计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(1):220−224 [doi: 10.15888/j.cnki.csa.007217]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于谱顶层分割的网络社区层次抽取方法①

熊　英

(江门开放大学 网络信息中心, 江门 529000)
通讯作者: 熊　英, E-mail: 18576781225@163.com

摘　要: 针对网络层次中不同尺度上社区内连接密度的异构性, 提出了基于谱顶层分割的网络社区层次抽取方法. 首先, 将网络的谱顶层分割定义为某个子网络的二分, 给出了顶层分割的期望划分; 然后, 引入队列的思想计算社区连接密度, 自顶向下逐层分解给定网络, 并提出了社区层次抽取算法; 最后, 通过实验表明: 所提出的方法比同步法和多尺度法在随机层次网络测试的性能更加优越, 为社区教育和大数据行为特征识别提供了相关技术基础支持.

关键词: 谱顶层分割; 网络社区; 层次抽取; 期望划分

引用格式: 熊英.基于谱顶层分割的网络社区层次抽取方法.计算机系统应用,2020,29(1):220–224. http://www.c-s-a.org.cn/1003-3254/7217.html

Extraction of Network Community Hierarchies Based on Spectrum Top-Segmentation

XIONG Ying

(Center for Network and Information, Jiangmen Open University, Jiangmen 529000, China)

Abstract: The network community hierarchies are defined by heterogeneous of different scales of link density in essence, it is necessary for network community to detect the dynamically changing information during hierarchies division. In view of this, a method of extraction of network community hierarchies based on spectrum top-segmentation is proposed. Firstly, the spectrum top-segmentation is defined as a dichotomy of subnetwork that no any top-level community can cross two parts, and an expected division top-level segmentation is presented. Then, the queue and link-density are introduced to decompose network, and an algorithm of network community levels extraction is presented. The simulation result shows that the performance of proposed method is better than that of synchronization and multi-scale in stochastic hierarchical networks, and the method is applicated in Email real-world network effectively.

Key words: spectrum top-segmentation; network community; extraction of hierarchies; expected division

网络层次设计进行网络社区检测的有效方式, 在社交网络、教育社区数据挖掘和犯罪特征识别等领域得到了广泛的应用. 网络层次本质上是由不同尺度上社区内连接密度的异构性定义[1], 若社区内的连接密度大于社区之间的连接密度, 则网络社区组织结构具有层次性. 将网络划分成若干个连接相对紧密的社区, 每个社区又可能包含若干个连接相对更紧密的子社区[2–4]. 例如: 存在一个具有 40 个节点的二层网络组织, 设网络的一个社区 C_i 内部连接密度为 p_1, 它到网络其余部分的密度为 p_0, 则有 $p_1>p_0$; 如果社区 C_i 由许多小的社区 C_{ik} 构成, C_{ik} 的连接密度为 p_2, 则有 $p_2>p_1$. 如何抽取网络层次社区结构是当前的一个重要研究热点[4].

抽取网络层次社区结构的主要方法是基于层次聚类方法[5,6], 其思想是采用谱顶层分割的算法将 k 最近邻图划分成大量较小的子社区, 并用相似的子社区反复地合并操作; 文献[7]利用谱顶层分割的方法, 提出了

① 基金项目: 广东远程开放教育科研基金 (YJ1613); 公安部技术研究计划 (2015JSYJC40)
Foundation item: Research Fund for Distant Open Education of Guangdong Province (YJ1613); Technology Research Program of the Ministry of Public Security (2015JSYJC40)
收稿时间: 2019-06-01; 修改时间: 2019-06-28; 采用时间: 2019-07-05; csa 在线出版时间: 2019-12-27

一种基于马尔可夫链的蒙特卡罗抽样技术预测丢失连接, 用于推导复杂网络的层次, 但由于其算法的抽取的空间较大, 容易导致数据的维度问题; 文献[8]提出了多层次节点相似的网络社区发现方法, 在改进节点相似度和团体连接紧密度的基础上构建社区发现模型, 从而更加准确地找到社区成员, 但这种方法未考虑网络的层次的异构特性, 且不能很好地适用于大型网络; 文献[9,10]提出了多尺度方法揭示不同尺度下的社区结构, 该方法对异构网络的检测具有较好的效果, 但未考虑社区内外连接密度的动态变化和社区间的异构性, 使该方法不能适用于动态演化的网络社区.

基于以上问题, 提出了一种基于谱顶层分割的网络社区层次抽取方法, 该方法将网络的顶层分割定义为某个子网络的二分使得没有任意一个顶层社区横跨两部分, 并给出了顶层分割的期望划分; 引入队列的思想计算社区连接密度, 自顶向下逐层分解给定网络, 提出社区层次抽取算法; 通过实验验证所提出方法的科学性和合理性.

1 谱顶层分割

1.1 顶层分割

存在一个具有内部结构的网络 N, 所有构成它的第一层的社区称为关于该网络 N 的顶层社区, 所有顶层社区的集合称为 N 的顶层分解, 而使 N 的顶层分解中所有节点只存在于一个分组中的方法则为 N 的谱顶层分割, 进而形成一个网络的二分, 使没有任何一个顶层分解跨越得到两个组. 如图 1 所示, 在具有两层网络组织结构中, P_1 和 P_2 为网络 N 的顶层社区, 社区 C_1 和 C_2 为 P_1 的顶层社区, 则 P_1-P_2 是网络的顶层分割, C_1-C_2 是 P_1 的顶层分割.

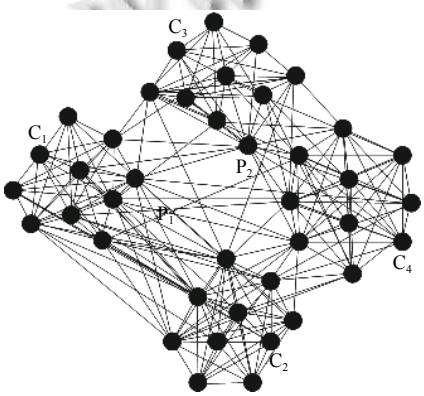

图 1 具有网络组织结构的顶层分割

谱顶层分割可以期望找到一个顶层分割或近似顶层分割, 由于每次分裂总是试图找到模块度最大或者增量最大的二分, 如果考虑更多的特征向量, 找到一个顶层分割的机会将进一步增强. 因此, 为使顶层分割得到较高的模块度, 需计算期望最高划分, 从而选择连接密度最小的返回顶层分割.

1.2 谱顶层期望划分

设存在具有 3 个社区 C_1、C_2 和 C_3 的随机网络, 假设连接概率的如表 1 所示且 $p_0 < p_1 < p_2$.

表 1 社区内和社区间的连接概率配置

连接概率	C_1	C_2	C_3
C_1	p_2	p_1	p_0
C_2	p_1	p_2	p_0
C_3	p_0	p_0	p_2

谱顶层分割设置了一个两层次网络, 即由 C_1 和 C_2 构成的社区和 C_3 形成了网络的第一层, 而 C_1 和 C_2 形成了第二层. 对于该网络, 存在 3 个二分, 即 π_1: (C_1, C_2)- (C_3), π_2: (C_1, C_3)- (C_2) 和 π_3: (C_1) - (C_2, C_3). 为进一步分析, 将 3 个连接概率参数设置为: p_0=0.1, p_n=p_0+k_n×r_n. 其中, p_0 作为社区与社区之间划分的初始连接概率, p_n 在 p_0 的计算基础上设置连接概率, 并以 k_n 和 r_n 取值[0, 1]中的随机数, 这里统一取 k_n=0.5 且 r_n=0.5, 以保持网络层次的稳定性, 以免在出现连接状态层次不统一问题. 在给定一个网络 N 的前提下, 设 Q 为期望划分值, 对两个层次的期望则定义为:

$$\begin{aligned} Q(p_1) &= m_1 \times m_2 \times p_0 - \frac{(k_1+k_2)^2+k_3^2}{2M} \\ Q(p_2) &= m_1 \times m_3 \times p_1 - \frac{(k_1+k_3)^2+k_2^2}{2M} \\ Q(p_2) &= m_1 \times m_3 \times p_1 - \frac{(k_1+k_3)^2+k_2^2}{2M} \end{aligned} \quad (1)$$

式 (1) 中, m_i 和 k_i 分别是社区 i 的尺寸和总度. 通过计算期望划分值可以将连接密度最小社区作为顶层分割, 进而对网络层次进行抽取.

2 网络层次社区抽取

2.1 连接密度计算

为获取连接密度最小的社区, 引入队列思想对网络 N 自顶向下逐层分解, 采用 q_curr 表示存储网络第 h 层的有待分析社区, q_work 表示当前工作队列, q_next 表示存储下一层社区. 当初始化时, q_curr 存储

包含网络中所有节点的唯一组, 从 *q_curr* 的队头中取出第一组并将其存入网络 N, 然后将其分解成两组网络数据 N_1 和 N_2, 并计算其连接密度:

$$\delta_0^* = \frac{|E(N_1,N_2)|}{|N_1|\times|N_2|} \tag{2}$$

式 (2) 中, $E(N_1, N_2)$ 表示网络 N_1 和 N_2 之间的边数, 而计算连接密度是关于 N 的顶层社区间的连接密度的一个估计.

当 N_1 和 N_2 进入工作队列 *q_work* 时, 都可能包含几个谱顶层分割. 如果当前 *q_work* 非空, 则可以从中取出第一组网络数据 N_1 并对它进行分解; 如果不可分, 则 N_1 被认为是一个顶层社区, 否则它被划分为两个更小的组 N_{11} 和 N_{12}, 并实时检查它们之间的连接密度δ_1, 计算是否超过谱顶层分割间连接密度的估计值δ_0^*. 如果计算结果大于估计值δ_0^*, 则此分割不属于 *h* 层分解, 而属于 *h*+1 层分解. 由此推知, N_1 是关于网络 N 的一个谱顶层分割, 则 N_1 进入 *q_next* 准备下一层分解, 否则, N_{11} 和 N_{12} 都可能是一个谱顶层分割或者几个谱顶层分割. 因此, 为进一步取代 N_1, 需进入 *q_work* 调整估计值δ_0^*, 调整估计值方法的思路是将原有的估计值在顶层分割次数的基础上对下一层网络分割后连接密度的预判, 可以实时检测顶层分割后的每一层网络的连接概况, 从而提高下一层分割的精准性. 其计算方法如下:

$$\delta_0^* = \frac{\delta_0^*\times n+\delta_1}{n+1} \tag{3}$$

式 (3) 中, *n* 表示网络 N 从 *q_curr* 中取出后, 执行顶层分割的次数, δ_0^*表示新的值. 当 *q_work* 为空时, 表示从 *q_curr* 中取出的第一个组 N_1 已经完全分解为它的顶层社区; 而从 *q_curr* 中取出下一个组直到 *q_curr* 为空, 将 *q_next* 中的组移到 *q_curr*, 并进行 *h*+1 层分析, 重复上述过程得到调整后的估计值.

2.2 算法实现

抽取网络层次由算法 1 实现, 在该算法中函数 subspaceMethod (G, N_1, N_2, δ_1, *d*) 为搜索一个顶层分割, 将 N 分解为两个组 N_1 和 G_2, δ_1为两部分间的连接密度, *d* 指示了 N 是来自于 *q_curr*(*d*=0) 还是来自于 *q_work* (*d*>0). 符号“←”和“→”对应队列的两个基本操作, 即“从队头取元素”和“存储数据到队尾”, 而 $q_a \Rightarrow q_b$ 表示将队列 q_a 的所有数据移到另一个队列 q_b, 算法实现如算法 1.

算法 1. 层次抽取算法 (伪代码)

输入: *q_curr*, *q_work*, *q_next*
输出: 新的层次社区

```
1) initialize q_curr, q_work, q_next
2) N→q_curr, h=0 //N 表示整个网络
3) while q_curr is not empty do
4)    while q_curr is not empty do
5)       N←q_curr, d=0
6)          v=subspaceMethod (N, N1, N2, δ1, d)
7)          if v>0 then //N 未被分解
8)             N1→q_work, N2→q_work, δ*=δ1
9)          end if
10)          while q_work is not empty do
11) N←q_work and v=subspaceMethod (N, N1, N2, δ1, d)
12)             if v<=0 then N→q_next
13)             else if δ*<=β×δ* //β 确定一个划分是否属于下一个层次
14)             N1→q_work, N2→q_work update δ*
15)             d=d+1, compute(Q)//计算期望划分值
16)             else N→q_next //N 为最顶层社区
17)             end if
18)          end if
19)       end while
20)       end while
21)    h=h+1
22)    if q_next is not empty then output the communities at h level from
qnext //返回新的层次社区
23)    q_next⇒q_curr
24)    end if
25)    end while
```

由于对网络分割的顺序取决于集合之间边的密度, 因此上述算法可看作为一种有序的谱方法, 首先搜索一个顶层分解并计算网络的特征向量, 判断网络 N 是否被分解状态, 然后进入队列进行顶层分割; 在$\delta^* \Leftarrow \beta\times\delta^*$中, 参数 β 的选择确定一个划分是否属于下一层次, 本文设置 β=1.6, 为实验测试设定一个稳定值, 以解决该值太小导致连接密度的同质性较高以及异质性较强的问题.

3 实验分析

仿真实验在 gephi 软件平台上验证本文方法的有效性, 数据来源于 Rovirai Virgili[11]大学 Email 数据中的教师联系网络, 该网络由 7 个主要学院的教师共 640 个节点构成的三层次网络, 自顶向下分别为学院、系和研究组, 网络第一层由 4 个 160 节点的社区构成, 每个类似的社区在第二层分解为 4 个 40 节点的小社区构成, 而每个小社区又在第三层包含了 4 个 10 节点的更小社区, 网络的边按照各层的不同的连接密度生

成的, 满足 $p_0<p_1<p_2<p_3$, 实验中设定 p_0=0.005, p_1=0.2, p_2=0.4, p_3=0.8. 通过比较同步法和多尺度法[12]测试层次随机网络, 同时在 Email 现实网络上测试本文方法的性能.

3.1 同质层次随机网络性能

对于数据源中学院、系、研究组, 其每个层次具有相似的层次分布结构, 在满足所有层次比例$\mu=k<h/k_l$[13] 的情形下, 它表示了层次的相对凝聚性, 若μ越小, 则凝聚性越强, 反之则越弱. 为说明本文方法的鲁棒性, 通过调整μ得到不同凝聚强度进行测试, 并采用规范化互信息[14]度量被划分的分割和已知的分割, 取 $0\le\mu\le1$ 表示无关匹配到完全匹配的状态. 如图 2 所示, 本文方法可以精确地识别 3 个层次上的社区, 其规范化互信息的匹配程度均在 90% 以上.

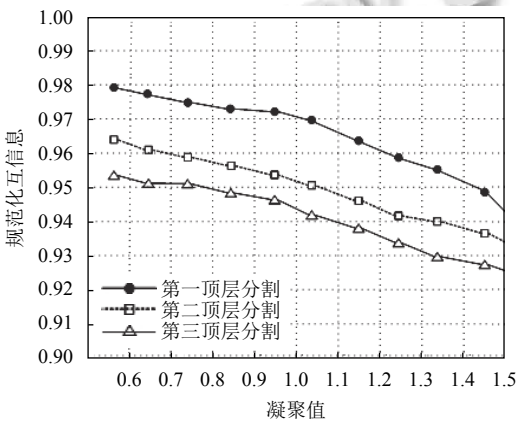

图 2 本文方法在层次随机网络的凝聚性精度

由图 2 可知, 对于具有 3 个同质层的随机网络上的性能, 每一个点是在 10 个实例网络上的平均, 一个稳定的分割可能对应于某一层次的划分.

由图 3 可知, 通过本文方法与多尺度法和同步法进行比较, 说明了多尺度法和同步法对 3 个层次分割的精度. 在最强凝聚情形 0.785 和 0.885 下, 虽然同步法在 3 个层次上都具有相当高的精度, 但它的精度随着凝聚强度的下降而快速下降; 多尺度法在精度和稳定性方面则更接近本文方法, 但其规范化互信息仍然低于本文方法且不适应于动态演化的网络.

3.2 异质层次随机网络性能

由于在更小的网络社区中, 其每一层所具有的社区的尺寸是完全不同的, 因此需要在通过本文方法验证是否适用于尺寸异质的情形. 由图 4 可知, 通过实验, 本文方法能够精确地抽取它的层次组织, 其中粗线表示第一层划分, 细线表示第二层划分, 在第二层中社区内的不同灰度节点表示第三层的网络组织, 可以显示层次抽取后的结果.

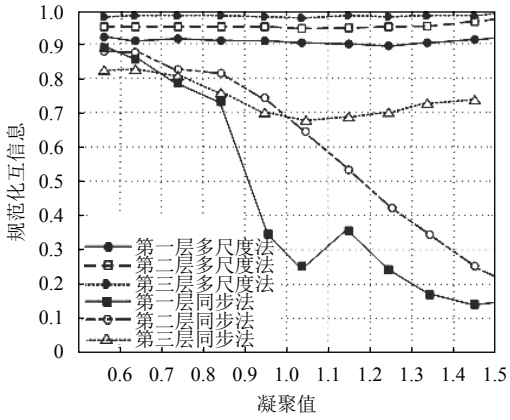

图 3 层次随机网络的凝聚性比较

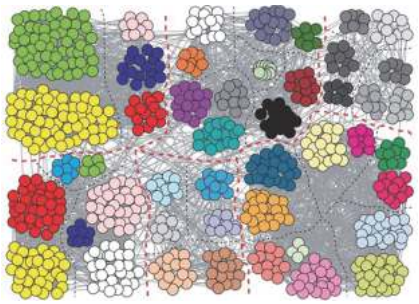

图 4 异质随机网络的社区层次抽取结果

由图 5 可知, 谱顶层划分与真实的划分完全一致, 第二层和第三层精确地逼近真实的划分, 按照互信息精度分别为 0.93 和 0.81. 同时, 将本文方法与同步法和多尺度法做了比较, 在第一层社区网络上, 本文方法比同步法和多尺度法在规范化互信息性能上分别高 0.05 和 0.15, 计算精度损失较小, 这是由于在第一层社区网络计算连接密度的数据量较大, 而第二层次和第三层次社区网络上, 本文方法比同步法和多尺度法在规范化互信息性能上的差距逐渐减少, 这是由于本文方法引入了队列的思想, 在空间不足的情形下通过不断调整连接密度的估计值, 实时预判和分析下一层分割网络社区的密度, 在期望划分的驱动下释放密度最小社区的相关节点, 然后再计算连接密度的估计值, 以此类推, 得出最小社区.

4 结论与展望

针对网络的层次社区检测问题, 提出了一种基于

谱顶层分割的网络社区层次抽取方法, 选取在线真实数据源作为实验数据, 说明了该方法的科学性和合理性, 为院校社区教育和大数据行为特征识别提供了相关技术基础支持, 下一步将从大数据的角度对社区的层次进行抽取, 利用语义特征检测的方法和大数据优化相关算法, 对社区层次检测的语义性进行探索研究, 以得出更好的实验效果.

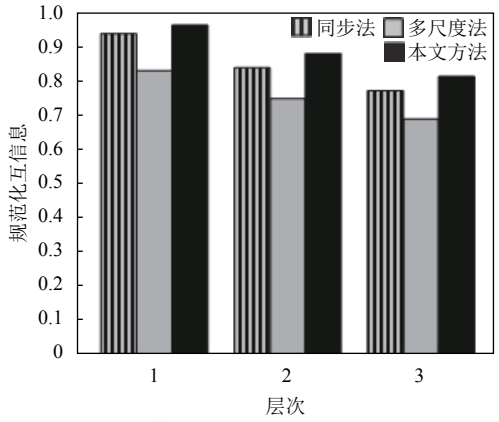

图 5 本文方法与同步法和多尺度法的规范化互信息比较

参考文献

1 Diwadkar A, Vaidya U. Synchronization in large-scale nonlinear network systems with uncertain links. Automatica, 2019, 100: 194–199. [doi: 10.1016/j.automatica.2018.06.002]

2 Arab M, Afsharchi M. Community detection in social networks using hybrid merging of sub-communities. Journal of Network and Computer Applications, 2014, 40: 73–84. [doi: 10.1016/j.jnca.2013.08.008]

3 赵建军, 汪清, 由磊, 等. 基于信息传递和峰值聚类的自适应社区发现算法. 重庆大学学报, 2018, 41(11): 76–83.

4 Li XM, Xu GQ, Tang MH. Community detection for multi-layer social network based on local random walk. Journal of Visual Communication and Image Representation, 2018, 57: 91–98. [doi: 10.1016/j.jvcir.2018.10.003]

5 Mondal SA. An improved approximation algorithm for hierarchical clustering. Pattern Recognition Letters, 2018, 104: 23–28. [doi: 10.1016/j.patrec.2018.01.015]

6 Everitt B. Cluster analysis. Quality and Quantity, 1980, 14(1): 75–100. [doi: 10.1007/BF00154794]

7 Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. Nature, 2008, 453(7191): 98–101. [doi: 10.1038/nature06830]

8 张虎, 吴永科, 杨陟卓, 等. 基于多层节点相似度的社区发现方法. 计算机科学, 2018, 45(1): 216–222. [doi: 10.11896/j.issn.1002-137X.2018.01.038]

9 Liang X, Du JB. Concurrent multi-scale and multi-material topological optimization of vibro-acoustic structures. Computer Methods in Applied Mechanics and Engineering, 2019, 349: 117–148. [doi: 10.1016/j.cma.2019.02.010]

10 Arenas A, Díaz-Guilera A, Pérez-Vicente CJ. Synchronization reveals topological scales in complex networks. Physical Review Letters, 2006, 96(11): 114102. [doi: 10.1103/PhysRevLett.96.114102]

11 http://www.urv.cat/en/about/directory/institutional/. [2017-07-01].

12 Lacerda J, Freitas C, Macau E. Multistable remote synchronization in a star-like network of non-identical oscillators. Applied Mathematical Modelling, 2019, 69: 453–465. [doi: 10.1016/j.apm.2018.12.026]

13 Arjmand D, Engblom S, Kreiss G. Temporal upscaling in micromagnetism via heterogeneous multiscale methods. Journal of Computational and Applied Mathematics, 2019, 345: 99–113. [doi: 10.1016/j.cam.2018.05.059]

14 王益文. 复杂网络节点影响力模型及其应用[博士学位论文]. 杭州: 浙江大学, 2015.