doi:10.3772/j.issn.1002-0470.2021.03.005

基于 Louvain 算法的作者合著网络社区划分研究①

褚叶祺②* 丁佳骏**

(*浙江工业大学图书馆 杭州 310014)

(**杭州电子科技大学计算机学院 复杂系统建模与仿真教育部重点实验室 杭州 310018)

摘 要 对作者合著网络进行社区划分有助于挖掘科研人员的合作和交流模式。采用 Louvain 算法将 C-DBLP 作者发文合作关系公开数据集进行了社区划分,并采用模块度对划分结果进行评估。结果表明,Louvain 算法能够快速高效地处理具有数千个节点的网络,与 LED 算法和 GN 算法相比,能更有效地进行社区划分。研究结果揭示了各个学科不同的合作交流模式,有助于挖掘潜在的合作团体,为学科合作研究提供帮助。

关键词 作者合著网络;社区划分;Louvain

0 引 言

随着科学技术不断发展,科学研究难度也日益增大,一个课题往往需要多位科研人员合作解决,因此研究者们开始相互合作来共同解决复杂问题。作者合著是指学者们共同署名发表科研论文,也是科研合作最直接的体现形式。基于合著关系构建的网络称为合著网络,该网络以作者为节点,以合著关系为链接[1]。

社区划分是指基于网络的属性,关系网络中的每个节点被划分为具有特殊意义的不同社区,其中社区内的点比社区的外部联系更紧密[2]。对作者合著网络进行社区划分,可以挖掘出具有相似研究兴趣的研究社团。区别于传统的学术团体和研究团队,利用复杂网络理论划分的研究社团更能体现研究者的合作和交流模式,有利于挖掘科研人员的研究兴趣和学术机构之间、科研人员之间的相互联系,对促进学科领域科研合作的发展具有重要意义。

1 相关研究

目前,网络社区划分的算法有多种,主要分为两类,一种是分离法[3],例如 Girvan 和 Newman[4] 提出的 GN 算法,该算法基于边介数并通过不断地从网络中移除介数最大的边来划分社区,但是该算法的缺点是计算速度缓慢而且无法预估最终能够分裂成几个社区。另一种通过聚合联系紧密的点成为一个社区的方法称为聚合法,可以采取优化相关变量的函数来进行聚合[5],该方法较分离法效率较高,也吸引了大量学者进行相关研究[6-7]。Newman[8] 在 2004 年提出了一种快速算法,该算法的目标是不断增大模块度,是一种基于贪婪思想的凝聚算法。该算法虽然提高了社区划分的效率,但遇到大型复杂网络时该算法耗时较长。Ma 等人[9] 提出了一种有效的重叠社区发现算法并称之为 LED(loop edges delete)算法,该算法基于结构聚类,将顶点之间的结构相似度转化为网络权值。谷瑞军等人[10] 提出基于 SALTON 方法构建作者合著网络,使用加权的链接聚类算法实现社区聚类划分,该方法能有效发现部分重叠的合著社区。苗蕊等人[11] 提出了社区-作者-主题模型来发现科研人员之间合作的隐性子社区,并且给出了基于 Gibbs 抽样的模型推断算法。李纲等人[12] 收集 WOS 中检索领域相关文献题录信

① 教育部人文社会科学研究项目(17YJA870003),浙江省社科联课题(21NDJC039YB),浙江工业大学校级科学研究基金(Z20160154)和浙江省图书馆学会学术研究课题(ZTXH2017A-08)资助项目。

② 女,1990 年生,硕士,馆员;研究方向:数据挖掘,复杂建模,深度学习;联系人,E-mail:cyq77@zjut.edu.cn

(收稿日期:2020-01-19)

息,构建作者合著网络并提出了 Jaccard 系数及余弦相似性系数的计算指标,定量研究在整个网络内和社区内研究兴趣的相似性。目前大部分社区划分的方法均是针对节点较少的小型网络,对于大规模网络不适用,且对社区划分合理性的定量评价较少。

本文采用一种 Louvain 启发式算法,该方法基于模块度最优化思想,能够快速高效处理数以亿计节点的网络,可用于大规模网络中的社区划分[3]。模块度是由 Newman 和 Girvan[13] 提出的用来衡量社区划分合理性的变量,其原理是通过对比某种方法与随机划分结果的内聚性,利用差值对划分结果进行评测,模块度值越大表明社区划分效果越好。Louvain 算法通过迭代合并邻近节点或节点群来划分社区,每一次合并过程伴随着模块度值的增加,若下一次合并的模块度增量不为正,那么此次合并后取得最优社区划分结果[13]。本文采用 C-DBLP 学者合作关系数据集作为研究对象,该数据集是 Datatang 上的公开数据集,包含了计算机、经济、法学和物理 4 个领域,每个领域都包含了 1000 名学者,该作者合著网络规模较大,故采用 Louvain 算法快速高效地进行社区划分。

2 算法介绍

Louvain 算法是一种基于模块度的社区发现算法,不同于普通的基于模块度和模块度增益,该算法对一些点多边少的图进行聚类的效果和效率都表现较好,并且能够发现层次性的社区结构,其优化目标是最大化整个社区网络的模块度[13]。

2.1 模块度介绍

模块度作为度量一个网络社区划分结果优劣的方法,具体由社区内节点的边数与随机划分下的边数之差来表示,取值范围为 $[-1/2,1)$,其定义公式如下:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i,c_j)$$

$$\delta(u,v) = \begin{cases} 1 & u == v \\ 0 & \text{其他} \end{cases} \tag{1}$$

其中,A_{ij} 表示节点 i 和节点 j 之间边的权重,若为无向图则所有边的权重为 1;$k_i = \sum_j A_{ij}$ 表示度数,即所有与节点 i 相连边的权重之和;c_i 表示 i 节点所在社区;$m = \frac{1}{2}\sum_{ij} A_{ij}$ 表示边的数目,即所有边的权重之和,模块度越高表明划分结果越好。

2.2 算法思想

Louvain 算法的思想是首先将每个节点看成一个独立的社区,然后将节点 i 依次分配到其他各个相邻节点所在的社区;若分配后模块度增量的最大值 $\max\Delta Q > 0$,则将节点 i 分配到该社区,否则不变。然后进行压缩,将同个社区的所有节点压缩成一个新节点,社区内节点间边的权重为新节点的权重,社区间边的权重为新节点之间边的权重。重复上述步骤直到整个网络的模块度都不再变化,则社区划分完成。

Louvain 算法可以产生分层的社区结构,除了对最底部社区划分计算时消耗时间较多,边和节点数随着节点按社区压缩而大大减少,当计算节点 i 分配到其邻居 j 时,模块度的变化只与节点 i 和 j 的社区有关,与其他社区无关,因此算法效率高、速度快。

合著网络数据的宏观表现同样存在层次结构,十分契合 Louvain 算法的思想。在科学研究等合著过程中,存在“作者个人-小型研究组-大型实验室-大规模研究合作”的层次包含关系,大型的合著圈往往由多个小型的合著圈组成,这一现象符合 Louvain 算法本身的特性。因此,合著网络对 Louvain 算法存在天然的适应性。此外,Louvain 算法更适合社区内高互联,社区间高稀疏的网络,合著网络恰恰满足这一特征。

Louvain 算法特殊的层次结构性质使它能够解决更广泛的问题,如对中间社区进行研究分析,对社区内部进行再聚类分析。结合前文,显然合著网络的中间层存在复杂且重要的宏观意义,因此中间社区的提取分析与再聚类分析对未来作进一步的研究具有巨大的价值。

本文采用算法原始思想的步骤对合著网络进行划分实验。在模块度计算中,由于合著网络无权的性质,A_{ij} 将恒等于 1;在中间层次的计算中,算法将使用自环边来记录大节点内部的原始节点度信息。

3 数据集

3.1 数据集的介绍

C-DBLP(http://www.cdblp.cn)是由中国人民大学网络与移动数据管理实验室开发的以作者为中心的中文文献数据库系统,能够提供权威的文献数据以及方便的查询服务。该实验室于 2012 年在数据运营平台 Datatang 上发布了 C-DBLP 作者发文合作关系数据集,包含计算机、经济、法学和物理 4 个领域[14]。每个领域都包含了 1000 名学者,通过找到该领域发文最多的学者,然后利用合作关系广度优先遍历来选取。

计算 4 个数据集具体统计指标如表 1 所示。

表 1 C-DBLP 数据集统计指标

数据集	结点数	边数	平均度	直径	聚类系数	平均路径长度
计算机	1000	2443	4.886	12	0.444	5.142
经济	1000	1435	2.870	27	0.297	11.039
法学	1000	2423	4.846	19	0.363	6.270
物理	1000	4051	8.102	13	0.585	5.230

3.2 数据集的应用及分析

Louvain 算法在网络划分中最主要的优势在于其复杂度呈线性,相较于大多数网络划分算法较低;但它的劣势在于层次结构占据了较高的储存容量。本文选用的数据集节点与边的数量级均在 1000 ~ 10 000,属于较大规模网络而并非超大容量网络。因此,在已选取数据集上使用 Louvain 算法既降低了算法复杂度,又无需担心储存空间爆满的问题,与大多数网络对比具有绝对优势。

在先前的研究中,Louvain 算法常被应用于微博网络与引文网络,但通过对比分析,发现 Louvain 算法在作者合著网络上更具优越性。首先,微博网络与引文网络数据多为有向图。因此,需要对网络中的双向情况做特殊处理,对 Louvain 算法进行改进,从而增加计算量;其次,微博网络中的节点存在多属性,大量的特征使 Louvain 算法的改进更加复杂,甚至需要一些额外的模型对特征做相关处理[15]。反观作者合著网络数据集,无权无向且节点仅存在度属性,可直接利用原始 Louvain 算法进行划分。显然,Louvain 算法用于划分合著网络存在明显的优势。

此外,Louvain 算法在计算复杂度方面的唯一弊端是对叶子节点的划分。吴祖峰等人[2]曾针对该问题改进 Louvain 算法,并表明当叶子节点数量超过 30% 时有必要进行此优化。但本文使用的数据集平均度均在 2.5 以上,这反映了网络中的叶子节点的数量较为稀少,无需使用剪枝相关的改进。这进一步说明了原始 Louvain 算法对合著网络的强适用性。

4 实证结果对比与讨论

4.1 划分结果对比

本文将 Louvain 算法与 LED 算法和 GN 算法的模块度进行了比较,4 个数据集的模块度计算结果如表 2 所示。从表中数据可以看出 Louvain 算法在 4 个数据集中均得到最好的模块度,且均高于其他两种算法。

表 2 Louvain、LED 和 GN 算法模块度比较

数据集	Louvain		LED		GN	
	社区数量	模块度	社区数量	模块度	社区数量	模块度
计算机	27	0.730	93	0.650	80	0.642
经济	33	0.869	27	0.824	29	0.832
法学	27	0.744	83	0.653	167	0.642
物理	25	0.786	59	0.741	43	0.745

图 1 ~ 图 4 为 4 个数据集采用 Louvain 算法得到的划分结果，相同颜色代表属于相同社区，不同颜色代表属于不同社区。值得注意的是，Louvain 算法划分的社区数量均在 20 ~ 30 之间，而 LED 算法划分的社区最小为 27，最大为 93，GN 算法划分的社区数量最小为 29，最大为 167，数据跨度大、稳定性较差，不利于学科之间的比较分析。

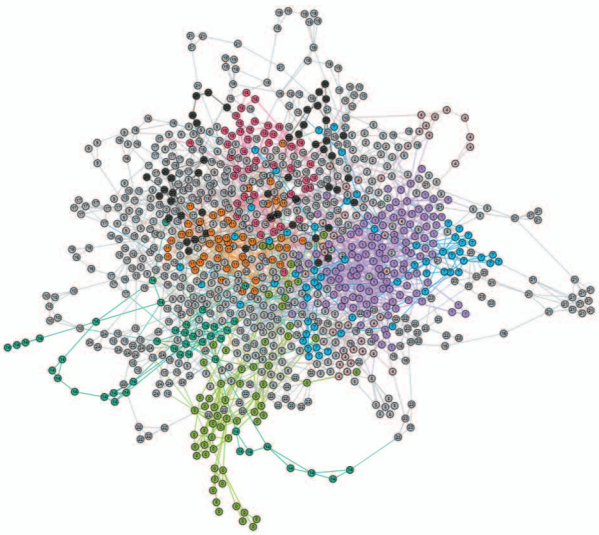

图 1　计算机数据集 Louvain 算法社区划分结果图

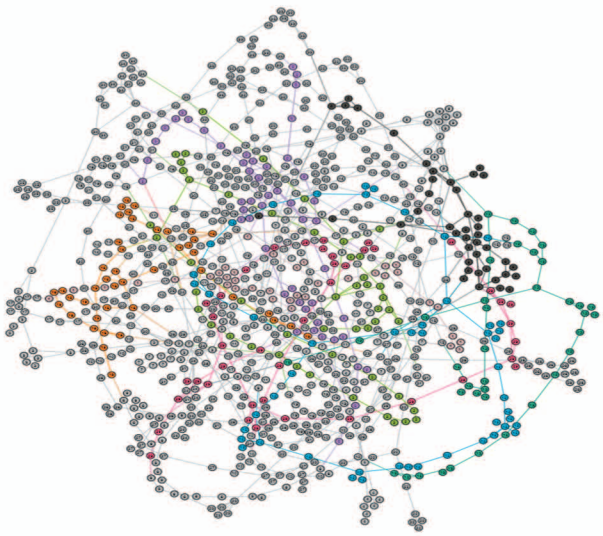

图 2　经济数据集 Louvain 算法社区划分结果图

图 3　法学数据集 Louvain 算法社区划分结果图

图 4　物理数据集 Louvain 算法社区划分结果图

对比结果表明，Louvain 算法的划分结果优于另外两种算法，且划分的社区稳定性更好。

4.2　划分结果讨论

Louvain 算法能对作者发文合作关系进行较好的划分。表 3 展示了该算法对 4 个数据集划分的统计结果（社区数量和社区所含人数），图 5 展示了各数据集社区划分结果对比。

计算机数据集经社区划分共得到 27 个社区，规模最大的社区有 1 个，由 121 位作者组成，规模最小的社区有 1 个，由 9 位作者组成。此外，9 ~ 30 位作者组成的社区共有 14 个，30 ~ 50 位作者组成的社区共有 7 个，50 ~ 100 位作者组成的社区共有 4 个。

经济数据集社区划分共得到 33 个社区，规模最大的社区有 1 个，由 60 位作者组成，规模最小的社区有 1 个，由 15 位作者组成。此外，10 ~ 20 位作者组成的社区共有5个，21 ~ 30位作者组成的社区共

表 3　C-DBLP 社区划分个数及所占比例

社区人数	计算机		经济		法学		物理	
	社区个数	所占比例	社区个数	所占比例	社区个数	所占比例	社区个数	所占比例
0 ~ 10	1	4%			3	11%		
11 ~ 20	3	11%	5	15%	2	7%	8	32%
21 ~ 30	11	41%	13	39%	5	19%	4	16%
31 ~ 40	3	11%	10	30%	7	26%	3	12%
41 ~ 50	4	15%	4	12%	5	19%	3	12%
51 ~ 60	3	11%	1	3%	2	7%	1	4%
61 ~ 70	1	4%			1	4%	2	8%
71 ~ 80							3	12%
81 ~ 90					2	7%		
90 以上	1	4%					1	4%
总计	27		33		27		25	

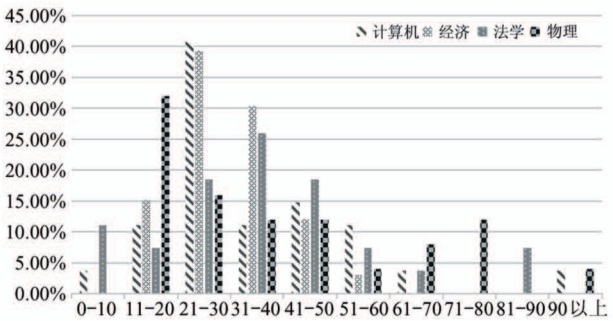

图 5　各数据集社区划分结果对比

有 13 个,占比最大为 39.4%,31 ~ 40 位作者组成的社区共有 10 个,41 ~ 60 位作者组成的社区共有 5 个。

法学数据集社区划分共得到 27 个社区,规模最大的社区由 88 位作者组成,规模最小的社区由 4 位作者组成。此外,4 ~ 10 位作者组成的社区共有 3 个,11 ~ 30 位作者组成的社区共有 7 个,31 ~ 40 位作者组成的社区共有 7 个,41 ~ 60 位作者组成的社区共有 7 个,61 ~ 90 位作者组成的社区共有 3 个。

物理数据集社区划分共得到 25 个社区,规模最大的社区由 114 位作者组成,规模最小的社区由 11 位作者组成。此外,11 ~ 20 位作者组成的社区共有 8 个,占比最大为 32%,21 ~ 30 位作者组成的社区共有 4 个,31 ~ 60 位作者组成的社区共有 7 个,61 ~ 80 位作者组成的社区共有 5 个。

对比分析 4 个数据集的社区划分结果,可以看出,计算机类和经济类均是社区人数在 21 ~ 30 人占比最大,法学类社区人数在 31 ~ 40 人占比最大,物理类社区人数 11 ~ 20 人占比最大。经济类没有 10 人以下或 60 人以上的社区,社区人数集中在 20 ~ 40 之间,大多数为中型社区,可见经济学学者研究课题更倾向于小范围之间合作,合作跨度比其他 3 个学科更小;而物理类 60 人以上社团占比 24%,仅次于 11 ~ 20 人社团的最大占比 32%,且最大社团人数达到了 114 人,由此可以看出物理学研究比其他 3 个学科更需要大群体合作来共同解决复杂问题,而且物理学者之间合作更加频繁。法学社区分布相对比较平均,不同法学类课题所需合作人数有所不同,但是 10 人以下的小社区较其他学科最多,可见法律学者在课题研究中对少数几人的团队合作也存在一定需求。计算机社区分布较其他学科更为集中,有近半数分布在 21 ~ 30 人规模大小,其他社区规模均较小。由此可明显看出,计算机学者往往更倾向于以项目组的形式合作,由 20 ~ 30 人左右的团队来共同研究解决课题。由以上对比可见,不同学科之间学者合作偏好和合著社区规模存在较大差别,而 Louvain 算法能够很好地揭示这一差异。

5　结　论

近年来科学合著网络已成为研究热点,各个学科领域都运用一些社会网络的方法进行社区划分。本文分析了 C-DBLP 公开数据集的特点,将 Louvain 算法应用到计算机、经济、法学和物理 4 个分别包含

1000 名学者的数据集的网络划分中，为解决作者合著网络问题提供了新的思路。对比结果表明，Louvain 算法的划分结果优于 LED 算法和 GN 算法，且划分的社区稳定性更好，对规模较大的数据集处理速度更快，能更高效地对大规模的作者合著网络进行划分。运用 Louvain 算法得出的科研合著网络有效揭示了各个学科研究者之间不同的合作和交流模式，有利于挖掘科研人员的研究兴趣以及学术机构、科研人员之间的相互联系，为学科合作研究提供帮助，促进学科领域科研合作的发展。

本文对单一学科的复杂网络进行社团划分，在下一步的工作中将对跨学科的复杂网络进行研究，挖掘不同学科之间的合作交流模式。

参考文献

[1] Newman M, Barabási A L, Watts D J. The Structure and Dynamics of Networks[M]. Princeton: Princeton University Press, 2006:206-207

[2] 吴祖峰，王鹏飞，秦志光，等. 改进的 Louvain 社团划分算法[J]. 电子科技大学学报，2013，42(1):105-108

[3] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004,101 (9): 2658-2663

[4] Girvan M, Newman M E J. Improved spectral algorithm for the detection of network communities[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2002(99): 7821-7826

[5] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 2006, 74 (3):36-104

[6] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 2004, 70 (6): 1-6

[7] Wu F, Huberman B A. Finding communities in linear time: a physics approach[J]. *European Physical Journal B*, 2004, 38(2):331-338

[8] Newman M E J. Fast algorithm for detecting community-structure in networks[J]. *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 2004, 69(6):1-5

[9] Ma T H, Wang Y, Tang M L, et al. LED: a fast overlapping communities detection algorithm basedon structural-clustering[J]. *Neurocomputing*, 2016, 207: 488-500

[10] 谷瑞军，陈圣磊，陈耿，等. 复杂合著网络中的重叠社团发现与可视化[J]. 图书情报工作，2012，56 (12):72-76，59

[11] 苗蕊，刘鲁. 科学家合作网络中的社区发现[J]. 情报学报，2011，30(12):1312-1318

[12] 李纲，李岚凤，毛进，等. 作者合著网络中研究兴趣相似性实证研究[J]. 图书情报工作，2015，59(2): 75-81

[13] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 2004,69(2): 1-15

[14] Ma T, Rong H, Ying C, et al. Detect structural-connected communities based on BSCHEF in C-DBLP[J]. *Concurrency and Computation Practice and Experience*, 2016, 28(2):311-330

[15] 胡健，薛龙龙. 基于用户特征和链接关系的 Louvain 算法研究[J]. 计算机与数字工程，2019，47(8): 1974-1978，2008

Research on community detection in co-authorship networks based on Louvain algorithm

Chu Yeqi*, Ding Jiajun**

(*Library of Zhejiang University of Technology, Hangzhou 310014)

(**Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018)

Abstract

The community detection in co-authorship networks is of great significance for understanding the cooperation and communication patterns of researchers. In this paper, Louvain algorithm is used in community detection in C-DBLP dataset, and the results are evaluated by modularity. Louvain algorithm can efficiently deal with networks with thousands of nodes, compared with the LED algorithm and G-N algorithm, it's more effectively in the community detection. The results reveal different disciplines' cooperation and communication mode, helping to mining potential cooperative groups and helping for collaborative research.

Key words: co-authorship network, community detection, Louvain