基于局部信息的复杂网络社团结构发现算法*

赵晓慧[1],刘 微[1],谢凤宏[1],赵凤霞[2]
(1.辽宁师范大学 计算机与信息技术学院,辽宁 大连 116081;
2.秦皇岛职业技术学院 信息工程系,河北 秦皇岛 066100)

摘 要:发现网络中的社团结构有助于更好地理解网络结构和分析网络属性。通过定义边的聚类系数和基于局部信息的方法,提出了一种寻找复杂网络中社团结构的算法。该算法首先在网络的剩余节点中寻找度最大的节点,然后利用该节点的局部信息、边的聚类系数和凝聚的思想,得到复杂网络的社团结构。在两个典型网络上的测试结果表明了该方法的可行性。

关键词:复杂网络;社团结构;边的聚类系数;节点的度
中图分类号:TP181 文献标识码:A 文章编号:1674-7720(2011)15-0043-04

An algorithm for detecting structure in complex network based on local information

Zhao Xiaohui[1], Liu Wei[1], Xie Fenghong[1], Zhao Fengxia[2]
(1.Department of Computer Science and Information Technology, Liaoning Normal University, Dalian 116081, China;
2.Department of Information Engineering, Qinhuangdao Vocational and Technical College, Qinghuangdao 066100, China)

Abstract: Detecting community structure in complex network contributes to understand the network structures and analyze the network properties better. Based on the local information, this article proposes an algorithm for discovering the communities in complex network by introducing the definition of the edge cluster coefficient. To obtain the community structure in complex network, the node with maximum degree in remainder network is first found. Then, some edge cluster coefficients are computed in terms of local information and the agglomeration idea. The tested results on two typical networks show the validity of this algorithm.

Key words: complex network; community structure; edge cluster coefficient; node degree

近年来,复杂网络已经成为众多领域的关注对象。例如,万维网、人类社会网、生物技术网络和科学家合作关系网[1-4]等。复杂网络已成为当前最重要的多学科交叉研究领域之一[5]。社团结构是复杂网络的一个重要特性,它把网络中的点分到不同的"组"或"团"之中。其中社团内部节点连接比较稠密,但是社团之间节点连接则比较稀疏[6]。发现网络中的社团结构,对于了解网络结构和网络性质具有非常重要的意义[7]。

复杂网络中社团结构的发现方法,根据向网络中添加边还是删除边,可以分为凝聚方法和分裂方法。具有代表性的是 GIRVAN 和 NEWMAN 提出的基于边介数的 GN 分裂算法[8]和 BREIGER 提出的 Concor 算法[9]。GN 算法是通过不断地从网络中移除介数最大的边对网络进行划分。而 Concor 算法则是利用对相关系数的重复迭代产生一个相关系数矩阵,进而对网络进行聚类。图形分割中比较著名的方法是基于贪婪方法的 Kernighan-Lin 算法[10]和基于 Laplace 图特征值谱平分法[11]。Kernighan-Lin 算法是一种基于贪婪思想的社团发现方法,该方法可以把一个复杂网络划分为两个大小已知的社团。谱平分法是利用网络 Laplace 矩阵的特征值近似相等的原理进行社团结构划分。Zhang 等人在 2010 年提出了一种复杂网络中社团结构的模糊分析方法[12]。该方法利用节点与社团核连接的紧密程度,判断将节点并入某个社团核中,从而实现了发现社团结构的目的。

一般来说,使用网络中的整体信息来划分网络得到的精度较高,但时间复杂度往往也很高;而利用局部信息划分网络,虽然可以得到较低的时间复杂度[13],但划

*基金项目:国家自然科学基金(10771092)

分精度往往不够理想。因此,如何利用局部信息而又能得到比较好的划分结果,是一个十分值得研究的问题。

本文通过定义边的聚类系数,提出了基于节点局部信息的社团发现算法。该方法通过不断地计算局部边的聚类系数,并利用凝聚算法的思想得到了网络的社团结构。通过对三个社团网络和 Zachary 网络的划分,表明了该算法的可行性。

1 预备知识

给定一个具有 n 个节点的无向无权网络 $G=<V,E>$,其中,$V=\{v_i|i=1\cdots n\}$,$E=\{(v_i,v_j)|v_i,v_j\in V\}$。$\boldsymbol{A}$ 是 G 的邻接矩阵,其中,若 i、j 两节点有边相连,则 $a_{ij}=1$;否则,$a_{ij}=0$。$N_i=\{j|a_{ij}=1\}$表示节点 i 的邻居集合,用 d_i 表示节点 i 的度。

节点 i 的聚类系数 X_i 为[6]:

$$X_i=2|S_i|/(d_i(d_i-1)) \quad (1)$$

其中,S_i 表示 d_i 个节点之间实际存在的边的集合,$(d_i(d_i-1)/2)$表示在 d_i 个节点之间最多可能存在的边数。从几何上看,节点 i 的聚类系数 X_i 表示包含节点 i 在内的实际的三角形数量和包含节点 i 在内的可能存在的三角形数量之比。如果 $a_{ij}=1$,那么可以定义边 e_{ij} 的聚类系数 C_{ij} 为:

$$C_{ij}=|N_{ij}|/(d_i+d_j-|N_{ij}|-2) \quad (2)$$

其中,$N_{ij}=N_i\cap N_j$,表示节点 i 和 j 的公共邻居集合[14],$|N_{ij}|$ 表示以 e_{ij} 为边组成的实际三角形的数量,$(d_i+d_j-|N_{ij}|-2)$表示包含节点 i 和 j 在内的可能存在的三角形数量。

边的聚类系数表示边所连接的两个节点的连接强度,值越大表明这两个节点在同一个社团的可能性越大。显然,$0\leqslant C_{ij}\leqslant 1$。

以图 1 所示的简单网络为例,计算边 e_{36} 和 e_{67} 的聚类系数。节点 3 的度 $d_3=4$,其邻居集合 $N_3=\{1,2,4,6\}$;节点 6 的度 $d_6=6$,$N_6=\{1,3,5,7,9\}$;节点 7 的度 $d_7=5$,$N_7=\{5,6,8,9,10\}$。因此,节点 3 和节点 6 的公共邻居集合 $N_{36}=\{1\}$,节点 6 和节点 7 的公共邻居集合 $N_{67}=\{5,8,9\}$。计算边 e_{36} 和 e_{67} 的聚类系数分别为 $C_{36}=|N_{36}|/(d_3+d_6-|N_{36}|-2)=1/(5+6-1-2)=1/9$,$C_{67}=|N_{67}|/(d_7+d_6-|N_{67}|-2)=3/(5+6-3-2)=1/2$。

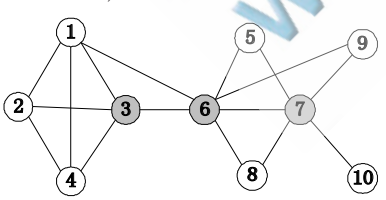


图 1 简单网络

由结果可以看出,边 C_{67} 较大,而 C_{36} 较小。说明节点 6 和节点 7 具有较强的凝聚性,即这两个节点在同一个社团内的可能性较大。

2 算法描述

给定一个具有 n 个节点的无向无权网络。首先选取度最大的节点作为社团的初始节点,通过邻接矩阵构造该社团的邻居集合。然后判断该集合中节点 v_i 与社团连接的紧密程度。如果节点 v_i 满足以下两个条件之一,说明 v_i 与该社团连接紧密,将该点加入到社团中去,更新社团及其邻居集合。

(1)v_i 有不小于一半的邻居节点在社团中;

(2)在与社团相连的所有边中,节点 v_i 的一条边 e_{ij} 的聚类系数在这些边中是最大的,并且 v_i 的其他边的聚类系数小于该边的聚类系数。

重复这个过程,直到社团的邻居节点中没有节点能够加入社团,标记所得到的社团。然后从其余节点中重复上述过程,直到整个网络划分完毕。

具体算法如下:

输入:一个无向网络,$G=<V,E>$,其中,$V=\{v_i|i=1\cdots n\}$,$E=\{(v_i,v_j)|v_i,v_j\in V\}$。

输出:网络的社团结构。

(1)初始社团 C_i 为空。

(2)选取剩余网络中度最大的节点 v_m 作为社团 C_i 的初始节点。令 $C_i=C_i+v_m$,$V=V-v_m$,并建立社团 C_i 的邻居集合。

(3)计算社团 C_i 的邻居集合,计算与社团 C_i 相连的所有边的聚类系数,找到聚类系数最大的边所对应的节点 v_n,计算节点 v_n 其他边的聚类系数。如果节点 v_n 满足条件(1)或(2),则将该节点并入社团 C_i,令 $C_i=C_i+v_n$、$V=V-v_n$,更新社团 C_i 的邻居集合。

(4)重复步骤(3),直到 C_i 的邻居集合中不再有新的节点加入到社团中为止,输出社团 C_i。令 $V=V-C_i$,若 V 不空,令 $i=i+1$,返回步骤(1)。

(5)输出结果。

3 实验与分析

3.1 三个社团网络

下面以计算机生成的三个社团网络为例,如图 2 所示,该网络包含 19 个节点和 37 条边。利用本文提出的算法详细分析该网络,结果如表 1 所示。表中①表示该节点具有当前网络中最大的度数值;②表示该节点有不小于一半的邻居节点与上一节点所在的社团相连;③表示该节点是上一节点所在社团的邻居并集中具有最大聚类系数的节点,并且该点的其他聚类系数小于该点与社团连接的边的聚类系数。③中内容指的是聚类系数值(具有最大聚类系数的边)。

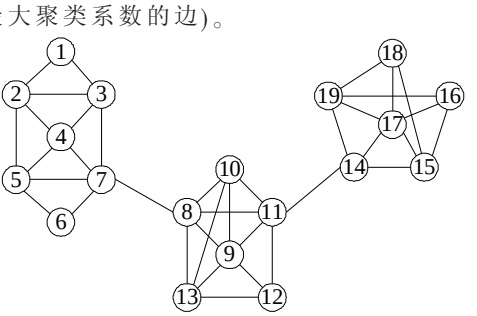


图 2 由 19 个点组成的三个社团网络

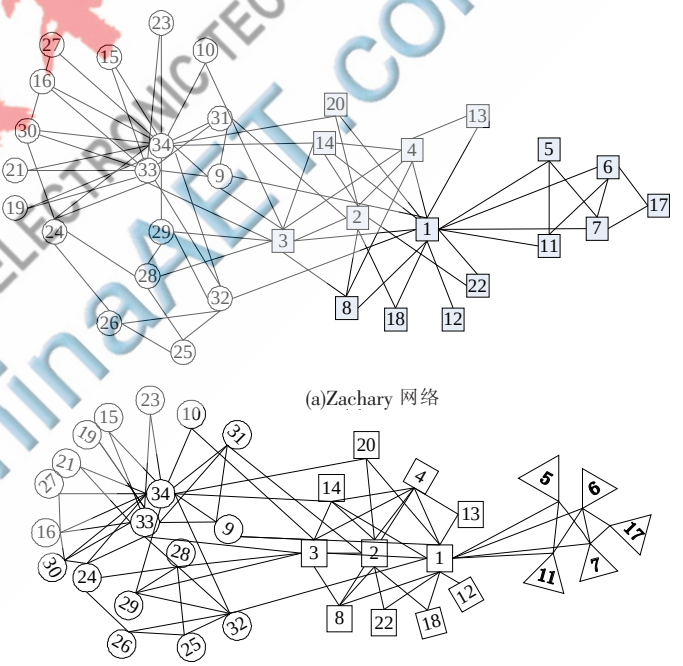
表 1 三社团网络算法分析表

节点	加入社团原因 ①	②	③	所在社团
7	√			C_1
6		√		C_1
5		√		C_1
4		√		C_1
2		√		C_1
3		√		C_1
1			0.333 3($e_{1,2}$)	C_1
8	√			C_2
10			0.750 0($e_{10,8}$)	C_2
13		√		C_2
9		√		C_2
12		√		C_2
11		√		C_2
17	√			C_3
15			0.750 0($e_{15,17}$)	C_3
16		√		C_3
18		√		C_3
19		√		C_3
14		√		C_3

首先选取网络中度最大的节点 7 作为社团 C_1 的初始节点，然后判断社团 C_1 的邻居集合{3,4,5,6,8}是否有点可以加入。发现节点 6 的度数值为 2，并且有一条边与社团 C_1 相连，因此有不小于一半的节点与社团 C_1 相连符合算法条件(1)，所以可以将节点 6 加入到社团 C_1 中去，得到社团 C_1 的邻居集合为{3,4,5,8}。经计算发现，与社团 C_1 相连的边聚类系数中 e_{74}=0.400 0，是所有与社团相连的边中聚类系数最大的。但是与节点 4 相连的边中聚类系数最大的是边 e_{42}，然而节点 2 不在社团 C_1 中，不符合算法条件(2)，因此不能将节点 4 加入到社团中去。观察到社团 C_1 的邻居集合中节点 5 的度数值为 4，与社团 C_1 相连的边数为 2，符合算法条件(1)，因此可将节点 5 加入到社团 C_1 中去，此时社团 C_1 的邻居集合为 {2,3,4,8}。发现与社团 C_1 相连的边中还是 e_{74} 的聚类系数最大，并且 e_{42} 是节点 4 聚类系数最大的边，节点 2 不在社团 C_1 内，故节点 4 不能加入社团 C_1。但是社团 C_1 的邻居集合中节点 4 的度数值为 4，与社团 C_1 相连的边数为 2，符合算法条件(1)，故将节点 4 加入到社团中去，则更新社团的邻居集合为{2,3,8}。经计算可知，e_{42}=0.500 0 是与社团 C_1 相连的聚类系数中最大的，而节点 2 的聚类系数最大的边是 e_{23} 而非 e_{24}，因此不能将节点 2 加入到社团 C_1 中。然而在更新后的社团邻居集合中，节点 2 和节点 3 的度数值都为 4，并且都有两条边与社团 C_1 相连，符合算法条件(1)，因此将节点 2 和节点 3 加入到社团 C_1 中，则社团的邻居集合变为{1,8}。此时得出与社团 C_1 相连的边中聚类系数最大的是 e_{21}=0.333 3，而且节点 1 的聚类系数最大的边也是 e_{21}，符合算法条件(2)，所以将节点 1 加入到社团 C_1 中去，更新社团的邻居集合为{8}，发现节点 8 的度数值为 5，与社团 C_1 有一条边相连，不符合算法条件(1)，而且节点 8 与社团 C_1 的聚类系数为 0，小于其他边的聚类系数，亦不符合算法条件(2)，因此不能将节点 8 加入到社团 C_1 中去。此时邻居集合中没有其他节点可以再加入到社团 C_1 中去，该社团发现完毕。将社团 C_1 从网络中移除，邻居集合清空。同理，分析剩余网络，分别得到社团 C_2 和社团 C_3，这时对应的结构被认为是网络的实际社团结构，实验结果与原图一致。

3.2 实例

20 世纪 70 年代初期，ZACHARY W 用了两年的时间观察美国一所大学空手道俱乐部成员间的相互社会关系。基于这些成员在俱乐部内部及外部的社会关系，ZACHARY W 构造了它们之间的关系网[15]，如图 3(a)所示。整个网络是由 34 个节点和 78 条边组成，节点代表俱乐部的成员，边代表成员之间的关系。



(a)Zachary 网络

(b)本文算法划分结果

图 3 空手道俱乐部内部成员的关系网络

根据本文的算法，以 Zachary 网络为例，Zachary 网络进行划分。由于篇幅有限，以第一社团为例分析该算法。首先选取当前网络中度最大的节点 34 作为社团 C_1 的初始节点，得到社团 C_1 的邻居集合为{9,10,14,15,16,19,20,21,23,24,27,28,29,30,31,32,33}。根据算法条件(1)将节点 10、15、16、19、21、23 和 27 号节点加入到社团 C_1 中，更新后社团的邻居节点为{9,14,20,24,28,29,30,31,32,33}。查找到社团 C_1 的最大聚类系数为 $e_{34,33}$=0.588 2，发现该聚类系数同时是节点 33 的最大的边聚类系数，因此将节点 33 加入到社团 C_1 中去，社团

C_1 的邻居节点变为{9,14,20,24,28,29,30,31,32}。根据算法条件 (1) 可以将节点 30 和 31 加入到社团 C_1 中去,然后更新邻居节点{9,14,20,24,28,29,32},再根据算法条件(2),得到最大聚类系数 $e_{30,24}$=0.400 0,判断可以将节点 24 加入到社团 C_1 中去,则社团 C_1 的邻居集合变为 {9,14,20,26,28,29,32}。接下来根据算法条件(1),将节点 9 和 28 加入到社团 C_1 中,更新邻居结合为{1,3,14,20,25,26,29,32}。计算发现最大聚类系数 e_{93}=0.181 8,但是 e_{93} 不是节点 3 聚类系数最大的边,故不能将节点 3 加入到社团 C_1 中去。然后根据算法条件(1)可陆续将节点 29、32、25 和 26 号节点加入到社团 C_1 中去。更新邻居集合为{1,3,14,20},发现其他邻居节点不能再加入到社团 C_1 中,至此社团 C_1 发现完毕。将社团 C_1 从网络中移除,并且清空邻居集合。同理对剩余网络进行判断,实验结果将网络划分为三个社团,如图 3(b)所示。

通过定义边的聚类系数,本文提出一个基于局部信息的社团结构发现算法。从网络中的节点和边出发,通过不断计算边的聚类系数进行节点合并。由于该算法是基于局部信息的,所以降低了时间复杂度。同时利用边聚类系数能够处理很多易混淆的节点,这样既节省了大量的计算时间又提高了计算的精度。通过对算法进行测试,实验结果证明了该方法的可行性和有效性。

参考文献

[1] ALBERT R, JEONG H, BARABÁSI A L. Diameter of the world-wide Web[J]. Nature (London), 1999, 401:130 - 131.

[2] SCOOT J. Social network analysis: a handbook [M]. London: Sage Publications, 2002.

[3] HOLME P, HUSS M, JEONG H. Subnetwork hierarchies of biochemical pathways [J]. Bioinformatics, 2003, 19:532 - 538.

[4] NEWMAN M E J. Scientific collaboration networks [J]. Physical. Review E, 2001, 64(1).

[5] 杨博,刘大有,Liu Jiming,等.复杂网络聚类算法[J].软件学报,2009,20(1):54-56.

[6] 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.

[7] 王立敏,高学东,马红权.基于最大节点接近度的局部社团结构探测算法[J].计算机工程,2010,36(1):25-29.

[8] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. Proceedings of the Natlional Academy Sciences of the United States of America, 2002,99(12):7821-7826.

[9] BREIGER R L, BOORMAN S A, ARABIE P. An algorithm for cluster relations data with applications to social network analysis and comparison with multidimensional scaling [J]. Journal of Mathematical Psychology, 1975,12: 328-383.

[10] KERNIGHAN B W, LIN S. A efficient beuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49:291-307.

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

[12] Zhang Dawei, Xie Fuding, Zhang Yong, et al. Fuzzy analysis of community detection in complex networks [J]. Physica a: Statistical Mechanics and its Applications, 2010,389(22): 5319-5327.

[13] 刘绍海,刘青昆,谢福鼎,等.复杂网络基于局部模块度的社团划分方法[J].计算机工程与设计,2009,3(20):4708-4714.

[14] 解伟,汪小帆.复杂网络的一种快速局部社团划分算法[J].计算机仿真,2007,24(11):82-85.

[15] ZACHARY W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33:452-473.

(收稿日期:2011-04-08)

作者简介:

赵晓慧,女,1987 年生,硕士研究生,主要研究方向:人工智能,数据挖掘。

刘微,女,1986 年生,硕士研究生,主要研究方向:人工智能,数据挖掘。