点.基于以上启发式信息,CPM 算法通过如下步骤识别出重叠网络簇结构:1) 对给定的参数 K,计算出网络中的全部 k-团($k\leq K$),并建立团-团重叠矩阵(clique-clique overlap matrix);2) 根据以上矩阵,计算出重叠网络簇

结构.CPM算法是第1种能够计算重叠网络簇结构的算法,但具有如下主要缺点:在实际应用中参数K难以确定,选取不同的K值往往得到差别较大的网络簇结构,但难以评判它们的优劣.

符号网络(signed network)是指包含正、负两种关系的二维复杂网络,是对一般复杂网络描述能力的一种推广.符号网络广泛存在于社会、生物等多种复杂系统中.例如,在社会系统中,“喜欢”、“尊重”和“表扬”属于正关系,而“厌恶”、“轻视”和“责备”属于负关系;再如,在神经系统中,神经元之间的“相互促进”属于正关系,而“相互抑制”属于负关系.符号网络簇结构具有簇内正关系稠密、同时簇间负关系也稠密的特点.针对符号网络聚类问题,杨博、Cheung 和 Liu 等人在 2007 年提出了基于马尔可夫随机游走模型的启发式符合网络聚类算法(FEC)[28].FEC 算法所采用的基本假设是:从任意给定的簇出发,网络中的随机游走过程达到起始簇内节点的期望概率将大于达到起始簇外节点的期望概率.基于该启发规则,FEC 算法首先计算出在给定时刻随机游走过程到达所有节点的期望转移概率分布,进而根据该分布的局部一致性——同簇节点具有近似相同的期望转移概率分布——识别出各个不同的网络簇.值得指出的是,FEC 算法是第 1 种综合考虑两种分簇标准(即连接密度和连接符号)的复杂网络聚类算法,既能有效处理符号网络(能够发现更加“自然”的符号网络簇结构),又能有效处理仅包含“正关系”的一般复杂网络.与现有方法相比,FEC 算法在时间和识别精度方面表现出了更好的性能,尤其适合于处理噪声高和网络簇结构不明显的复杂网络.该算法的参数是随机游走的步长,步长的设置会影响最终的聚类结果.通过实验分析,FEC 算法给出了步长设置的经验值,建议取值区间为[6,20].其中,6 表示复杂网络中两点间的平均距离(大多数网络都满足六度分离理论),20 表示网络的直径(WWW 是迄今最大的复杂网络,研究表明其直径为 19).但是,FEC 算法没有从理论上给出一种针对不同网络设置最优参数的方法.

1.3 其他复杂网络聚类方法

除了以上两类主要方法以外,还存在其他复杂网络聚类方法.例如,基于相似度的层次聚类方法.在这类方法中,节点间的相似度根据网络拓扑结构定义,如基于结构全等的相关系数(correlation coefficient)[51]、基于随机游走的相似度[52]和节点聚类中心度(clustering centrality)[53]等.

研究发现,WWW 呈现的全局拓扑结构是由多个分散、自治实体的局部行为通过多种自组织方式涌现而成的.针对具有自组织特点的 WWW 聚类问题,文献[53]分析了复杂网络的宏观拓扑结构和网络节点的局部信息之间的关系,发现隐藏在网络中的全局簇结构能够从评价各个节点重要程度的局部中心度(local centrality)推断出来.据此,提出了节点聚类中心度概念和基于节点聚类中心度的复杂网络层次聚类算法(identifying community structure,简称 ICS),并给出了该算法在搜索引擎中的应用实例.

此外,聚类复杂网络的另一个思路是:将网络聚类转化为向量聚类.通过给每个网络节点分配一个合理的K-维坐标,我们可以把网络聚类问题转换为传统的空间点聚类问题,然后采用 K-means 等经典聚类算法聚类这些新生成的空间点.实际上,这个思想最早可以追溯到 1970 年 Hall 针对图分割问题提出的加权二次型变换算法[54].该算法能够将网络投影到一维空间,使得网络中连接紧密的节点在一维空间中的位置相对较近,而连接稀疏的节点在一维空间中的位置相对较远.基于相似的思想,Donetti 和 Munoz 在 2004 年提出了一种结合谱方法和空间点聚类方法的复杂网络聚类算法[55].他们首先通过计算拉普拉斯矩阵的 K 个最小特征向量将网络映射到K-维空间中,然后采用某种基于距离的空间点聚类算法聚类网络节点.

2 实 验

为了定量地分析和比较不同复杂网络聚类方法的性能,我们分别从优化方法和启发式方法中选择了具有代表性的 7 种算法,针对不同的基准数据集,从聚类精度和聚类速度两个方面进行对比实验.实验环境为:处理器 Intel(R)Core(TM)2 4400 2.0GHz,内存 2G,硬盘 160G,操作系统为 Windows XP,编程语言为 Matlab 7.0.相关算法的代码可以从 http://datastructure.jlu.edu.cn/www/网站下载.

2.1 聚类精度比较

首先采用已知簇结构的随机网络测试所选择算法的聚类精度.该实验方法被相关工作广泛采用,已成为测

试复杂网络聚类算法准确性的一种基准方法[5,43,44].已知簇结构的随机网络定义为 $RN(C,s,d,p_{in})$,其中,C 表示网络簇的个数,s 表示每个簇包含节点的个数,d 表示网络中节点的平均度,p_{in} 表示簇内连接密度(即簇内连接总数与网络连接总数的比值).p_{in} 值越大,随机网络的簇结构越明显;反之,簇结构越模糊.特别地,当 p_{in}<0.5 时,认为该随机网络不具有簇结构.一个随机网络被正确聚类当且仅当预定义的 C 个网络簇被全部正确识别,且没有某个簇被进一步分割为多个子簇.图 3 给出了实验结果,这里所采用的随机网络是被普遍采用的基准随机网络 $RN(4,32,16,p_{in})$.图 3 中,y-轴表示聚类精度,曲线上的每个数据点是采用不同算法聚类 100 个随机网络得到的平均准确率.

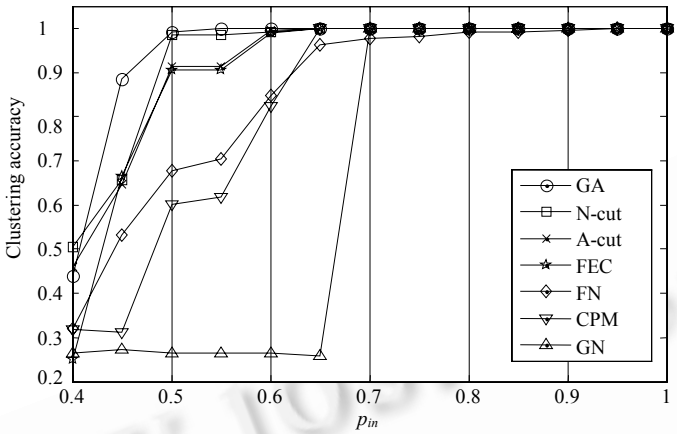


Fig.3 Testing the clustering accuracy of different algorithms against random networks

图 3 使用随机网络测试不同算法的聚类精度

由图3可得:(1) 各种算法的聚类精度随 p_{in} 的增加而增加.在网络簇结构非常明晰的情况下(p_{in}>0.7),全部算法都具有很好的聚类精度(>97%);但在网络簇结构不明显的情况下(p_{in}<0.5),大部分算法的聚类精度都比较低;(2) 对于已知簇结构的随机网络模型,基于优化的聚类方法(FN 算法除外)比启发式算法(FEC,CPM,GN)展现出更好的聚类精度.在所有的方法中,GA 算法具有最高的聚类精度,即使在 p_{in}=0.45 的情况下,其精度也接近 90%.FEC 是启发式方法中聚类精度最高的算法,其性能接近采用平均截(A-cut)的谱方法.在所有参加比较的方法中,GN 算法的聚类精度最低,该方法只适合于处理网络簇结构比较明显(p_{in}>0.65)的网络.

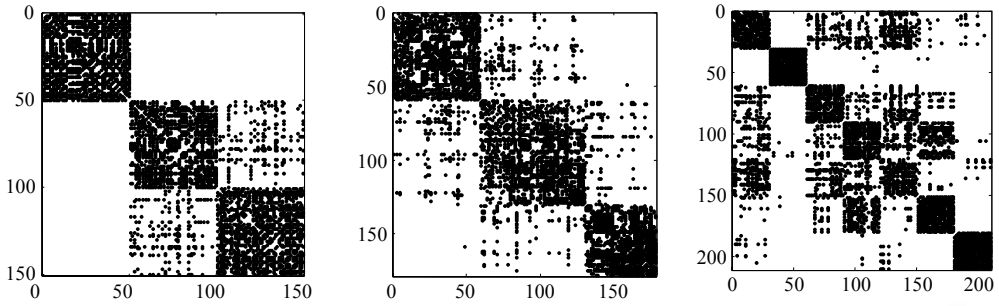
值得说明的是,由于谱方法不具备自动识别网络簇个数的能力,在该实验中,谱方法(N-cut 和 A-cut)的聚类精度是在事先知道“网络簇个数”的前提下计算的,否则,谱方法将会把已识别出的正确网络簇进一步划分为更小的子网络簇,从而降低聚类精度.此外,CPM 算法的计算结果依赖于其参数 K 的设置,不同的 K 值会产生差别较大的聚类精度.在该实验中,对每一个 p_{in},我们分别测试采用不同 K 值的 CPM 算法,并选择其中聚类精度的最大值作为 CMP 在该 p_{in} 下的聚类准确率.在以下实验中,对谱方法和 CPM 算法的测试也采用了相同的方法.

为进一步测试以上算法的聚类精度,我们选取 UCI 的 3 个数据集 Iris,Wine 和 Image 作为测试数据.这些数据集包含有手工标注的类标识,常被用于测试和评价空间聚类算法的聚类性能.Iris 包含 3 个类,150 个样本,每类代表一种类型的鸢尾花,每类各有 50 个样本,每个样本包含 3 个属性.Wine 数据集包含 3 个类,178 个样本,每类代表由一种植物酿制而成的酒,每类的样本数不同,每个样本包含 13 个属性.Image 包含 7 个类,210 个样本,每个样本分别从 7 个户外图像集合中随机选取,每类的样本数不同,每个样本包含 19 个属性.

为采用复杂网络聚类算法,首先采用 kNN(k nearest neighbors)方法将空间数据集转化为 kNN 网络.kNN 网络中的每个节点表示一个样本数据,其 k 个邻居表示在欧式空间中距离其最近的 k 个样本数据.图 4 给出了以上 3 个数据集对应的 20-NN 网络的邻接矩阵,矩阵中的黑点表示“1”.如果将同簇节点排列在一起,网络簇则以“块”的形式分布在邻接矩阵的正对角线上.容易看出,相对于 Image 而言,Iris 和 Wine 具有更好的网络簇结构.

图 5 给出了实验结果,y-轴表示聚类精度.为消除参数 k 对聚类精度的影响,对每个数据集,我们分别生成 21

个 *k*NN 网络(从 5-NN~25-NN),复杂网络聚类算法对每个数据集的聚类精度是对 21 个 *k*NN 网络聚类准确率的平均值.



(a) Iris (b) Wine (c) Image

Fig.4 Adjacency matrices of the *k*NN networks of three datasets

图 4 3 个数据集对应 *k*NN 网络的邻接矩阵

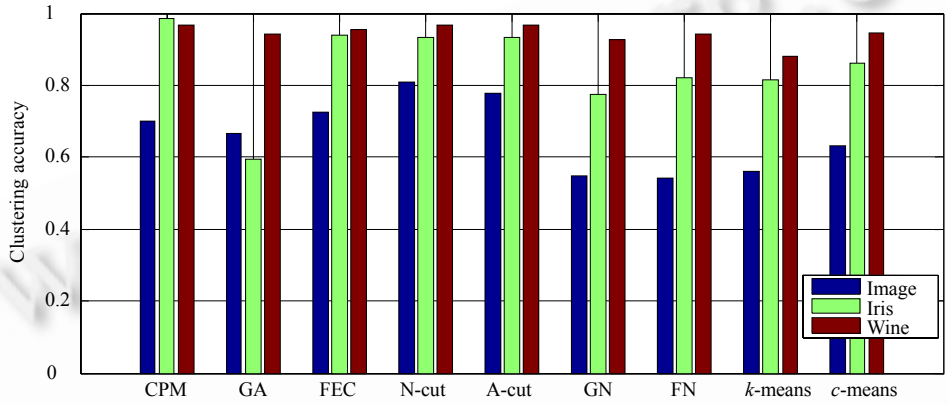


Fig.5 Testing the clustering accuracy of different algorithms against real datasets

图 5 使用真实数据库测试不同算法的聚类精度

由图 5 可知:(1) 与经典的空间聚类算法(*k*-means,*c*-means)相比,复杂网络聚类算法的聚类精度不差,部分算法(CPM,FEC,谱方法)还明显占优.这表明,复杂网络聚类算法不仅适合于处理社会网、生物网等关系型数据聚类问题,也能有效地解决传统的空间聚类问题;(2) 各种算法的聚类精度和数据集密切相关,对类簇结构好的数据集(Iris,Wine),聚类精度普遍偏高,而对于类簇结构较差的数据集(Image),聚类精度普遍偏低;(3) 对于具有较好类簇结构的 Iris 数据集,GA 算法的聚类精度最低(不足 60%),再次验证了 Q 函数的有偏性,导致以优化 Q 函数为目标的聚类方法对某些网络具有较低的聚类精度.

2.2 计算速度比较

计算速度是评价聚类算法性能的重要指标.第 2.1 节给出了不同复杂网络聚类算法在理论上的时间复杂性分析,本节从实验角度给出典型算法的实际运行时间,作为比较和评价各算法性能的重要依据.图 6 给出了 6 种典型算法的实际运行时间.该实验所采用的测试网络是随机网络 $RN(4,s,16,0.7)$.该网络的类簇结构确定,但其规模可由 s 值来调节,共包括 $4s$ 个网络节点,$64s$ 条网络连接.图 6 中,y-轴表示以秒为单位的实际运行时间,x-轴表示网络规模(节点数+连接数).值得说明的是,GA 算法运行非常耗时,即使对规模为 128 个节点、8 192 条连接的小型网络,其运行时间也在 2 小时以上,因此,其实际运行时间曲线没有显示在图 6 中.

由图 6 可得,启发式算法 FEC 和 CPM 具有最快的运行速度,其实际运行时间与网络规模呈近似线性关系,

其次是两种谱方法、FN 算法,而 GN 算法的速度较慢.

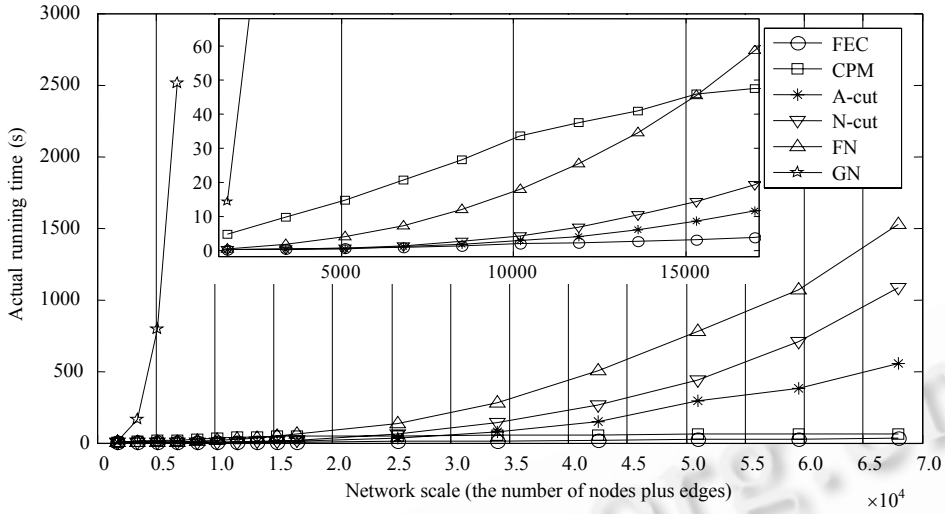


Fig.6 Actual running time of algorithms against networks with different scales

图 6 不同规模网络下各算法的实际运行时间

结合图 3、图 5 和图 6,我们可以评价各种典型算法的综合性能.GA 算法对同构随机网络(网络簇规模和节点度近似相同)具有最好的聚类精度,但其计算速度非常缓慢,只适合处理规模很小的复杂网络;FEC,CPM 和两种谱方法(A-cut,N-cut)的聚类精度(针对随机网络和真实数据集)和运行效率都较高,具有很好的综合性能,适合处理中、大规模的复杂网络.

3 结 论

复杂网络聚类是最重要的复杂网络分析方法之一,应用广泛.本文主要从以下 3 个方面对现有工作进行了综述:(1) 根据所采用的基本求解策略将现有的复杂网络聚类方法分为两大类:基于优化的方法和启发式方法;(2) 从基本原理、关键技术和优缺点等方面对现有方法进行了分析和评价;(3) 采用基准数据集,从聚类精度和运行效率两方面定量分析和比较了典型方法的性能.通过以上几个方面的分析和总结,本文得出如下结论:尽管已经投入了大量的、艰苦的研究工作并取得诸多令人鼓舞的研究结果,但复杂网络聚类问题还远未被很好地解决,集中体现在以下几个方面:

第一,我们还没有从客观上认清网络簇结构的本质含义.目前我们还无法回答类似如下的基本问题:网络簇结构是怎么形成的?它与网络的其他复杂现象有什么必然联系?它与网络自身的哪些内在属性有关?因此,现阶段我们不得不通过观察有簇网络所展示出的“外在”现象去理解网络簇概念,借助“主观”定义的目标函数或启发式规则去刻画和计算网络簇结构.如前分析,基于这些目标函数或启发式规则的方法常常会导致“有偏”的计算结果(即计算出的网络簇结构与真实存在的网络簇结构不一致),并且采用不同的目标函数或启发式规则常常会计算出不同的网络簇结构.因此,一个基本问题是:从网络的“内在”属性出发,我们能否给出一种“客观”的理论模型去理解、刻画和计算复杂网络簇结构.

第二,现有的复杂网络聚类方法都具有局限性,不能同时满足计算速度快、聚类精度高和无监督(即不依赖先验知识、对参数不敏感)等基本要求.通过定性和定量的分析、比较现有的主要方法后可以发现,聚类精度高的方法往往具有很高的时间复杂性(高于 $O(n^2)$),而快速的聚类方法往往以牺牲精度为代价,并且需要较多的参数和先验知识.另外,特别需要指出的是,如何在没有任何先验信息的情况下识别出真实的网络簇总数仍是一个未解决的难题.因此,如何设计出快速、高精度和无监督的复杂网络聚类方法仍是当前最期待解决的问题之一.

第三,除了以上未解决的理论问题之外,随着应用领域的拓展、网络聚类问题的多样化,现有的复杂网络聚类方法已难以胜任,需要针对特殊类型的复杂网络研究新型的复杂网络聚类方法,典型问题包括动态复杂网络聚类、高维复杂网络聚类和分布式复杂网络聚类等.动态复杂网络旨在分析随时间进化的网络结构和动力性,如通过“电子邮件流”或“通话记录流”逐步形成的社会通信网络,通过“论文流”逐步形成的科学家协作网络,自组织动态演化的 WWW 网络及发育动物体中逐步进化而成的蛋白质交互网络和基因调控网络都可以建模为动态网络.此外,基于自主计算(autonomy oriented computing,简称 AOC)[56]的普式网络簇结构挖掘(distributed ubiquitous community mining)已成为该领域的一个研究热点.上述类型的复杂网络聚类方法具有广泛的应用领域,但相关理论和方法还远未成熟,因此,如何针对特殊类型的复杂网络设计出新型的复杂网络聚类方法也是当前面临的主要问题之一.

References:

[1] Watts DJ, Strogatz SH. Collective dynamics of Small-World networks. Nature, 1998,393(6638):440−442.

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

[3] Barabási AL, Albert R, Jeong H, Bianconi G. Power-Law distribution of the World Wide Web. Science, 2000,287(5461):2115a.

[4] Albert R, Barabási AL, Jeong H. The Internet’s Achilles heel: Error and attack tolerance of complex networks. Nature, 2000, 406(2115):378−382.

[5] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Science, 2002,9(12):7821−7826.

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

[7] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structures of complex networks in nature and society. Nature, 2005,435(7043):814−818.

[8] Wilkinson DM, Huberman BA. A method for finding communities of related genes. Proc. of the National Academy of Science, 2004,101(Suppl.1):5241−5248.

[9] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. Proc. of the National Academy of Science, 2004,101(9):2658−2663.

[10] Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. Nature, 2007,446(7136):664−667.

[11] Newman MEJ. Coauthorship networks and patterns of scientific collaboration. Proc. of the National Academy of Science, 2004,101(1):5200−5205.

[12] Tyler JR, Wilkinson DM, Huberman BA. Email as spectroscopy: Automated discovery of community structure within organizations. In: Huysman M, Wenger E, Wulf V, eds. Proc. of the 1st Int’l Conf. on Communities and Technologies. Dordrecht: Kluwer Academic Publishers, 2003. 81−96.

[13] Ravasz E, Somera AL, Mongru DA. Hierarchical organization of modularity in metabolic networks. Science, 2002,297(5586): 1551−1555.

[14] Wang Z, Zhang J. In serach of the biological significance of modular structures in protein networks. PLOS Computational Biology, 2007,3(6):e107.

[15] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc. of the National Academy of Science, 2003,100(21):12123−12128.

[16] Farutin V, Robison K, Lightcap E, Dancik V, Ruttenberg A, Letovsky S, Pradines J. Edge-Count probabilities for the identification of local protein communities and their organization. Proteins: Structure, Function, and Bioinformatics, 2006,62(3):800−818.

[17] Flake GW, Lawrence S, Giles CL, Coetzee FM. Self-Organization and identification of Web communities. IEEE Computer, 2002,35(3):66−71.

[18] Li X, Liu B, Yu PS. Discovering overlapping communities of named entities. In: Fürnkranz J, Scheffer T, Spiliopoulou M, eds. Proc. of the 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2006. 593−600.

[19] Ino H, Kudo M, Nakamura A. Partitioning of Web graphs by community topology. In: Ellis A, Hagino T, eds. Proc. of the 14th Int'l Conf. on World Wide Web. New York: ACM Press, 2005. 661−669.

[20] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604−632.

[21] Almeida RB, Almeida VAF. A community-aware search engine. In: Feldman SI, Uretsky M, Najork M, Wills CE, eds. Proc. of the 13th Int'l Conf. on World Wide Web. New York: ACM Press, 2004. 413−421.

[22] Sidiropoulos A, Pallis G, Katsaros D, Stamos K, Vakali A, Manolopoulos Y. Prefetching in content distribution networks via Web communities identification and outsourcing. World Wide Web, 2008,11(1):39−70.

[23] Newman MEJ. Modularity and communities structure in networks. Proc. of the National Academy of Science, 2006,103(23): 8577−8582.

[24] Fortunato S, Barthelemy M. Resolution limit in community detection. Proc. of the National Academy of Science, 2007,104(1): 36−41.

[25] Hopcroft J, Khan O, Kulis B, Selman B. Tracking evolving communities in large linked networks. Proc. of the National Academy of Science, 2004,101(1):5249−5253.

[26] Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a potts model. Physical Review Letters, 2004,93(19):218701.

[27] Garlaschelli D, Loffredo MI. Patterns of link reciprocity in directed networks. Physical Review Letters, 2004,93(26):268701.

[28] Yang B, Cheung WK, Liu J. Community mining from signed social networks. IEEE Trans. on Knowledge and Data Engineering, 2007,19(10):1333−1348.

[29] Brandes U, Delling D, Gaertler M, Görke R, Hoefer M, Nikoloski Z, Wagner D. On modularity clustering. IEEE Trans. on Knowledge and Data Engineering, 2008,20(2):172−188.

[30] Cartozo CC, Rios PDL, Piazza F, Lio P. Bottleneck genes and community structure in the cell cycle network of S.pombe. PLOS Computational Biology, 2007,3(6):e103.

[31] Shiga M, Takigawa I, Mamitsuka H. A spectral clustering approach to optimally combining numerical vectors with a modular network. In: Berkhin P, Caruana R, Wu X, eds. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 647−656.

[32] Zhou D, Councill I, Zha H, Giles CL. Discovering temporal communities from social network documents. In: Shi Y, Clifton CW, eds. Proc. of the 7th IEEE Int'l Conf. on Data Mining. New York: IEEE Society, 2007. 745−750.

[33] White S, Smyth P. A spectral clustering approach to finding communities in graphs. In: Kamath C, Goodman A, eds. Proc. of the 5th SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2005. 76−84.

[34] Donetti L, Munoz MA. Improved spectral algorithm for the detection of network communities. In: Garrido PL, Muñoz MA, Marro J, eds. Proc. of the 8th Int'l Conf. on Modeling Cooperative Behavior in the Social Sciences. New York: American Institute of Physics, 2005,779:104−107.

[35] Fiedler M. Algebraic connectivity of graphs. Czechoslovakian Mathematical Journal, 1973,23(2):298−305.

[36] Fiedler M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslovakian Mathematical Journal, 1975,25(4):619−637.

[37] Wei YC, Cheng CK. Ratio cut partitioning for hierarchical designs. IEEE Trans. on Computer-Aided Design, 1991,10(7):911−921.

[38] Hagen L, Kahng AB. New spectral methods for ratio cut partition and clustering. IEEE Trans. on Computer-Aided Design, 1992, 11(9):1074−1085.

[39] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligent, 2000,22(8): 888−904.

[40] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W.H. Freeman&Co., 1990. 60−63.

[41] Golub GH, Loan CFV. Matrix Computations. Baltimore: Johns Hopkins University Press, 1989.

[42] Newman MEJ. Detecting community structure in networks. European Physical Journal (B), 2004,38(2):321−330.

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

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

[45] Pujol JM, Béjar J, Delgado J. Clustering algorithm for determining community structure in large networks. Physical Review E, 2006,74(1):016107.

[46] Duch J, Arenas A. Community detection in complex networks using extreme optimization. Physical Review E, 2005,72(2):027104.

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

[48] Guimera R, Sales M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Physical Review E, 2004,70(2):025101.

[49] Wu F, Huberman BA. Finding communities in linear time. A physics approach. European Physical Journal B, 2004,38(2):331−338.

[50] Goldberg AV. Recent developments in maximum flow algorithms. In: Arnborg S, Ivansson L, eds. Proc. of the 6th Scandinavian Workshop on Algorithm Theory. Berlin: Springer-Verlag, 1998. 1−10.

[51] Wasserman S, Faust K. Social Network Analysis. Cambridge: Cambridge University Press, 1994.

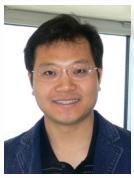
[52] Pons P, Latapy M. Computing communities in large networks using random walks. In: Yolum P, ed. Proc. of the 20th Int'l Symp. on Computer and Information Sciences. Berlin: Springer-Verlag, 2005. 284−293.

[53] Yang B, Liu J. Discovering global network communities based on local centralities. ACM Trans. on the Web, 2008,2(1):article 9:1−32.

[54] Hall KM. An *r*-dimensional quadratic placement algorithm. Management Science, 1970,17(3):219−229.

[55] Donetti L, Munoz MA. Detecting network communities: A new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004,10:P10012. http://www.iop.org/EJ/abstract/1742-5468/2004/10/P10012

[56] Liu J, Jin XL, Tsui KC. Autonomy Oriented Computing-From Problem Solving to Complex Systems Modeling. Boston: Kluwer Academic Publishers, 2004.



杨博(1974－),男,河南新乡人,博士,副教授,主要研究领域为 Agent 系统,数据挖掘,复杂网络分析.

金弟(1982－),男,博士生,主要研究领域为数据挖掘.

刘大有(1942－),男,教授,博士生导师,CCF 高级会员,主要研究领域为知识工程与专家系统,Agent 系统,时空推理.

马海宾(1984－),男,硕士生,主要研究领域为数据挖掘.

LIU Jiming(1962－),男,博士,教授,博士生导师,主要研究领域为 Agent 系统,自主计算,Web 智能.