《电子技术应用》 www.ChinaAET.com

基于密度分布的社区发现算法研究*

常富蓉
(喀什师范学院 信息工程技术系,新疆 喀什 844006)

摘 要:基于密度吸引点和其对相邻节点的影响度,提出了一种密度分布社区发现算法。该算法以节点度数最大的密度吸引点为初始社区,访问社区的相邻节点,把对社区影响度最大的节点加入到社区中,如果有些节点对多个社区都有影响,则把它归属为影响度最大的那个社区中,同时如果两个社区之间的相互影响度很大,可以将这两个社区合并为一个社区。将该算法应用到 Zachary 空手道俱乐部网络和随机无标度网络中,实验表明该算法能够很好地分出网络中的社区,同时实验还发现社区的收敛速度与幂率分布特性近似成反比。

关键词:复杂网络;幂率分布;社区发现;密度分布;影响度
中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2014)06-0081-03

Community finding algorithm based on density distribution

Chang Furong
(Information Engineering Department, Kashgar Teachers College, Kashgar 844006, China)

Abstract: Based on the density attractor and its effect on adjacent nodes, this paper proposes a community discovery algorithm of density distribution. The algorithm takes the highest density attractor of the node degree as the initial community, and then visits the adjacent nodes of communities, adding the most influential node to the initial community. If some nodes have influence on several communities, then it belongs to the most influential community. At the same time, two communities can be merged into one community, if the influence is big enough between these two communities. Applying the algorithm to Zachary karate club network and random scale-free network, the experiments show that this algorithm can divide communities of network; meanwhile, it is also found in the experiences that the rate of convergence speed of the community is inversely proportional to the power distribution characteristics.

Key words: complex networks; power law distribution; community; density distribution; influence

所谓社区,是指具有共同兴趣、爱好的人或者学有专精的专业人士,通过一定的方式聚集在一起,彼此之间可以沟通、交流、分享相关信息。在现实世界中,存在着很多这样的社区,例如社会关系网络[1]、神经网络、食物链网络、城市交通网络等。在这些社区中,有着复杂的内部结构,用节点表示实体,用连线表示实体间的联系,社区内部节点之间的联系非常紧密,社区之间的联系相对稀疏。近几年,随着网络的急速发展,网络社区也成为一个研究热点,同时取得了很重要的进展,并且发现了网络社区的很多特点,其中包括小世界特性(即网络中节点之间的平均距离很短,对数依赖于网络中的节点数)、无标度特性(即网络中节点的度分布右偏斜,具备幂函数或指数函数的形式)、聚集性或网络传递性以及社区结构特性,大量实证研究表明,许多网络是异构的,即复杂网络不是大批性质相同节点的随机连接,而是许多类型节点的组合,其中相同类型的节点存在较多的连接,而不同类型节点的连接则相对较少。把同一类型节点以及这些节点之间的边所构成的子图称为网络中的社区。社区如图 1 所示,图中的小型网络中包含 3 个社区,对应图中的 3 个椭圆,在这些社区内部,节点与节点之间的联系非常紧密,而社区之间的联系则比较稀疏[2-4]。

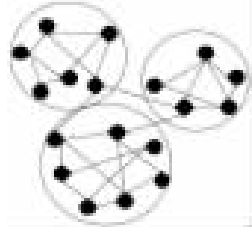


图 1 三社区网络

*基金项目:喀什师范学院校内课题——青年专项((13)2488)

《电子技术应用》 www.ChinaAET.com

网络社区发掘对于了解网络结构和分析网络特征具有重要的意义，目前已经提出了具有基于节点集划分的社区发现算法，例如基于贪婪算法原理的 Kernighan-Lin 算法、基于谱思想的谱平分算法、基于分裂思想的 GN 算法和基于凝聚思想的 Newman 快速算法等[5-7]。网络社区发现的最终目的就是将网络划分为若干个互相独立的社区，这些算法对研究社区发现都产生了重大的影响。

1 算法设计

由于网络社区中节点之间的联系相对紧密，社区之间节点的联系相对稀疏，根据这一特性，本文提出了基于密度分布的社区发现算法。在实际网络中，存在一些节点与其他节点的联系非常紧密，即该节点的度最大，称为“密度吸引点”，其他节点以“密度吸引点”为中心，从而形成社区。该算法的基本思想是以密度吸引点为初始社区，然后找出对该吸引点影响最大的节点依次加入到社区，一个网络中存在可能不止一个吸引点，因此在网络中可能存在多个社区，在计算节点影响度时要考虑对多个社区的影响。直到所有的节点都被分到了各自的社区中。同时要考虑不同社区之间的影响度，如果两个社区之间的相互影响度非常大，就认为这两个社区为一个社区，则合并这两个社区。

1.1 相关知识

为了清晰说明提出的算法，首先定义几个相关概念[8]：

(1)影响函数：每个节点的影响可以用一个数学函数来形式化模拟，它描述了一个节点在领域内的影响。假设节点 x，y 是 d 维网络特征空间 f^d 中的对象，对象 y 对 x 的影响函数是 $f_B^y: F^d \to R_0^+$，则可以根据一个基本函数定义影响函数：

$$f_B^y(x)=f_B(x,y) \tag{1}$$

在这里影响函数可以由某个网络邻域内两个节点的距离决定。

(2)密度函数：在网络中，一个节点 x 的密度函数被定义为在社区中所有节点影响函数的平均值，给定 n 个节点的网络 $D=\{x_1,\cdots,x_n\}\subset F^d$，在 x 上密度函数定义为：

$$f_B^D(x)=\frac{1}{n}\sum_{i=1}^{n} f_B^{x_i}(x) \tag{2}$$

(3)密度吸引点：指在整个网络中密度最大的节点。一个节点 x 是被一个密度吸引点 x^* 吸引的，如果存在一组点 $(x_0,x_1,\cdots,x_k)$，$x_0=x$，$x_k=x^*$，对于 $0<i<k$，x_{i-1} 的梯度在 x_i 的方向上。在这里密度吸引点为度数最大的节点。

1.2 算法描述

本算法中，假设社区内部度数越大则社区密度越大；密度越大的社区对其周围节点的吸引力越大；对在同一个层次即相对密度吸引点来说密度相同的节点的吸引力相同。

(1)初始化网络，将网络中的每个节点看成是独立的社区，计算社区的密度，每个社区的初始密度为节点的度数；

(2)找出网络中密度最大的社区，该社区为密度吸引点；

(3)根据影响函数，计算与密度吸引点相邻社区对密度吸引点的影响度，找出影响度最大的社区，此处认为这个社区与密度吸引点为同一个社区（影响度最大的社区可能不止一个）；

(4)再次计算网络中所有社区的密度，此时应用密度函数进行计算，并找出密度最大的社区作为密度吸引点；

(5)重复步骤(3)、步骤(4)，直到所有的社区都合并完成，算法结束。

算法流程图如图 2 所示。

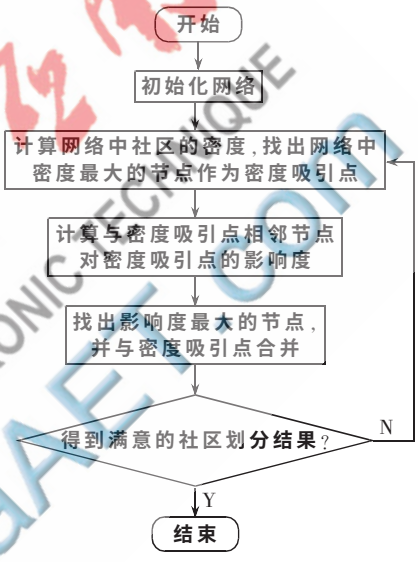


图 2 算法流程图

2 算法验证

为了验证本算法的有效性，将本算法应用到 Zachary's Karate Club Network[9]和随机的无标度网络。实验结果表明，本算法可以将网络划分为大小不同的社区，且具有较好的效果。

2.1 Zachary's Karate Club Network

在复杂网络社区结构的研究中，Zachary's Karate Club Network 是经常被使用的经典数据集，它是 20 世纪 70 年代初期 Zachary 用了两年的时间观察得到的美国一所大学中空手道俱乐部成员的相互社会关系网络。在这个网络数据集中包含了 34 个节点和 78 条边，其中节点表示俱乐部成员，边表示成员之间的社会关系，网络结构如图 3 所示。

通过应用密度分布社区发现算法，可以将该经典网络准确地分为两个社区，社区分布节点为：第一个社区(1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22)，第二个社区(9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34)。同时还发现由于网络节点的分布遵循幂率分布定律，因此社区合并的速度正好与幂率分布呈反比，即社区密度越小，社区收敛的越快。比较结果如图 4

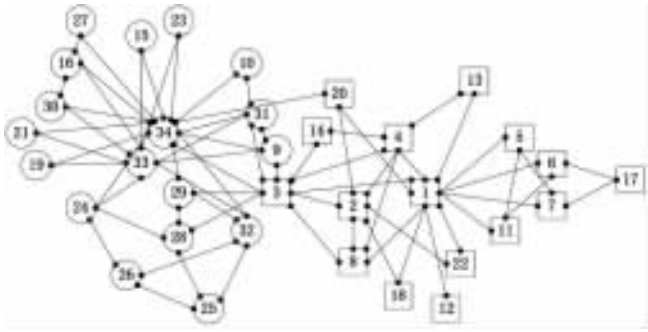
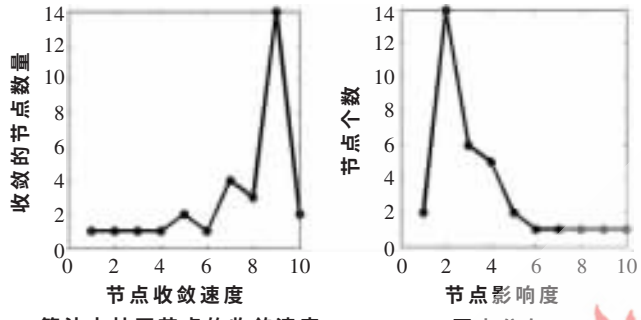


图 3 Zachary's Karate Club Network



(a)算法中社区节点的收敛速度 (b)幂率分布

图 4 社区收敛速度与幂率分布对比图

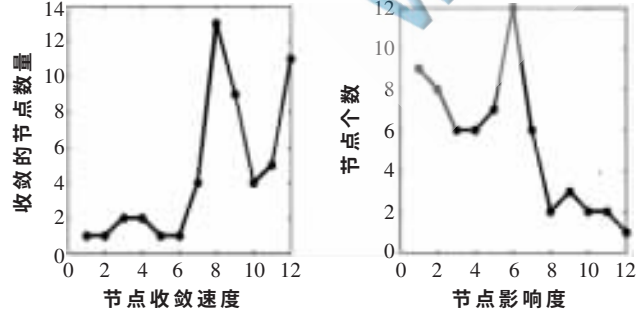
所示。

2.2 随机无标度网络

在现实网络中，大部分网络都符合复杂网络特性，因此在第二个实验中，随机截取了网络中节点相互关联的一部分随机的无标度网络作为实验对象，该网络中共有 62 个节点，400 条边，节点代表其相应的网页，边代表网页之间相互的连接关系，其网络结构如图 5 所示。

图 5 随机无标度网络图

在实验过程中，对该网络节点用 1~62 进行了随机的标注，通过应用密度分布社区发现算法，成功地将社区分成了两个部分，社区收敛速度与幂率分布对比如图 6 所示。



(a)算法中社区节点收敛速度 (b)幂率分布

图 6 随机网络社区收敛速度与幂率分布对比图

实验结果表明本算法可以准确地将随机无标度网络分为两个社区，实验结果如图 7 所示。

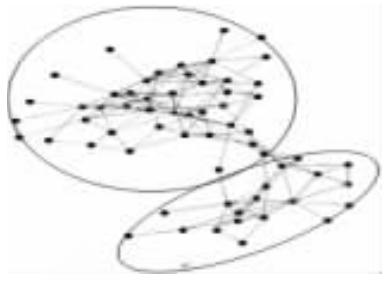


图 7 社区划分结果图

本文提出的基于密度分布的社区发现算法，根据聚类算法中的密度分布函数，利用密度吸引点，通过计算影响函数，对网络进行聚类，完成网络社区的划分。利用两个数据集对本算法进行了有效性验证，结果表明，该算法能准确地找出网络中存在的社区，并且发现社区收敛时与幂率分布的规律。本算法仅对部分社区进行了实验，关于时间复杂度等并没有进行精确的计算，以后还需要进一步对该算法进行验证和改进。

参考文献

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

[2] 马兴福，王红. 一种新的重叠社区发现算法[J]. 计算机应用研究，2012，29(3)：844-846.

[3] 林友芳，王天宇，唐锐，等. 一种有效的社会网络社区发现模型和算法[J]. 计算机研究与发展，2012，49(2)：337-345.

[4] 李峻金，向阳，牛鹏，等. 一种新的复杂网络聚类算法[J]. 计算机应用研究，2010，27(6)：2097-2099.

[5] CAPOCCI A, SERVEDIO V D P, CALDARELLI G, et al. Detecting communities in large networks[J]. Physica A, 2005, 352(2-4): 669-676.

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

[7] DUCH J, ARENAS A. Community detection in complex networks using extreme optimization[J]. Phys. Rev. E, 2005, 72(2): 027104.

[8] 韩家炜，坎伯. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2001.

[9] 汪小帆，李翔，陈关荣. 复杂网络及其应用[M]. 北京：清华大学出版社，2006.

(收稿日期：2013-11-18)

作者简介：

常富蓉，女，1985 年生，硕士研究生，助教，主要研究方向：复杂网络。