

基于先验知识与模块性的网络社区结构探测算法

杜海峰^{1,2}, 李树苗¹, Marcus W. Feldman², 悦中山^{1,3}, 杨绪松^{1,3}

(1. 西安交通大学人口与发展研究所, 710049, 西安; 2. 斯坦福大学人口与资源研究所, 94305, 美国斯坦福;
3. 西安交通大学管理学院, 710049, 西安)

摘要: 在分析模块性指标和 Newman 有关网络社区结构探测算法的基础上, 提出了一种基于先验知识与模块性的社区结构探测算法. 利用节点度等社会网络结构先验知识, 获得一个社区结构的基本划分, 然后进行社区的合并, 以此获得一个清晰的社区结构. 经计算机模拟网络、Ucinet 软件网络和中国农民工社会网络的社区结构探测, 结果表明所提算法比 Newman 的迭代次数减少近 50%, 并且可以获得更好的模块性指标.

关键词: 社会网络; 社区结构; 模块性; 探测算法

中图分类号: N94 **文献标识码:** A **文章编号:** 0253-987X(2007)06-0750-05

Detecting Algorithm Based on Prior Knowledge and Modularity for Networked Community Structure

Du Haifeng^{1,2}, Li Shuzhuo¹, Marcus W. Feldman², Yue Zhongshan^{1,3}, Yang Xusong^{1,3}

(1. Institute for Population and Development Studies, Xi'an Jiaotong University, Xi'an 710049, China;
2. Morrison Institute for Population and Resource Studies, Stanford University, CA 94305, USA;
3. School of Management, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: On the basis of analyzing the modularity and Newman detecting algorithm for network community structure, an algorithm based on prior knowledge and modularity (PKM) is put forward to detect community structure. An original community division is acquired by using the prior knowledge of the structure of social networks, such as the degree of the node, and then the communities are combined so as to get a clarified partition. Through calculation of computer simulation networks, Ucinet networks and Chinese rural-urban migrants social networks, the results indicate that the number of iterations of the proposed algorithm is reduced nearly by 50% compared to that of Newman's, and the higher modularity can be yielded.

Keywords: social network; community structure; modularity; detecting algorithm

自然界中存在的大量复杂系统都可以通过网络加以描述, 然而这些网络是不均匀的, 由许多子网络构成^[1-2], 这一现象在社会网络中尤其常见^[3]. Newman 等人将异构网络中的由不同性质、类型节点组成的关系丰富的结构称之为“社区”(子网络), 并指出社区内的节点关系稠密, 不同社区节点之间的、关系稀疏的结构(社区结构)是复杂网络的特征之

一^[1]. 社会网络中凝聚子群等概念与社区结构的含义基本一致, 但定义过于严格^[4], 缺乏相应的评价指标, 计算的时间复杂度也比较高.

随着复杂网络研究的不断深入, 社区结构探测方法受到了广泛关注^[5], 典型的包括 Newman-Girvan(NG)算法、基于模块化指标的分级聚类算法^[6-7], 这些算法忽略了网络存在结构先验知识, 而

且在确定社区数目、计算复杂度、计算边中间性值上还有待于完善。

本文用社区结构这一概念代替了社会网络中的凝聚子群,这种社区结构是在对网络进行划分以后形成的新的网络结构,而社区结构探测是获得社区结构的过程。在分析 Newman 模块性指标特点以及相关社会网络社区结构探测算法的基础上,本文还提出了一种适用于社会网络的社区结构探测算法。

1 模块性指标

如果忽略节点之间的关系强弱和作用方向,社会网络可以用无向无权图 $G(V, E)$ 表示,其中 V 表示节点(网络成员)集合, E 表示边(网络成员关系)集合。邻接矩阵 A 是社会网络的另一种表示方法,对于有 n 个节点的网络,定义 $A = (a_{ij})$, $i, j = 1, 2, \dots, n$, 如果节点 i 与 j 相连,则 $a_{ij} = 1$, 否则 $a_{ij} = 0$ 。一般认为,节点不存在自身连接,所以 $a_{ii} = 0$, $i = 1, 2, \dots, n$ 。

设网络节点集合 V_p, V_q 是 V 的真子集,即 $V_p \neq \emptyset, V_q \neq \emptyset$, 且 $V_p \subset V, V_q \subset V$ 。若 $V_p \cap V_q = \emptyset$, 则有 $A_{pq} = \{a_{ij}\}, i \in V_p, j \in V_q$, 且 $A_{pq} \subset A$ 。记 $\|A\| = \sum_{i=1}^n \sum_{j=1}^n a_{ij}$, 因此当 $p \neq q$ 时, $\|A_{pq}\| = \sum_{i \in V_p} \sum_{j \in V_q} a_{ij}$, $a_{ij} \in A$ 为子集 V_p, V_q 之间的关系数量,而 $p = q$ 时, $\|A_{pp}\| = \sum_{i \in V_p} \sum_{j \in V_p} a_{ij}$, $a_{ij} \in A$ 为子集内部的关系数量。网络社区结构划分就是将节点集合 $V = \{v_1, v_2, \dots, v_n\}$ 划分为 m 个子集合 $V_1, V_2, \dots, V_m$, 使其满足条件:① $V_p \neq \emptyset, p = 1, 2, \dots, m$; ② $\bigcup_{p=1}^m V_p = V$; ③ $V_p \cap V_q = \emptyset, p \neq q, p, q = 1, 2, \dots, m$ 。由于社会网络关系复杂,很难将其划分为互不联系的社区组合,因此条件③很难满足。在应用中,若定义 $e_{pq} = \|A_{pq}\| / \|A\|$, 则网络社区结构划分只要同时保证 $\max(\sum_p e_{pp}), p = 1, 2, \dots, m$ 和 $\min(\sum_{p,q} e_{pq}), p \neq q$ 且 $p, q = 1, 2, \dots, m$, 就可以满足社区内部的关系密集而社区之间的关系稀疏。因此,Newman 将度量社区结构划分的模块性指标定义为^[1]

$$Q = \sum_{p=1}^m [e_{pp} - (\sum_{q=1}^m e_{pq})^2] \quad (1)$$

并进一步指出,如果 Q 大于 0.3, 网络存在明显的社区结构^[1]。因为, $G(V, E) = G(\bigcup_{i=1}^m V_i, E)$, 所以网

络社区结构并没有改变网络本身的基本特征,只是对原网络按社区结构的方式重新表达。但是,一个规模为 n 的网络,可能存在很多种社区结构划分,寻找合适的网络社区结构是一个复杂的过程。在给定目标函数之后,网络社区结构探测本质上就是一个聚类 and 分类的问题,也是一个优化问题^[2], 因此目标函数的定义就显得非常重要。

2 算法实现

本文在 Aaron-Newman(AN) 等级聚类算法的基础上^[7], 强调了社会网络结构的先验知识, 尤其是节点度对寻找社区结构的影响。具体地, 本文算法首先基于网络节点度获得一个初始社区网络结构, 即以度值较大的节点为中心实现社会网络的初始聚类, 然后采用与 AN 算法一样的思想, 将社区结构探测视为子网络合并的聚类过程, 以获得最大模块性指标。与 AN 算法相比, 本文算法的聚类不是从单个节点开始, 而是从子网络(初始社区)开始, 这样可以减少算法的迭代次数。虽然获得初始网络社区结构需要一定的计算量, 但是比 AN 从节点开始合并迭代所需的计算量要小得多。更为重要的是, 本文算法考虑了社会网络特征, 其将节点度的先验知识引入到了初始网络社区探测之中, 由此增加了算法的适用性。

不失一般性, 考虑一次只合并一个基本单位(可能是单个网络成员, 也可能是已经获得的社区), 如果合并是从 m 个社区开始, 并用 $U_{m-k}^{(k|m)}$ 表示对这 m 个社区进行了 k 次合并, 获得了 $m-k$ 个社区, 则这一过程可以表示为

$$U_m^{(0|m)} \rightarrow U_{m-1}^{(1|m)} \rightarrow \dots \rightarrow U_{m-k}^{(k|m)} \rightarrow \dots \rightarrow U_1^{(m-1|m)} \quad (2)$$

记 $Q(U_{m-k}^{(k|m)})$ 是合并 $U_{m-k}^{(k|m)}$ 对应社区结构的模块性指标值, 则有

$$Q(U_{m-k}^{(k|m)}) = \sum_{c=1}^{m-k} e_{cc}^{(k)} - \sum_{c=1}^{m-k} (\sum_{q=1}^{m-k} e_{cq}^{(k)})^2 \quad (3)$$

式中: $e_{cq}^{(k)}$ 是第 k 次合并后获得的社区 c, q 之间的关系度量。设 $U_{m-k-1}^{(k+1|m)}$ 是由 $U_{m-k}^{(k|m)}$ 所对应的社区结构合并社区 p, q 获得的, 因此有

$$Q(U_{m-k-1}^{(k+1|m)}) = Q(U_{m-k}^{(k|m)}) + 2e_{pq}^{(k)} - 2 \sum_{j=1}^{m-k} e_{qj}^{(k)} \sum_{i=1}^{m-k} e_{pi}^{(k)} \quad (4)$$

记 $\phi_{pq}^{(k)} = 2e_{pq}^{(k)} - 2 \sum_{j=1}^{m-k} e_{qj}^{(k)} \sum_{i=1}^{m-k} e_{pi}^{(k)}$, 显然只要满足 $\phi_{pq}^{(k)} > 0$, 就可保证 $Q(U_{m-k-1}^{(k+1|m)}) > Q(U_{m-k}^{(k|m)})$, 即在

$U_{m-k}^{(k|m)}$ 基础上进行一次合并 $U_{m-k-1}^{(k+1|m)}$ 是有效的. 进一步地, 如果 $\max_{p,q}(\phi_{pq}^{(k)}) > 0$, 那么就能保证本次合并是最优的.

上述合并过程的计算复杂度与 m 有关, 如果能够减小 m , 则可以减小计算复杂度. 社会网络一般都具有特殊的结构特征, 例如小世界现象和无标度特征^[8-9], 即网络中的局部连接丰富且可能存在核心节点, 因此可以利用这些网络结构的先验知识, 先获得一个网络的初始社区结构, 使 m 尽量小, 然后再用上述合并步骤, 从而达到减少计算复杂度的目的. 在社会网络中, 网络节点的度是一个重要的结构参数, 是指与某节点直接相连的关系数目, 那么节点 i 的度

$$d_i = \sum_{j=1}^n a_{ij} \quad (5)$$

网络无标度特征表明, 度值大的节点是核心节点, 也应该是相应社区的重要节点. 因此, 可以借助节点度的特征, 将网络节点按照度进行归类, 并由此获得网络初始划分, 进行社区合并.

综合考虑上述因素, 本文提出基于模块性指标与网络结构先验知识的社区结构探测算法(PKM), 步骤描述如下.

步骤 1: 计算网络中节点的度.

步骤 2: 找到度为最大的节点 i , 将该节点以及与其直接相连的节点划为 1 个社区.

步骤 3: 将邻接矩阵中节点 i 所在的行、列全部赋 0.

步骤 4: 如果邻接矩阵中的所有元素都为 0, 执行步骤 5, 否则执行步骤 1.

步骤 5: 计算任意 2 个社区合并引起的指标变化 $\phi_{pq}^{(k)}$.

步骤 6: 选取最大 $\phi_{pq}^{(k)}$, 并将相应的子网络 p, q 进行合并.

步骤 7: 如果所有的网络已经被合并成一个网络, 算法结束, 否则执行步骤 5.

步骤 1~步骤 4, PKM 算法是基于节点度获得一个网络的基本社区结构. 对于步骤 2, 如果多个节点度相同, 则选择聚类系数较大的节点. 在最差的情况下, PKM 算法需对每一个节点进行操作, 因为求取 n 元素序列最大值的时间复杂度为 $O(n)$, 所以获得初始划分的时间复杂度应为 $O\left(\frac{n(n-1)}{2}\right) \leq O(n^2)$. 步骤 5~步骤 7 是 PKM 算法基于模块性指标的变化情况, 对相应的社区进行合并以获得最佳社区结构, 整个合并过程的计算复杂度为

$O\left(\frac{m(m-1)}{2}\right) \leq O(n^2)$ ^[7]. 综上所述, PKM 算法的计算复杂度为 $O(n^2)$, 如果是稀疏矩阵, 或者网络社区结构明显, 算法的复杂度还会大大降低^[7]. 由于网络规模是有限的, 循环合并的次数一定小于网络规模, 所以算法必然收敛.

3 应用

PKM 算法分别应用于 3 种数据的社区结构探测, 即计算机生成数据、Ucinet 软件数据以及西安交通大学人口与发展研究所深圳市流动人口专项调查数据. 计算机生成数据主要是验证算法对人为预设网路社区结构的识别能力; Ucinet 数据是公用数据, 主要用于比较算法的探测结果; 人口专项调查数据则为 PKM 算法提供实际应用平台, 以检验算法对不同类型网络社区结构的识别能力. PKM 算法、NG 算法^[10]、AN 算法^[7]对不同网络进行了探测, 即探测了社区结构的模块性指标 Q 及迭代次数 T . 显然, Q 值越大, 算法获得的社区结构越显著, 而迭代次数越少, 算法的计算效率越高. 需要说明的是, AN、PKM 算法由 Matlab 软件实现, 而 NG 算法的计算结果由 Ucinet 软件的相关程序获得.

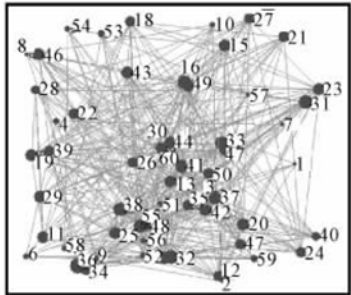
3.1 计算机生成数据

按照如下规则生成网络: 规模为 $\{n_1, n_2, \dots, n_m\}$ 的 m 个网络, 各网络内部节点间的连接概率为 $\{p_1, p_2, \dots, p_m\}$, 不同网络间的节点按照概率 p_c 连接, 构成一个新网络, 如果 $p_i \gg p_c$, $i=1, 2, \dots, m$, 则该网络具有明显的社区结构特征. 本文的规模为 $\{10, 20, 20, 10\}$ 的 4 个网络, 网络内部节点间的连接概率均为 0.5, 网络间节点的连接概率为 0.05, 构成图 1a 所示网络, 而图 1b 是网络社区结构的探测结果. 可以看出, PKM 算法可以很好地发现预设的社区结构, 当其迭代 46 次便获得了最优值 0.607, 而 AN 算法需要迭代 56 次.

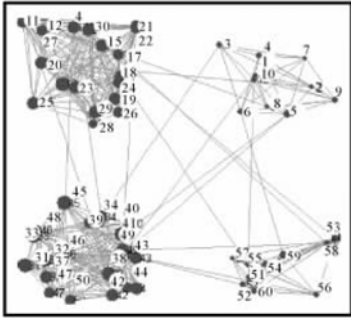
3.2 Ucinet 网络

Zachary 空手道俱乐部网络以及 Drugnet 网络是 Ucinet 软件提供的, 前者的网络规模为 34, 平均度为 2.294 1, 是一个对称 0-1 网络, 广泛用于网络分析研究, 而后的网络规模为 293, 是一个非对称 0-1 网络. 由于本文只探讨无向网络的社区结构, 所以应将 Drugnet 网络转化为对称网络, 转化后的平均度为 0.969 3.

针对 Zachary 空手道俱乐部、Drugnet 这 2 种网络, 对比了 PKM 算法和 AN 算法的模块化指标 Q



(a) 网络原始拓扑结构



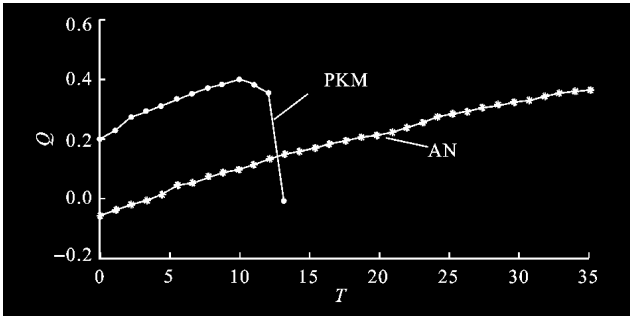
(b) 社区结构探测结果

图1 计算机模拟数据

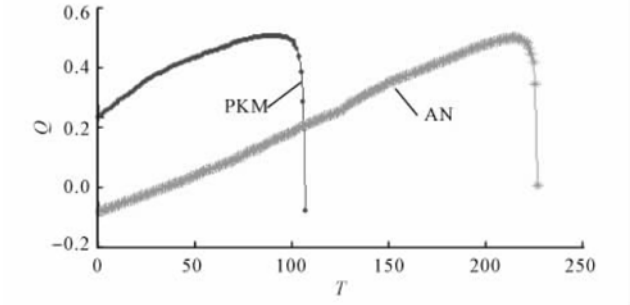
随迭代次数 T 变化的情况,如图2所示.从中可以看出,PKM算法获得最优值的迭代次数远小于AN算法.另外,针对Zachary空手道俱乐部网络,AN算法的最优 Q 值为0.383,而PKM算法为0.412;针对Drugnet网络,AN算法的最优 Q 值为0.745,PKM算法为0.751.由此可见,针对这2个网络,PKM算法无论是迭代次数还是社区结构探测结果,均优于AN算法.

3.3 农民工社会网络

基于西安交通大学人口与发展研究所于2005年4月组织进行的深圳农村流动人口调查,本文利用深圳市3个不同性质的聚居点,采样了以男性为



(a) Zachary 空手道俱乐部网络



(b) Drugnet 网络

图2 模块性指标 Q 随迭代次数的变化情况

主的CZ公司,以女性为主的HM公司,以及男女混合的XYX公司等整体网络.访问的问题包括社会支持和社会讨论,社会支持包括实际支持、情感支持和社会交往支持,而社会讨论包括婚姻讨论、生育及子女教育讨论、避孕讨论和养老讨论.有关本次抽样调查的目的、内容、执行和数据质量的具体描述请参见西安交通大学2005年关于深圳市外来农村流动人口调查报告.本文利用该调查获得了21个整体网络数据(3个聚居点,每个聚居点有7个网络),并对比了PKM、Newman相关算法在社区结构探测中的性能.

表1列出了3种算法的社区结构探测最优模块

表1 不同算法的模块性与迭代次数比较

指标	调查地点	算法	实际支持网	情感支持网	社会交往支持网	婚姻讨论网	生育讨论网	避孕讨论网	养老讨论网
Q	HM	NG	0.437	0.480	0.463	0.553	0.544	0.389	0.465
		AN	0.539	0.578	0.546	0.638	0.680	0.740	0.678
		PKM	0.547	0.578	0.545	0.643	0.673	0.740	0.676
	XYX	NG	0.421	0.455	0.351	0.474	0.556	0.371	0.556
		AN	0.464	0.507	0.413	0.496	0.603	0.609	0.587
		PKM	0.462	0.498	0.422	0.526	0.591	0.605	0.589
	CZ	NG	0.692	0.696	0.696	0.658	0.660	0.612	0.761
		AN	0.728	0.713	0.701	0.762	0.744	0.783	0.789
		PKM	0.711	0.717	0.702	0.760	0.742	0.783	0.787
T	HM	AN	176	181	187	170	139	101	114
		PKM	88	84	100	88	87	29	43
	XYX	AN	85	82	82	76	73	45	74
		PKM	47	43	48	33	26	8	26
	CZ	AN	122	119	125	92	98	44	93
		PKM	68	62	72	36	39	8	41

性指标 Q 和迭代次数 T . 可以看出, AN 和 PKM 算法的最优 Q 值基本相同, 均优于 NG 算法. 同时, 在获得最优 Q 值的情况下, 对比了算法的子网络合并迭代次数 T , 从中发现, PKM 算法的迭代次数只是 AN 算法的 $1/2$. 由于 NG 算法的计算复杂度比较高, 且其原理及实现方法均不同于 PKM、AN 算法, 因此表 1 中没有比较 NG 算法的迭代次数.

进一步地考察不同性质网络下的城乡流动人口最优社区结构的模块性指标 Q , 可以发现: ①所有网络的 Q 值均大于 0.3, 有的甚至超过了 0.7, 这说明社区结构在中国农民工网络中广泛存在, 且有研究者认为^[3], 社区现象(小团体)在中国尤其重要, 如果这种现象过分严重, 可能会损坏社会利益; ②不同性质网络的社区结构存在差异, 而且社会讨论网络的 Q 值一般都要大于社会支持网络(少数避孕讨论网有异常, 可能与成员较少、分散有关), 这表明对涉及个人隐私的问题, 城乡流动人口往往只与社区集团内部的人进行讨论, 不愿意与其他社区成员交流; ③不同调查地之间的社区结构也存在差异, 以男性为主的 CZ 的 Q 值大于以女性为主的 HM, 以及男女比例接近的 XYX. 由此表明, 在城乡流动人口中, 男性比女性更容易团体化.

4 结 论

网络分析, 特别是 20 世纪末兴起的复杂网络研究, 已受到物理学、社会学、生物学、计算机等众多领域学者的关注^[8,11], 而以社区结构为代表的网络结构研究已经成为主要方向之一. 本文将社区结构探测过程视为一种基于目标函数的聚类优化过程, 并考虑了网络结构具备的先验知识, 由此提出了一种网络社区结构探测方法. 理论分析表明, 本文算法的时间复杂度不会超过 $O(n^2)$. 针对网络数据的社区结构探测表明, 本文算法可以在模块性指标体系下获得更好的社区结构.

后续研究主要集中在两个方面: 一是网络结构的先验知识, 即除了本文考虑的节点度, 还应考虑网

络的距离等; 二是将社区结构探测算法应用于实际问题, 以进行有关网络的结构分析, 研究网络的社区组成特点等.

参考文献:

- [1] Newman M E J. Detecting community structure in networks [J]. Eur Phys J:B, 2004, 38(2):321-330.
- [2] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Phys Rev:E, 2004, 69(6):066133.
- [3] 罗家德. 社会网分析讲义 [M]. 北京: 社会科学文献出版社, 2005.
- [4] Wasserman S, Faust K. Social network analysis: methods and applications [M]. New York: Cambridge University Press, 1994.
- [5] 王林, 戴冠中. 复杂网络中的社区发现——理论与应用 [J]. 科技导报, 2005, 23(8):62-66.
Wang Lin, Dai Guanzhong. Community finding in complex networks — theory and applications [J]. Science & Technology Review, 2005, 23(8):62-66.
- [6] Scott W, Smyth P. A spectral clustering approach to finding communities in graphs [EB/OL]. [2006-04-11]. http://www.datalab.uci.edu/papers/siam_graph_clustering.pdf.
- [7] Aaron C, Newman M E J, Moore C. Finding community structure in very large networks [J]. Phys Rev:E, 2004, 74(3):36104.
- [8] Albert R, Barabási A L R. Statistical mechanics of complex networks [J]. Reviews of Modern Physics, 2002, 74(1):48-97.
- [9] Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ networks [J]. Nature, 1998, 393(4):440-442.
- [10] Newman M E J, Girvan M. Statistical mechanics of complex networks [M]. Berlin: Springer, 2003.
- [11] Dorogovtsev S N, Mendes J F. Scaling properties of scale-free evolving networks: continuous approach [J]. Phys Rev:E, 2001, 63(5):056125.

(编辑 苗凌)