

基于科研贡献度加权的作者合作网络对比研究*

明均仁^{1,2}, 党永杰³

(1. 中国科学技术信息研究所, 北京 100038; 2. 武汉工程大学管理学院, 武汉 430205; 3. 华中师范大学计算机学院, 武汉 430079)

摘要: 针对作者合作网络分析中高估作者对论文的科研贡献度和忽略单篇论文对整个网络边权的贡献度问题, 分别利用作者署名顺序对网络节点加权和引用排列组合的思想通过遍历作者关系对边加权。以《情报科学》2008-2013年刊发论文中的核心作者群构建作者合作加权网络, 从节点强度、最短路径、聚集系数、“富人俱乐部”系数等4个角度分析网络特征, 并通过与频次加权网络相关特征对比分析验证基于科研贡献度加权方法的必要性和可行性。结果表明科研贡献度加权方法能够更加真实地反映作者的个体特征及其网络结构特征。

关键词: 加权网络; 网络特征; 作者合作网络; 复杂网络

中图分类号: G201

DOI: 10.3772/j.issn.1673-2286.2016.1.005

1 引言

随着科学研究的多元化及学科的交叉、融合趋势的发展, 科学研究的形式已从个体研究发展为由课题负责人领导下的群体研究^[1], 科研合作成为学术交流的重要表现形式, 其成果主要表现为研究人员合作发表期刊论文^[2], 这种共同署名一篇论文的多位科研人员之间的关系就是作者合作关系。由于研究领域的专业化、研究内容的深入化, 跨学科、跨机构、跨地区的寻求相关专业领域有共同研究方向的科研合作对象进行科研合作, 已成为科研项目取得重大突破的智力保障。

随着近年来科研产出的爆发式增长, 作者合作现象引起了相关学者的密切关注。最初, 对作者合作现象的研究主要表现为利用传统文献计量学方法对合作论文的比重、不同作者数论文的数量和比例等问题进行分析, 如汪冰^[3]将作者合作的类型概括为师生、同行、同事等类型, 并统计了论文分布状况。随着社会网络分析方法的兴起, 相关学者通过该方法从网络视角分析作者合作关系, 如邱均平、王菲菲^[4]对我国竞争情报领域的作者合作关系进行了深层次探索, 郑曦等^[5]分析了链接分析领域的作者合作网络特征, 朱庆华等等^[6]对《情

报学报》期刊的合作网络进行了实证研究。然而, 这些研究往往利用二值化数据构建无权网络或者利用绝对发文量和作者间合作发文的绝对耦合频次对节点和边简单加权, 忽视了作者对论文的科研贡献度问题和单篇论文对整个网络的贡献度问题。

本文基于单篇论文对整个网络贡献一个单位点权和边权的思想, 构建基于科研贡献度加权的作者合作网络, 通过与耦合频次加权的网络进行对比分析, 结果表明本文所述加权方法有利于构建科学合理的作者合作网络, 有利于客观真实地衡量作者间合作关系强弱, 有利于正确分析作者在网络中的地位。

2 基于科研贡献度的加权网络构建

2.1 加权网络的模型表示

本文对加权网络的建模通过加权图 G^W 表示, $G^W = (N, V)$, 其中 $N = \{(N_1, W_1), (N_2, W_2), \dots, (N_n, W_n)\}$ 表示节点的集合, N_m 表示待标记的作者, W_m 表示节点 N_m 的权值, 即加权后作者的相对发文量, n 代表集合中的作者数目, 用于表示作者合作网络规模的大小;

* 本研究得到湖北省教育厅科学技术研究项目“基于威客模式的知识共享平台及其运营机制设计”(编号: Q20151515)资助。

$V = \{(L_1, W_1), (L_2, W_2), \dots, (L_i, W_i)\}$ 表示边的集合, L_m 表示待标记的边, W_m 表示边 L_m 的权值, 即对应节点间共同发文数的加权值, i 代表集合中的边数目。

2.2 科研贡献度加权方法概述

在“点一线一面(网)”的递进层次关系结构中, 节点与边作为网络的构成要素, 个体特征权值如何设置直接影响网络整体结构特征。对节点与边基于科研贡献度加权的方法如下。

2.2.1 节点加权方法

作者合作网络中往往采用绝对发文量作为节点的权值, 这种不区分作者对论文的科研贡献度的方法不仅会高估作者的科研水平, 导致一些经常“挂名”或署名顺序靠后的科研人员的科研贡献度“虚高”的问题, 而且会高估署名顺序靠后的作者在整个作者合作网络中的地位。孙丽娟^[7]提出根据作者署名的先后顺序衡量相关作者的科研贡献度, 对作者顺序位置的倒数作为该作者贡献度大小的系数; 苏新宁^[8]在《中国人文社会科学学术影响力报告》中将每篇论文的作者总权重设为1, 按照参与合作的作者数固定分配每位作者的权重。本文比较认同单篇论文贡献一个单位权值的观点, 在分析过程中引用《中国人文社会科学学术影响力报告》中提出的计算作者发文量的标准, 将其应用到对作者科研贡献度的衡量, 具体计算标准如表1所示。

表1 合作者科研贡献度

作者数	第一作者	第二作者	第三作者	第四作者	第五作者	第六作者
1	1					
2	0.6	0.4				
3	0.6	0.25	0.15			
4	0.6	0.2	0.1	0.1		
5	0.6	0.1	0.1	0.1	0.1	
6	0.6	0.1	0.1	0.1	0.05	0.05

2.2.2 边加权方法

如果采用作者间合作频次作为作者合作网络中边的权重, 则多作者署名论文会对整个网络贡献多个单位的边权重, 且署名人数越多虚假贡献越大。本文认为单

篇论文对整个作者合作网络贡献一个单位的边权重, 而且不同作者数量论文中的个体边对整个网络中边的权重贡献具有差异性, 暂不考虑同一论文中不同边之间权值的大小差异的前提下, 边的权重^[9]定义为公式(1):

$$W_{vw} = \sum_k \frac{\delta_v^k \delta_w^k}{\binom{2}{n_k}} \quad (1)$$

由公式(1)可知边的累积权值不只由对应的两个作者共同发表的论文数量决定, 还受论文署名人数多少的影响。其中 W_{vw} 为作者V和W之间的边的权重; n_k 表示论文k署名的作者数目(假设单篇论文不存在同名作者问题); 当作者V在论文M中出现时 δ_v^k 等于1, 否则为0, 当作者W在论文M中出现时 δ_w^k 等于1, 否则为0。具体处理过程为: (1)运用公式(1)对每篇论文进行分析并生成一个边集合s和与之对应的加权边向量d; (2)对每篇论文生成的边集合去重处理, 得到一个关于边的总集合S: $S = \{L_1, L_2, L_3, \dots, L_i\}$, L代表具体的边, i代表样本论文库中存在作者合作关系即边的总数量; (3)根据每篇论文的加权边向量d, 统计出所有边在整个样本论文库中的累积加权值, 从而得到所有边的加权边向量D: $D = \{W_1, W_2, W_3, \dots, W_i\}$, i 仍代表边的总数量, W_j 代表第j ($0 < j \leq i$) 条边 L_j 的累积加权值。

2.3 网络特征指标

2.3.1 节点强度

无权网络中节点度指与节点相连接节点数的总和, 反映了节点在网络中的资源获取能力, 是衡量节点网络地位的重要特征。度分布用来描述节点度的分布情况, 即网络中度数为某一特定值的节点数与节点总数的比值。加权网络中对节点度进行自然推广为节点强度 S_i , 节点度分布可以推广为节点强度分布。节点强度的值为与节点相连的所有边权值之和, 定义为公式(2):

$$S_i = \sum_{j \in N} W_{ij} \quad (2)$$

其中, N为网络中与节点连接的边的总数, W_{ij} 为边的权值。一般情况下节点强度越大的节点在网络中的“地位”越高。平均节点强度反映的是网络中各节点间连接关系的频繁程度和紧密程度, 平均节点强度越大,

说明网络中节点间的关系越紧密^[10]。平均节点强度 $\bar{S}$ 定义为公式(3)：

$$\bar{S} = \frac{\sum_{j \in N} W_{ij}}{N} \quad (3)$$

2.3.2 最短路径长度

最短路径长度作为描述网络特征的重要参数之一,指的是两点之间所有连通的路径中距离之和最小的一条或几条路径。无权网络中的最短路径即为条数最少的路径,加权网络因为距离不满足三角不等式,两边距离之和不一定大于第三边,所以边数最少的路径不一定是最短路径,此外加权网络中节点之间的连接根据权重不同具有“倾向性”,因此不能根据边权重直接选择路径,一般利用平均权重对边权进行归一化处理:

(1) 将所有边的权值之和除以边数,即求出网络中边的平均权值;(2) 将每条边的权值除以边的平均权值;

(3) 对步骤(2)中结果求倒,并将其作为节点间的路径长度。例如,在由A、B、C三者相互连接组成的无权网络中,A、B间的最短路径即为连接二者的边;图1(a)所示的加权网络中AB(BA)、AC(CA)、BC(CB)三条边的权重分别为1、3、5,利用前文所述方法对其归一化处理后结果如图1(c),A、B间的直接路径长度为3(1/(1/3)),经过C点的A、B间的间接路径长度为1.6(1/(3/3)+1/(5/3)),虽然A、B间存在权重为1直接连接,但是通过中间节点C连接A、B的路径有很高的权重,因此A、B间通过C合作的概率可能比A、B间直接合作的概率大。对加权网络的边权重进行归一化处理不仅有助于明确统计量的物理意义,还可以比较不同权重范围网络的平均路径长度^[11]。

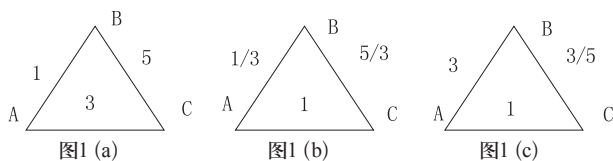


图1 边权归一化处理演示图

2.3.3 聚集系数

聚集系数是统计网络聚类特性的物理量,反映了

节点邻居节点间联系的紧密程度,值越大说明邻接点之间的联系越紧密。无权网络中节点i的聚类系数 C_i 定义为公式(4)：

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

t_i 为与节点i的邻接点间实际相连的边数, k_i 为节点i的邻接点的总数,这些邻接点间最多可能有 $k_i(k_i-1)/2$ 条边。整个网络的聚集系数即为所有节点i的聚集系数的平均值。加权网络应该根据边权重对比哪些邻接点更加重要,节点之间更有可能形成连接关系,为此Onnela将加权网络的聚集系数定义为公式(5)：

$$C_w(i) = \frac{2}{k_i(k_i-1)} \sum_{j,k} (W_{ij} W_{jk} W_{ki})^{\frac{1}{3}} \quad (5)$$

其中 W_{ij} 、 W_{jk} 、 W_{ki} 均为利用网络中的最大权重标准化后的权重,如果所有权重相同,则公式(5)与公式

(4)的运算结果相同。加权网络聚集系数在运算过程中既考虑了邻接点中边的数量,又考虑在网络中的相对权重,可以反映网络拓扑中边权重高低的相对水平。加权网络的聚集系数从不同粒度反映网络的聚集程度,最小粒度角度能够反映单节点及其邻居节点间的紧密程度,整个加权网络的聚集系数能够求出更粗粒度下网络的紧密程度,一般而言,连接越紧密、连接强度越高,则聚集系数越高^[12]。

2.3.4 “富人俱乐部”系数

网络中度数较大的节点称为“富节点”,这些节点之间的相互连接称之为“富人俱乐部”,描述的是网络中节点度较大的核心层节点的连接情况,“富人俱乐部”系数属于聚集系数的一种特殊情况,从不同角度描述了网络的整体结构层次性。富人俱乐部系数 $\Phi(r)$ 表示最富有r个节点的总链路数 E_r 与这r个节点之间总的可能存在的共享链路数的最大值 $r(r-1)/2$ 的比值,当 $\Phi(r)=0$,不存在富人俱乐部现象;当 $\Phi(r)=1$,网络所有节点与其他节点都直接相连,形成一个完全连通图。加权网络的“富人俱乐部”系数计算过程中首先确定核心节点的数量,然后求出这些节点间的边权重之和与同样数量的权重最大的边的权重之和,二者之比即

为加权网络的“富人俱乐部”系数。

3 数据来源与实证分析

3.1 数据来源及数据预处理

本文选择CNKI中国期刊全文数据库为数据来源, 选定2008-2013年间《情报科学》期刊所有论文作为样本数据, 共有2397篇论文, 1612篇合作论文, 合作率为67.3%, 说明作者合作现象的普遍性及其分析的必要性。不同作者数的论文分布情况如图2所示, 合作论文中97.5%的论文是2-4位作者间的合作, 4位以上作者合作论文数较少, 说明本文引用的科研贡献度衡量方法对样本数据具有适用性。

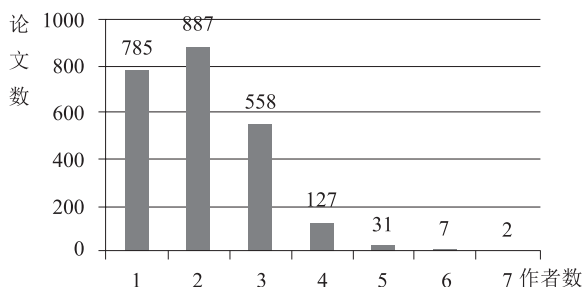


图2 论文分布图

1612篇合作论文中包含2722位合作作者, 根据普赖斯定律选定发文数在4篇及以上的183位作者进行分析, 去除其中的60位孤立节点, 最终确定123位作者组成的核心作者群作为本文的具体研究对象, 利用耦合频次加权和前文所述根据贡献度加权的作者合作网络展示分别如图3 (a)、图3 (b) 所示, 二者均为无向连接图, 节点代表作者, 节点间的连线代表二者具有合作关系, 边的权值用于衡量作者间合作关系的强弱。对比可知两个图中的节点数和边数总量相同, 但是边的权值有所不同, 例如图3 (a) 中余肖生、刘向与马费成教授分别有1和3次的合作频次, 而图3 (b) 中对应两条边的权重分别为1和0.8, 二者反映的作者合作关系强弱相互矛盾, 究其原因原始数据中余肖生与马费成教授存在一次合作且论文中仅有他们二人署名, 而刘向与马费成教授存在三次合作关系但论文中署名作者较多, 因此图3 (b) 更能客观真实地反映作者间合作关系的强弱, 更有利于科学合理分析网络结构特征。在科研产出总量同为1篇论文的情况下, 多作者间的合作必然会影响作

者间的合作深度, 合作者较少时作者间的关系更加紧密, 因此利用耦合频次加权存在一定程度的不合理性, 同时验证了本文所述边加权方法的合理性。

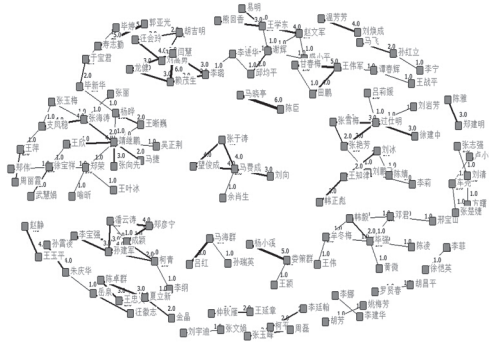


图3 (a) 频率加权作者合作网络图

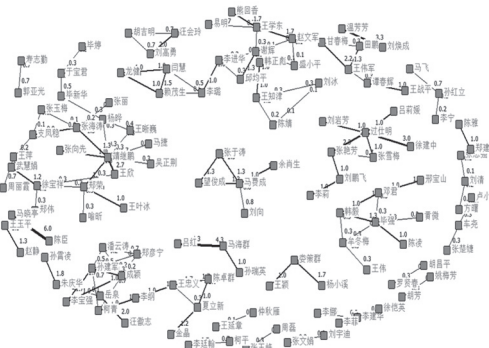


图3 (b) 贡献度加权作者合作网络图

3.2 网络拓扑结构分析

网络拓扑结构是指网络形状, 即节点间物理连接的特征。图3 (a)、图3 (b) 中的网络节点间连接较为分散、连通性不好, 没有形成一个大规模相互连接的网络, 主要包含几个规模相对较大的子网和多个双核子网, 前文所述两种加权方式构建的网络拓扑结构一致, 具体包含以下几种类型:

(1) 双核型拓扑。即网络中仅有两个节点之间互相连接, 一方面可能仅有两个作者合作发文, 这种情况前文所述的两种加权方式的边权值相同; 或者由于其他合作作者不在样本数据选择的核心作者群范围内导致。图3 (a)、图3 (b) 中都含有13个双核型子网, 涉及26个节点, 约占总节点数的1/5。

(2) 环形拓扑。即网络中的节点首尾相连形成一个闭合的环, 节点只与前后两个节点首尾连接。如图中刘高勇、汪会玲、胡吉明三者形成一个闭合的子网, 他

们之间相互存在合作关系。在大于3个节点的环形网络中节点间不存在相互连接,否则即为网状拓扑。

(3) 星形拓扑。即各节点都与中心节点连接,而这些普通节点间没有连接。在作者合作网络中这种结构基本不存在,普通节点间一般会存在一定程度的合作关系。

(4) 网状拓扑。节点间的连接不受限制,没有规律,结构较为复杂。图3(a)、图3(b)中以毕强、靖继鹏等为中心的节点数较多的子网都属于网状拓扑,这类网络节点间存在多条路径相连。

3.3 个体特征对比分析

3.3.1 节点权重对比分析

频次加权和贡献度加权分别采用作者参与发表论文的绝对发文量和按照作者署名顺序加权后的相对发文量作为节点权值,相对发文量 $\leq$ 绝对发文量,二者虽然都是反映作者科研能力的指标,但相对发文量更能客观真实地反映作者学术造诣的“真实”水平。图4为样本数据中123位作者的发文量对比分析结果,二者均呈现“长尾分布”现象,绝大多数作者的发文量都处于较低水平。利用SPSS统计样本数据中绝对发文量和相对发文量的相关系数为0.936,呈显著正相关关系,因为虽然按照作者署名顺序确定每位作者的权重,在绝对发文量提高的情况下,相对发文量也根据署名顺序有不同程度的提高。

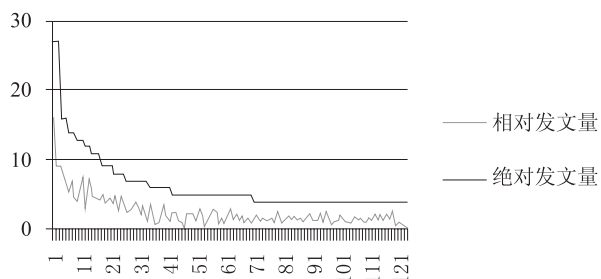


图4 发文量对比分析图

3.3.2 边权重对比分析

作者合作网络的边连接分析从最初的二值化数据到根据合作频次设置边权值,有效地确定了边链接强度,挖掘了更多的网络特征^[13]。对边的频次加权和贡献度加权后权值的关系如同绝对发文量与相对发文量,

加权规则有所不同,同样二者也具有显著正相关关系。合作次数越多合作关系越紧密,但是随着次数的增加,新的合作关系对作者间关系紧密程度的贡献递减,呈边际效用递减规律,且合作次数与紧密程度存在非线性关系。

3.4 网络静态特征对比分析

首先,分析样本数据的节点强度及强度分布。样本数据中123个节点间共存在231条边,频次加权与贡献度加权两种方式的节点强度分布分别如图5(a)和5(b)所示,二者均呈现幂律分布的特性,大部分作者与其他作者存在弱合作关系,仅有少量作者与其他作者存在强合作关系。图5(a)中靖继鹏和孙建军的节点强度都为15,其中靖继鹏与8个作者存在合作关系,孙建军与5个作者存在合作关系,整个网络的平均节点强度约为3.89,说明每个作者平均与其他作者存在3.89次绝对合作次数。图5(b)中由于节点强度的精度较高,分布比较分散,大部分节点的节点强度仍较低。整个网络的节点平均强度约为1.66,说明每个作者平均与其合作对象有1.66次的加权合作次数,样本数据源有限的情况下该值属于中高水平,主要是因为作者的合作对象较少和绝对合作次数较多。此外,对节点度、节点度分布、节点强度、节点强度分布等定义进行推广还可以定义边权、边权分布等内容,由此可以得出整个加权网络的拓扑特征,分析方法类似,本文不再赘述。

其次,分析样本数据的最短路径。利用前文所述方法将两种加权方式的边权归一化处理后的距离可视化展示分别如图6(a)、图6(b)所示,与图3(a)、图3(b)中利用结构相似性衡量作者关系强弱的方式相反,二者从相异性的角度衡量作者间的路径长度,对比

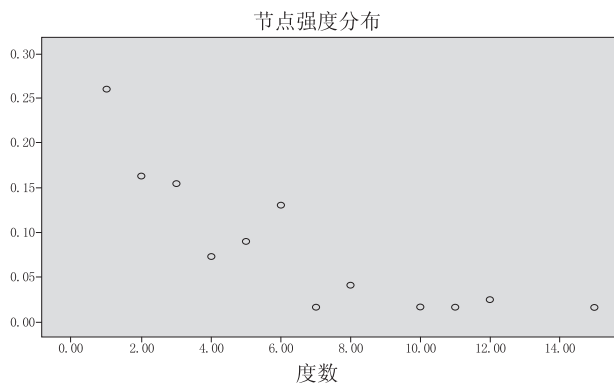


图5 (a) 频次统计加权节点强度分布图

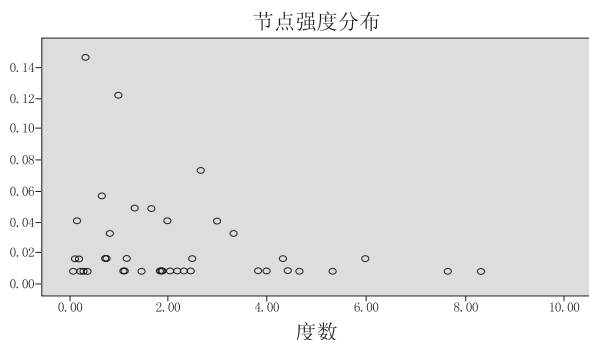


图5 (b) 科研贡献度加权节点强度分布图

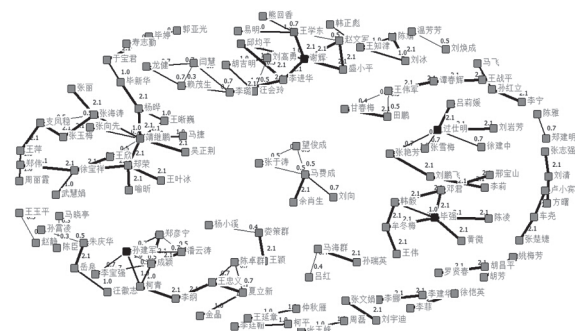


图6 (a) 频次加权距离图

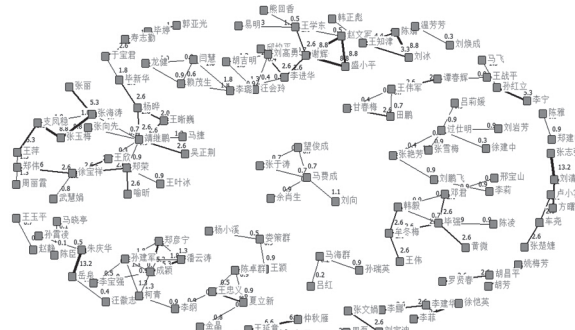


图6 (b) 贡献度加权距离图

可知图6 (b) 更能准确反映作者间的合作路径长度, 进而可以求出任意二者间的最短路径长度以及在最短路径的基础上计算平均路径长度、介数、效率等网络中的其他全局变量。

再次, 分析样本数据的聚集系数。利用公式 (5) 计算两种加权方式中整个网络 (G) 和以毕强 (G1)、过仕鹏 (G2)、马费成 (G3) 为核心节点的三个子网的聚集系数如图7所示, 两种加权方式中整个网络的聚集系数都很低, 主要是因为网络中存在大量双核型子网和大量节点的连接边较少, 说明整个网络中节点间的紧密程度不高; 与G1、G2网络相比, G3网络的聚集程度较高,

说明G3网络中内部连接较为紧密, 因为G3网络规模较小且单边连接的节点相对较少, 而G1、G2网络呈现很强的星型网络特征, 因而网络聚集系数较小。

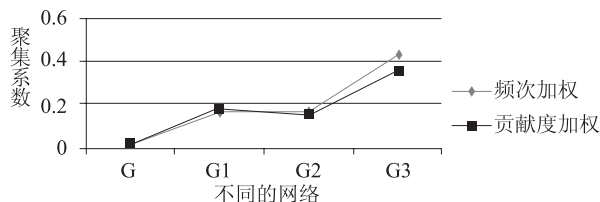


图7 聚集系数对比分析图

最后, 分析样本数据的“富人俱乐部”系数。由图3 (a)、图3 (b) 可知, 样本数据中几大核心节点间均无直接合作关系, 说明不存在富人俱乐部现象。高速公路、动车、高铁等新型交通运输方式的推广缩短了科研人员间的相对距离, 有线、无线等网络技术的发展促使科研人员交流方式多样化, 客观上为跨区域、跨机构的合作提供了便利, 在一定程度上提高了跨区域、跨机构合作的论文比例, 但是机构内部合作仍然占主导地位, 特别是网络地位较高的“富节点”也往往选择机构内师生间、同事间的合作方式。

4 结语

社会网络分析方法为作者合作关系分析提供了新的研究视角, 目前已普遍应用于各学科领域的作者合作网络分析中, 并已经形成了固定的研究思路和研究范式。本文通过实证对比分析验证, 在作者合作网络分析中利用科研贡献度加权能够在一定程度上解决耦合频次加权造成的虚假贡献问题, 然而本文在分析过程中仅分析了加权网络中节点的个体特征和静态特征, 针对作者合作网络的动态特征及动态网络演化机制, 需要对其进行进一步的后续研究。

参考文献

- [1] 孙丽娟. 科技论文作者署名排序与通讯作者[J]. 中国科技期刊研究, 2005(2): 242-244.
- [2] 沈耕宇, 黄水清, 王东波. 以作者合作共现源数据的科研团队发掘方法研究[J]. 现代图书情报技术, 2013(1): 57-62.
- [3] 汪冰. 中国情报学期刊论文合著现象研究与思考[J]. 情报科学, 1992(2): 8-16.

- [4] 邱均平,王菲菲.基于SNA的国内竞争情报领域作者合作关系研究[J].图书馆论坛,2010(6):34-40.
- [5] 郑曦,孙建军.链接分析领域的作者合作网络及其分析[J].图书情报工作,2009(2):29-33.
- [6] 李亮,朱庆华.社会网络分析方法在合著分析中的实证研究[J].情报科学,2008(4):549-556.
- [7] 孙丽娟.科技论文作者署名排序与通讯作者[J].中国科技期刊研究,2005(2):242-244.
- [8] 苏新宁.中国人文社会科学学术影响力报告[M].北京:中国社会科学出版社,2007.
- [9] 王星.大数据分析:方法与应用[M].北京:清华大学出版社,2013.
- [10] 苏凯,汪李峰,等.一种灵活的加权复杂网络演化模型及其仿真[J].系统仿真学报,2009(1):266-272.
- [11] 姚尊强,尚可可,等.加权网络的常用统计量[J].上海理工大学学报,2012(1):18-26.
- [12] 吕天阳,谢文艳,等.加权复杂网络社团的评价指标及其发现算法分析[J].物理学报,2012(21):1-10.
- [13] 孙海生.基于加权网络模型的国际科学合作研究[J].情报杂志,2014(7):91-97.

作者简介

明均仁,男,1981年生,博士,副教授,研究方向:数字信息服务与文本挖掘等,E-mail: ming_junren@qq.com。
党永杰,男,1988年生,博士研究生,研究方向:电子商务、数据挖掘。

Comparative Study of Author Collaboration Network Based on Scientific Contribution Degree

MING JunRen^{1,2}, DANG YongJie³

(1. Institute of Scientific and Technical Information of China, Beijing 100038, China; 2. School of Management, Wuhan Institute of Technology, Wuhan 430205, China;
3. Department of Computer, Central China Normal University, Wuhan 430079, China)

Abstract: For the problems in the analysis of co-authors network, like over-estimation of authors' scientific contributions to the articles and the ignorance of single article's contribution to the whole network edge weight, we separately use author signature in order to measure network nodes and thoroughly cover authors' relationship for edge weight by citing permutation and combination ideology. We construct the author cooperation weight network on the core authors in Information Science published papers in 2008-2013, and analyze the network's characteristics from four perspectives such as node strength, shortest path, clustering factor and "rich man's club". Then we analyze and validate the necessity and feasibility to construct author cooperation weight network model by contrast to entitled network related features. The results show that weight network model can more truly reflect authors' personal and their network structural characteristics.

Keywords: Weighted Network; Network Character; Author Collaboration Network; Complex Network

(收稿日期: 2015-12-31)