基于边图的重叠社团检测研究*

桂 春, 林 强
(西北民族大学 数学与计算机科学学院, 兰州 730030)

摘 要: 社团检测吸引了大量的研究。在真实网络中,社团的层次性和重叠性交织在一起,然而到目前为止大部分工作只研究网络的层次性或重叠性。2010 年 Ahn 等人在《Nature》上发表的文章证明层次性和重叠性是网络相同现象的两个方面,针对社团检测在 Ahn 方法的基础上提出了新算法——边图谱分析。将谱分析方法应用到边社团发现上,进行了兼顾层次性和重叠性的社团检测研究。实验使用两个真实网络来对比边图谱分析、Ahn 和派系过滤算法,结果表明提出的边图谱分析算法实现了网络的重叠社团检测,并且社团划分结果比较满意。

关键词: 社团检测; 边图; 谱分析; 角距离
中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2018)05-1342-03
doi:10.3969/j.issn.1001-3695.2018.05.013

Overlapping community detection research based on edge graph

Gui Chun, Lin Qiang
(*College of Mathematics & Computer Science*, *Northwest Minzu University*, *Lanzhou* 730030, *China*)

Abstract: Community detection has attracted much attention in the past years. Overlap and hierarchical structure are usually closely linked in real networks. Previous methods investigate these two properties of community structure separately. In 2010, Ahn et al published an article on *Nature* indicated that overlap and hierarchy structure were the two aspects of the same phenomenon. This paper proposed a new method for community detection named as SAEG (spectral analysis of edge graph) to achieve a balance between overlap and hierarchy by applying spectral analysis to edge community detection. It used two real social networks to compare the proposed algorithm with Ahn et al's and CPM algorithms. The results of SAEG algorithm are satisfactory.

Key words: community detection; edge graph; spectral analysis; angle distance

社团结构是网络的重要特征之一。社团结构是一个外松内紧的网络拓扑结构,它由许多网络子图组成,子图的内部边有很高的密度,而子图之间的连接边是稀疏的,这些子图叫做社团[1]。重叠性和层次性是社团的两个属性。仅仅研究重叠性或者层次性不足以准确地描述社团的结构,因为重叠性和层次性是网络相同现象的两个方面[2]。能够同时发现社团重叠性和层次结构的算法是 2008 年 Lancichinetti 等人[3]提出来的通过一个局部最优化函数寻找重叠社团算法。这种算法的缺点是随机种子节点的选择可能会影响所有重叠社团的层次性的检测结果。2009 年 Evans 等人[4]提出将流行的 k 派系过滤方法应用到原图的边图上来实现重叠社团检测。2010 年 Ahn 等人[2]使用边划分来代替节点的划分实现了网络重叠社团检测,该实验通过边社团检测成功地将网络的重叠性和层次性统一起来。Evants 和 Ahn 发现,当把非重叠社团检测算法应用到边社团上时都可以检测到重叠社团。边社团合理地解释了重叠性和层次性在网络中并存的现象。

2004 年 Donetti 等人[5]将谱分析与层次结构划分结合提出一种新算法。谱聚类通过使用一个矩阵的特征向量将数据进行聚类,使用特征向量使得原始数据集的聚类属性更为明显。该方法优于传统的聚类算法如 K-均值算法。迄今为止,谱聚类中使用最多的矩阵是拉普拉斯矩阵。拉普拉斯矩阵的最小非平凡特征向量的每个因子对应网络中的每个元素,可以使用非平凡特征向量构建网络元素的向量空间。本文的主要研究成果为:a)提出了一种新的重叠社团检测算法——边图谱分析,尝试将谱分析应用到网络的边图上;b)定义了边图谱分析所需要的边图度矩阵及线图拉普拉斯矩阵;c)基于边图拉普拉斯矩阵谱分析,在二维特征向量空间基础上进行了兼顾社团层次性和重叠性的复杂网络重叠社团发现算法的研究;d)实验发现 Zhang 等人提出的划分密度函数的不收敛并对其进行了修正,修正后的密度函数得到了较好的实验结果。

1 方法描述

本文使用的图是简单的无向无权图。$G=(V,E)$ 是一个图,$V(G)$ 是顶点的集合,$E(G)$ 是边的集合。

令 $n=|V(G)|$, $m=|E(G)|$, $i=\{1,2,\cdots,n\}$, $v_i\in V(G)$, 顶点 v_i 的度记为 $d_G(v_i)$, E_{v_i} 是 G 中与 v_i 关联的边的集合。

1.1 边图

根据图论的知识,每个图可以转换成相应的边图来表示。图 G 的边图记为 $C(G)$,其关联矩阵 $B(G)$ 是一个 $n\times m$ 的矩阵。关联矩阵中的元素记为 $(b_{i,j})$, n 和 m 分别是顶点和边的数目。如果顶点 v_i 与边 e_j 关联,那么 $b_{ij}=1$,否则 $b_{ij}=0$。边图 $C(G)$ 中顶点的数目对应图 G 中边的数目,即对每一条边 $e\in E(G)$,相应地都有顶点 $v_e\in V(C(G))$。$C(G)$ 的两个顶点相

收稿日期: 2017-01-03; **修回日期**: 2017-02-21 **基金项目**: 国家自然科学基金资助项目(61562075)

作者简介:桂春(1981-),女,甘肃兰州人,副教授,主要研究方向为复杂网络、数据挖掘(guichun2103@aliyun.com);林强(1979-),男,甘肃兰州人,副教授,博士,主要研究方向为复杂网络、数据流.

邻当且仅当对应的 G 中两条边拥有共同的顶点。边图 $C(G)$ 的邻接矩阵与图 G 的关联矩阵的关系如式(1)所示。

$$A(C(G)) = B(G)^{\mathrm{T}}B(G) - 2I_m \tag{1}$$

其中:$A(C(G))$是边图 $C(G)$的邻接矩阵;$B(G)$是图 G 的关联矩阵;I_m 是维数为 m 的单位矩阵[6]。如果一个网络的顶点数为 n,那么可以用一个对称的 $n \times n$ 阶拉普拉斯矩阵表示这个网络的拓扑结构,图 $G=(V,E)$的拉普拉斯矩阵 $C(G)$用式(2)计算。

$$C(G) = D(G) - A(G) \tag{2}$$

其中:$D(G)$、$A(G)$分别为图 G 的度矩阵和邻接矩阵。令图 G 中顶点 v_i 的度为 $d_G(v_i)$。在无向无权图中,不需要考虑顶点的入度和出度,拉普拉斯矩阵主对角线元素 C_{ii}的值就是顶点 v_i 的度,非主对角线元素 C_{ij}的值等于邻接矩阵中相应元素的负值 $-a_{ij}$。当 $i=j$ 时,度矩阵 $D(G)$中元素的值 $d_{ij}=d_G(v_i)$,否则为 0,每个顶点的 $d_G(v_i)$值是与该顶点相连的边的数目。图 G 的邻接矩阵是一个 n 阶对称矩阵 $A(G)=[a_{ij}]$。如果顶点 v_i 与 v_j 相邻,则 $a_{ij}=1$,否则 $a_{ij}=0$。所以一个网络的拉普拉斯矩阵为

$$C_{ij} = \begin{cases} d_G(v_i) & \text{if } i=j \\ -a_{ij} & \text{otherwise} \end{cases} \tag{3}$$

其中:$d_G(v_i)$是节点 v_i 的度。由式(1)可以得到边图的邻接矩阵。为了计算出边图的拉普拉斯矩阵,还需要原图中边的度。原图中边的度可以从节点的度类推得到:一个节点的度是与之相连的边的数目,也可以说一个节点的度是直接与之相连的节点的数目。由此给出了网络中边的度的定义。

定义 1　边的度。网络中边的度是与之直接相连的边的个数。因此边的度 m 可以由式(4)来计算。

$$\deg(e_m) = \sum_n e_{mn} \tag{4}$$

其中:e_{mn}是邻接矩阵 $A(C(G))$的元素。边图 $C(G)$顶点的度矩阵如式(5)所示。

$$D(C(G)) = \begin{pmatrix} \deg(e_1) & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \deg(e_m) \end{pmatrix} \tag{5}$$

主对角线元素的值就是原图中边 e_i 的度,m 是原图 G 中边的数目,非主对角线元素的值为 0。根据式(1)和(5)得到边图 $C(G)$的拉普拉斯矩阵 $L(C(G))$的定义。

定义 2　边图的拉普拉斯矩阵。

$$L(C(G)) = D(C(G)) - A(C(G)) \tag{6}$$

其中:$D(C(G))$和 $A(C(G))$分别是边图 $C(G)$的度矩阵和邻接矩阵。在得到边图的拉普拉斯矩阵的定义后,就可以根据谱分析的方法进行边图的操作。

1.2　算法原理

边图谱分析算法将谱分析应用到边图的层次聚类。其具体步骤如下:

a)读取网络 G,计算关联矩阵 $B(G)$,并定义 G 的边图为 $C(G)$;

b)根据式(1)得到边图 $C(G)$的邻接矩阵 $A(C(G))$;

c)由邻接矩阵 $A(C(G))$得到边图 $C(G)$的度矩阵 $D(C(G))$;

d)根据式(6)计算边图 $C(G)$的拉普拉斯矩阵 $L(C(G))$以及 $L(C(G))$的特征值及特征向量;

e)选择两列非平凡特征向量组成二维非平凡特征向量空间;

f)根据相似度计算方法生成相似度矩阵,并根据相似度矩阵进行层次聚类,生成层次聚类龙骨图;

g)利用截断函数 D 切割龙骨图,得到网络的最优化社团结构。

2　基于边图谱分析算法的重叠社团检测实验

2.1　截断函数的选择

社团检测的目的不是寻找简单的社团层次结构,而是找到网络中所有有意义的社团,但社团个数通常是未知的,因此如何截断龙骨图是一个很重要的问题。使用 Ahn 等人 2010 年在《Nature》上发表的文章中定义的截断龙骨图来划分函数 D。

$$D = \frac{2}{M}\sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \tag{7}$$

其中:M 是网络中边的数目;m_c 是社团 c 中边的数目;n_c 是社团 c 中顶点的数目;D 的最大值是 1,最小值是 $-2/3$,划分密度 D 越大越好。Zhang 等人[7]对 Ahn 等人提出的划分密度作改进,在 D 的分母处添加一个控制节点活动度的惩罚因子 q_c,令 Zhang 等人的划分密度为 D_Z,如式(8)所示。

$$D_Z = \frac{2}{M}\sum_c \frac{m_c}{q_c}\frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \tag{8}$$

其中:$q_c = \max_{j \in c} l_j, j \in c$ 表示节点 j 属于社团 c;l_j 是节点 j 所属社团的标签代码,即节点 j 属于第几个社团。但是实验结果表明这个划分密度可能出现不收敛的情况,即可能会出现网络中的所有节点聚集成为一个社团,此时本文令划分密度等于零,这个划分密度定义为 D_{XZ}。

2.2　网络

在这个实验中共使用了空手道俱乐部网络和海豚社交网络,这两个数据集可以从网址 http://www-personal.umich.edu/~mejn/netdata/下载。这两个网络适合作分类,可以用来测试社团检测算法的准确性。

2.3　实验

2.3.1　空手道网络

空手道网络边图的拉普拉斯矩阵的最小、次小非平凡特征值分别为 $\lambda_2 \approx 0.900506243293657$,$\lambda_3 \approx 1.74628971847267$,基于边图谱分析算法得到的最优社团检测结果如图 1 所示。第一个社团与第二个社团之间的重叠节点是 9、10、14、20、28、29、31、32 和 33;第一个社团与第三个社团之间的重叠节点是 5、6、7 和 11;第二个社团与第三个社团之间没有重叠节点。

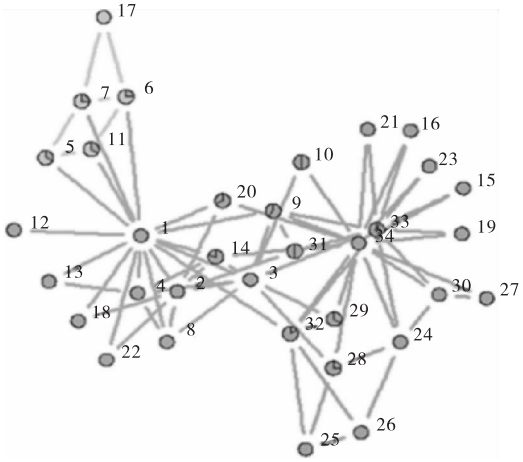

图1　基于边图谱分析算法的空手道俱乐部网络最优社团检测结果

图 2 是基于 Ahn 算法的最优社团检测结果:共发现了八个社团。这与空手道网络的传统划分结果差距很大,因此这个结果是不可接受的。基于边图谱分析的空手道网络社团检测结果如表 1 所示。

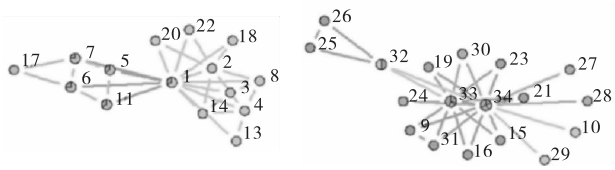

图2 基于Ahn算法的空手道俱乐部网络最优社团检测结果

表 1 基于边图谱分析的空手道网络社团检测结果

社团数目	社团检测结果
1	1,2,3,4,5,6,7,8,9,10,11,12,13,14,18,20,22,28,29,31,32,33
2	9,10,14,15,16,19,20,21,23,24,25,26,27,28,29,30,31,32,33,34,
3	5,6,7,11,17

2.3.2 海豚社交网络

海豚社交网络边图的拉普拉斯矩阵最小、次小非平凡特征值分别为 $\lambda_2 \approx 0.229628750199307$，$\lambda_3 \approx 1.38812933722022$，基于边图谱分析算法和 Ahn 算法得到的最优社团检测结果如图 3 和 4 所示。从图 3 可以看出，基于边图谱分析算法的海豚社交网络的最优社团检测结果为三个社团：第一个社团与第二个社团之间的重叠节点是 Stripes 和 Shmuddel；第二个社团与第三个社团之间的重叠节点是 SN100、Oscar、DN63、Knit 和 SN89；第一个社团与第三个社团之间没有重叠节点。这个结果与传统划分结果几乎一致，可以接受。基于 Ahn 算法的海豚网络的最优社团检测结果（图 4）为 13 个社团，其结果是不可接受的。基于边图谱分析的海豚社交网络社团检测结果如表 2 所示。

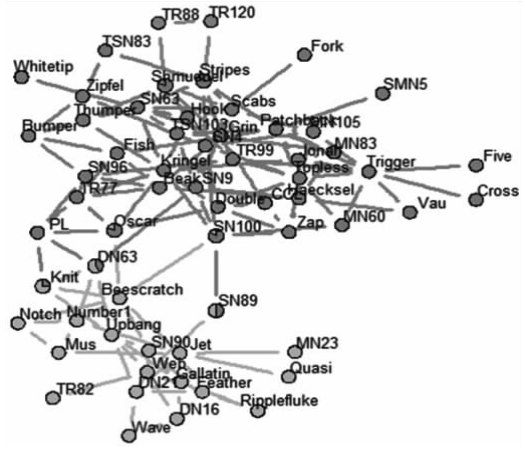

图3 基于边图谱分析算法的海豚社交网络最优社团检测结果

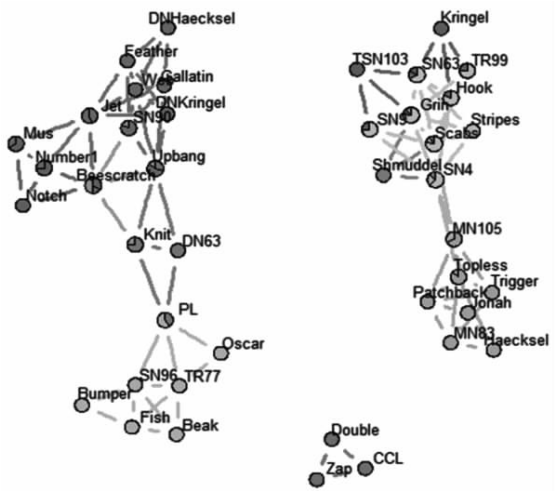

图4 基于Ahn算法的海豚社交网络最优社团检测结果

表 2 基于边图谱分析的海豚社交网络社团检测结果

社团数目	社团检测结果
1	Feather, Gallatin, Jet, Knit, MN23, Mus, Notch, Number1, Oscar, Quasi, Ripplefluke, SN100, SN90, Upbang, Wave, Web, DN16, DN21, Beescratch, DN63, SN89, TR82
2	Double, DN63, Fish, Five, For, Grin, Haecksel, Hook, Jonah, Knit, Kringel, MN105, MN60, MN83, Oscar, Patchback, PL, Scabs, Shmuddel, SMN5, SN100, SN4, SN63, SN89, SN9, SN96, Stripes, Thumper, Topless, TR77, TR99, Trigger, TSN103, TSN83, Vau, Whitetip, Zap, Zipfel, CCL, Beak, Bumper, DN63, Fork, Cross, Five
3	TR120, TR88, Stripes, Shmuddel

3 实验结果讨论

谱分析在实验中被尝试应用到边图领域，并被改进为能够发现重叠社团的算法。层次聚类通过谱分析实现，其主要思想是使用拉普拉斯矩阵的两列非平凡特征向量构建向量空间，最后根据截断函数得到最优社团划分结果。实验结果证明在计算相似度时角距离比 Jaccard 指数的结果更加适合于社团检测，这与 Donetti 的结论一致。实验使用的评价标准有社团划分个数、覆盖率和未覆盖节点个数三个，并与社团发现一种非常重要的方法——派系过滤算法进行了对比，实验结果如表 3 所示。

表 3 三种社团检测算法实验对比结果

网络	边图谱分析算法			Ahn 算法			派系过滤算法		
	CN	CR/%	UN	CN	CR/%	UN	CN	CR/%	UN
空手道	3	100	0	8	97	1	3	94	1
海豚	3	100	0	13	68	20	4	74	16

注：加粗的数字表示最优实验结果；CN 表示社团划分个数；CR 表示覆盖率；UN 表示未覆盖节点数

该算法的复杂度依赖于计算拉普拉斯矩阵的特征向量的复杂度，计算 $n \times n$ 矩阵的所有特征向量的算法复杂度通常为 $O(n^3)$，然而真实网络的拉普拉斯矩阵是稀疏矩阵，因此可以用 Lanczos[8] 方法快速计算主要特征向量，使得算法的复杂度降低至 $m/(\lambda_3 - \lambda_2)$。其中：$m$ 是网络中边的数目；λ_3 和 λ_2 分别为拉普拉斯矩阵的第三小和次小特征值。

参考文献：

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

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

[3] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2008, 11(3): 19-44.

[4] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(1): 016105.

[5] Donetti L, Munoz M A. Detecting network communities: a new systematic and efficient algorithm[J]. Journal of Statistical Mechanics Theory & Experiment, 2004, 2004(10): 10012.

[6] 柳柏濂. 组合矩阵论[M]. 2 版. 北京：科学出版社, 2005.

[7] Zhang Zhongyuan, Wang Yong, Ahn Y Y. Overlapping community detection in complex network using symmetric binary matrix factorization[J]. Physical Review E, 2013, 87(6-1): 062803.

[8] Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators[J]. Journal of Research of the National Bureau of Standars, 1950, 45(4): 255-282.