一个基于中心度的社团结构发现新算法 *

戴爱明[1a,2]，高学东[1a]，王立敏[1b]
(1. 北京科技大学 a. 经济管理学院；b. 中国教育经济信息网管理中心，北京 100083；2. 南昌航空大学 制造业发展研究所，南昌 330063)

摘　要：针对 GN 算法在社团结构发现中时间复杂度高等问题，提出一种基于中心度的 GN 改进算法(DCGN)。该算法根据节点中心度以及节点之间的最短路径首先确定社团结构中心节点集，然后逐步删除社团结构中心节点之间的最大边介数连边，完成社团结构划分。DCGN 算法避免了 GN 算法边介数计算开销大的问题，算法的时间复杂度约为 $O(cmn)$，其中 c 为常数，n 为网络成员数，m 为网络连边数。将 DCGN 和 GN 算法同时应用到 Zachary 网络及计算机随机生成网络中并进行了比较。实验结果表明，所提出的 DCGN 算法在运行效率和效果方面较之 GN 算法均具有一定的优势。

关键词：社团结构；节点中心度；GN 算法；DCGN 算法
中图分类号：TP393　　**文献标志码**：A　　**文章编号**：1001-3695(2011)08-2909-03
doi:10.3969/j.issn.1001-3695.2011.08.030

New algorithm of detecting community structure based on degree centrality

DAI Ai-ming[1a,2], GAO Xue-dong[1a], WANG Li-min[1b]
(1. *a. School of Economics & Management, b. Management Center of China Education Economy Information Net, University of Science & Technology Beijing, Beijing* 100083, *China*; 2. *Institute of Manufacture Industry Development, Nanchang Hangkong University, Nanchang* 330063, *China*)

Abstract: Using GN algorithm to detect the community structure, there will be high time complexity. This paper proposed a new GN algorithm based on degree centrality(DCGN). According to node degree centrality and the shortest path among them, the algorithm first confirmed the community structure central nodes, then deleted edges with the biggest betweenness among the community structure central nodes by step, to finish the community structure dividing. This algorithm got rid of high cost of parameter calculating when using GN algorithm, the algorithm ran in time $O(cmn)$ when c was a constant, n was the number of network member, m was the number of network edge. Applied both this algorithm and GN algorithm to Zachary net and the net generated randomly by computer, and then compared them. Experiment results shows the proposed algorithm has advantage in feasibility and effectiveness.

Key words: community structure; degree centrality; GN algorithm; DCGN algorithm

0　引言

社团结构发现对理解复杂网络结构特征属性至关重要。目前社团结构发现方法主要分为两大类，即计算机科学中的图分割算法和社会学领域的层级聚类算法[1]。针对图分割算法中 NP 难问题，典型局部搜索策略算法是基于 Laplace 图特征值的谱平分法[2,3]和 Kernighan-Lin 算法(简称 KL 算法)[4]。谱平分法最坏情况下时间复杂度为 $O(n^3)$，KL 算法的时间复杂性是 $O(tn^2)$，其中 n 表示网络节点数，t 表示迭代次数。谱平分法和 KL 算法均需要事先知道该网络的两社团特征，否则可能得不到正确结果[5]。层级聚类典型算法是 Girvan-Newman 算法(简称 GN 算法)[6]。GN 算法引入边介数(edge-betweenness)概念，其基本思想是网络中边介数最大的边有可能是社团之间的连边，边介数可以通过最短路径来进行计算。GN 算法由于边介数计算开销过大而导致时间复杂性为 $O(m^2n)$，其中 n 为节点数，m 为节点连边数。Tyler 等人[7]提出一种近似 GN 算法，该算法用蒙特卡罗方法(Monte Carlo method)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，Radicchi 等人[8]提出了连接聚类系数(link clustering coefficient)取代 GN 算法的边介数，这一算法的时间复杂性为 $O(m^4/n^2)$。对于稀疏网络，算法时间复杂度为 $O(n^2)$，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。

本文针对 GN 算法节点边介数计算开销大的问题提出一种基于中心度的启发式 GN 改进算法(简称 DCGN 算法)。DCGN 算法基本思想是根据节点中心度(degree centrality)以及节点之间的最短路径确定不同社团结构中心节点，然后以社团中心节点之间最短路径需经过的节点作为源节点计算边介数，逐步删除最大边介数的连边，直到全部成员节点完成社团划分。DCGN 算法避免了计算网络中所有节点的边介数，从而降

收稿日期：2011-01-06；**修回日期**：2011-02-28　　**基金项目**：国家自然科学基金资助项目(70962008)；国家航空科学基金资助项目(2009ZG56022)

作者简介：戴爱明(1975-)，男，江西南丰人，副教授，博士研究生，主要研究方向为复杂网络(daiming 1795@yahoo.com.cn)；高学东(1963-)，男，河北唐山人，副院长，教授，博导，主要研究方向为管理过程优化.

低了 GN 算法的时间复杂度。

1　相关定义及算法提出

复杂网络节点度是指网络中任意节点与此节点连接边的数量。节点中心度是描述网络局部特性的基本参数,可以反映网络的宏观统计特征,而节点最短路径具有重要几何性质,对衡量节点之间的跨度比较有实际意义。为了简便起见,本文研究的是无方向、无权重的网络图,结合节点中心度和最短路径特点,对 GN 算法进行了改进。

1.1　相关定义

给定图 $G(V,E)$,$V=\{v_1,v_2,\cdots,v_n\}$ 表示图中节点集合,$E\subseteq\{(v_i,v_j):v_i,v_j\in V\}$ 表示节点连接集合,令无序偶对 (v_i,v_j) 表示节点 $v_i\in V$ 与 $v_j\in V$ 之间的边,(v_i,v_j) 可以用邻接矩阵描述。

定义 1　邻接矩阵(adjacency matrix)。邻接矩阵描述图中各节点两两之间的关系。邻接矩阵 $A=(a_{ij})$ 的元素 a_{ij} 定义为 $a_{ij}=\begin{cases}1\\0\end{cases}$。其中,如果 $(v_i,v_j)\in E$,元素 $a_{ij}=1$;否则,$a_{ij}=0$。

定义 2　节点中心度(degree centrality)。节点中心度基本思想是重要节点是那些拥有与其他节点有较多连接边数的节点。一个网络图中节点的重要性可依据它们度分布的大小进行排序。相应地,节点 v_i 的中心度定义[9]为

$$C_{v_i}^D=\frac{D_{v_i}}{n-1}=\frac{\sum_{v_i\in V}a_{ij}}{n-1}\tag{1}$$

其中:D_{v_i} 表示节点 v_i 的度,n 表示网络节点数。

定义 3　边介数(edge-betweenness)。对图 $G=(V,E)$,对 G 中的连边 $e\in E$,$\sigma_{v_iv_j}$ 表示节点 v_i 到 v_j 的最短路径数量,用 $\sigma_{v_iv_j}(e)$ 表示 v_i 到 v_j 的最短路径中包含边 e 的数量(边介数),则边 e 的边介中心度值可以定义[10]为

$$C_{EB}(e)=\sum_{v_i\in V,v_j\in V}\frac{\sigma_{v_iv_j}(e)}{\sigma_{v_iv_j}}\tag{2}$$

边介数基本思想是在包含不同社团的网络中所有最短路径经过次数最多的边,也就是介数最大的边必然是连接两个社团之间的边。

1.2　算法提出

节点中心度描述了节点在网络中处于中心的程度。如果将网络看做一个社团,那么节点中心度最大的节点应该是社团中心节点[11]。在确定了第一个中心节点之后,假设网络进一步分裂,准确找到第二个社团中心节点是关键问题。寻找社团第二中心节点过程中除了需要考虑节点本身的节点中心度之外,还需考虑该节点与其他中心节点之间的距离。因为即使某个节点中心度很高,但因为距离其他中心节点太近而不可能成为下一个中心节点。鉴于此,可以构造一个新的社团结构探测方法,那就是在通过节点中心度确定第一个社团中心节点之后,其他节点能否成为下一个社团中心节点,需要结合该节点本身的节点中心度和该节点到已知社团中心节点之间的最短路径。社团中心节点最优评价方法可由式(3)计算:

$$M_{\max}=\sum((C_{v_i}{}^D)^{\alpha}\times(d_G(v_i,v_j))^{\beta})\tag{3}$$

其中:$\alpha+\beta=1$,α、β 分别表示距离中心度和节点间最短路径的权重系数。本文经过多次实验,得到较为合理的权重分配是 $\alpha=(1/3)$,$\beta=(2/3)$。$d_G(v_i,v_j)$ 表示节点 v_i 到 v_j 的最短路径长度。考虑到 $C_{v_i}^D$ 和 $d_G(v_i,v_j)$ 计量单位的不同,在实际计算过程中需作标准化处理。

给定无向连通图 $G=(V,E)$,用 n、m 和 h 分别表示 G 中的节点、连边和社团划分数量。该算法本质是网络由第一个社团中心节点开始不断向外扩展,直到找到满足社团结构数的中心节点,然后利用边介中心度删除社团结构之间的连边,从而实现社团成员的划分。

基于距离中心度的社团结构探测算法通过如下步骤实现:

输入:社团数目 h,权重系数 α、β。

输出:网络 G 社团结构划分。

a)找出网络节点中心度最大节点作为初始节点 v_0 加入社团中心节点集 C。

b)取网络中其余节点按式(3)计算 M 值。

c)找出 $M_{\max}$ 对应节点 v_i 合并到初始社团中心集合 C 中。

d)重复步骤 b)c),直到找到 h 个社团中心节点。

e)找出集合 C 的任意节点对 (v_i,v_j),计算 $d_G(v_i,v_j)$。

f)找到 $d_G(v_i,v_j)_{\min}$,记录 $d_G(v_i,v_j)_{\min}$ 路径上的边(或称节点对),放入集合 E,然后删除 $d_G(v_i,v_j)_{\min}$ 路径上的边和节点(不删除社团中心节点)。

g)重复步骤 e)f),直到网络划分为 h 个社团。

h)计算节点集 E 中节点边介度。

i)找出边介度最大的边,在原始网络中把这条边删除。

j)重复步骤 h)i),直到得到 h 个社团。

1.3　算法时间复杂度分析

计算网络每个节点度时间复杂度为 $O(m)$,生成 h 个社团中心节点集的时间复杂度为 $O(m)$。围绕单个节点计算边介数时间复杂度为 $O(m)$,计算社团中心节点集边介数并截断所有最短路径时间复杂度变为 $O(hm)$。假设社团之间有 z 条连边,将社团之间的边全部删除,那么该步骤总时间复杂度为 $O(zhm)$。网络最大可能分裂为 n 个社团(n 为网络节点数),算法复杂度为 $O(zmn)$。综合以上步骤,算法总体时间复杂度为 $O(m+m+zmn)$。z 和 h 相对 m、n 来说是个很小的值,这样改进的 DCGN 算法的时间复杂度可以约等于 $O(cmn)$。显然,本算法的时间复杂度比传统 GN 算法时间复杂度 $O(m^2n)$ 有较大改善。

2　算法测试

目前,Zachary 俱乐部等经典网络和计算机生成网络已经成为人们公认的测试网络。为了验证本算法的性能和计算的准确性,本文将 DCGN 和 GN 算法同时应用在 Zachary 网络和计算机生成网络中并进行比较,用 Python 2.6 编写算法程序,在 IBM 笔记本电脑(CPU 1.80 GHz,内存 1.0 GB,硬盘 160 GB,Windows XP 操作系统)上进行了实现。比较的结果显示,DCGN 算法寻找的社团结构在正确率和运算速度方面均优于 GN 算法。

2.1　对 Zachary 网络的测试

在复杂网络的社团结构分析中,Zachary 网络是一个用于比较探测网络社团结构算法准确性的经典网络。Zachary 网络包含有 34 个节点,基于成员在俱乐部内部以及外部的社会关系,他们形成两个小俱乐部。应用传统 GN 和 DCGN 算法对 Zachary 网络进行 2 社团结构探测,圆形和方块节点分别表示

俱乐部分裂后的成员,结果如图 1 和 2 所示。在 Zachary 网络中,节点中心度从大到小前三名依次是 34、1 和 3 节点,在确定第一中心节点 34 后通过对式(3)进行计算,找到网络第二中心是节点 1,然后通过边界中心度完成节点社团的划分。GN 算法将节点 3 错误划分,而 DCGN 算法节点划分结果完全正确,因为 DCGN 算法考虑了节点间最短路径参数,从节点 3 到 34 和 1 的最短路径来看,节点 3 被划入以 1 为中心的社团是合理的。

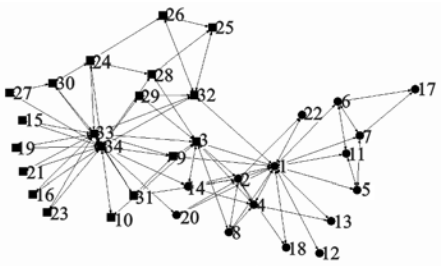

图1　GN算法对Zachary网络社团划分效果

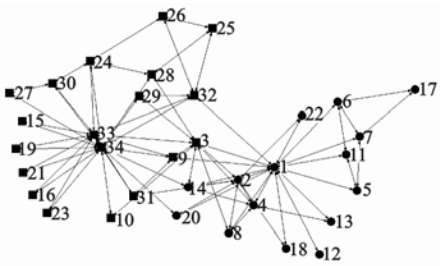

图2　DCGN算法对Zachary网络社团划分效果

2.2　对计算机生成网络测试

将 DCGN 算法应用到一组已知社团结构的计算机生成的随机网络图中[8]。图 3 显示了一个由 128 个节点组成的网络结构分布,整个网络分成四个社团,每个社团有 32 个节点,四个社团分别由圆形、方形、三角形和方格形表示。随机网络选取 $D_{v_i}=16$,考虑到网络社团结构的尖锐性,设定 $D_{v_i}^{\text{out}}=1$,网络中每个节点的节点度 D_{v_i} 均被分成两部分,分别为与社团内部相连的边数 $D_{v_i}^{\text{in}}$ 和与社团外部相连的边数 $D_{v_i}^{\text{out}}$,即 $D_{v_i}=D_{v_i}^{\text{in}}+D_{v_i}^{\text{out}}$。节点之间连接的边都是独立的和随机的,因此 $D_{v_i}^{\text{in}}$ 和 $D_{v_i}^{\text{out}}$ 的值是网络图中所有节点的平均值。该方法可以随机生成一组已知社团结构的无向网络图。

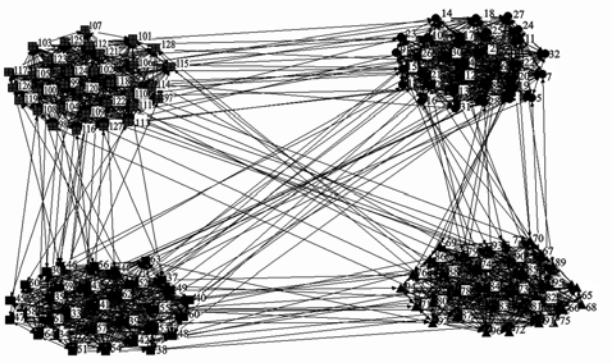

图3　100成员网络四社团结构

用上述方法生成不同成员数的复杂网络,分别应用 DCGN 和 GN 算法对它们进行社团结构探测。探测正确率和运行时间对比如图 4 所示。

从图 4 中可以看出,DCGN 算法在搜索效果和计算速度上均优于 GN 算法,特别是随着社团成员规模的扩大,DCGN 算法在计算速度上的优势越明显。

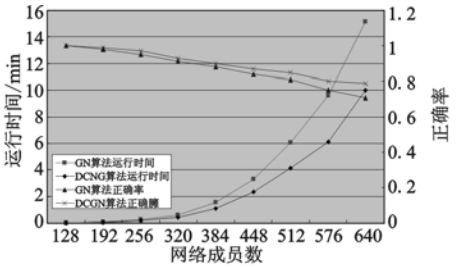

图4　DCGN和GN算法对不同规模网络的运行时间和正确率对比

3　结束语

本文基于节点中心度的思想,结合节点间最短路径,通过合理分配其权重,提出了一种新的探测网络社团结构算法(DCGN),算法的时间复杂度约为 $O(cmn)$,其中 c 为常数,n 为网络成员数,m 为网络连边数。把该算法应用到 Zachary 网络及计算机随机生成不同规模网络,并与 GN 算法进行比较,实验表明本文提出的算法在运行效率和效果方面较之 GN 算法均具有一定的优势。但从算法思路来看,本算法较为适合社团结构特征较为明显的网络,若社团结构不明显,对社团中心节点集的搜索可能会存在一定偏差,如何解决此问题将是本文进一步要做的工作。

参考文献:

[1] 高学东,王立敏,马红权,等. 基于共享最近邻探测社团结构的算法[J]. 系统工程理论与实践,2009,29(10):102-109.

[2] FIEDLER M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal,1973,23(98):298-305.

[3] POTHEN A, SIMON H D, LIOU K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis Applications,1990,11(3):430-452.

[4] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal,1970,49(2):291-307.

[5] 杨博,刘大有,LIU Ji-ming,等. 复杂网络聚类方法[J]. 软件学报,2009,20(1):54-66.

[6] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proc of the National Academy of Science,2002,99(12):7821-7826.

[7] TYLER J R,WILKINSON D M, HUBERMAN B A. Email as spectroscopy: automated discovery of community structure within organizations[C]//Proc of the 1st International Conference on Communities and Technologies. Dordrecht:Kluwer Academic Publishers,2003:81-96.

[8] RADICCHI F,CASTELLANO C, CECCONI F, *et al*. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Science,2004,101(9):2658-2663.

[9] WASSERMAN S, FAUST K. Social networks analysis:methods and applications[M]. Cambridge:Cambridge University Press,1994.

[10] SHIMBEL A. Structural parameters of communication networks[J]. Bulletin of Mathematical Biophysics,1953,15(4):501-507.

[11] 刘盂. 复杂网络中挖掘社团结构的新聚类算法研究[D]. 长春:吉林大学,2010.

[12] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E,2004,69(2):026113.