

基于复杂网络视角的科学文献数据分析

周建林, 牛琪锴, 曾安, 樊瑛, 狄增如

北京师范大学系统科学学院, 北京 100875

摘要 综述了科学文献数据的复杂网络表现形式, 介绍了科学家合作网络和科学引文网络的拓扑结构性质及演化模式和演化机制, 概述了论文和科学家的相关评价方法。分析表明, 基于复杂网络视角对科学文献数据的分析, 能解释很多有意义的研究问题及有趣的现象。

关键词 科学文献数据; 复杂网络; 演化; 评价

科学文献数据是指由论文信息(DOI、题目、分类号、期刊杂志、参考文献等)及其作者信息(通信作者、工作单位等)构成的数据集。早期研究者对于科学文献数据的分析方法主要是针对论文、作者做一些描述性统计分析。除此之外, 科学文献数据还可以被抽象为一个科学文献系统, 其中系统中的个体由论文、作者、杂志、研究领域、科研机构等元素组成, 个体间通过相互作用关系(论文间的引用关系、作者间的合作关系等)联系在一起, 进而通过分析该系统挖掘出隐藏在数据背后的更多信息。而科学文献系统可以用不同形式的网络结构进行描述, 研究者可以通过网络分析方法研究网络的结构、功能和演化, 从而更加深入地了解系统性质。近年来, 随着复杂网络研究的迅速发展为科学文献系统分析提供了有效的方法和工具, 并开发了相关分析软件, 如 CiteSpace、SCI2 等。目前, 基于复杂网络视角的科学文献数据分析已经取得了较为丰富的研究成果。本文着重从以下 3 个方面对已有研究工作进行梳理和综述: 1) 科学文献数据的网络表现形式;

2) 科学文献网络的拓扑性质及演化; 3) 科学文献网络中的评价。

1 科学文献数据的复杂网络表现形式

科学文献数据提供了论文的参考文献以及作者信息, 基于此可以建立由科学家、论文构成的二分网, 将该二分网投影到单顶点网络可以得到科学家合作网络; 基于论文之间的相互引用关系, 可以建立科学引文网络。由于科学文献数据中还包含了论文的其他属性信息, 如杂志、作者的工作单位、分类号等。这些属性信息通过论文之间的引用关系耦合在一起, 可以建立不同类型的耦合网络, 如杂志-论文耦合网络、科研单位-论文耦合网络等。下面介绍几种典型的网络表现形式。

1.1 科学家-论文二分网络和科学家合作网络

科学家-论文二分网络是指含有科学家、论文两类节点并且两类节点之间带有连边, 连边表示写作关系的一种二分无向网络, 并且该网络在同类节点之间不

收稿日期: 2017-05-11; 修回日期: 2017-12-05

基金项目: 国家自然科学基金项目(61374175, 61573065); 北京市自然科学基金项目(L160008)

作者简介: 周建林, 博士, 研究方向为复杂网络, 电子信箱: jianlinzhou@mail.bnu.edu.cn; 樊瑛(通信作者), 教授, 研究方向为复杂性分析, 电子信箱: yfan@bnu.edu.cn

引用格式: 周建林, 牛琪锴, 曾安, 等. 基于复杂网络视角的科学文献数据分析[J]. 科技导报, 2018, 36(8): 55-64; doi: 10.3981/j.issn.1000-7857.2018.08.006

存在连边,如图1(a)所示。科学家合作网络是以科学家为节点,以科学家之间的合作关系为连边的无向网络,如图1(b)所示。通常,可以把这种合作关系看成是两个科学家合作发表一篇论文或共同参与过某个科研项目。当仅考虑科学家之间有无合作关系时,科学家合作网络是一种无权网络,而考虑科学家之间的合作强度时,它则是一种加权网络并且科学家合作网络也可以通过科学家-论文二分网络投影生成^[1]。此外,科学家合作网络还可以用超图表示,如图1(c)所示,超图的节点代表科学家,而超图的边为每一条闭合的曲线,表示科学家们合著的论文。

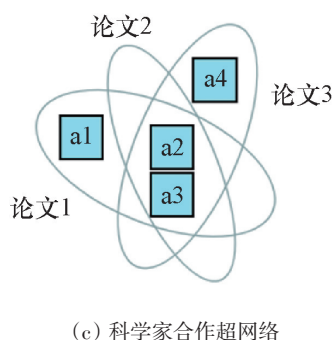
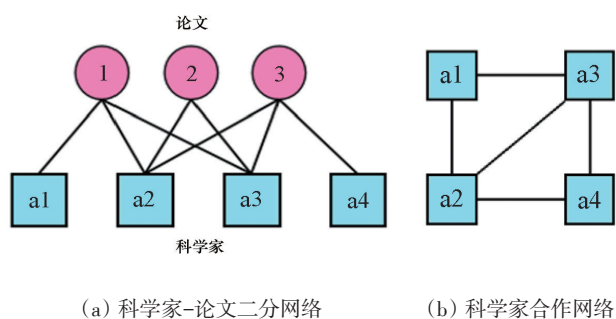


图1 科学家-论文二分网络与科学家合作网络示意

Fig. 1 Bipartite scientist-paper network and scientific collaboration network

1.2 科学引文网络

科学引文网络是基于科学论文之间的引用关系而构建的有向无环网络,如图2(a)所示。科学引文网络中的节点代表论文,而连边代表论文之间的引用关系,即若论文 i 引用论文 j ,则在它们之间存在一条由论文 i 指向论文 j 的连边。

在引文网络中,若两篇论文同时被一篇或多篇论文引用,则这两篇论文之间存有共引关系。若两篇论

文同时引用一篇或多篇论文,则称这两篇论文之间存有文献耦合^[2]关系。共引和文献耦合都在一定程度上反映了两篇论文之间研究的相关性。基于科学引文网络,可以得到共引网络与文献耦合网络,并且这两种网络均为无向网络。在共引网络中,节点表示论文,边表示论文间的共引关系,边权大小表示两篇论文被其他论文共引的次数,如图2(b)所示。在文献耦合网络中,节点表示论文,边代表文献耦合关系,边权大小表示两篇论文引用相同文献的数目,如图2(c)所示。

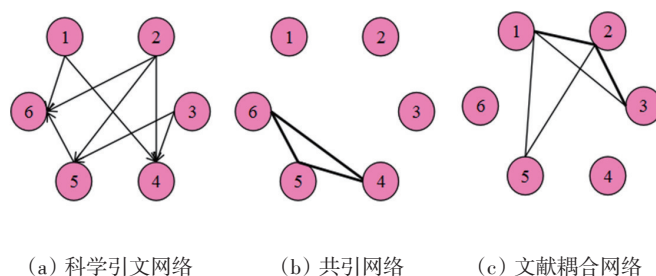


图2 科学引文网络与共引、文献耦合网络示意

Fig. 2 Scientific citation network, co-citation network and bibliographic coupling network

1.3 科学家-论文耦合网络

在信息计量学中,耦合主要是指1.2节中提到的文献耦合,这一概念最早由美国学者 Kessler在对《Physical Review》期刊进行引文分析研究时提出^[2]。而本小节中的耦合网络是指具有相互作用或相互依赖的网络。科学家-论文耦合网络由科学家和论文这两类节点组成,这两类节点之间的连边表示该名科学家是这篇论文的作者,若两名科学家共同连向同一篇论文,则表明这两名科学家之间存在合作关系,并且在论文类节点之间存在有向边,表示论文之间的引用关系,如图3所示。科学家-论文耦合网络可以被看作由一个科学家-论文二分网络和一个科学引文网络组成。

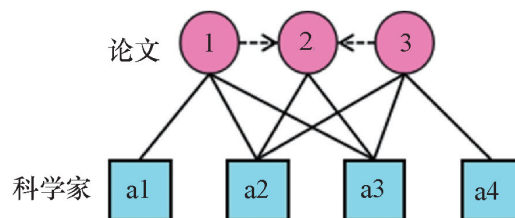


Fig. 3 Scientist-paper coupling network

2 科学文献网络的拓扑性质

科学文献网络表现出很多统计特性,通过研究这些统计特性有利于更好的认识科学文献系统。由于科学家合作网络以及科学引文网络作为最重要的科学文献网络形式,因此以这两种网络为例,研究其拓扑性质。各种科学文献网络形式均可以通过邻接矩阵表示。若科学家合作网络的邻接矩阵 $A=(a_{ij})_{(N \times N)}$,则

$$a_{ij} = a_{ji} = \begin{cases} 1 & \text{如果科学家} i \text{与科学家} j \text{之间有合作} \\ 0 & \text{如果科学家} i \text{与科学家} j \text{之间无合作} \end{cases} \quad (1)$$

式中, N 为科学家合作网络中总科学家数。

若科学引文网络的邻接矩阵 $B=(b_{ij})_{(M \times M)}$,则

$$b_{ij} = \begin{cases} 1 & \text{如果论文} i \text{引用论文} j \\ 0 & \text{如果论文} i \text{与论文} j \text{之间没有引用关系} \end{cases} \quad (2)$$

式中, M 为科学引文网络中总论文数。

基于网络的邻接矩阵,可以量化网络的拓扑结构特性。

2.1 度与度分布

科学家合作网络是一种无向网络,其节点 i 的度表示科学家 i 拥有的合作者数目,定义为

$$k_i = \sum_j a_{ij} \quad (3)$$

科学引文网络是一种有向网络,其节点 i 的入度表示论文 i 被其他论文引用的次数,节点 i 的出度则表示论文 i 引用其他论文的数目。

很多实证研究发现,合作网络的度分布服从幂律分布,例如:Newman研究了物理、生物医学、高能物理学及计算机科学领域的合作网络,发现这几个领域合作网络的度分布近似为幂律分布^[3-5]。Barabási等^[6]研究了数学和神经科学领域的合作网络,发现它们的度分布也表现出幂律尾分布特征。樊瑛等^[7]在经济物理学家合作网络中也发现其度分布近似为幂律分布。研究者们发现在科学引文网络中其入度分布也同样表现出幂律特性^[8-11]。例如:Redner通过研究被美国科学信息研究所(Institute of Scientific Information)收集的783339篇论文以及在1975—1994期间期刊《Physical Review D》出版的24296篇论文的引文分布,发现其入度分布近似为幂律分布^[8]。Lehmann等^[9]分析了高能物理数据集(SPIRES)的引文分布,发现论文的引用近似服从幂律分布,且小于50次引用的论文幂指数约为1.2,至少引用50次的论文幂指数约为2.3。

2.2 集聚系数与平均路径长度

节点 i 的集聚系数衡量了节点 i 的邻居节点相互连接的程度,定义为

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \quad (4)$$

式中, E_i 为节点 i 的 k_i 个邻居节点之间实际存在的连边数。

节点的集聚系数越大,节点与邻居节点之间连接越紧密。将所有节点的集聚系数平均可以得到整个网络的集聚系数 C 。科学家合作网络的集聚系数反映了科学家的合作者之间合作的紧密程度。李梦辉等^[12]还介绍了加权合作网的集聚系数计算。而对于有向的科学引文网络来说,集聚系数反映了论文的间接引用情况,其集聚系数的计算可以将引文网络转换为无向网络进行计算。

在科学家合作网络中,定义2名科学家 i 和 j 之间的距离 d_{ij} 为连接这2个节点的最短路径上的边数,它表示一个合作者关联另一个合作者需要经过的最少步数,则任意2个节点间距离的最大值为网络的直径,而网络的平均路径长度定义为任意两个节点间距离的平均值,即

$$L = \frac{1}{C_N^2} \sum_{i \neq j} d_{ij} \quad (5)$$

在科学引文网络中,定义2篇文章的距离为由一篇文章沿着同一方向到达另一篇文章所要经过边的最少数目,它表示了一篇文章要引用另一篇文章所需要看的最少文章数目。科学引文网络的平均路径长度定义与科学家合作网络的定义相同。

实证研究^[3-5,13-14]表明,不同领域的科学家合作网络均表现出集聚性,在某些领域如数学、物理领域表现出较高的集聚特性,某些领域,如生物领域则表现出较低的集聚特性,而且这些合作网络都具有较小的平均路径长度。同样,在一些引文网络中也发现它们表现出较高的集聚特性以及相对较低的路径长度^[15]。

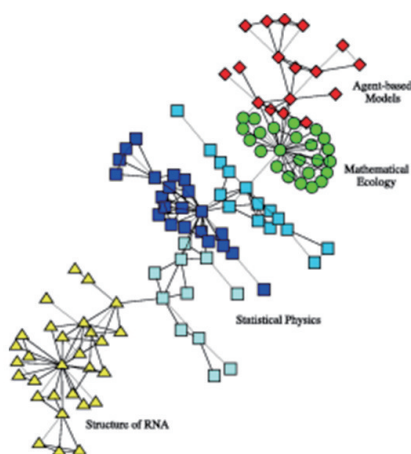
2.3 社团结构

网络中的社团结构是指网络中的节点可以分成组,组内节点的连接比较紧密而组间节点的连接比较稀疏^[16]。Newman和Girvan提出了模块度并用来定量衡量社团划分的质量,一般来讲,网络的模块度越大,说明其社团结构越明显^[17]。常用的模块度定义为

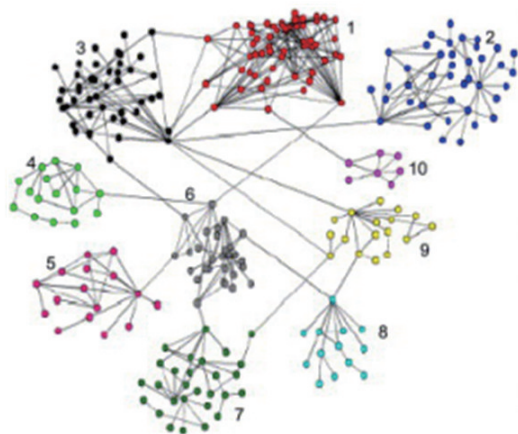
$$Q = \frac{1}{2M} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2M} \right) \delta(C_i, C_j) \quad (6)$$

式中, M 为网络中总边数, C_i 和 C_j 分别表示节点 i 与节点 j 所属的社团。

一系列的研究都表明在科学家合作网络中存在社团结构^[16-21], 例如: Girvan 等^[16]发现在圣塔菲研究所科学家合作网络中存在明显社团结构(图 4(a))。张鹏等^[16]发现在经济物理学家合作网络中也存在明显的社团结构(图 4(b))。徐玲等^[22]研究了由过去 20 年间发表在《科学通报》上论文的作者所构建的科学家合作网, 发



(a) 带有社团结构的圣塔菲研究所科学家合作网络



(b) 经济物理学家合作网络

图 4 带有社团结构的圣塔菲研究所科学家合作网络^[16]和经济物理学家合作网络^[21]

Fig. 4 Santa Fe Institute collaboration network within community structure and econophysicists collaboration network

现该合作网的最大连通集团呈现出社团结构。同样在引文网络中也存在社团结构, Chen 等^[23]基于模块化函数最大化的方法对在 1893—2007 年 8 月出版的所有物理评论论文组成的引文网络进行社团探测, 发现该引文网络存在对应于明显的物理学研究子领域的潜在社团结构。Gopalan 等^[24]运用一种基于贝叶斯网络模型的社团探测算法分析了含有 575000 篇物理学论文的 arXiv 引文网络, 探测到该网络可划分为 200 个社团。

2.4 度度相关性

在网络中, 如果大度节点倾向于连接其他大度的节点, 则称网络是正相关的或同配的; 如果大度节点倾向于连接小度节点, 则称网络是负相关的或异配的。对于科学家合作网络, 可以通过计算同配系数^[25]定量判断网络的度度相关性。对于观测网络, 其同配系数 r 计算如下:

$$r = \frac{M^{-1} \sum_i j_i k_i - \left[M^{-1} \sum_i \frac{1}{2} (j_i + k_i) \right]^2}{M^{-1} \sum_i \frac{1}{2} (j_i^2 + k_i^2) - \left[M^{-1} \sum_i \frac{1}{2} (j_i + k_i) \right]^2} \quad (7)$$

式中, M 是网络中的总边数, j_i 和 k_i 分别为第 i 条边两端节点的度值, $r \in [-1, 1]$ 。若 $r > 0$, 则网络是正相关或同配的, 反之, 网络是负相关或异配。

而对于有向的科学引文网络, 由于网络中的节点同时有入度和出度, 因此该网络有 4 种度度相关性, 并且它们可以通过 Foster 等^[26]提出的有向同配系数来测量。实证研究表明, 大多数科学家合作网络是正相关的, 即合作者较多的科学家倾向与同样拥有较多合作者的科学家合作^[13, 25, 27-28]; 而在很多科学引文网络却表现出负相关^[15]。

在本文中, 分别搜集了 6 个数据集所构建的科学家合作网络的基本统计量(表 1), 这 6 个合作网分别为: 生物医学文献数据库 MEDLINE 中的合作网^[3-5, 14]、高能物理文献数据集 SPIRES 中的合作网^[3-5, 14]、网络计算机参考图书馆 NCSTRL 中的合作网^[3-5, 14]、遗传编程数据集 (GP) 中的合作网^[14]、数学家合作网 (Mathematics)^[13-14]、物理学家合作网 (Physics)^[13-14]。本文还分别搜集 6 个数据集所构建的引文网络的基本统计量^[15](表 2), 这 6 个数据集分别为 Web of Science (WoS) 数据库、学术论文数字图书馆 CiteSeer、计算机科学文献搜索引擎 Cora 数据集、引文分析软件 HistSite 数据集、计算机科学文献库 DBLP、论文预印本网站 arXiv 数据集。

表1 一些科学家合作网络的基本统计

Table 1 Basic statistics of several scientific collaboration networks

数据集	文章总数	作者总数	作者平均文章数	文章平均作者数	作者的平均合作者数	最大连通集团的规模/%	集聚系数	平均路径长度	直径	度相关系数
MEDLINE	2163923	1520251	6.4	3.8	18.1	92.6	0.1	4.6	24	0.1
SPIRES	66652	56627	11.6	9.0	173.0	88.7	0.7	4.0	19	—
NCSTRL	13169	11994	2.6	2.2	3.6	57.2	0.5	9.7	31	—
GP	4564	2809	3.2	2.3	4.2	36.5	0.7	4.7	12	—
Mathematics	1600 000	253339	7.0	1.5	2.9	82.0	0.2	7.7	27	0.1
Physics	98502	52909	5.1	2.5	9.7	85.0	0.4	5.9	20	0.4

表2 几个科学引文网络的基本统计

Table 2 Basic statistics of several scientific citation networks

数据集	节点数	连边数	最大连通集团的规模/%	平均度	幂律指数			无向的同配系数	有向的同配系数				平均集聚系数
					度分布	入度分布	出度分布		$r_{(in,in)}$	$r_{(in,out)}$	$r_{(out,in)}$	$r_{(out,out)}$	
WoS	140362	639110	97.0	9.11	2.70	2.39	3.88	-0.06	0.04	-0.02	-0.03	0.09	0.14
CiteSeer	384413	1744619	95.0	9.08	2.65	2.28	3.82	-0.06	0.05	0	0	0.12	0.18
Cora	23166	91500	100.0	7.90	2.88	2.60	4.00	-0.06	0.07	0.02	0	0.17	0.27
HistSite	4324	41595	98.7	9.99	2.55	3.50	2.37	-0.10	0.11	0.01	-0.13	0	0.31
DBLP	12591	49744	99.2	7.90	2.42	2.64	2.75	-0.05	0	-0.02	-0.05	-0.02	0.12
arXiv	34536	421534	99.6	24.40	2.67	2.54	3.45	-0.01	0.08	-0.04	0	0.11	0.28

3 科学文献网络的演化

科学文献网络作为增长型的网络,随着时间的推移,其网络拓扑结构会发生变化,相应的网络统计性质也会发生变化,表现出不同的网络演化模式。为了刻画网络不同的演化模式,研究者试图从机制分析来阐述网络的演化过程。

3.1 科学文献网络的演化模式

研究者已经对多个领域的科学家合作网络做了动态演化分析^[6,14,19,29-32],例如:Barabási等^[6]研究了数学和神经科学领域的科学家合作网络在8年内的动态演化过程,发现网络的平均路径长度以及平均集聚系数随时间逐年减小,而最大连通集团的相对规模以及网络的平均度随时间不断增加。Martin等^[29]利用1893—2009年的物理评论数据集构建的科学家合作网络研究物理学领域内的合作模式,发现论文的发表数量以及作者数量随时间变化近似指数增长,并且论文的平均作者数量以及科学家平均合作者数量随时间变化均表现出明显的增长趋势。Liu等^[32]分析了合作演化研究领域的科学家合作网络的动态演化,发现最大连通集团随着时间的变化存在由一小集群演化至社团结构再达到小

世界结构的过程。

随着新的文章被添加到引文网络中,科学引文网络的拓扑结构同样发生变化。Martin等^[29]研究了物理评论杂志中的文章引用规律,发现文章的引用随时间近似指数递减。同时在科学引文网络的演化过程中也会出现首发优势及睡美人等现象,Newman通过理论分析得出在科学引文网络中存在首发优势现象^[33]。即最先在一个领域内的发表文章要比后面发表在该领域的文章得到的引用更高,并在网络理论研究领域进行了实证研究。van Raan在研究中提出了睡美人的概念,即长期得不到关注的文章突然吸引了人们的注意力得到大量的引用,并指出衡量睡美人现象的3个因素:睡眠时长、睡眠深度、觉醒强度^[34]。

3.2 科学文献网络演化的机制模型

自从2001年Newman研究了大型合作网络的宏观和微观特征后,人们对于科学家合作网络的研究表现出浓厚的兴趣。对于合作网络的演化,研究者提出了一系列机制模型^[6,28,35-38]。Barabási等^[6]提出了基于优先连接机制的动态网络演化模型研究科学家合作网络的演化。Ramasco等^[28]为了解释在科学家合作网络中展

现的高集聚以及同配特性,考虑社会网络的二分结构特性以及科研工作者的科研年限提出了合作网络的自组织模型;Börner等^[37]考虑到科学家合作网络以及科学引文网络的相互作用,分析了作者和引文这两个网络的同步演化。李梦辉等^[38]提出了关于加权合作网的演化模型,在该模型中老节点之间允许建立新连边以及已存在的连边可以重复产生。

对于科学引文网络的幂律度分布,Price最早提出了累积优势的思想解释该特性,他认为一篇文献得到新引用的概率应该与它之前得到的引用次数成比例^[39]。在此基础上,一些新机制又被陆续提出,一些研究者考虑了受引文年龄影响的优先连接模型,例如:考虑到节点的吸引力随其年龄增长逐渐衰退以及引文网络中三角形的数量,吴枝喜等^[40]提出了一种随机演化网络模型。Medo等综合考虑论文影响力的异质性及其随时间衰减的特性来研究引文网络的演化,提出了对于度演化模拟更适合的模型^[41]。另外还有一些研究者从其他角度考虑了引文网络的生成机制,例如:Golosovsky等^[42]发现文章的引用增长服从超线性偏好依附,并且其引用过程不能被描述成无记忆的马尔可夫链,他们提出了一种将节点度的演化看作是非齐次的自激点过程的随机增长模型。Krapivsky等^[43-44]提出了生成引文网络的复制模型,他们认为一篇文献在引用一篇目标文献时,也会以一定的概率引用该目标文献的参考文献。Peterson等考虑了一篇新论文的作者在引用一篇老论文时存在直接和间接两种机制,分别为直接找到目标文献A并引用它和先找到一篇论文B,然后在论文B的参考文献中找到目标文献A,他们将这两种机制结合起来研究科学引文的分布^[45]。

4 科学文献网络中的评价

基于科学文献数据的评价工作一直是研究重点和热点,如评价论文或科学家的影响力,评价科研工作者的科研绩效或者是对国家科研表现的评价。而如何基于科研文献网络的局域信息或全局信息来对论文或科学家进行公平评价是最重要的研究问题之一。具体相关研究方法可参见曾安等在2017年发表在《Physics Reports》上的关于科学学研究的综述文章^[45]。下面对科学家影响力评价以及论文影响力评价的相关工作进行简要介绍。

4.1 科学家影响力的评价

传统的评价科学家影响力的指标主要有总论文数、总引文数、篇均引文数、重要论文数量以及重要论文的引文数量。以上评价指标优缺点并存^[46],例如:总论文数能够反映科学家的生产力但忽略了论文自身的质量;总引文数能够反映科学家发表论文的总体影响力,但有些论文很可能是由多名科学家合作完成,不能反映个人的真实影响力。2005年,Hirsch提出了兼顾科学家科研产出的数量以及质量的评价指标—— h 指数^[46],该指标不仅克服了上述传统的评价指标的种种缺点,而且还具有很强的鲁棒性,不易受恶意自引等行为的影响。但 h 指数也存在很多不足,例如:不利于对发表论文数量少但被引次数高的科学家评价;对具有相同 h 指数的科学家不能很好区分;由于存在累积优势, h 指数不利于年轻学者的评价。研究者们针对这些不足,提出了很多变型指标^[47]。

不同于以上指标,基于复杂网络的评价指标提供了新的视角及方法。基于科学文献网络(如科学引文网络,作者引文网络,科学家合作网络)评价科学家影响力,一方面可以基于科学家合作者的学术影响力间接评价科学家的影响力;另一方面可以从知识传播扩散的角度评价科学家影响力。运用网络分析方法评价科学家影响力主要有度、接近度、介数、PageRank及其变型^[48]等中心性指标。研究者们已经利用这些指标在不同领域的科学文献网络中评价科学家的影响力。Yan等^[49]将接近中心性、介数中心性、度中心性及PageRank算法应用在图书情报学合作网络中来对作者排名,发现这4种度量指标均与引文计数之间呈明显正相关关系,认为中心性指标可以用来评价科学家的影响力。他们还为了衡量合著网络中的作者影响力,提出了考虑作者的被引次数以及合著网络拓扑结构的加权PageRank算法,认为有影响力的作者被随机浏览的概率更大,并测试了带有不同阻尼因子下的PageRank算法的评价效果^[50]。相类似的加权PageRank算法还应用在作者引文网络以及作者共引网络中^[51-52]。除了上述常见指标外,还有研究者基于科学文献网络提出了一些新评价指标,例如:Radicchi等^[53]基于加权的作者引文网络提出了一个评价科学家影响力的SARA指标,该指标根据与边权重比例的扩散概率在全局网络上模拟了科学家的声誉扩散过程。

4.2 论文影响力的评价

依据被引次数评价一篇论文的影响力是一种典型的基于引文网络局域信息的排序算法。Garfield在1955年首次提出利用一篇学术论文的被引次数评价一篇论文的科学影响力^[54]。在引文网络中,节点的入度即论文的被引次数,通过对所有节点的入度排序实现对论文影响力的评价。通常,一篇论文的被引次数在一定程度上能够反映一篇论文的质量,被引次数越高,则论文的质量越高。这种评价方式的最大优点在于它计算简单方便,但是也存在很多不足。基于被引次数的论文评价仅考虑引文的数量而忽视了引文的质量,而通常引文的质量越高,被引论文的科学影响力也越大。而且这种基于被引次数的评价方式还可能引发人为操纵及恶意自引现象。此外,不同学科的引用存在差异,被引次数不适宜直接用于比较学科间论文的学术影响力。Radicchi等在2008年提出了评价论文影响力的重新标度的被引次数,不仅可以用来比较不同领域论文的科学影响力而且也可以用来比较不同领域的科学家的科研绩效^[55]。

PageRank算法是一种经典的基于引文网络全局信息的论文排序算法。该算法最早应用于网页的排序^[56],后来被研究者引入到科学计量领域^[57-60]。PageRank算法在对论文排序时不仅考虑了论文的被引次数而且还考虑了引文的质量,为论文排序提供了新的视角。同时,PageRank算法也存在不足,它自身带有参数,参数的取值不同会影响论文的排序结果。而且论文的引用存在累积优势,PageRank算法中未考虑时间因素的影响更有利于发表时间久的论文评价。研究者们陆续提出了很多修正算法^[61-66]。Walker等^[64]提出了CiteRank算法,该算法模拟了大量的研究者如何沿着引用链寻找新信息的过程并以通过论文的交通流来衡量一篇论文的重要性。该模型中考虑了引文网络的老化特性,使得近期发表的文章得到更多的关注,使得评价更加公正。姚丽阳等^[65]将非线性思想引入到PageRank算法中,改变了原有的PageRank算法在对论文重要性评价时的线性叠加方式,提高了重要文章的排序。周建林等^[66]在PageRank算法的基础上引入相似优先的扩散机制对论文进行排序,算法的有效性在美国物理学会引文网络上得以验证并且对于人为操作等行为具有很强的鲁棒性。

5 结论与展望

随着科技的发展,科学文献数据的体量在不断增大,正朝着大数据方向发展。对于科学文献数据的分析方法,不仅要依赖统计分析的方法,还需要使用其他多元化的分析方法,而复杂网络恰好为科学文献数据分析提供了一个全新的视角及分析方法。本文简要综述了基于复杂网络视角下的科学文献数据分析工作。首先介绍了几种典型的科学文献数据复杂网络表现形式,然后重点介绍了科学家合作网络、科学引文网络的拓朴性质以及演化的模式及机制。最后探讨了科学文献网络中的相关评价问题,主要是对论文及科学家的评价,给出了一些评价方法。

研究者对于科学文献网络最初的研究多关注网络的拓朴结构特征以及网络的演化过程,但随着复杂网络理论的深入发展以及数据处理技术的进步,研究者更倾向于去挖掘科学文献数据背后的现象及规律,解决实际问题,例如:通过观测领域演化的规律以及评价领域的发展状况来更加公平合理地分配项目经费。对于评价问题可以推广到评价杂志、科研机构、国家的科研表现力及发展状况,还可以推广到评价项目、基金的质量。除此之外,基于科学文献数据及复杂网络的分析方法,可以预测一名科学家在未来的职业发展趋势,还可以探究科研工作者的行为模式,例如科学家科研兴趣的转移,科学家工作单位的迁移,科学家如何选择合作者等。总之,科学文献数据的背后隐藏着非常多有意义的研究问题,等待人们去充分挖掘;而基于复杂网络的分析方法会揭示出更多有趣的现象。

参考文献(References)

- [1] 吴亚晶, 张鹏, 狄增如, 等. 二分网络研究[J]. 复杂系统与复杂性科学, 2010, 7(1): 1-12.
Wu Yajing, Zhang Peng, Di Zengru, et al. Study on bipartite networks[J]. Complex Systems and Complexity Science, 2010, 7(1): 1-12.
- [2] Kessler M M. Bibliographic coupling between scientific papers [J]. American Documentation, 1963, 14(1): 10-25.
- [3] Newman M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences, 2001, 98(2): 404-409.
- [4] Newman M E J. Scientific collaboration networks. I. Network construction and fundamental results[J]. Physical Review E, 2001, 64(1): 016131.

- [5] Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality[J]. *Physical Review E*, 2001, 64(1): 016132.
- [6] Barabási A L, Jeong H, Neda Z, et al. Evolution of the social network of scientific collaborations[J]. *Physica A: Statistical Mechanics and its Applications*, 2002, 311(3): 590–614.
- [7] Fan Y, Li M, Chen J, et al. Network of econophysicists: A weighted network to investigate the development of econophysics[J]. *International Journal of Modern Physics B*, 2004, 18(17–19): 2505–2511.
- [8] Redner S. How popular is your paper? An empirical study of the citation distribution[J]. *The European Physical Journal B–Condensed Matter and Complex Systems*, 1998, 4(2): 131–134.
- [9] Lehmann S, Lautrup B, Jackson A D. Citation networks in high energy physics[J]. *Physical Review E*, 2003, 68(2): 026113.
- [10] Golosovsky M, Solomon S. Runaway events dominate the heavy tail of citation distributions[J]. *The European Physical Journal Special Topics*, 2012, 205(1): 303–311.
- [11] Peterson G J, Pressé S, Dill K A. Nonuniversal power law scaling in the probability distribution of scientific citations[J]. *PNAS*, 2010, 107(37): 16023–16027.
- [12] Li M, Fan Y, Chen J, et al. Weighted networks of scientific communication: The measurement and topological role of weight[J]. *Physica A: Statistical Mechanics and its Applications*, 2005, 350(2): 643–656.
- [13] Newman M E J. Coauthorship networks and patterns of scientific collaboration[J]. *PNAS*, 2004, 101(Suppl 1): 5200–5205.
- [14] Tomassini M, Luthi L. Empirical analysis of the evolution of a scientific collaboration network[J]. *Physica A: Statistical Mechanics and its Applications*, 2007, 385(2): 750–764.
- [15] Šubelj L, Fiala D, Bajec M. Network-based statistical comparison of citation topology of bibliographic databases[J]. *Scientific Reports*, 2014, 4: 6496.
- [16] Girvan M, Newman M E J. Community structure in social and biological networks[J]. *PNAS*, 2002, 99(12): 7821–7826.
- [17] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69(2): 026113.
- [18] Lužar B, Levnajic Z, Povh J, et al. Community structure and the evolution of interdisciplinarity in Slovenia's scientific collaboration network[J]. *PloS One*, 2014, 9(4): e94429.
- [19] Evans T S, Lambiotte R, Panzarasa P. Community structure and patterns of scientific collaboration in business and management[J]. *Scientometrics*, 2011, 89(1): 381–396.
- [20] Velden T, Lagoze C. The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication[J]. *Journal of the American Society for Information Science and Technology*, 2013, 64(12): 2405–2427.
- [21] Zhang P, Li M, Wu J, et al. The analysis and dissimilarity comparison of community structure[J]. *Physica A: Statistical Mechanics and its Applications*, 2006, 367: 577–585.
- [22] 徐玲, 胡海波, 汪小帆. 一个中国科学家合作网的实证分析[J]. *复杂系统与复杂性科学*, 2009, 6(1): 20–28.
Xu Ling, Hu Haibo, Wang Xiaofan. Empirical analysis of a China scientists collaboration network[J]. *Complex Systems and Complexity Science*, 2009, 6(1): 20–28.
- [23] Chen P, Redner S. Community structure of the physical review citation network[J]. *Journal of Informetrics*, 2010, 4(3): 278–290.
- [24] Gopalan P K, Blei D M. Efficient discovery of overlapping communities in massive networks[J]. *Proceedings of the National Academy of Sciences*, 2013, 110(36): 14534–14539.
- [25] Newman M E J. Assortative mixing in networks[J]. *Physical Review Letters*, 2002, 89(20): 208701.
- [26] Foster J G, Foster D V, Grassberger P, et al. Edge direction and the structure of networks[J]. *PNAS*, 2010, 107(24): 10815–10820.
- [27] Newman M E J. Mixing patterns in networks[J]. *Physical Review E*, 2003, 67(2): 026126.
- [28] Ramasco J J, Dorogovtsev S N, Pastor-Satorras R. Self-organization of collaboration networks[J]. *Physical Review E*, 2004, 70(3): 036106.
- [29] Martin T, Ball B, Karrer B, et al. Coauthorship and citation patterns in the Physical Review[J]. *Physical Review E*, 2013, 88(1): 012814.
- [30] Zhai L, Li X, Yan X, et al. Evolutionary analysis of collaboration networks in the field of information systems[J]. *Scientometrics*, 2014, 101(3): 1657–1677.
- [31] Li J, Li Y. Patterns and evolution of coauthorship in China's humanities and social sciences[J]. *Scientometrics*, 2015, 102(3): 1997–2010.
- [32] Liu P, Xia H. Structure and evolution of co-authorship network in an interdisciplinary research field[J]. *Scientometrics*, 2015, 103(1): 101–134.
- [33] Newman M E J. The first-mover advantage in scientific publication[J]. *Europhysics Letters*, 2009, 86(6): 68001.
- [34] van Raan A F J. Sleeping beauties in science[J]. *Scientometrics*, 2004, 59(3): 467–472.
- [35] Newman M E J. Clustering and preferential attachment in growing networks[J]. *Physical Review E*, 2001, 64(2): 025102.
- [36] Jeong H, Neda Z, Barabási A L. Measuring preferential attachment in evolving networks[J]. *Europhysics Letters*, 2003, 61(4): 567.

- [37] Börner K, Maru J T, Goldstone R L. The simultaneous evolution of author and paper networks[J]. PNAS, 2004, 101(Suppl 1): 5266–5273.
- [38] Li M, Wu J, Wang D, et al. Evolving model of weighted networks inspired by scientific collaboration networks[J]. Physica A: Statistical Mechanics and its Applications, 2007, 375(1): 355–364.
- [39] Price D S. A general theory of bibliometric and other cumulative advantage processes[J]. Journal of the American society for Information science, 1976, 27(5): 292–306.
- [40] Wu Z X, Holme P. Modeling scientific-citation patterns and other triangle-rich acyclic networks[J]. Physical Review E, 2009, 80(3): 037101.
- [41] Medo M, Cimini G, Gualdi S. Temporal effects in the growth of networks[J]. Physical Review Letters, 2011, 107(23): 238701.
- [42] Golosovsky M, Solomon S. Stochastic dynamical model of a growing citation network based on a self-exciting point process[J]. Physical Review Letters, 2012, 109(9): 098701.
- [43] Krapivsky P L, Redner S. Organization of growing random networks[J]. Physical Review E, 2001, 63(6): 066123.
- [44] Krapivsky P L, Redner S. Network growth by copying[J]. Physical Review E, 2005, 71(3): 036118.
- [45] Zeng A, Shen Z, Zhou J, et al. The science of science: From the perspective of complex systems[J]. Physics Reports, 2017, (714/715): 1–73.
- [46] Hirsch J E. An index to quantify an individual's scientific research output[J]. PNAS, 2005, 102(46): 16569–16572.
- [47] Alonso S, Cabrerizo F J, Herrera-Viedma E, et al. h-Index: A review focused in its variants, computation and standardization for different scientific fields[J]. Journal of Informetrics, 2009, 3(4): 273–289.
- [48] Nykl M, Campr M, Ježek K. Author ranking based on personalized PageRank[J]. Journal of Informetrics, 2015, 9(4): 777–799.
- [49] Yan E, Ding Y. Applying centrality measures to impact analysis: A coauthorship network analysis[J]. Journal of the American Society for Information Science and Technology, 2009, 60(10): 2107–2118.
- [50] Yan E, Ding Y. Discovering author impact: A PageRank perspective[J]. Information Processing & Management, 2011, 47(1): 125–134.
- [51] Ding Y, Yan E, Frazho A, et al. PageRank for ranking authors in co-citation networks[J]. Journal of the American Society for Information Science and Technology, 2009, 60(11): 2229–2243.
- [52] Ding Y. Applying weighted PageRank to author citation networks[J]. Journal of the American Society for Information Science and Technology, 2011, 62(2): 236–245.
- [53] Radicchi F, Fortunato S, Markines B, et al. Diffusion of scientific credits and the ranking of scientists[J]. Physical Review E, 2009, 80(5): 056103.
- [54] Garfield E. Citation Indexes for Science: A new dimension in documentation through association of ideas[J]. Science, 1955, 122(3159): 108–111.
- [55] Radicchi F, Fortunato S, Castellano C. Universality of citation distributions: Toward an objective measure of scientific impact [J]. PNAS, 2008, 105(45): 17268–17272.
- [56] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks & Isdn Systems, 1998, 30(98): 107–117.
- [57] Bollen J, Rodriguez M A, van de Sompel H. Journal status[J]. Scientometrics, 2006, 69(3): 669–687.
- [58] Chen P, Xie H, Maslov S, et al. Finding scientific gems with Google's PageRank algorithm[J]. Journal of Informetrics, 2007, 1(1): 8–15.
- [59] Fiala D, Rousselot F, Ježek K. PageRank for bibliographic networks[J]. Scientometrics, 2008, 76(1): 135–158.
- [60] Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis[J]. Information Processing & Management, 2008, 44(2): 800–810.
- [61] Su C, Pan Y T, Zhen Y N, et al. PrestigeRank: A new evaluation method for papers and journals[J]. Journal of Informetrics, 2011, 5(1): 1–13.
- [62] Fiala D. Time-aware PageRank for bibliographic networks[J]. Journal of Informetrics, 2012, 6(3): 370–388.
- [63] Nykl M, Ježek K, Fiala D, et al. PageRank variants in the evaluation of citation networks[J]. Journal of Informetrics, 2014, 8(3): 683–692.
- [64] Walker D, Xie H, Yan K K, et al. Ranking scientific publications using a model of network traffic[J]. Journal of Statistical Mechanics: Theory and Experiment, 2007, 2007(6): P06010.
- [65] Yao L, Wei T, Zeng A, et al. Ranking scientific publications: The effect of nonlinearity[J]. Scientific Reports, 2014, 4: 6663.
- [66] Zhou J, Zeng A, Fan Y, et al. Ranking scientific publications with similarity-preferential mechanism[J]. Scientometrics, 2016, 106(2): 805–816.

Analysis of scientific literature database from a perspective of complex network

ZHOU Jianlin, NIU Qikai, ZENG An, FAN Ying, DI Zengru

School of Systems Science, Beijing Normal University, Beijing 100875, China

Abstract Scientific literature data cover the complete information of papers and authors. Facing the massive scientific literature data, traditional statistical analysis methods cannot fully explore the information hidden behind the data without the help of other analysis methods. The interactions in scientific literature data, such as citation between papers and co-authorship between scientists, allow for the construction of different forms of complex networks (citation networks, collaboration networks, etc.), which can allow us to distinguish the effective information hidden in the scientific literature data based on network analysis. This paper summarizes the complex network forms of scientific literature data and highlights the topological properties, evolution patterns as well as evolution mechanisms of scientific collaboration networks and scientific citation networks. As impact evaluation of papers and scientists has attracted so much attention from researchers for a long time, we also briefly summarize the related evaluation methods of papers and scientists. From the perspective of complex network, it can also explain many meaningful questions and interesting phenomena in scientific literature data, such as the shift of scientists' research interests and the sleeping beauties. In the future, the method of network analysis must be able to achieve more abundant research results in mining scientific literature data.

Keywords scientific literature data; complex network; evolution; evaluation ●



(责任编辑 田恬)