

# 面向大规模在线社交网络的社团抽取算法

张锡哲, 张聿博, 陈章禄, 张 斌

(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

**摘 要:** 针对现有的社团分析算法无法在大规模网络上应用的问题, 提出一种社团抽取算法, 可以高效地分析网络的社团特征. 该方法无需事先获取网络的全部拓扑结构, 采用网络搜索与社团判定相结合的思路, 可有效地抽取结构未知的社交网络上的某个特定社团, 从而使分析超大规模网络社团结构成为可能. 在仿真数据集上进行实验, 分析抽取准确率的影响因素, 得出网络平均度越大抽取准确率越高. 进一步实验结果表明, 社团抽取算法的准确率与现有方法接近, 并且执行效率明显高于现有方法, 验证了该算法的可行性和有效性.

**关 键 词:** 社交网络; 社团抽取; 社团检测; 社团结构; 网络搜索

中图分类号: TP 391

文献标志码: A

文章编号: 1005-3026(2015)03-0342-04

## Community Extraction Algorithm for Large-Scale Online Social Networks

ZHANG Xi-zhe, ZHANG Yu-bo, CHEN Zhang-lu, ZHANG Bin

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHANG Xi-zhe, E-mail: zhangxizhe@ise.neu.edu.cn)

**Abstract:** Since the existing community analysis methods cannot be applied in large-scale networks, a community extraction algorithm is proposed. The community structure can be analyzed effectively with the algorithm. The topology of the network is not needed, with the combination of network search and community detection capabilities, the structure of the particular community can be effectively extracted from the social network with unknown topology. The analyzing of the community structure of large scale network is possible with the algorithm. Experiments on simulation data are performed to analyze the influence factor of accuracy, and it is concluded that the accuracy increases with the average degree. Furthermore, it is found that the accuracy of community extraction algorithm is close to existing methods, and the efficiency is much better, the results show the algorithm is feasible and effective.

**Key words:** social network; community extraction; community detection; community structure; network search

随着在线社交网络的应用范围不断扩大, 其已成为目前产业界和学术界的研究热点. 社交网络成员之间依据社交关联形成了复杂的网络结构. 对社交网络结构和动力学的分析, 例如社团发现<sup>[1]</sup>、链接预测<sup>[2]</sup>、传播建模等<sup>[3-4]</sup>, 具有重要的应用价值.

然而, 急剧增加的网络规模, 给网络结构分析带来了巨大挑战. 例如, 截止到2012年12月, Facebook在全球的用户节点总量已经超过10

亿, Twitter的节点数也超过了5亿. 现有的社团分析算法的复杂度多为 $O(n^2)$ 或 $O(n^3)$ , 只能用于分析中等规模(十万或百万节点级别)的网络. 现有的解决方法是寻找快速算法<sup>[5]</sup>, 或者采用新的图计算框架, 例如google的Pregel<sup>[6]</sup>. 但是对于超大规模复杂网络, 当前仍然没有高效的解决方法.

社团结构是社会网络中普遍存在的结构之一. 人们通过社会关联形成若干社会团体, 社团间

的连边比较稀疏,社团内部的连边比较紧密. 社团发现的目的是找出网络中密切交互的模块化子网,对更准确地理解复杂系统的组织原则、拓扑结构与动力学特性,具有十分重要的意义.

目前已有很多针对复杂网络的社团发现算法. GN 算法<sup>[7]</sup>是一种自顶向下的分裂方法,其基本思想是不断地从网络中移除介数最大的边,直到网络中每个节点就是一个退化的社团为止. 在 GN 算法的基础上,Newman 等<sup>[8]</sup>提出了基于模块度优化的快速算法,克服了传统的 GN 算法只能处理中等规模网络的缺点,可以用于分析大型复杂网络. Ahn 等<sup>[9]</sup>提出了一种从边的角度出发,同时考虑社团结构的层次性和重叠性的单链接层次聚类算法,把一个社团看作一组有密切相关的链接集,而不是假设一个社团是一个节点之间有许多链接的节点集.

上述社团结构方法,都有一个共同点,即需要知道整个社团的网络拓扑结构. 然而,有时只关心网络中的某个特定社团,例如,个性化营销希望找到与产品相符合的潜在用户群;舆情分析希望了解特定人群对某一话题的看法. 在这种应用场景下,只需要找到关注的社团,而没必要给出网络中全部的社团结构.

本文提出了一种用于超大规模社交网络的社团抽取方法. 与传统的社团发现不同,该方法是寻找网络中某个特定社团,而不是全部的社团结构. 因此,不要求已知网络的全部拓扑结构,而是边爬行获取网络结构,边抽取社团,这会显著地提高算法在超大规模网络上的执行效率. 算法的一个基本动机是,社团内部的连接比较紧密,而社团之间的连接相对稀疏. 那么,从若干少量已知的社团内部节点出发,进行启发式的搜索,将连接紧密的邻居加入社团内部,将连接稀疏的邻居排除在社团外,这样就可以获得某个期望的特定社团的范围和结构.

## 1 社团抽取算法

考虑网络  $G$ , 社团抽取的目标是找出网络  $G$  中一个已知规模的特定社团, 假设网络全局拓扑的信息是未知的(即网络中节点和边的数量是未知的).

算法的基本思路是,从某些节点出发,沿着边扩展子图,从而不断扩大社团. 在扩展的过程中,根据社团的性质,同时进行节点的筛选. 因此,需要解决几个问题:1) 起始节点的选取;2) 社团评

价的标准;3) 节点的评价. 社会网络上的节点,一般具有若干基本属性,可以用于标识其所属社团. 如某用户的学校为“东北大学”,可以判定其属于东北大学社团. 但是,一般只有少数节点具有明确的社团标签. 本文视这些节点为核心种子节点,从这些节点进行扩展,找到属于该社团的其他节点.

对于社团的期望规模,根据经验选择社团的实际大小,例如在判定“东北大学”社团的过程中,选择东北大学的总人数作为社团的期望规模. 对于社团的评价,一种通用的标准是社团内部关联紧密,外部关联稀疏. 因此,通过社团内外连边的比值,可以有效地度量社团的“内聚”程度. 如果在社团中加入一个节点使得社团的“内聚”程度增加,那么该节点属于社团的可能性较大,反之较小. 因此,可以将社团抽取的过程视为从若干已知核心点出发的搜索过程,通过扩展当前社团中的节点,即将当前社团中节点的邻居节点作为被筛选节点,将属于社团内部的节点加入社团,筛掉不属于社团的节点,直到得到期望的社团规模为止.

根据以上思路,本文将社团抽取视为在隐含网络  $G$  上的搜索过程,目标是使搜索图的适应度函数最大化. 一个子图的社团性质是否明显,取决于其内部关联是否大于外部关联. 因此,定义评价子图的适应度函数为

$$f_G = \frac{|e_G|}{|e_G| + |e'_G|},$$

其中:  $|e_G|$  为子图  $G$  边的个数;  $|e'_G|$  为  $G$  中节点指向外部节点的边的个数. 显然,如果  $G$  的社团性质较明显,  $f$  的值会较大,反之较小.

另一个需要考虑的问题是如何选取初始节点. 对于一个节点来说,当其连边大多都在社团  $G$  内部时,其属于社团的概率较大;当节点大部分的连边都指向社团外部节点时,其有很大可能不属于社团. 因此,本文定义节点  $n$  的评价函数  $h$  为

$$h_n = \frac{k_n^G}{k_n}.$$

其中:  $k_n^G$  为节点  $n$  在子图  $G$  中的度;  $k_n$  为节点  $n$  在整个网络中的度.

基于以上两个函数,社团抽取算法主要包括两个过程:一个是寻找最有可能属于社团的节点,扩展该节点从而扩大子图;另一个是判断是否将该节点加入子图,这是通过判定加入该节点是否能增加子图适应度函数来决定的. 重复上述步骤直到子图达到指定规模  $N$  为止,具体算法步骤如下:

1) 令初始节点集为  $S$ , 初始社团  $G = \{S\}$ ; 令社团期望规模为  $N$ ;

2) While  $\text{Size}(G) < N$  do;

3) 计算  $f_G$  及  $G$  中所有节点的  $h$  值;

4) 令  $G$  中  $h$  值最大的节点为  $n$ ;

5) 令节点  $n$  的不在  $G$  中的邻居节点集为  $M$ ;

6) selected = true;

7) While selected and  $\text{Size}(M) > 0$  do;

8) 令  $M$  中  $h$  值最大节点为  $m$ ,  $M = M - \{m\}$ ;

9) 令  $G' = G + \{m\}$ , 计算  $f_{G'}$ ;

10) If  $f_{G'} > f_G$ ;

11) then  $G = G'$ , selected = false;

12) End If;

13) End while;

14) End while.

算法中初始节点集可以由节点的社会标签确定, 社团期望规模  $N$  一般通过经验判定. 算法首先寻找社团  $G$  中最可能扩展的节点  $n$ , 然后选择其中评价函数  $h$  值最大的节点加入社团  $G$ , 重复上述步骤直到达到指定规模为止.

## 2 实验与分析

为了验证社团抽取算法的有效性, 本文选用具有社团结构的仿真网络<sup>[10]</sup>进行实验. 每一个仿真网络包含 4 个子社团, 每个社团的节点数均为  $n/4$ . 对于网络中任意节点对, 以一定的概率来放置一条连边. 如果节点对在同一个社团内, 则连边概率为  $p_{\text{in}}$ , 否则为  $p_{\text{out}}$ . 该方法可以生成具有明显社团结构的大规模网络, 目前已成为社团发现算法的标准测试集<sup>[11]</sup>.

通常, 具有明显社团特性的网络社团内部连边紧密, 社团之间连边稀疏. 为了保证仿真网络具有明显社团特性, 本文取  $p_{\text{in}} > p_{\text{out}}$ . 设一个节点的社团内连边数  $l_{\text{in}} = n \cdot p_{\text{in}}/4$ , 社团间连边数为  $l_{\text{out}} = 3n \cdot p_{\text{out}}/4$ , 则有网络平均度 (average degree) 为  $d_{\text{ave}} = l_{\text{in}} + l_{\text{out}}$ . 图 1 给出了 3 个平均度为 12 的仿真网络. 其中图 1a 中  $l_{\text{in}} = 11$ , 图 1b 中  $l_{\text{in}} = 10$ , 图 1c 中  $l_{\text{in}} = 9$ , 对应模块度分别为 0.667, 0.577, 0.502. 模块度越大, 网络具有的社团结构越明显. 本文所采用的仿真网络数据如表 1 所示.

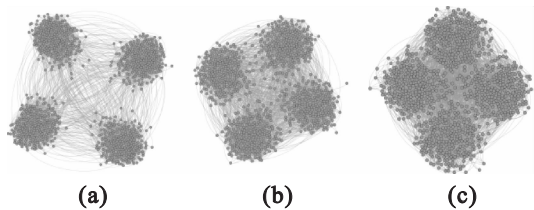


图 1 四社团随机网络的拓扑示意图  
Fig. 1 The topological diagram of ER-network with four communities  
(a)— $l_{\text{in}} = 11$ ; (b)— $l_{\text{in}} = 10$ ; (c)— $l_{\text{in}} = 9$ .

表 1 四社团随机网络  
Table 1 ER-network with four communities

名称	节点	边	模块度	$p_{\text{in}}$	$p_{\text{out}}$
Net1	1 000	5 945	0.667	0.022 00	0.000 670
Net2	1 000	5 944	0.577	0.020 30	0.001 300
Net3	1 000	5 995	0.502	0.018 00	0.002 000
Net4	1 000	7 990	0.495	0.024 00	0.002 700
Net5	1 000	10 131	0.518	0.030 00	0.003 300
Net6	1 000	11 999	0.499	0.036 00	0.004 000
Net7	1 000	13 984	0.500	0.042 00	0.004 700
Net8	1 000	6 201	0.413	0.016 00	0.002 700
Net9	20 000	51 946	0.601	0.000 40	0.000 040
Net10	40 000	104 209	0.630	0.000 20	0.000 020
Net11	60 000	151 276	0.599	0.000 13	0.000 013
Net12	80 000	207 524	0.605	0.000 10	0.000 010

首先, 分析算法准确度的影响因素. 将社团抽取准确率定义为  $A = |N_{\text{extract}} \cup N|/|N|$ , 其中  $N_{\text{extract}}$  为社团抽取算法得到的点集,  $N$  为目标社团的点集. 选取模块度相近但平均度不同的 5 个网络 Net5 ~ Net9, 分别从每个网络中随机选取 100 组个数为 1 ~ 9 的初始节点集合, 取 100 组社团抽取结果的均值, 结果如图 2 所示. 结果表明, 初始节点的个数对社团抽取准确率的影响不大, 这是因为本文所提出的社团抽取方法是从某些节点开始, 对网络进行逐步搜索的过程. 同时, 对于不同平均度的网络, 平均度越大, 其抽取准确率越高. 原因是对于平均度较大的网络, 每一次进行网络搜索时, 都可以有更多的社团内节点被找到.

其次, 本文分析了具有不同社团结构的网络上社团抽取算法准确率的变化. 图 3 给出具有不同模块度的网络 Net1, Net2, Net3, Net8 上社团抽取算法的结果. 当模块度较高, 即社团结构较明显时, 社团抽取算法可以准确地得到网络的目标社团; 当模块度较低时, 社团抽取准确率仍然可以达到 75% 以上, 验证了该算法的可行性. 与基于模块度优化的社团发现算法<sup>[8]</sup>相比, 在具有明显社团结构的网络上, 社团抽取算法达到了与社团发

现相当的准确率。

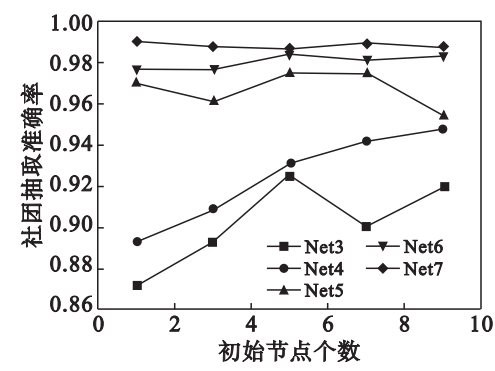


图 2 算法准确率的影响因素分析  
Fig. 2 Analysis for the influence factors of algorithm accuracy

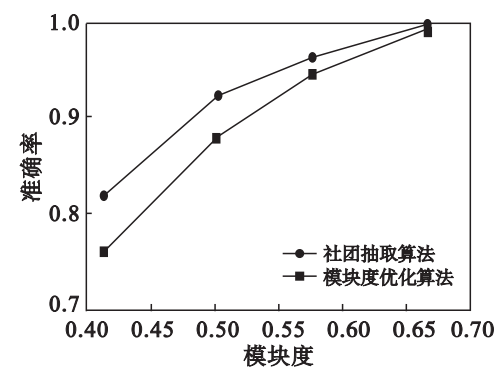


图 3 社团抽取算法与模块度优化算法的准确率  
Fig. 3 The accuracy of community extraction algorithm and modularity optimal algorithm

最后,为了验证社团抽取算法的效率,选用经典的 Newman 模块度优化算法<sup>[8]</sup>进行对比实验,图 4 给出了两种方法在 Net9 ~ Net12 上运行时间的对比结果.可以看出,在运行时间方面,社团抽取算法明显优于模块度优化算法.原因是社团发现算法考虑全局拓扑,得到整个网络的社团划分结果;而社团抽取算法考虑局部社团,通过节点扩展抽取网络局部,因此其效率较高。

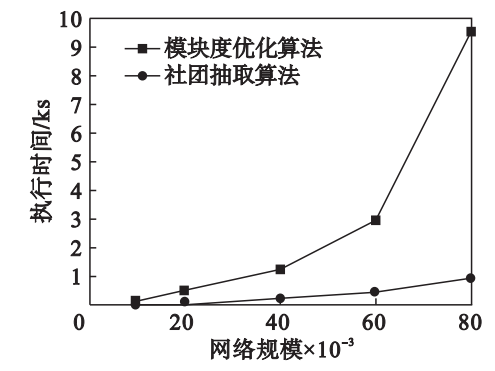


图 4 社团抽取算法与模块度优化算法的效率比较  
Fig. 4 Comparison between community extraction algorithm and modularity optimal algorithm

### 3 结 语

本文分析了社交网络分析的社团发现需求,提出了社团抽取问题,并为大规模网络上的社团抽取问题提出了一个高效算法。

社团作为社交网络的基本结构特征,在网络分析领域具有重要的意义.社交网络上的个人行为与其所在社团具有紧密联系.因此,给出某节点所处的社团结构,具有重要意义.此外,社团抽取算法也可用于社团发现问题.如果已知网络中各社团的初始节点,那么就从网络的各个社团同时进行社团抽取。

下一步的工作将探索社团抽取算法在社团发现问题上的应用,同时进一步提高社团抽取算法的性能和准确度。

#### 参考文献:

[1] 解肖,汪小帆. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学, 2005 (3): 1-12.  
(Xie Zhou, Wang Xiao-fan. An overview of algorithms for analyzing community structure in complex networks [J]. *Complex System and Complexity Science*, 2005 (3): 1-12.)

[2] Sarukkai R R. Link prediction and path analysis using Markov chains[J]. *Computer Networks*, 2000, 33 (1): 377-386.

[3] Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ networks[J]. *Nature*, 1998, 393 (6684): 440-442.

[4] Barabási A L, Albert R. Emergence of scaling in random networks[J]. *Science*, 1999, 286 (5439): 509-512.

[5] Newman M E J. Fast algorithm for detecting community structure in networks[J]. *Physical Review E*, 2004, 69 (6): 066133.

[6] Malewicz G, Austern M H, Bik A J C, et al. Pregel: a system for large-scale graph processing [C]//Proceedings of the 2010 ACM SIGMOD. Indianapolis, 2010: 135-146.

[7] Girvan M, Newman M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences*, 2002, 99 (12): 7821-7826.

[8] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69 (2): 026113.

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

[10] Newman M E J. Modularity and community structure in networks[J]. *Proceedings of the National Academy of Sciences*, 2006, 103 (23): 8577-8582.

[11] Fortunato S. Community detection in graphs[J]. *Physics Reports*, 2010, 486 (3): 75-174.