计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(10): 2140-2154, 2013

复杂网络社区挖掘综述

刘大有[1,2] 金 弟[3] 何东晓[1,2] 黄 晶[1,2] 杨建宁[1,2] 杨 博[1,2]

[1](吉林大学计算机科学与技术学院 长春 130012)

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

[3](天津大学计算机科学与技术学院 天津 300072)

(liudy@jlu.edu.cn)

Community Mining in Complex Networks

Liu Dayou[1,2], Jin Di[3], He Dongxiao[1,2], Huang Jing[1,2], Yang Jianning[1,2], and Yang Bo[1,2]

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

[2](*Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun* 130012)

[3](*School of Computer Science and Technology, Tianjin University, Tianjin* 300072)

Abstract Many complex systems in the real world exist in the form of networks, such as social networks, biological networks, Web networks, etc, which are collectively named complex networks. The research of complex networks has attracted many researchers from different fields such as physics, mathematics, computer science, among others. One of the main problems in the study of complex network is the detection of community structure, i.e. the division of a network into groups of nodes with dense intra-connections and sparse inter-connections, which has recently triggered great interest. The ability to detect community structure has a large amount of usefulness in many aspects. Furthermore, community structure may provide insights in understanding some uncharacteristic properties of a complex network system. For instance, in the World Wide Web, community analysis has uncovered thematic clusters; in biochemical or neural networks, communities may be functional groups and separating the network into such groups could simplify functional analysis considerably. This paper reviews the background, the significance and the state-of-the-art in discovering (overlapping) communities. Also, it raises several open issues as the conclusion. Here we try to draw a comprehensive overview for this emerging scientific area, with the purpose of offering some beneficial suggestions for related researchers.

Key words complex network; community structure; community mining; overlapping community mining; network clustering

摘 要 复杂网络社区挖掘是近10年来多学科交叉的前沿研究热点之一,其研究不仅有重要的理论意义,而且有广泛的应用前景.介绍了社区挖掘及重叠社区挖掘的研究背景和研究意义,分析了研究现状,讨论了该研究所面临的一些主要问题及未来的发展方向.同时,为了对不同的社区挖掘算法进行更好地

收稿日期:2012-05-09;**修回日期**:2012-09-20

基金项目:国家自然科学基金项目(61133011,61202308,61303110,61373053);教育部新世纪优秀人才支持计划基金项目(NCET-11-0204);符号计算与知识工程教育部重点实验室开放基金项目(93K172013K02);天津大学自主创新基金项目(2013XQ-0136);吉林大学科学前沿与交叉学科创新项目(450060481084).

通信作者:杨 博(ybo@jlu.edu.cn)

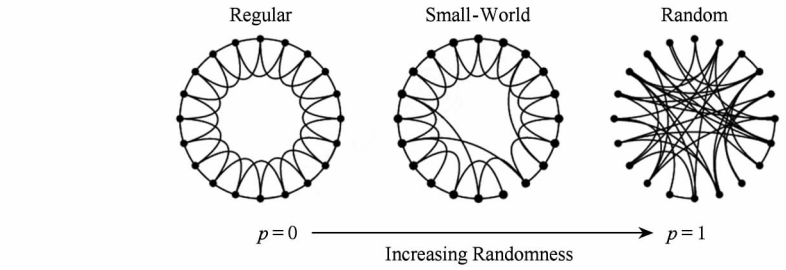
评估,选择了有代表性的6个社区挖掘算法和3个重叠社区挖掘算法进行测试,并给出了对比分析结果,试图为这个新兴研究领域勾画出一个较为全面和清晰的轮廓.

关键词 复杂网络;社区结构;社区挖掘;重叠社区挖掘;网络聚类

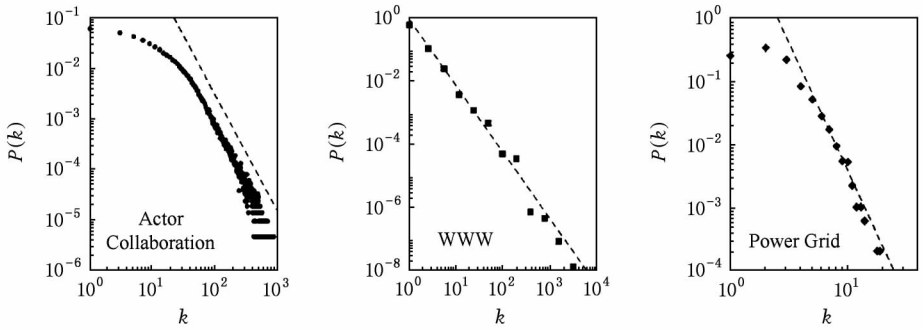
中图法分类号 TP18

现实世界中的许多复杂系统或以复杂网络的形式存在、或能被转化成复杂网络.例如: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、因特网和万维网等等.复杂网络普遍存在着一些基本统计特性,如反映复杂网络具有短路径长度和高聚类系数之特点的"小世界效应"[1];又如表达复杂网络中结点之度服从幂率分布特征的"无标度特性"[2];再如描述复杂网络中普遍存在着"同一社区内结点连接紧密、不同社区间结点连接稀疏"之特点的"社区结构特性"[3].目前,关于复杂网络基本统计特性的研究已吸引了不同领域的众多研究者,复杂网络分析已成为最重要的多学科交叉研究领域之一[1-3].图1中给出了上述统计特性的直观描述.

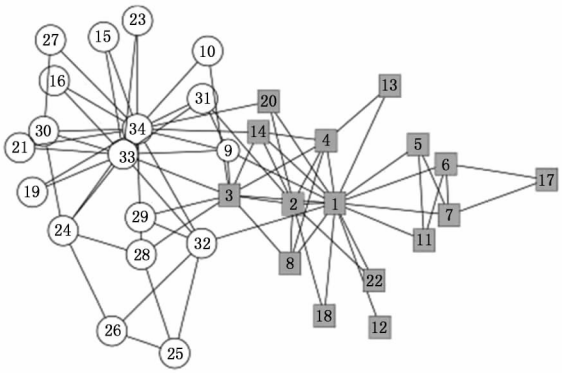
本文主要涉及复杂网络社区结构特性.随着应用



(a) The existence of "small world effect"[1]



(b) The presence of "power laws" in the link distribution[2]



(c) The community structure with a high density of within-group edges and a lower density of between-group edges[3]

Fig. 1 Some basic statistical properties of complex networks in the real world.

图1 真实世界中复杂网络具备的一些基本统计特性

领域的不同,社区结构具有不同的内涵.譬如,社会网中的社区代表了具有某些相近特征的人群、生物网络中的功能组揭示了具有相似功能的生物组织模块、Web 网络中的文档类簇包含了大量具有相关主题的 Web 文档等等[4].社区挖掘就是要探测并揭示出这些不同类型复杂网络中固有的社区结构,它可被用来帮助人们理解复杂网络的功能、发现复杂网络中隐藏的规律和预测复杂网络的行为.自 2002 年 Girven 和 Newman[3] 提出社区挖掘的概念至今,新的理论、方法层出不穷,新的应用领域不断涌现.它不仅吸引了来自计算机科学、物理、数学、生物和社会学等领域的研究者,成为最具挑战性的多学科交叉研究课题之一;而且已被应用于社会网络分析(如恐怖组织识别和组织结构管理等)、生物网络分析(如新陈代谢网络分析、蛋白质交互网络分析、未知蛋白质功能预测、基因调控网络分析和主控基因识别等)、Web 社区挖掘、基于主题词的 Web 文档聚类和搜索引擎等众多领域[5].

近年来,社区挖掘方面的研究工作不断被多个领域的著名国际杂志和国际会议所报道.这些刊物主要有:综合科学期刊《Nature》,《Science》,《Proceedings of National Academy of Sciences》等;物理数学领域期刊《Physics Review Letter》,《Physics Review E》,《Journal of Statistical Mechanics》,《New Journal of Physics》,《Europhysics Letters》,《European Physical Journal B》,《Physica A》等;计算机领域期刊《IEEE Transactions on Pattern Analysis and Machine Intelligence》,《ACM Transactions on Knowledge Discovery from Data》,《ACM Transactions on the Web》,《IEEE Transactions on Knowledge and Data Engineering》,《IEEE Transactions on Systems, Man, and Cybernetics》,《Machine Learning Journal》等;复杂系统领域期刊《Advances in Complex Systems》,《Complexity》,《Complex Systems》等;生物信息学期刊《BioInformatics》,《PLOS Computational Biology》,《PLOS One》等;计算机重要国际会议 SIGKDD, WWW, ICDE, ICDM 等.

此外,复杂网络社区挖掘方法已成为图论、复杂系统和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,例如密歇根大学物理系开设的"Complex Systems"课程、斯坦福大学计算机系开设的"Social and Information Network Analysis"课程、康奈尔大学计算机系开设的"The Structure of Information Networks"课程、麻省理工电子工程和计算机系开设的"Networks and Dynamics"课程等等.

最近人们还发现,在许多复杂网络中,结点可能同时属于多个社区.例如:在人际关系网中,每个人都可能会属于家庭、工作单位、朋友圈子等多个社会团体;在科学家协作网中,一个研究者可能会活跃在几个不同的研究领域;一个人按照自己的意愿,可能会对在线社会网中的多个组织感兴趣,那么他(或她)就能不受限制地加入这些组织;在生物网络中,一个结点可能会有多种生物功能,比如大部分蛋白质都同时属于多个蛋白质联合体(protein complexes)[6].为此,2005 年 Palla 等人[7] 对传统社区模式(每个结点仅属于一个社区)进行扩展,允许结点同时属于多个社区,从而开辟了重叠社区挖掘研究新领域.重叠社区模式不仅在现实世界中具有普遍性,而且具有十分重要的应用价值.因此,它迅速引起不同领域研究者们的广泛关注,并很快成为社区挖掘的前沿研究热点.

在上述背景下,本文综述了复杂网络社区挖掘当前的研究现状、面临的主要问题以及未来的发展前景,试图为社区挖掘研究勾画出一个较为全面和清晰的轮廓,为该领域的研究提供有益的参考.

1 复杂网络社区挖掘方法介绍

1.1 社区挖掘方法

近 10 年来,已有很多复杂网络社区挖掘方法被提出,它们分别采用了来自物理学、数学和计算机科学等领域的理论和技术,就其依据的原理可分为基于划分、基于模块性优化、基于标签传播、基于动力学和基于仿生计算的方法等.下面我们具体介绍各类代表性方法.

1.1.1 基于划分的社区挖掘方法

在复杂网络中发现社区的一个最简单的方法是:先找出社区间的所有链接,接着将它们全部删除,最后每个连通分支对应着一个社区.这就是基于划分的社区挖掘方法的基本原理,其关键点在于如何合理确定社区间的链接,从而产生合适的社区划分.

基于上述思想,2002 年,Girvan 和 Newman[3] 提出了最著名的社区挖掘方法 GN(Girvan-Newman).该算法采用的启发式规则为:社区间链接的边介数(edge betweenness)应大于社区内链接的边介数,其中每个链接的边介数被定义为"网络中经过该链

接的任意两点间最短路径的条数”.算法 GN 通过反复计算边介数,识别社区间链接,删除社区间链接,以自顶向下的方式建立一棵层次聚类树(dendrogram).该算法最大的缺点是计算速度慢.由于边介数的计算开销过大($O(mn)$),使其具有很高的时间复杂性($O(m^2n)$),其中 m 表示链接数,n 表示结点数.因此,该算法只适合处理中小规模网络(其包含的结点数通常$<10^3$).

算法 GN 在复杂网络社区挖掘研究中占有十分重要的地位,Girvan 和 Newman 工作[3]的重要意义在于:他们首次发现了复杂网络中普遍存在的社区结构,启发了其他研究者对这个问题进行深入研究,进而掀起了复杂网络社区挖掘研究的热潮.针对算法 GN 计算速度慢的不足,研究者们提出了一些改进方法.

2003 年,Tyler 等人[8]将统计方法引入算法 GN 中,提出一种近似的 GN 算法.他们的策略是:采用蒙特卡洛方法估算出部分链接的近似边介数,而不去计算全部链接的精确边介数.很显然,这种方法计算速度的提高是以牺牲聚类精度为代价的.

考虑到算法 GN 效率低是因为边介数计算开销过大引起的,2004 年,Radicchi 等人[9]提出了用链接聚类系数(link clustering coefficient)取代算法 GN 中链接的边介数.他们认为:社区间链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链接也会成为社区间链接,从而显著增加社区间的链接密度.基于上述直观假设,他们把链接聚类系数定义为包含该链接的短回路数目,并以“社区间边的链接聚类系数应小于社区内边的链接聚类系数”作为启发式规则.在算法的每次迭代中,具有最小链接聚类系数的边被删除.由于链接聚类系数的平均计算时间是 $O(m^3/n^2)$,所以该算法的时间复杂性为 $O(m^4/n^2)$.对于稀疏网络而言,该算法的时间复杂度为 $O(n^2)$,要优于算法 GN 的时间复杂度($O(n^3)$).该算法的局限性是,它不适合处理短回路很少甚至没有短回路的复杂网络.

2010 年,刘大有等人[10]采用结构相似度(structural similarity)取代了算法 GN 中的边介数.结构相似度是根据社会学中熟人模型建立的:如果将 2 个人看作是社会网中的 2 个结点,若他们共享的朋友圈越大则他们就可能越熟悉,从而他们位于同一社区内的可能性也就越大.基于上述直观假设,该算法采用的启发式规则如下:社区间链接的结构相似度应小于社区内链接的结构相似度.在算法的每次迭代中,具有最小结构相似度的边被删除.由链接结构相似度的平均计算时间为 $O(k^2)$,其中 k 表示平均度,可知该算法的时间复杂性为 $O(mk^2)$.对于稀疏网络,该算法甚至有近似线性的时间复杂度.因为该方法主要是针对社会网络结构设计的,所以它不一定适合所有类型的复杂网络.

1.1.2 基于模块性优化的社区挖掘方法

2004 年,Newman 和 Girvan[11]提出了一个用于刻画网络社区结构优劣的量化标准,被称之为模块性函数 Q.函数 Q 给出了社区结构的清晰定义,并在实际应用中获得很大成功.因此,函数 Q 逐渐被相关领域的研究者广泛接受,同时以函数 Q 为目标函数的优化方法也成为复杂网络社区挖掘领域的主流方法之一.下面对基于模块性优化的一些典型方法进行介绍.

2004 年,Newman[12]提出了第 1 个基于模块性优化的社区挖掘方法(fast Newman, FN).该算法中候选解的搜索策略为:选择并合并两个现有的社区.初始化时,候选解中每个社区仅包含一个结点;在每次迭代时,算法 FN 选择使函数 Q 值增加最大(或减小最少)的社区对进行合并;当候选解只对应一个社区时算法结束.通过这种自底向上的层次聚类过程,算法 FN 输出一棵层次聚类树(dendrogram),然后将对应的函数 Q 值最大的社区划分作为最终聚类结果.该算法的时间复杂性为 $O(mn)$.

2005 年,Guimera 和 Amaral[13]提出了基于模拟退火的模块性优化算法(simulated annealing, SA).该算法首先随机生成一个初始解;在每次迭代中,在当前解的基础上产生一个新的候选解,由函数 Q 判断其优劣,并采用模拟退火策略中的 Metropolis 准则决定是否接受该候选解.SA 算法产生新候选解的策略是:将结点移动到其他社区、交换不同社区的结点、分解社区或合并社区.该算法具有非常好的聚类质量,但其缺点是运行效率低.据报道,在普通配置的计算机上采用 SA 算法处理包含 3 885 个结点和 7 260 条边的酵母菌蛋白质交互网络需要 3 d 时间[14].

2006 年,Newman[15]将谱图理论引入模块性优化中.他将模块性函数 Q 表达成一个图拉普拉斯矩阵,又称为模块性矩阵;并证明了模块性矩阵的第二大特征值之特征向量的正负二分结果,正好对应了模块性优化的二分结果;基于此,他提出了一个优化函数 Q 的谱方法.该算法具有很高的聚类质量,但其时间复杂度仍稍高,在 $O(n^2\log n)$ 和 $O(n\log n)$ 之间.

2008 年,Blondel 等人[16]提出了快速模块性优化方法(fast unfolding algorithm, FUA).该算法结合了局部优化与多层次聚类技术.它首先使每个结点在其邻居区域内局部优化模块性函数 Q,获得一个社区划分结果;然后将得到每个社区作为一个超级结点、社区间的链接作为加权边,构建一个上层网络;不断迭代上述两步,直到函数 Q 的值不再增加为止.该算法对于稀疏网络具有线性时间复杂度 $O(m)$,同时可获得非常高的聚类质量,被著名社区挖掘专家 Fortunato 等人[17]认为是当前性能最佳的模块性优化算法.

2011 年,刘大有等人[18]针对目前复杂网络的规模越来越庞大、且呈天然分布式特性的特点,从局部观点出发,提出了一个快速社区挖掘算法(fast network clustering algorithm, FNCA).他们基于对模块性函数 Q 的分析,提出一个针对单个结点的局部目标函数 f,并证明了函数 Q 随网络中任一结点的函数 f 呈现出单调递增特性;进而提出一个基于局部优化的近似线性的社区挖掘算法.在该算法中,每个结点只利用网络局部社区结构信息优化自身的目标函数 f,所有结点通过相互协同实现整个网络的聚类.该算法不仅有相当高的运行效率和聚类质量,而且适用于分布式网络社区挖掘.

值得注意的是,尽管模块性函数 Q 已被人们广泛接受,但它仍然存在一些不足.例如,人们最近发现函数 Q 存在分辨率限制(resolution limit)[19]和极端退化(extreme degeneracies)现象[20]等.Blondel 等人[16]认为,设计多层次、多粒度社区挖掘算法可缓解分辨率限制问题.Khadivi 等人[21]认为,在应用社区挖掘算法之前采用链接加权的预处理机制,可缓解函数 Q 的分辨率限制问题和极端退化现象.目前,如何有效解决上述两个问题仍然是模块性优化社区挖掘方法所面临的挑战性问题.

1.1.3 基于标签传播的社区挖掘方法

标签传播方法是一类启发式算法,和模块性优化算法不同,它没有特定的目标函数,而是通过一种直觉、富有启发的思想推断社区结构和设计算法.标签传播类方法的启发式规则为"在具有社区结构的网络中,任一结点都应当与其大多数邻居在同一个社区内".

2007 年,Raghavan 等人[22]提出了著名的标签传播算法(label propagation algorithm, LPA).该算法的流程为:初始化时,为每个结点赋一个唯一标签;每次迭代中,每个结点采用大多数邻居的标签来更新自身标签;当所有结点的标签都与其多数邻居的标签相同时,算法结束.此时,稠密子图中所有结点通过标签达成共识,即形成了社区.为保证收敛和避免循环,他们建议在每次迭代前对结点随机排序,并异步更新结点标签.算法 LPA 每次迭代消耗的时间为线性 $O(m)$,且只需极少的迭代次数即可收敛.Leung 等人[23]指出,算法 LPA 一般在 5 步之内就能使 95%以上的结点达到稳定状态.

2008 年,Tibély 等人[24]发现标签传播算法 LPA 等价于最小化哈密尔敦函数,并指出该算法对一些网络是无效的.进而,他们通过两个真实网络进行评估,发现算法 LPA 对同一网络会产生大量不同的聚类结果,该数目甚至超过网络中的结点数.因此,他们认为标签传播算法 LPA 的聚类性能有待改善.

2009 年,Leung 等人[23]将算法 LPA 作为分析大规模在线社会网的工具.他们通过研究算法 LPA 的优势和限制,讨论了其扩展和优化方面的一些问题,进而对算法 LPA 进行了修正.实验表明,在普通 PC 机上该修正算法可在几分钟内有效聚类包含上千万条边的网络.他们的研究结果预示了:标签传播算法作为一种快速有效的社区挖掘算法,在处理超大规模网络方面蕴含着巨大的潜力.

2009 年,Barber 等人[25]将算法 LPA 等价为一个优化问题,并给出对应的目标函数.通过研究该目标函数的特性,他们揭示出算法 LPA 在原理及实际应用方面的缺陷.最重要的一点是,在该算法运行过程中,目标函数值的增加并不一定意味着社区质量的改善.为了克服这一缺陷,他们对目标函数进行修正,并设计了一个带约束的标签传播算法(modularity-specialized label propagation algorithm, LPAm).特别令人感兴趣的是,这个修正的目标函数正好是模块性函数 Q,改进后的算法 LPAm 也就自然对应了模块性函数优化,从而他们解决了文献[24]提出的问题.

2010 年,Liu 等人[26]发现算法 LPAm 得到的社区划分具有"每个社区内结点的度之和都相似"的特性,就是说该算法有陷入局部最优解的倾向.为跳出局部最优解,他们给出一种多步层次贪婪算法(multistep greedy agglomerative algorithm, MSG),每次可合并多个社区对.进而他们将算法 LPAm 与 MSG 相结合,提出了一个基于模块性优化、层次化标签传播算法 LPAm+,使标签传播类算法的聚类性能得到进一步改善.

1.1.4 基于动力学的社区挖掘方法

基于 Markov 随机游走理论的启发式求解策略也被广泛应用于复杂网络社区挖掘领域.下面介绍这方面的代表性工作.

2000 年,van Dongen[27]提出了 Markov 聚类算法(Markov cluster algorithm, MCL).该算法主要是基于 Markov 动力学理论,通过改变和调节 Markov 链呈现出网络社区结构.MCL 模拟了网络中许多的随机游走流.它通过将转移概率值提升到一个大于 1 的幂率,强化已经很强的流,弱化较弱的流.通过不断重复这一过程使社区结构逐渐清晰.当一些具有强内部流的区域被那些流几乎不可见的边界所划分开时,也就是说高链接密度区域被低链接密度区域分开时,这一迭代过程就结束了.值得指出的是,MCL 算法已被广泛应用于社区挖掘领域[28-29].

2007 年,杨博等人[30]针对符号网络社区挖掘问题(包括正负权值的网络),提出了基于 Markov 随机游走模型的启发式社区挖掘算法(finding and extracting communities, FEC).该算法所采用的基本假设是:从任意给定的社区出发,网络中的随机游走过程到达起始社区内结点的期望概率将大于到达起始社区外结点的期望概率.基于该启发式规则,FEC 算法首先计算出在给定时刻随机游走过程到达所有结点的期望转移概率分布,进而根据该分布的局部一致性(同社区结点具有近似相同的期望转移概率分布)识别出各个不同的社区.FEC 算法是第 1 种综合考虑两种社区标准(即连接密度和连接符号)的复杂网络社区挖掘算法.它既能有效处理符号网络,又能有效处理仅包含"正关系"的一般复杂网络.

2008 年,Rosvall 等人[31]提出了映射平衡算法 infomap.该方法基于最小描述长度(MDL)原理[32],通过信息传播扩散技术探测网络社区结构.MDL 原理的主要思想为:数据中包含的任何规律性都可用来压缩数据长度.如果我们对某 agent 在网络中随机游走的路径进行编码,并用社区结构这一规律压缩数据,那么该路径的最小描述长度所对应的网络划分就构成了一个有效的社区结构.值得注意的是,infomap 算法不仅可从网络中发现社区结构(同配结构),还可找出多方结构(异配结构),因此它更具一般性.

2011 年,Morārescu 等人[33]研究了一类离散时间的多 agent 系统,基于信任度衰减的观点建立动力学模型.他们将复杂网络视为一个 agent 网络,其中每个 agent 拥有一个信念值.在每一个时间步,这些 agent 与邻居结点进行通信,并用它信任的邻居结点之信念值来更新自身信念(当两个结点的信念值之差小于某阈值时表示相互信任).由于信念阈值是随时间衰减的,因此每个 agent 只会将它的信任给予那些与它迅速达成共识的邻居.本质上来说,只有当 agents 信念互相逼近的速度大于阈值衰减的速度时,它们才会达成共识.可见,在信念阈值的约束下,很难达成全局共识,一般只能形成局部共识,这些达成局部共识的 agents 就形成了一个社区.最后他们采用上述信念动力学模型设计了社区挖掘算法.

2012 年,杨博等人[34]给出了一个采用 Markov 转移矩阵的特征值来评估亚稳态之进出时间的方法,揭示了网络内在属性与社区结构的数学联系,提出了分析复杂网络社区结构的谱理论.基于此,定义了 3 个刻画社区结构的量,分别为社区之间的分离度、每个社区的凝聚度和刻画社区结构的谱特征.进而他们给出了基于特征系统刻画和识别网络社区的理论框架.然而,上述谱分析理论需要计算网络的特征值,特征值的计算时间一般为 $O(n^3)$,因此它无法有效处理现实世界中的大规模网络.针对该问题,他们又基于 Markov 随机游走亚稳态的两个基本特性(局部一致性和暂时稳定性),为上述谱分析理论提出一种快速的近似算法.

1.1.5 基于仿生计算的社区挖掘方法

近年来,仿生算法也逐渐成为一种有竞争力的社区挖掘方法,已引起许多研究者关注.这方面的研究工作主要侧重于蚁群算法和遗传算法.

这里先介绍基于蚁群算法的社区挖掘方法.2007 年,Liu 等人[35]基于每个蚂蚁个体的行为,提出了一个用于探测邮件社会网社区结构的蚁群聚类算法.2010 年,他们在聚类目标与群体相似度之距离的计算方面,对其先前工作进行了改进[36].2009 年,Sadi 等人[37]采用蚁群优化技术发现网络中的团,并将这些团视为新结点而构建一个简化网络,然后通过传统社区挖掘算法来探测社区结构.2010 年,他们又通过将团扩展为准团,进一步改进其前面的算法[38].2010 年,刘大有等人[39]从仿生角度出发提出一个基于 Markov 随机游走的蚁群算法(ant colony optimization based on random walk, RWACO).RWACO 将蚁群算法框架作为基本框架.以 Markov 随机游走模型作为启发式规则,通过集成学习的思想将蚂蚁的局部解融合为全局解,并用其更新信息

素矩阵. 通过“强化社区内连接,弱化社区间连接”这一进化策略逐渐呈现出网络的社区结构. 2011 年,他们基于集成网络理论提出了一个受约束的随机游走新模型,对上述蚁群算法进行改进,进一步提高了该算法的聚类质量和运行效率[40]. 2012 年,何东晓等人[41]提出一个多层蚁群算法(multi-layer ant-based algorithm, MABA),它包含单层蚁群子算法(single layer ant-based algorithm, SABA). SABA 结合自避免标签传播技术与模拟退火策略,通过使每只蚂蚁局部优化模块性函数 Q 实现社区探测. MABA 首先执行子算法 SABA,以获得网络社区结构;之后通过将社区视为结点、社区间的链接视为加权边,构建一个上层网络;接着再将 SABA 用于新的上层网络. 这一过程不断迭代,直到函数 Q 值不再增加为止. MABA 不仅能探测复杂网络蕴含的层次社区结构,而且能缓解模块性优化带来的分辨率限制问题.

下面介绍基于遗传算法的社区挖掘方法. 它们主要分为两类,分别采用字符串编码方式和基于位置的邻接编码方式.

针对社区挖掘问题,研究者们首先注意到了字符串编码方式. 字符串编码简单、直观,但缺点是它不适于传统的交叉操作符. 2007 年,Tasgin 等人[42]提出了第 1 个用于社区挖掘的遗传算法,其主要贡献是给出了一种适合字符串编码的单路交叉操作. 他们采用一些小规模网络对算法进行验证. 2010 年,何东晓等人[43]提出了一种基于聚类融合的社区挖掘遗传算法. 该算法将聚类融合引入到交叉算子中,利用父个体的聚类信息辅以网络拓扑结构的局部信息产生新个体,避免了传统交叉算子单纯交换字符块而忽略了聚类内容所带来的问题. 该算法可获得很高质量的聚类结果,但缺点是运行效率较低. 2011 年,刘大有等人[44]在分析模块性函数 Q 之局部单调性的基础上,提出一种快速、有效的局部搜索变异策略;进而结合单路交叉算子和 $\mu+\lambda$ 选择算子,给出了一个遗传型社区挖掘算法. 实验表明,该算法可有效分析大规模网络. 2011 年,公茂果等人[45]结合了遗传算法和爬山法局部搜索,提出了一种用于社区探测的密母算法. 实验结果表明,该算法能有效发现复杂网络蕴含的层次社区结构.

后来人们注意到另外一种编码方式——基于位置的邻接编码,亦被称为基于图的编码. 与字符串编码不同,这种基于图的编码方式适用于传统交叉算子,因此人们无需再为其专门设计新的交叉方法. 2008 年,Pizzuti[46]首次将基于图的编码策略引入到社区挖掘领域. 他们采用均匀交叉、随机变异和精英选择 3 个传统的遗传算子,设计了探测复杂网络社区结构的遗传算法. 但验证算法时仅使用了一些小规模网络. 2009 年,他们采用了两种社区函数,将其先前的方法扩展为多目标遗传算法[47]. 2010 年,石川等人[48]对 Tasgin 等人[42]针对字符串编码设计的单路交叉算子进行修正,进而将其引入基于图编码的遗传算法中,设计了一个能有效处理大规模网络的遗传算法. 2010 年,刘大有等人[49]发现传统变异方法对于图编码策略的不足,给出了边缘基因的概念,并推导出模块性函数 Q 的局部单调性;进而将上述两点有机结合,提出了一个快速有效的局部搜索变异方法;最后结合均匀交叉与 $\mu+\lambda$ 选择算子,设计了一个可有效处理大规模网络的遗传算法.

1.2 重叠社区挖掘方法

复杂网络中的重叠社区模式是对传统社区概念的一个重要拓展. 近年来,重叠社区挖掘已成为社区挖掘领域的研究热点,受到广泛关注. 目前已提出诸如团渗方法、线图方法和局部扩展方法等一些重叠社区挖掘算法. 下面我们将具体分析 3 类代表性方法.

1.2.1 基于团渗理论的方法

基于团渗理论的方法是一类被最早提出的重叠社区挖掘方法. 它的主要思想是将社区视为由一些由团(全连通子图)构成的集合,这些团之间通过共享结点而紧密连接.

2005 年,Palla 等人[7]发现复杂网络中普遍存在重叠社区结构,提出了著名的团渗算法(clique percolation method, CPM). Palla 等人认为社区是由若干重叠的团构成的,通过搜索邻接的团可探测网络社区. CPM 首先从网络中找出所有大小为 k 的团,然后将每个 k 团作为顶点构建一个新的图,当两个 k 团共享 $k-1$ 个结点时,新图中两个对应的顶点间才会有边;新图中的每个连通子图所对应的 k 团集合便构成了一个社区. 由于一个结点可能会同时属于多个 k 团,所以 CPM 找到的社区自然会出现重叠现象. Palla 等人指出:当 $k=4$ 时,CPM 获得的聚类结果最好. Palla 等人的工作掀起了重叠社区挖掘的研究热潮.

CFinder[50]是基于 CPM 算法实现的一个软件工具,虽然其时间复杂度为非多项式级,但它在很多情况下的运行效率还是很高的.

2009 年,沈华伟等人[51]将极大团(极大全连通子图)视为社区的基本构成单元,提出了一个基于极

大团的凝聚式层次聚类方法,用于发现网络中的重叠、层次社区结构.该算法首先找出网络中所有的极大团,并作为初始社区;然后将具有最大相似度的社区对不断合并,从而形成一棵层次聚类树;最后他们将模块性函数 Q 扩展为重叠社区函数 EQ,作为层次聚类树的截断标准,从而获得最优的重叠社区结构.该算法的时间复杂性为 $O(n^2+(h+n)s)$,仍然较高,其中 s 是极大团的数目(理论上界为 $3^{n/3}$),h 为相邻极大团对的数目.

2010 年,Evans 等人[52]为了发现重叠社区,先将原网络转化为一个团图,然后用一个合适的结点划分算法聚类该团图.他们的出发点为:团图不仅使研究者能避免因"以结点为中心研究网络"所导致的偏好,而且还能使他们灵活选择合适的结点分析技术分析团图.

1.2.2 基于链接划分的方法

基于链接划分的方法主要是将链接而不是结点作为考虑对象,通过设计适当的划分策略来获取链接社区结构.由于链接社区间可能会出现一些重叠的结点,所以它天然的对应了结点的重叠社区结构.同时,链接社区挖掘本身也是一个非常值得研究的课题.

2009 年,Evans 等人[53]首次对网络中的链接进行划分,以产生结点的重叠社区结构.他们首先通过"用边表示结点,用结点形成边"的方法,将原始网络转化为线图;然后选择已有的结点划分算法以获取链接社区结构.此后,他们又考虑了结点度分布的异构性,提出了一个加权线图构建方法,进一步改善了原算法的性能.

2010 年,Ahn 等人[54]提出一个基于边相似性划分链接的层次聚类方法.给定由结点 k 连接的一对边 e_{ik} 和 e_{jk},用 Jaccard 指数计算它们的相似度 $S(e_{ik},e_{jk})$,$S(e_{ik},e_{jk})=|n_+(i)\cap n_+(j)|/|n_+(i)\cup n_+(j)|$,其中 $n_+(i)$ 表示包含结点 i 及其邻居结点的集合.然后用单链层次聚类算法产生一棵链接社区结构的层次树.最后定义一个用于截断该层次树的密度函数 D,从而获得最优链接划分.该算法的时间复杂性为 $O(nk_{\max}^2)$,其中 $k_{\max}$ 是网络中最大的结点度.

2011 年,Kim 等人[55]将结点社区定义为链接社区,进而把结点社区挖掘算法 infomap[31]有效扩展到链接社区挖掘领域.他们工作的另一个意义在于,针对任意网络,人们可以在仅了解网络拓扑结构的情况下,定量的判断是结点社区更好还是链接社区更好.在实际应用中,基于此人们就能决定究竟是采用结点社区还是链接社区.

2011 年,Ball 等人[56]提出一个用于探测链接社区结构,采用生成网络模型(随机块模型)的统计学方法,并给出了一个快速的期望最大化算法.实验表明,该算法可在合理时间内处理包含上百万个结点的大规模网络.但它的最大缺点是社区数目 k 需事先指定.此外,该工作还给出了一种重要的针对链接社区的基准网络生成方法.

1.2.3 基于局部扩展的方法

基于局部扩展的方法主要通过采用具有局部特性的方法,不断从网络中探测出局部社区.如果网络具有重叠社区结构,那么这些被发现的社区则天然呈现重叠现象.

2009 年,Lancichinetti 等人[57]提出了基于局部扩展的重叠社区挖掘算法(local fitness measure,LFM).LFM 先从随机选择的一个种子结点出发,通过不断向外扩张构建社区,直至社区函数 $f(c)=k_{\text{in}}^c/(k_{\text{in}}^c+k_{\text{out}}^c)^\alpha$ 达到局部最优为止,其中 k_{in}^c 和 k_{out}^c 分别表示两个顶点均在和只有一个顶点在社区 c 内的链接数,α 表示分辨率参数,它控制所产生社区之大小.当 LFM 获得一个社区后,它将从已生成的社区外随机选取某结点作为新种子,迭代执行上述局部扩张操作.该算法不断探测局部社区,直至网络的所有结点均被分配给某一或某些社区为止.LFM 的最坏时间复杂性为 $O(n^2\log n)$.该算法是第 1 个能同时探测重叠和层次两种结构的社区挖掘算法.

2010 年,Lee 等人[58]针对当前重叠社区挖掘算法大都无法有效发现高度重叠社区结构的问题,提出了一个贪婪的团扩张算法(greedy clique expansion,GCE).该算法首先找出一些明显的团作为种子;然后从这些种子出发,通过贪婪搜索策略对社区函数 $f(c)$ 进行局部优化,以扩充这些种子构建局部社区;算法收敛时即得到重叠社区结构.GCE 与 LFM 的重要差别在于它们选择种子的方法不同.将团作为种子使得 GCE 具有更好地发现高度重叠社区结构的能力.

不同于以往只考虑原始网络的社区挖掘方法,2011 年刘大有等人[59]将原始网络与相应的退火网络[60]结合成一个集成网络,进而在集成网络的基础上提出重叠社区挖掘算法(unfold and extract overlapping communities,UEOC).该算法首先给出一个基于集成网络的约束 Markov 游走方法,使网络中的社区逐渐得以呈现;然后基于局部社区函数"导

电率”[61]，设计了一个有效的截方法，将已呈现的固有社区全部被抽取出来. 如果该网络具有重叠结构，则被抽取出的社区将自然呈现结点重叠现象. 特别是，UEOC 仅有一个无需先验知识很容易被确定的参数. UEOC 的时间复杂性为 $O(Kn^2)$，其中 K 表示重叠社区数.

2 实 验

为定量评估不同算法的性能，这里选择了具有代表性的 6 个社区挖掘算法和 3 个重叠社区挖掘算法，并进行实验比较. 算法实验环境为：处理器 Intel® Xeon® CPU 5130@2.00 GHz 2.00 GHz，内存 4.00 GB，硬盘 160 GB，操作系统 Microsoft Windows Server 2003. 文中使用的所有算法的源代码都是从作者获取的.

2.1 社区挖掘算法的比较

这里选择了 6 个优秀的社区挖掘方法进行测试，它们分别为：FN 算法[12]、FEC 算法[30]、LPA 算法[22]、infomap 算法[31]、FUA 算法[16]和 MABA 算法[41]，1.1 节已给出这些算法的介绍. 我们采用两种常用的人工生成网络（GN 基准网络[3]和 LFR 基准网络[62]）评估不同社区挖掘算法的性能. 此外，我们采用一种基于信息论的公制 NMI（normalized mutual information）[63]作为精度度量标准.

GN 基准网络由 Newman 等人提出[3]. 对于该基准网络，每个图包含了 128 个结点，分为 4 个由 32 个结点构成的社区. 每个结点平均有 z_{in} 条边与同社区内结点连接，z_{out} 条边与社区外结点连接. 其中 $z_{in}+z_{out}=16$，作为每个结点期望的度. 随着 z_{out} 的增大，所产生的随机网络给社区探测算法带了更大的挑战. 特别是当 $z_{out}>8$ 时，意味着每个结点在社区内的边都要小于社区外的边，这时网络的社区结构非常模糊[3]. 图 2 显示了每种算法获得的聚类精度随 z_{out} 的变化趋势. 可见，在 Newman 基准网络上，算法 FUA 与 MABA 的聚类精度是最高的.

LFR 基准网络由 Lancichinetti 等人[62]提出. 和 GN 基准网络不同，它具有“结点度和社区规模呈幂率分布”的异构特性，这一点与真实网络相似. 与 Lancichinetti 等人[62]设计的实验相同，这里如下设置 LFR 基准网络的参数：网络规模 n 为 1 000 或 5 000；最小社区大小为 10 或 20；混合参数 u（任一结点 p 与社区外结点形成的链接占结点 p 之度的比例）从 0 增加至 0.8，间隔为 0.05. 同时保持其他参数不变：平均度 $d=20$；最大度 $d_{max}=2.5\times d$；最大社区规模 $c_{max}=5\times c_{min}$；结点度的幂率分布系数 $\tau_1=-2$；社区规模的幂率分布系数 $\tau_2=-1$. 该参数设置方法产生了 4 种基准网络. 图 3 显示了每种算法获得的聚类精度随混合参数 u 的变化趋势. 可见，在这个更具挑战性的基准网络上，算法 infomap 的聚类精度是最高的，算法 FUA 与 MABA 稍逊于 infomap.

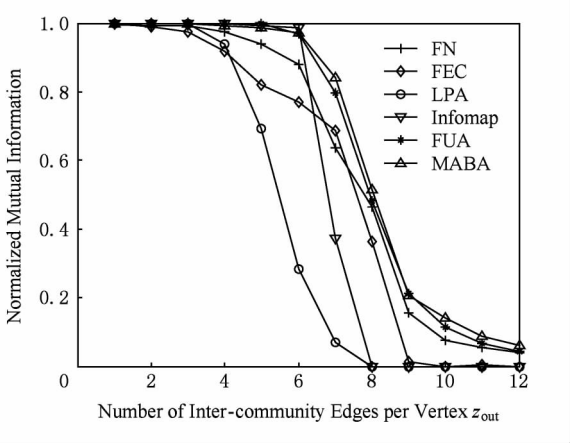


Fig. 2 Comparison of different algorithms in terms of accuracy on the GN benchmarks.

图 2 在 GN 基准网络上不同算法聚类精度的比较

在上述算法中，FEC 与 LPA 为启发式算法，其启发式规则与网络社区结构类型的匹配程度决定了它们的精度. 其中，FEC 以“Markov 随机随机过程的亚稳性”作为启发式规则，LPA 以“结点应与多数邻居在同一社区内”作为启发式规则. 从实验结果可见，FEC 的启发式规则更适合于 Newman 基准网络，而 LPA 的启发式规则在 LFR 基准网络上表现更好. FN，infomap，FUA 与 MABA 均为基于优化的算法，其精度性能依赖于优化目标，同时也与算法的寻优能力相关. 其中，infomap 采用基于信息流压缩的目标函数 L，而 FN，FUA 与 MABA 均采用了著名的模块性函数 Q. 从实验结果可见，函数 Q 更适合于 Newman 基准网络，而函数 L 则在 LFR 基准网络上表现更好；在以函数 Q 为目标函数的算法中，FUA 与 MABA 的寻优能力要明显强于 FN. 此外，基于优化方法的精度整体优于启发式方法的精度. 总体上看，上述算法中的 infomap，FUA 与 MABA 是精度最高的.

我们选择的这 6 种算法运行效率都比较高，就时间复杂性而言，FN 为 $O(mn)$，FEC 为 $O(m+n\log n)$，LPA 为 $O(m)$，infomap 为 $O(m)$，FUA 为 $O(m)$，MABA 为 $O(cn)$，其中 n 表示网络结点数，m

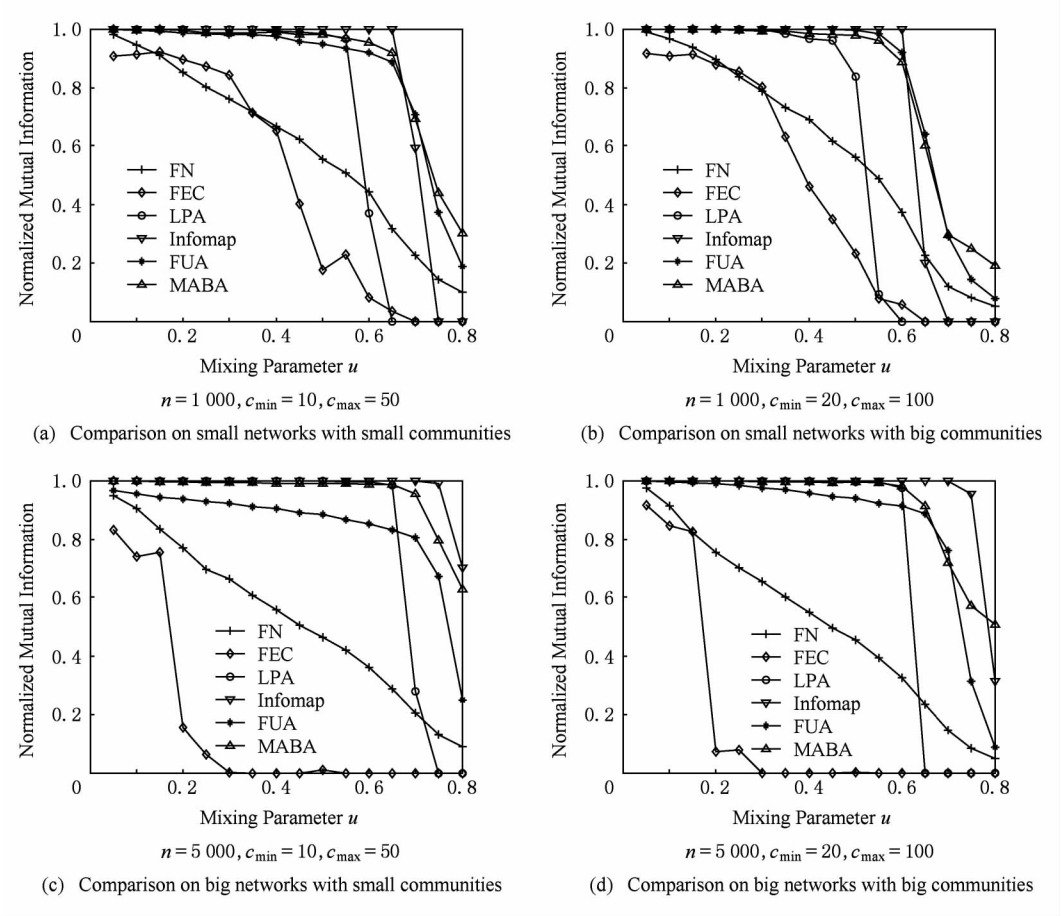


Fig. 3 Comparison of different algorithms in terms of accuracy on the LFR benchmarks.

图3 在LFR基准网络上不同算法聚类精度的比较

表示链接数,c 为社区平均规模.可看出,LPA,infomap和FUA均为线性时间,FEC与MABA在线性时间与二次多项式时间之间,而FN为二次多项式时间算法.那么,若综合考虑聚类精度和运行效率两个方面,算法infomap与FUA应排在前面.

2.2 重叠社区挖掘算法的比较

我们选择了3个具有代表性的重叠社区挖掘方法进行测试,它们分别为CPM算法[7]、LFM[57]算法和UEOC算法[59],2.2节已给出算法介绍.这里采用Lancichinetti等人[64]提出的重叠基准网络Lancichinetti评估不同重叠社区挖掘算法的性能.Lancichinetti模型生成的网络不仅具有结点度和社区规模呈幂率分布的异构特性,而且呈现出重叠社区结构特性.

目前尽管已有一些评估已知社区结构与算法获得的社区结构间相似程度[63]的精度度量标准,但这些标准的大多数并不适合于重叠社区.幸运的是,最近文献[57]对被广泛应用的标准共享信息(normalized mutual information, NMI)度量进行了扩展,使其能处理重叠社区.因此,在实验中我们采用了NMI之扩展,即扩展的标准共享信息(extended normalized mutual information, ENMI)作为精度度量标准.

与Lancichinetti等人[64]设计的用于测试CPM探测重叠社区之能力的实验相同,我们对Lancichinetti基准网络的参数设置如下:网络规模 $n=1\,000$,最小社区大小 $c_{\min}=10$ 或20,混合参数 u(任一结点 p 与社区外结点形成的链接占结点 p 之度的比例)设置为0.1或0.3,重叠结点所占的比例(o_n/n)按间隔0.05从0增加至0.5.同时,保持其他参数不变:平均度 $d=20$,最大度 $d_{\max}=2.5\times d$,最大社区规模 $c_{\max}=5\,c_{\min}$,每个重叠结点所属的社区数 $o_m=2$,结点度的幂率分布系数 $\tau_1=-2$,社区规模的幂率分布系数 $\tau_2=-1$.该设置方法产生了4种基准网络.

图4显示出针对具有重叠社区结构的异构人工网络,算法CPM,LFM和UEOC聚类精度的比较.从图4可以看出,对于社区规模相对较大的网络,算法UEOC最有效;对于社区规模相对较小的网络,算法CPM最有效;对于这两种情况,算法LFM的

性能都属于中等. 此外,随着混合参数 u 的变大,算法 UEOC 的聚类质量相对稳定,即退化不明显;然而,在该情况下,其他两种算法的聚类精度则迅速下降.

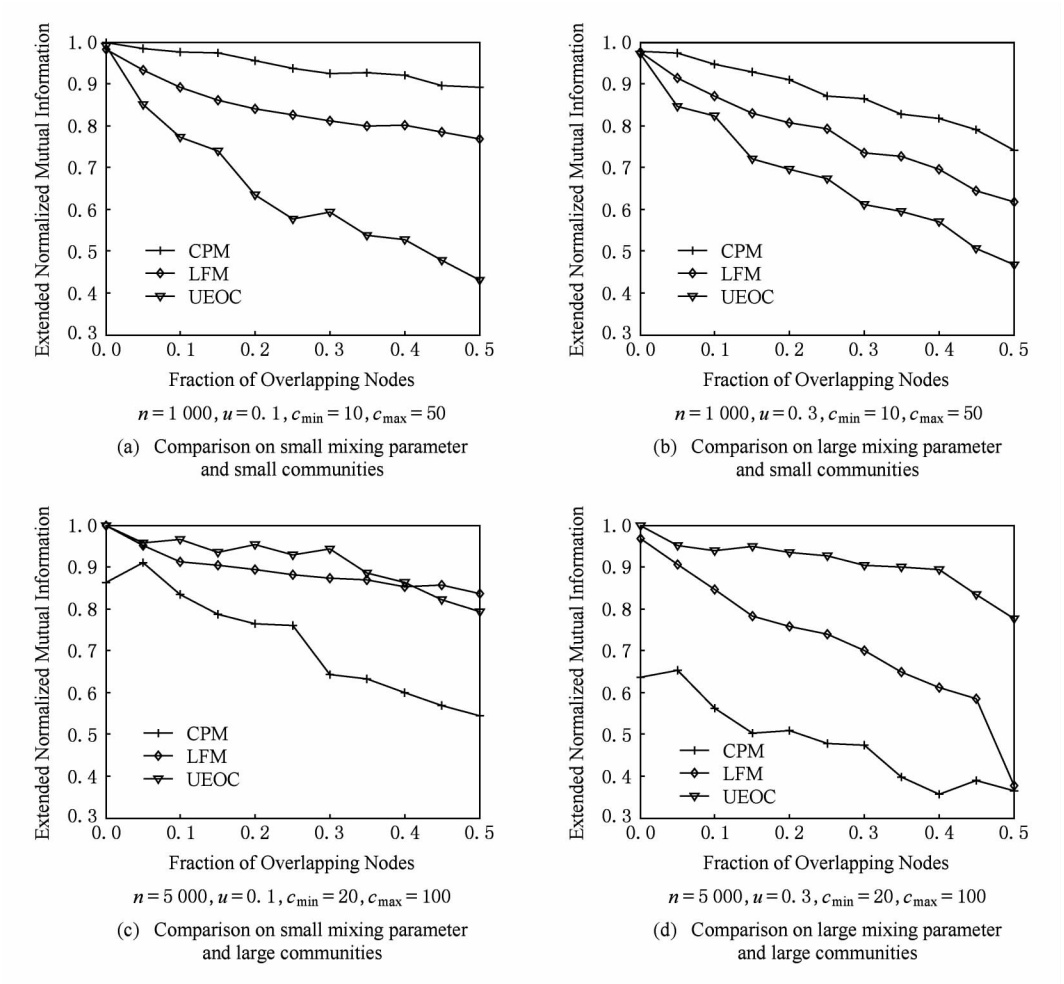


Fig. 4 Accuracy of each algorithm as a function of the fraction of overlapping nodes.

图 4 诸算法之聚类精度随重叠结点所占比例(o_n/n)的变化趋势

上述算法中,CPM 是基于团渗理论的扩张方法,LFM 是基于局部社区函数优化的扩张方法,而 UEOC 为基于约束 Markov 动力学的算法. 其中 CPM 与 LFM 均属于局部算法,更容易找到较小的社区;而 UEOC 为全局算法,在探测过程中需要利用网络的全局拓扑信息,更善于发现较大的社区. 此外,CPM 的 k 团扩张较 LFM 的贪婪优化扩张具有更强的局部特性,因此,它更倾向于找到规模较小的社区. 总体来说,上述 3 种算法在精度方面性能相当.

算法 CPM 的时间复杂性为非多项式级,即使它通常的运行效率还是很高的;算法 LFM 的时间复杂度为 $O(n^2 \log n)$;算法 UEOC 的时间复杂度为 $O(Kn^2)$. 其中 n 表示网络结点数,K 表示社区数目. 可见,若综合考虑聚类精度和运行效率两个方面,算法 UEOC 和 LFM 要稍优于算法 CPM.

3 结 论

本文主要从以下两方面对现有工作进行了综述:1)根据所采用的基本原理,分别对非重叠社区挖掘方法和重叠社区挖掘方法进行分类,并从中选择代表性方法进行介绍;2)从聚类精度及时间复杂性两方面定量分析、比较了一些典型方法的性能.

尽管社区挖掘已取得了许多令人鼓舞的成果,但该问题还远未被很好地解决. 下面我们给出一些当前亟待解决的问题:

第一,关于何为社区,目前有许多各具特色的观点,尚难达成共识. 因此,人们不仅无法判断哪个算法是最好的,而且也无法控制大量新方法的盲目产生. 针对该问题,人们已开展了一些工作,譬如:Lancichinetti 等人[62]定义了基准网络生成器,使人

们可通过已知社区结构评估算法优劣,然而现实世界中的网络是复杂的,人工网络很难代替真实网络;Newman 等人[11]定义了模块性函数 Q,使人们可通过函数 Q 值来评估算法的优劣,但函数 Q 是有偏的,譬如它存在分辨率限制、极端退化等问题.因此,该领域亟需给出一个能精确定义社区挖掘算法之目标的理论框架.

第二,许多真实复杂网络中的社区包含层次结构、具有重叠现象.譬如:一个大学拥有多个学院,学院又由多个系所组成,每个系又可能包含多个实验室;通常由于交叉研究的需要,学院或系之间可能会组建联合实验室.然而到目前为止,既可探测层次结构又能发现重叠结构的社区挖掘算法并不多见,而且人们在设计此类算法时,还未能综合考虑层次与重叠结构之间的内在联系,这可能导致其发现的层次与重叠结构之间存在矛盾[54].此外,当前还没有一种能合理度量层次化重叠社区结构优劣的量化标准.

第三,在一些复杂网络中,每个结点都可能属于多个不同社区,例如一个人可能属于家庭、学校、工作单位和朋友圈子等多个社区;然而,每对结点间的链接却常常是由一个唯一原因产生的,例如亲人关系、工作关系和朋友关系等.因此,由链接构成的社区(即链接社区)就构成了一种新的社区模式.此外,链接社区天然反映了复杂网络中社区结构高度重叠的特点,它从另一个角度体现重叠社区模式.然而,由于链接社区模式刚刚被提出,还有大量亟待解决的问题,例如设计包含链接社区模式的基准网络生成器、建立链接社区质量评价标准、给出更有效的链接社区挖掘算法等等.

参 考 文 献

[1] Watts D J, Strogatz S H. Collective dynamics of small-world networks [J]. Nature, 1998, 393(6638): 440-442

[2] Barabási A-L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286(5439): 509-512

[3] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proc of National Academy of Science, 2002, 9(12): 7821-7826

[4] Fortunato S. Community detection in graphs [J]. Physics Reports, 2010, 486(3/4/5): 75-174

[5] Yang Bo, Liu Dayou, Liu Jiming, et al. Complex network clustering algorithms [J]. Journal of Software, 2009, 20(1): 54-66 (in Chinese)
(杨博, 刘大有, Liu Jiming, 等. 复杂网络聚类方法 [J]. 软件学报, 2009, 20(1): 54-66)

[6] Gavin A C, Bosche M, Krause R. Functional organization of the yeast proteome by systematic analysis of protein complexes [J]. Nature, 2002, 415(6868): 141-147

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

[8] Tyler J R, Wilkinson D M, Huberman B A. Email as spectroscopy: Automated discovery of community structure within organizations [C] //Proc of the 1st Int Conf on Communities and Technologies. Amterdam, Netherlands: Kluwer Academic Publishers, 2003: 81-96

[9] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks [J]. Proc of National Academy of Science, 2004, 101(9): 2658-2663

[10] Jin Di, Liu Jie, Jia Zhengxue, et al. Study on k-nearest-neighbor network for data clustering algorithm [J]. Pattern Recognition and Artificial Intelligence, 2010, 23(4): 546-551 (in Chinese)
(金弟, 刘杰, 贾正雪, 等. 基于 k 最近邻网络的数据聚类算法[J]. 模式识别与人工智能, 2010, 23(4): 546-551)

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

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

[13] Guimera R, Amaral L A N. Functional cartography of complex metabolic networks [J]. Nature, 2005, 433(7028): 895-900

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

[15] Newman M E J. Modularity and community structure in networks [J]. Proc of National Academy of Science, 2006, 103(23): 8577-8582

[16] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, P10008. [2012-05-01] http://iopscience.iop.org/1742-5468/2008/10/P10008

[17] Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis [J]. Physical Review E, 2009, 80(5): 056117

[18] Jin Di, Liu Dayou, Yang Bo, et al. Fast complex network clustering algorithm using local detection [J]. Acta Electronica Sinica, 2011, 39(11): 2540-2546 (in Chinese)
(金弟, 刘大有, 杨博, 等. 基于局部探测的快速复杂网络聚类算法[J]. 电子学报, 2011, 39(11): 2540-2546)

[19] Fortunato S, Barthélemy M. Resolution limit in community detection [J]. Proc of National Academy of Science, 2007, 104(1): 36-41

[20] Good B H, de Montjoye Y-A, Clauset A. The performance of modularity maximization in practical contexts [J]. Physical Review E, 2010, 81(4): 046106

[21] Khadivi A, Ajdari Rad A, Hasler M. Network community-detection enhancement by proper weighting [J]. Physical Review E, 2011, 83(4): 046104

[22] Raghavan U N, Albert R, Kumara S. Near linear-time algorithm to detect community structures in large-scale networks[J], Physical Review E, 2007, 76(3): 036106

[23] Leung I X Y, Hui P, Liò P, et al. Towards real time community detection in large networks [J]. Physical Review E, 2009,79(6): 066107

[24] Tibély G, Kertész J. On the equivalence of the label propagation method of community detection and a Potts model approach [J]. Physica A, 2008, 387(19/20): 4982-4984

[25] Barber M J, Clark J W. Detecting network communities by propagating labels under constraints [J]. Physical Review E, 2009, 80(2): 026129

[26] Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks [J]. Physica A, 2010, 389(7): 1493-1500

[27] van Dongen S. Graph clustering by flow simulation [D]. Utrecht, Netherlands: University of Utrecht, 2000

[28] Capocci A, Rao F, Caldarelli G. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia Wikipedia [J]. Europhysics Letters, 2008, 81(2): 28006

[29] Vlasblom J, Wodak S J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs [J/OL]. BMC Bioinformatics, 2009, 10: 99-112. [2012-05-01]. http://www.biomedcentral.com/1471-2105/10/99/

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

[31] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure [J]. Proc of National Academy of Science, 2008, 105(4): 1118-1123

[32] Grünwald P D, Myung I J, Pitt M A. Advances in Minimum Description Length: Theory and Applications [M]. Cambridge: MIT Press, 2005

[33] Morārescu C I, Girard A. Opinion dynamics with decaying confidence: Application to community detection in graphs [J]. IEEE Trans on Automatic Control, 2011, 56(8): 1862-1873

[34] Yang B, Liu J, Feng J. On the spectral characterization and scalable mining of network communities [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(2): 326-337

[35] Liu Y, Wang Q X, Wang Q, et al. Email community detection using artificial ant colony clustering [C]//Proc of Joint the 9th Asia-Pacific Web Conf and the 8th Int Conf on Web-Age Information Management Workshops. Berlin: Springer, 2007: 287-298

[36] Liu Y, Luo J Y, Yang H J, et al. Finding closely communicating community based on ant colony clustering model [C] //Proc of the 2010 Conf on Artificial Intelligence and Computational Intelligence (AICI'10). Piscataway, NJ: IEEE, 2010: 127-131

[37] Sadi S, Etaner-Uyar Ş, Öğüdücü Ş. Community detection using ant colony optimization techniques [C] //Proc of the 2009 Conf on Soft Computing (MENDEL'09). Piscataway, NJ: IEEE, 2009: 206-213

[38] Sadi S, Öğüdücü Ş, Uyar A Ş. An efficient community detection method using parallel clique-finding ants [C] //Proc of the 2010 IEEE Congress on Evolutionary Computation (CEC'10). Piscataway, NJ: IEEE, 2010: 1-7

[39] Jin Di, Yang Bo, Liu Jie, et al. Ant colony optimization based on random walk for community detection in complex networks [J]. Journal of Software, 2012, 23(3): 451-464 (in Chinese)

(**金弟, 杨博, 刘杰, 等. 复杂网络簇结构探测—基于随机游走的蚁群算法**[J]. **软件学报**, 2012, 23(3): 451-464)

[40] Jin D, Liu D, Yang B, et al. Ant colony optimization with a new random walk model for community detection in complex networks [J]. Advances in Complex Systems, 2011, 14(5): 795-815

[41] He D, Liu J, Liu D, et al. An ant-based algorithm with local optimization for community detection in large-scale networks [J]. Advances in Complex Systems, 2012, 15(8): 1250036-1-26

[42] Tasgin M, Herdagdelen A, Bingol H. Community detection in complex networks using genetic algorithms [EB/OL]. 2007 [2012-09-10]. http://arxiv.org/abs/0711.0491v1

[43] He Dongxiao, Zhou Xu, Wang Zuo, et al. Community mining in complex networks—Clustering combination based genetic algorithm [J]. Acta Automatica Sinica, 2010, 36(8): 1160-1170 (in Chinese)

(**何东晓, 周栩, 王佐, 等. 复杂网络社区挖掘——基于聚类融合的遗传算法**[J]. **自动化学报**, 2010, 36(8): 1160-1170)

[44] Jin Di, Liu Jie, Yang Bo, et al. Genetic algorithm with local search for community detection in large-scale complex networks [J]. Acta Automatica Sinica, 2011, 37(7): 873-882 (in Chinese)

(**金弟, 刘杰, 杨博, 等. 局部搜索与遗传算法结合的大规模复杂网络社区探测**[J]. **自动化学报**, 2011, 37(7): 873-882)

[45] Gong M, Fu B, Jiao L. Memetic algorithm for community detection in networks [J]. Physical Review E, 2011, 84(5): 056101

[46] Pizzuti C. Community detection in social networks with genetic algorithms//Proc of Genetic and Evolutionary Computation Conf (GECCO'08). New York: ACM, 2008: 1137-1138

[47] Pizzuti C. A multi-objective genetic algorithm for community detection in networks [C] //Proc of the 21th IEEE Int Conf on Tools with Artificial Intelligence (ICTAI'09), Piscataway, NJ: IEEE, 2009: 379-386

[48] Shi C, Yan Z, Wang Y, et al. A genetic algorithm for detecting communities in large-scale complex networks [J]. Advances in Complex Systems, 2010, 13(1): 3-17

[49] Jin D, He D, Liu D, et al. Genetic algorithm with local search for community mining in complex networks [C] //Proc of the 22th IEEE Int Conf on Tools with Artificial Intelligence (ICTAI'10). Piscataway, NJ: IEEE, 2010: 105-112

[50] Palla G, Derényi I, Farkas I, et al. The software of clique percolation method [CP/OL]. 2005 [2012-05-01]. http://www.cfinder.org/

[51] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks [J]. Physica A, 2009, 388(8): 1706-1712

[52] Evans T S. Clique graphs and overlapping communities [J/OL]. Journal of Statistical Mechanics: Theory and Experiment, 2010: P12037. [2012-05-01] http://iopscience.iop.org/1742-5468/2010/12/P12037

[53] Evans T S, Lambiotte R. Line graphs, link partitions and overlapping communities [J]. Physics Review E, 2009, 80(1): 016105

[54] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks [J]. Nature, 2010, 466(7307): 761-764

[55] Kim Y, Jeong H. The map equation for link community [J]. Physical Review E, 2011, 84(2): 026110

[56] Ball B, Karrer B, Newman M E J. An efficient and principled method for detecting communities in networks [J]. Physical Review E, 2011, 84(3): 036103

[57] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 033015

[58] Lee C, Reid F, McDaid A, et al. Detecting highly overlapping community structure by greedy clique expansion [C] //Proc of the 4th Int Workshop on Social Network Mining and Analysis (SNA-KDD'10). New York: ACM, 2010: 33-42

[59] Jin D, Yang B, Baquero C, et al. A Markov random walk under constraint for discovering overlapping communities in complex networks [J/OL]. Journal of Statistical Mechanics: Theory and Experiment, 2011: P05031. [2012-05-01]. http://iopscience.iop.org/1742-5468/2011/05/P05031/figures

[60] Newman M E J, Strogatz S H, Watts D J. Random graphs with arbitrary degree distributions and their applications [J]. Physics Review E, 2001, 64(2): 026118

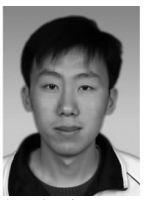
[61] Leskovec J, Lang K J, Mahoney M W. Empirical comparison of algorithms for network community detection [C] //Proc of the 19th Int World Wide Web Conf (WWW'10). New York: ACM, 2010: 631-640

[62] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms [J]. Physics Review E, 2008, 78(4): 046110

[63] Danon L, Duch J, Diaz-Guilera A, et al. Comparing community structure identification [J/OL]. Journal of Statistical Mechanics: Theory and Experiment, 2005: P09008. [2012-05-01]. http://iopscience.iop.org/1742-5468/2005/09/P09008/

[64] Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities [J]. Physics Review E, 2009, 80(1): 016118

Liu Dayou, born in 1942. Received his MSc degree in computer science from Jilin University, Changchun, China, in 1981. Currently professor with the College of Computer Science and Technology, Jilin University. Senior member of China Computer Federation. His current research interests include data mining, analysis of complex networks, spatial-temporal reasoning, distributed intelligence, knowledge engineering, and expert system (dyliu@jlu.edu.cn).

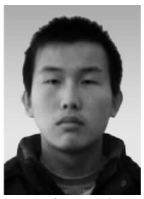


Jin Di, born in 1981. Received his BSc, MSc, and PhD degrees in computer science from Jilin University, Changchun, China, in 2005, 2008, and 2012, respectively. He is currently assistant professor with the School of Computer Science and Technology, Tianjin University. His current research interests include evolutionary computation, data mining, and complex network analysis (jindi@tju.edu.cn).

He Dongxiao, born in 1984. Received her BSc, and MSc degrees in computer science from Jilin University, Changchun, China, in 2007, and 2010, respectively. Currently a PhD candidate with the College of Computer Science and Technology, Jilin University. Her current research interests include evolutionary computation, data mining, and complex network analysis (hedongxiaojlu@gmail.com).

Huang Jing, born in 1975. Received her PhD degree in technology of computer application in 2008. Now associate professor in the College of Computer Science and Technology, Jilin University. Member of China Computer Federation. Her current

research interests include multi-agent system, distributed constraint optimization problem solving, complex network, information fusion, etc (huangjing@jlu.edu.cn).



Yang Jianning, born in 1987. MSc candidate with the College of Computer Science and Technology, Jilin University, Changchun, China. His current research interests include data mining, and complex network analysis (yangjianning0000@126.com).

Yang Bo, born in 1974. Received his BSc, MSc, and PhD degrees in computer science from Jilin University, Changchun, China, in 1997, 2000, and 2003, respectively. Currently professor with the College of Computer Science and Technology, Jilin University. His current research interests include data mining, complex/social network, self-organized and self-adaptive multiagent systems, knowledge engineering, applications in Web intelligence, and social computing.

研发动态

荷兰科学家发明一项可阅读人类思维新系统

据报道,来自荷兰奈梅亨大学的科学家发明一项新系统,可通过对大脑的扫描,来确定一个人所看到的事物.在实验中,志愿者面前的屏幕上出现字母 B,R,A,I,N 和 S,科学家会观察他们大脑的反应.通过观察,研究人员要确定实验对象在某一时刻看到了哪个字母.为实现这一结果,研究人员通过扫描志愿者在看到不同字母时的大脑信息,创建了大脑看到不同图像的反应变化基数.之后,获得的数据会通过一种特殊算法,将大脑的变化变成图像,从而得知实验对象所看到的字母.该研究的主要目的并不是阅读人类思维,科学家最感兴趣的是大脑如何处理视觉图像.同时,未来读取人类思维的技术或许会实现,甚至出现有心灵感应的计算机程序(摘自:http://tech.huanqiu.com/,2013-08-30,环球网科技).

哈佛大学研制出可拉伸全透明离子导体

美国哈佛大学一个研究小组日前宣布,他们在实验室中造出了可拉伸且全透明的离子导体,能在 10 000 Hz 以上的频率及 10 000 V 以上的电压下工作.这种离子导体在医学、柔性机器人和柔性光学器件等领域有着广阔的应用前景.伸缩性导体可应用在生物医学等多个领域,成为业界研究热点.但现有伸缩性导体大多是电子导体,尽管它们具有良好的导电性与拉伸性,但在高频高压或在高度变形的情况下工作,导电性能将受较大影响,此外透明度与生物相容性都不及离子导体.离子导体的多样性将为各种应用提供更多选择.比如,现代医学需要电子器件直接接触皮肤、心脏及大脑,可拉伸、透明且具有生物相容性的离子导体就可能比电子导体更适合(摘自:http://www.stdaily.com/,2013-09-02,《科技日报》).

科学家发现细菌 DNA 序列可作信息"存储器"

阿根廷科学家近日成功将该国国歌旋律以人工基因编码形式植入某种细菌染色体中.这一方法不仅可以用来存储音乐旋律,还可能发展为一种拥有巨大应用潜力的信息存储方式.生物的 DNA(脱氧核糖核酸)由 4 种脱氧核苷酸组成,即腺嘌呤、胸腺嘧啶、胞嘧啶和鸟嘌呤,分别用字母 A,T,C,G 表示.研究人员通过不同组合对 4 种核苷酸进行编码,使之对应不同的音符,然后将音符按照阿根廷国歌的旋律排列,并植入某种细菌的染色体中.被植入细菌染色体的基因片段以一条 DNA 链形式存在,这条 DNA 链虽然经过人为修改,但是在细菌繁殖过程中也会被精确复制.这是一种有巨大应用潜力的信息存储方式.首先细菌繁殖能力很强,一般 20 min 就可以复制一次,如同一个"生物复印机".更重要的是,这种方法可以带来强大的存储能力,一个拥有 6 万册藏书的图书馆,如果以人工基因编码形式存放在细菌染色体中,将只有 0.01 g,存储的信息还可以随时提取(摘自:http://www.xinhuanet.com/,2013-09-03,新华网).

美研制出新型"MIIM"二极管

据报道,美国俄勒冈州立大学(OSU)的研究人员在提高金属-绝缘体-金属(MIM)二极管的功能方面取得了显著进步,他们研制出了一种性能更加优异的金属-绝缘体-绝缘体-金属(MIIM)二极管.传统的硅电子设备虽然成本低廉,但其运行速度目前正接近极限.新的"MIIM"二极管是一块由 2 块金属中间夹着 2 块绝缘体组成的"三明治",这一结构使电子不会通过材料而是隧穿过绝缘体并且几乎同时出现在另一边.对于电子设备来说,这是一个完全迥异的制备方法.最新研究证明,添加第 2 块绝缘体使电子的"步隧穿"成为可能,在"步隧穿"这种情况下,一个电子仅仅隧穿过一个绝缘体而非两个绝缘体,这一点使二极管的非对称性、非线性得到精确的控制,而且,也能在更低的电压下整流(摘自:http://www.stdaily.com/,2013-09-13,《科技日报》).