机构合作网络的特征挖掘及演化分析*

□ 温婉婷 吴斌 王柏 / 北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876

摘要：文章应用复杂网络的相关知识，对国内的医药学文献数据进行数据挖掘。以机构为研究对象，通过构建机构科研合作网络，对网络的静态参数、拓扑结构、动态演化进行挖掘分析，找出机构间科研合作网的静态特征，并以年为单位切分时间片，分析网络的动态演化特征。通过研究得出机构合作网络的静态参数，同时发现，机构科研合作网络有明显的局部化特征，它的主网络是一个小世界网络，具有无标度特性。机构的影响力和活跃度不仅体现在发文量上，同时也体现在与其他机构的合作程度上。

关键字：复杂网络，科研合作网，机构
DOI：10.3772/j.issn.1673—2286.2010.08.006

1 引言

自然界中存在的大量复杂系统都可以通过形形色色的网络加以描述。一个典型的网络是由许多节点与连接两个节点之间的一些边组成的，其中节点用来代表真实系统中不同的个体，而边则用来表示个体间的关系，规则是两个节点之间具有某种特定的关系则连一条边，反之则不连边，有边相连的两个节点在网络中被看作是相邻的。近几年，由于计算机数据处理和运算能力的飞速发展，科学家们发现大量的真实网络既不是规则网络，也不是随机网络，而是具有与前两者皆不同的统计特征的网络。这样的网络被科学家们称作复杂网络。

现在的科研形式越来越倾向于合作而不是分裂，这种合作关系可以被抽象成一个复杂网络。合作网络的形式有很多种，常见的有作者合作网络和机构合作网络。作者合作网络体现了作者之间的合作与交流。与作者合作网络分析不同，对机构合作网络进行分析，可以从更宏观的角度把握科研合作网的发展方向。当今时代，学科之间渗透日益扩展，使科学技术领域日益扩大，研究开发向纵深发展，使得一些大科学研究项目也越来越具有广域性和交叉性。越来越多的机构意识到科技合作对科研发展有利，科研机构十分重视研究合作，特别是近十年来，科技合作与交流空前活跃，且走向深入。因此，对机构之间的合作网络的分析研究是必要的，并可以更深层次地理解研究合作的意义，对合作的方向和方式起指导作用。

在本文中，一个机构是指一个医药类的科研单位，包括医学院、研究所、医院单位以及其他有关医药科学的单位。在文献领域里，机构像作者一样，可以作为一个发文单位而存在，机构的发文情况可以用机构内作者的发文情况予以统计。每个作者发表的文章都将隶属于机构的发文，两个不同机构的作者的合作关系可以看作是这两个机构之间的合作。

本文以机构为研究主体，以国内十年的医药学文献数据为研究对象，运用复杂网络的分析方法，着重研究机构本身的量化特征以及机构间的科研合作关系，并分析随着时间演化，机构自身和机构的合作发生的变化。

2 相关工作

近年来，通过引入网络分析的方法对文献数据进行分析正受到越来越多专家学者的青睐。人们研究作者合作网络已经有很长时间，最早可以追溯到1960年Price和Beaver[1]使用作者合作网络研究社会结构和科研合作网的影响。Newman通过分析不同学科的科研合作网络，发现数学家喜欢独自或和很少几个人合作发表论文，而

* 本文得到国家自然科学基金项目（90924029）和国家“十一五”科技支撑计划项目（2006BAH03B05）资助。

物理学家更喜欢和更多的同行合作[2,3]。

Erjia Yan等人针对中国LIS（library and information science）数据的研究发现，作者合作网络是一个小世界网络，并具有无尺度特性，作者的中心度与其被引用次数有很大关联[4]。

彭奇志针对的科研机构的评估结果认为，科研机构与其他单位合作发表的论文比例愈高，说明其作者横向科研能力愈强。在国外合作中，涉及到的国家和地区愈多，说明作者在国际横向科研合作方面的能力愈强[5]。

然而，针对机构的科研合作网的研究还有待进一步加深。张鹏、王继民、王建冬针对我国数字图书馆的机构合作网络分析得出，机构合作网络松散，机构间的合作关系显现出很强的地域性特征[6]。

Zhang Jian、Chen chaomei等人通过研究科研合作网络的网络尺寸、聚集系数、子图规模、平均距离等参数已得出结论：最近几年，机构之间的两两合作越来越普遍[7]。因此，研究机构之间的合作关系网络也有了更广泛的意义。

3 数据预处理

实验数据集是从1999年到2008年国内发表的医药学文献。在总计2787111篇文献中，出现机构239991个，网络内含477113条边。其中，1999年和2008年数据不完整，因此只展示结果，不分析。

数据中每条记录代表一篇文献数据中的一个作者，包含文献编号、作者姓名、所属机构。文献编号是一篇文献的唯一标识，包含期刊编号、年份信息和文章编号，从这里可以提取出文献所属期刊以及发表日期，供分析所用。数据中所属机构一项所填均为规范的机构名称，但由于机构名称在10年内或有改动，因此做初步机构名称的规范。例如同济医科大学在2000年改为华中科技大学同济医学院，两个机构名称实属同一机构，而数据中包含1999年数据，当时机构名称并未改

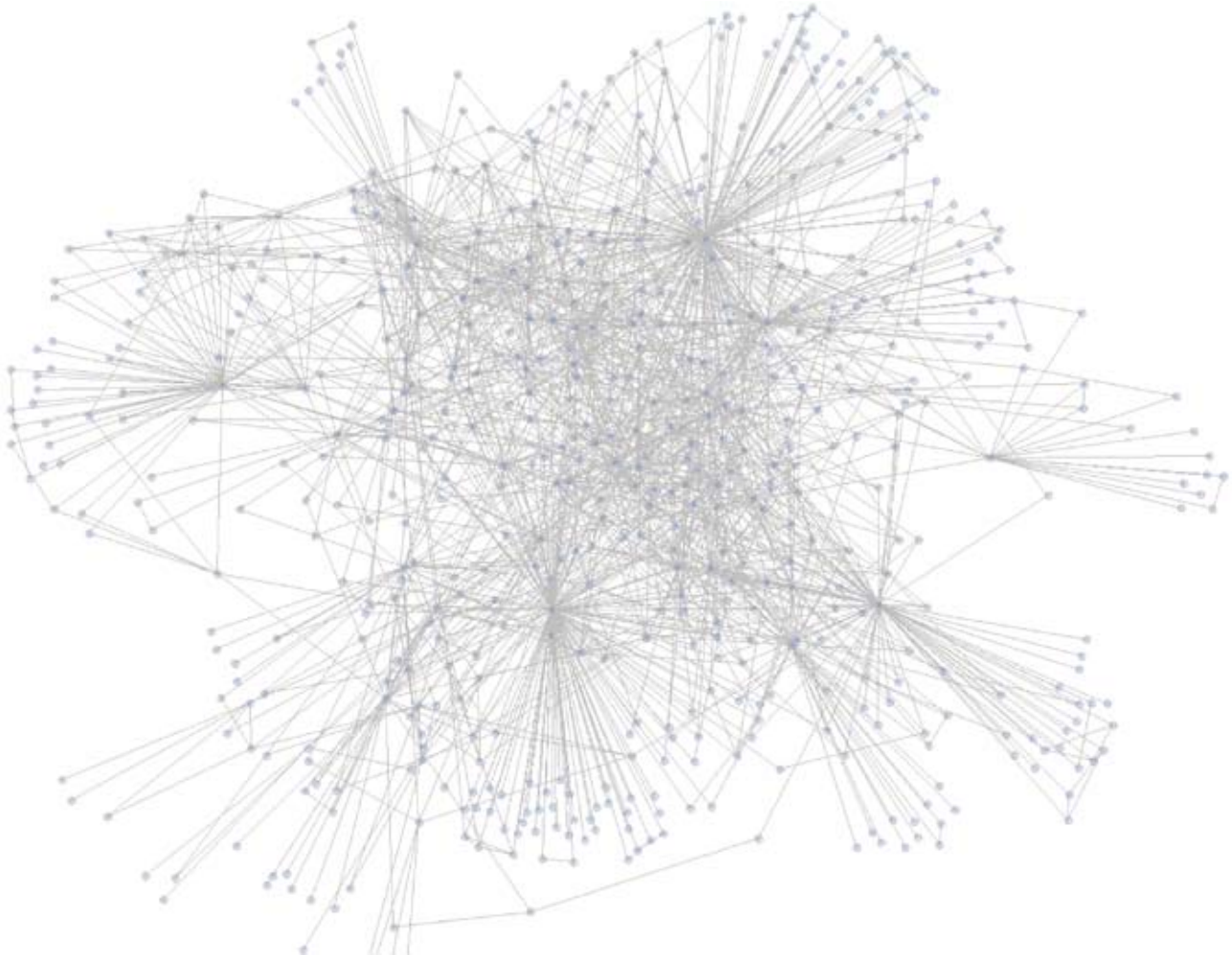

图1 机构科研合作网络

变，因此在统计时需要将这两项数据进行合并。

首先提取每一篇文献的年份信息，按年份分为十个子集，每个子集包含一年的文献数据。然后，在每一年的数据集中，如果两个作者所属的机构合写了同一篇文章，则这两个机构之间建立了合作关系。以此合作关系可以形成一个机构的科研合作网络。按时间建立十个机构的科研合作网，可分析网络的特性和演化状况。图1是一个机构科研合作网络，每个点代表一个机构，机构之间的连线代表机构之间有过合作。

4 机构合作网络的演化分析

表1总结了机构科研合作网的基本统计特征。从最后的全部数据来看，在聚集系数方面，全网的平均聚集系数和最大子图的平均聚集系数都比各年的平均聚集系数要高，体现了全网的合作要比各年份的合作普遍。在子图的分布上，由于机构数量较多，其中很多在网络中为孤立点，因此连通分量的个数较多。但从孤立点所占比例来看，比大多数年份的孤立点比例要小，而且从最大连通分量的规模来看，也比大部分年份的最大连通分量规模大，说明十年内，机构之间的合作是有所变化和扩展的，每年合作的机构不会完全相同，慢慢地演化成了全网的合作。全网的最大的极大团计算的空间复杂度和时间复杂度太高，因此未作统计。

1) 静态特征分析

表1 基本统计特征

Y	Total	V	E	Cc	MCc	Com.	ISOL	GCC	Max.clq
1999	44638	13818	14085	0.17273413	0.303645697	5862	36.68%	49.97%	16
2000	125834	29476	37414	0.18330929	0.305683227	11618	35.17%	55.15%	20
2001	264042	42459	62392	0.17429398	0.295811606	16846	35.75%	55.46%	18
2002	286842	40966	65012	0.181843141	0.298212704	15425	34.29%	58.16%	22
2003	330240	41964	70614	0.18410263	0.296676627	15384	33.29%	59.11%	19
2004	351996	41888	72558	0.187017572	0.298394818	15226	33.09%	59.55%	17
2005	359778	51488	82737	0.179287358	0.307515532	20769	36.68%	55.10%	30
2006	395643	55993	89916	0.186287072	0.317550515	22518	36.56%	55.10%	23
2007	390691	65712	90057	0.16653266	0.316518696	29884	41.60%	49.71%	23
2008	237407	51117	59061	0.158223467	0.32081235	25168	44.96%	45.30%	22
全部	2787111	239991	477113	0.203219791	0.330949752	85454	33.31%	56.71 %	

注：Y表示年份，Total表示文章总数，V表示点，文中表示机构数，E表示边。Cc表示聚集系数，MCc表示最大子图聚集系数，Com.表示连通分量个数，ISOL表示网络中的孤立点占整个网络节点的比重，GCC表示最大连通分量包含的点占整个网络节点的比重，Max.clq表示最大极大团大小。

A. 网络结构分析

i.聚集系数（cluster coefficient）经常被用来描述网络的传递性[3]。举例来说，在你的朋友关系网络中，你的各个朋友很可能彼此也是朋友，这种属性成为网络的聚集特性。从实际上讲，它表示假如AB间有一条边，BC间有一条边，则AC间有一条边的概率。公式如下：

$$C_i=\frac{2\times E_i}{d_i\times(d_{i-1})}$$

从表中可看出，机构科研合作网络的整体聚集系数维持在0.15到0.19之间，但最大子图的聚集系数维持在0.3左右，相较其他网络略高。

ii. 极大团的大小和数量能表现网络的局部化特征。极大团的定义如下：对于图G =(V,E)，$\exists V' \subseteq V$，如果顶点集V'导出的子图G′ =(V′ ,E′)是完全图（完

全图是指每对顶点之间都恰连有一条边的简单图），则称G'为图G中的团；如果∃v∈V且v ∉V'，使顶点集{v}的导出子图G''=(V' ∪ {v}, E'')是完全图，则称G'为图G中的极大团[8]。换句话说，极大团是指网络中不能再被分割为子团体的最大节点集[9]。在10年的医药文献数据集的机构合作网络中，最大的极大团规模为30，最小的极大团规模也有16，显著高于其他网络，例如电信网络。这表现出机构科研合作网络的局部化特征明显，部分机构之间互相合作非常紧密。

iii. 从机构科研合作网络的子图分布与最大子图规模来看，基本保持在50%的机构是相互关联的，这个程度并不算高。而从网络的孤立点比重来看，相当多的机构则几乎孤立或形成各自的小团体，游离在主团体之外。

从上述三个特征可以分析出，机构科研合作网络整体上联系比其他网络紧密，但是只限于主要机构间的联系，仍有将近半数的机构习惯于单独发文，不与或很少与其他机构合作。

2) 动态特征分析

A. 个体演化特性

i.节点度分布：节点的度是最直观的描述一个节点的重要程度的指标之一。结合机构合作网络，它代表了一个机构的学术交流情况。图2和图3以每个机构作为分析节点，展示节点的度分布情况。图2为机构的度与相应的机构数量关系图。图3为机构的合作率与相应机构数量的关系图。因为每年的机构数量不同，因此将每个节点的度除以总节点数当作机构的合作率，作为统一的衡量指标。在两幅图中都可看出，大部分的点的合作关系非常少。在左图中，节点的度在逐年增加，度大的点也相应增加。而右图中，机构的合作率反而随着年份的推移而减小。这表明了整个网络中，机构的总数是在呈上升趋势，然而机构之间的合作紧密程度却跟不上机构的增加速度。每年有大量新的机构发文，但它们不与其他机构合作，或只是在局部范围内合作。同时，通过左图的观察，发现机构合作网络的连接度分布函数也近似呈现幂率分布的特性，表现出它也是一个无标度网络。

ii. 发文量分布：机构的发文量很大程度上代表了机构的学术能力。图4和图5为机构的发文量分布图。图4为机构的发文量与机构数量的关系图，图5为各机构的发文比率分布图。从图4中可以看出，各年份发文量在10篇以下的机构占了大多数，而2000年到2001年

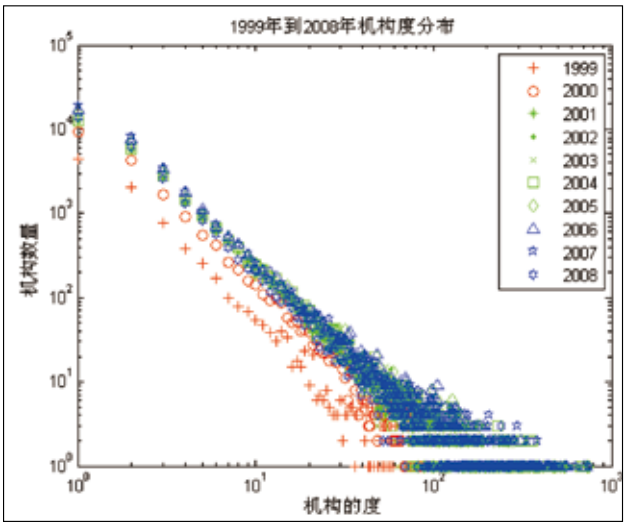

图2 节点度分布

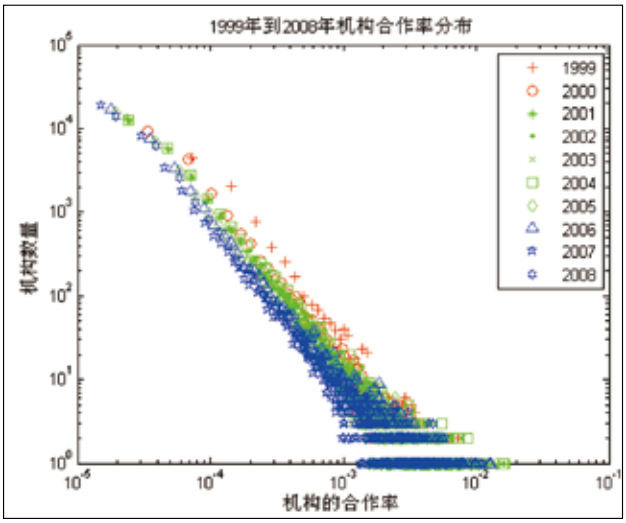

图3 节点合作率分布

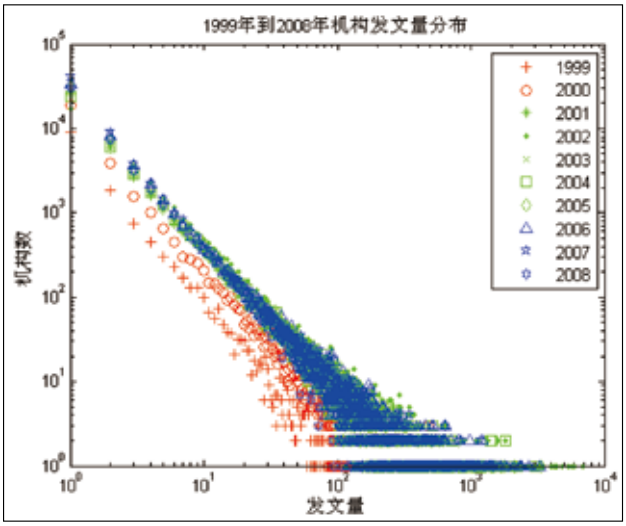

图4 机构发文量分布图

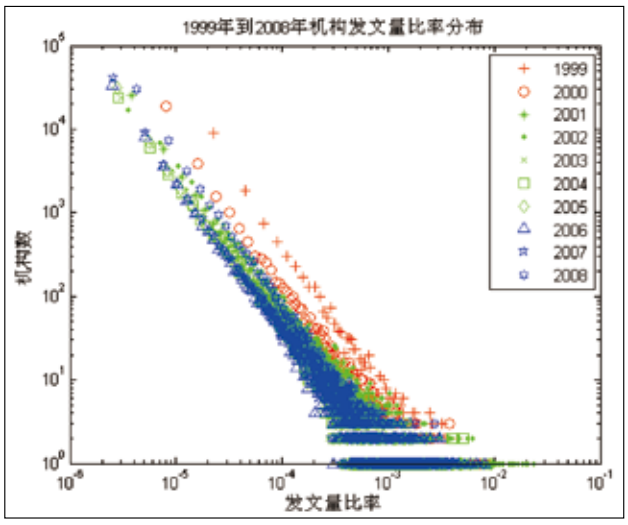

图5 机构发文量比率分布

间，机构的发文量有了比较明显的跃变。从图5来看，每年的机构发文量比率的分布大致相同，且都呈现幂率的分布特征。

iii. 机构与发布刊物分布：机构所投期刊的广泛程度从一个侧面反映了机构的学术影响力。图6为机构所发文章的期刊分布。从图中来看，多数机构所投期刊数在十个以下，而真正在很多期刊上广泛发表文章的机构屈指可数。

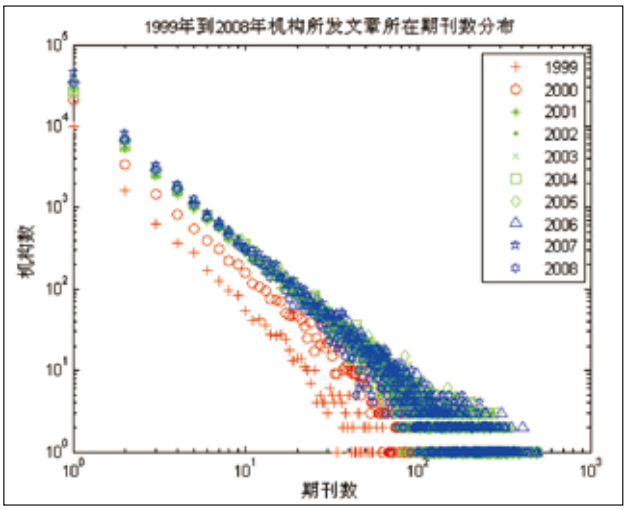

图6 期刊分布

综上三点，我们可以看出，随着时间的变化，发文的机构数量越来越多，发文量越来越大，所发文献的期刊也越来越广泛。而且三组图的分布都呈现出“重尾（heavy-tailed）”分布特征[10-12]，即存在相对少数极大的孤立点坐落于分布的尾端，但其对总量有很大的贡献。但是机构间的合作关系并没有跟上机构数量的增加，每年增加大量的机构单独发文，这些新进入网络的机构与其他机构联系很少，显示出机构科研合作网的特征是习惯以个体进入网络，再随时间的推移，慢慢与其他个体相互联系。

B. 群体演化特性

i. 合作紧密度演化：由前文统计可得，每年的机构合作网络中存在一个大的子图和若干孤立点以及小子图。本文以每年最大的子图为研究对象，此子图中度最大的点作为中心点，统计此点至少经过几跳，即经过几次合作可以到达子图中的任意节点，以此来作为衡量机构合作紧密度的指标。由图7统计显示可得，在经过3跳之后，合作率达到80%，在经过4跳之后，合作率达到90%以上，接近100%，而且有逐渐缩短距离的趋势，体现出大部分机构已经开始重视研究的相互交流和合作性。

同时，因为通过此中心点，有90%的节点可以在四跳内达到，则说明在最坏的情况下，最大子图的90%节点可以在8跳内互相到达，符合一个小世界网络的特征。

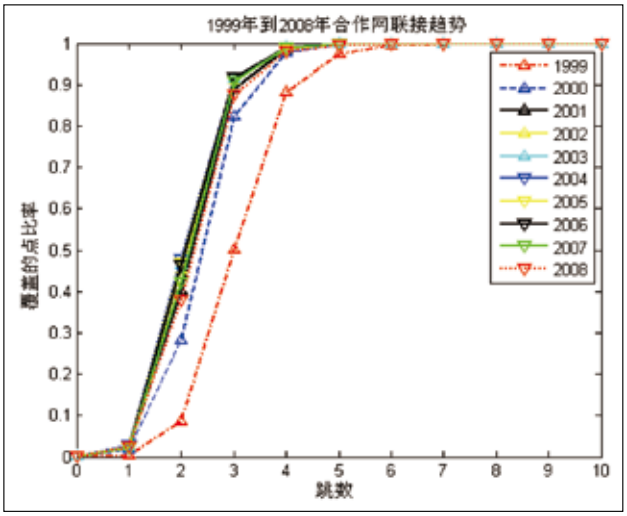

图7 合作紧密度

ii. 极大团分布图：从拓扑结构来看，极大团的大小和数量表现出网络的局部化特征。如下图所示，在2000年，多数极大团规模都在10个节点之下，大型的极大团数量极少。而从2001年开始，规模大于10的极大团有增长趋势，最大的极大团规模也越来越大，显示出网络的局部化特征越来越明显。

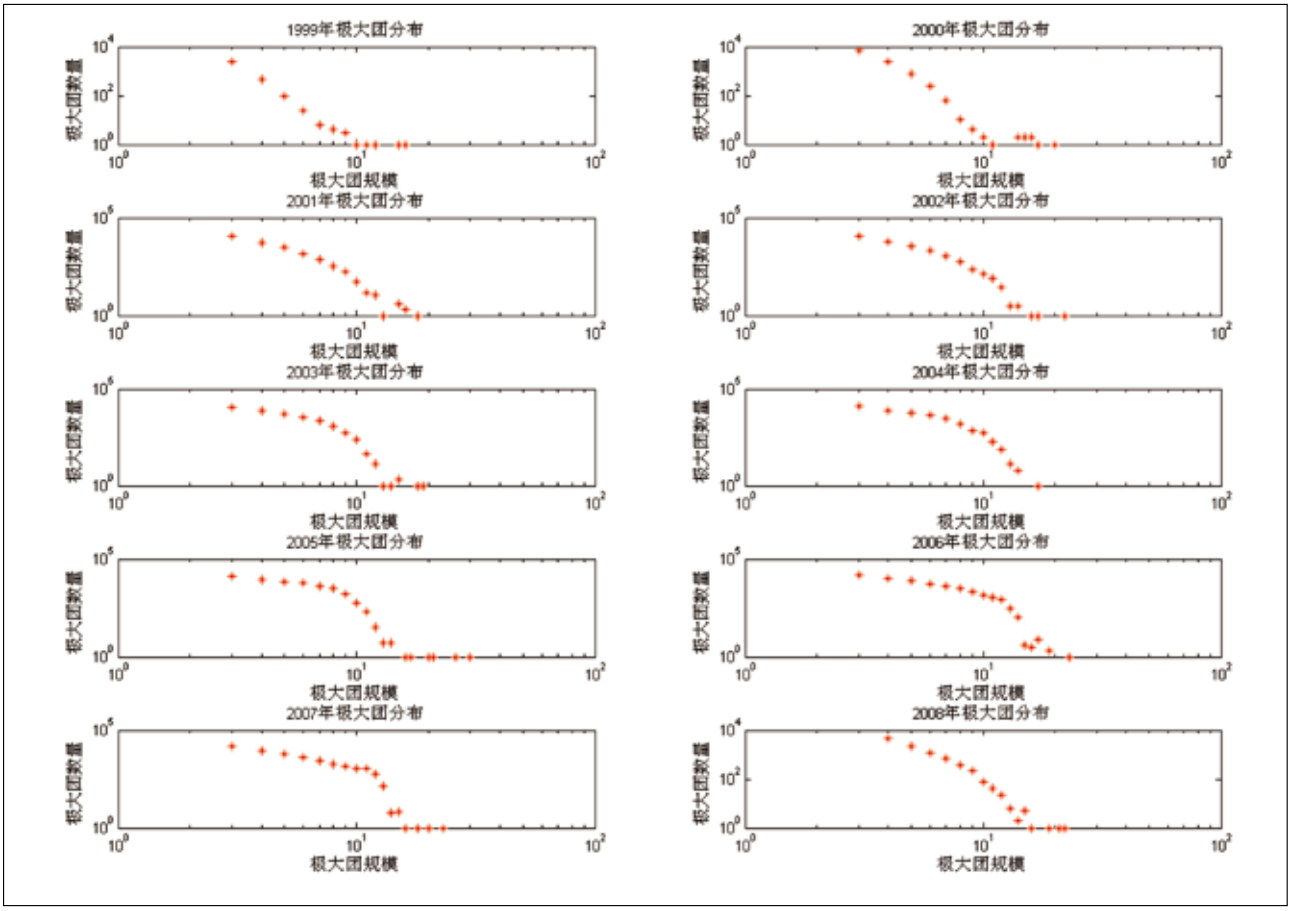

图8 极大团分布

iii. 子图规模演化：本文将孤立节点、最大子图以及剩余的零散子图当做三个节点集合，统计了三个集合节点数以及它们占全部节点的比例，并作出对比，以此找出机构科研合作网络的网络特性。

如图9所示，可以看出，十年内零散子图的规模比例十分小，占全部节点的10%左右。而有40%左右的节点为孤立点，即有40%左右的机构为单独发文，不与其他任何机构合作，余下50%左右的节点形成一个大的机构合作网络。这表明，机构之间的合作只占整体的一半，而另一半机构完全不与其他机构交流及合作。这将对科研机构之间的交流有一定负面影响。

然而，从图10来看，孤立节点的总体数量和规模最大的子图节点数量同步上升，说明每年有大量机构加入发文行列，旧有发文机构也慢慢从独立发文变成与其他机构合作发文，这种网络特征是机构间科研合作网的一大特性，既显示出机构间的合作与差距，又保证机构科研合作网的科研能力的提升。

5 典型机构分析

本节选取典型机构从个体和网络两方面进行分析。分析年份从2001年到2007年。1999年、2000年及2008年数据不全，因此不作分析。个体分析包括发文量和发文期刊数分析，网络分析包括度和聚集系数分析。选取机构为解放军总医院、北京大学、中山大学、华中科技大学同济医学院附属同济医院。

下表为全部数据的特定机构的文章数排名、期刊数排名和度的排名以及聚集系数的统计。

从文章数、期刊数和度的排名可以看出，解放军总医院和北京大学的排名都比较稳定且排在前面，说明其发文数量与合作程度都比较高。中山医院的文章数较低，但是期刊数和度排名则较高，显示出其发表的文章与较多的机构合作并且广泛发表在各个期刊上。华中科技大学同济医学院附属同济医院则与中山大学相反，文章数比较高，但是期刊数和度相对较

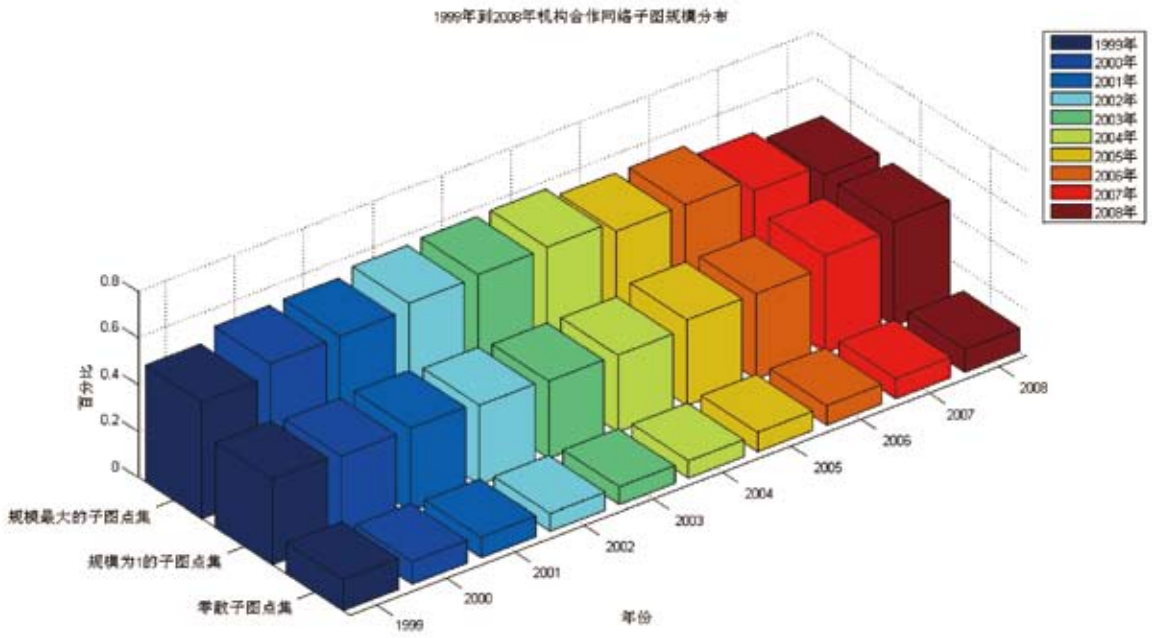

图9 子图比例

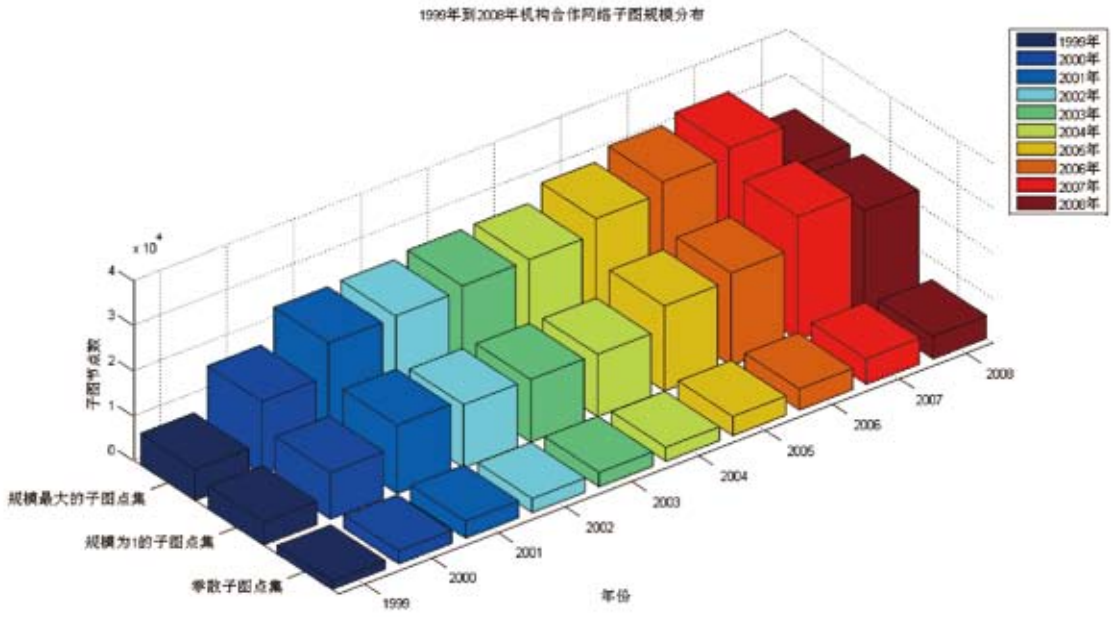

图10 子图规模

低，体现了这个机构更习惯在机构内部合作发文，与其他机构的合作相对较少。

同时观察各机构的聚集系数可以看出，度越大的机构其聚集系数反而越小，两者结合说明度大的点在其合作的机构中更容易处于核心地位。

A. 个体分析

图11为各年典型机构的文章数统计。观察可以得出各机构的发文数在2005年达到峰值。2005年前呈上升趋势，而2005年后呈下降趋势。解放军总医院在各年发文数量上都远远超过其他机构，北京大学和中山大学在排名和数量上都比较稳定，而华中科技大学同济医学院附属同济医院则呈逐年上升趋势，在2006、

表2 典型机构的整体统计数据

	解放军总医院	北京大学	中山医院	华中科技大学同济医学院附属同济医院
文章数（排名）	1	3	10	2
期刊数（排名）	1	2	3	5
度（排名）	1	2	4	7
聚集系数	0.012954684	0.014071624	0.017649624	0.025003178

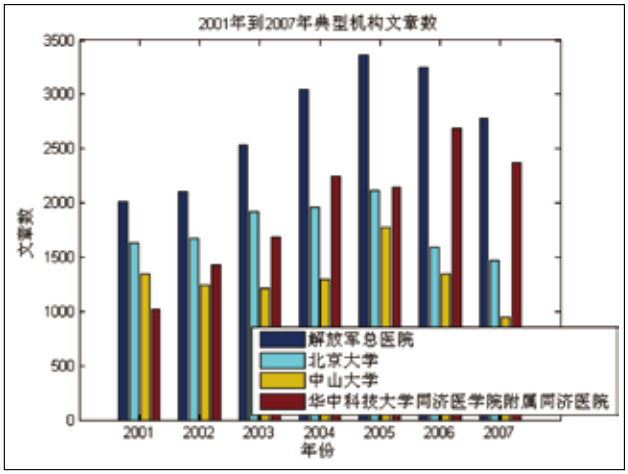

图11 2001年到2007年典型机构文章数

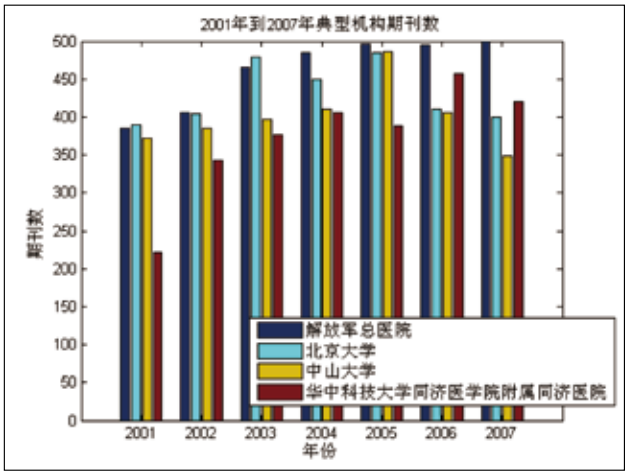

图12 2001年到2007年典型机构期刊数

2007两年已远远超过北京大学和中山大学。

图12为各年典型机构的期刊数统计。北京大学在2003年之前期刊分布数量居第一位，而解放军总医院则后来居上，在2004年之后超越北京大学。中山大学基本处于稳定状态。华中科技大学同济医学院附属同济医院也在2006、2007两年赶超北京大学和中山大学。

B. 网络分析

图13为机构的度统计。可以看出，2005年之前，北京大学的度最高，其次为解放军总医院。对比个体分析中的发文量统计和期刊数统计，可见发文量高的机构并不一定度也大，例如解放军总医院。综合来看，同时有大的发文量和广泛的发文期刊的机构，则其度的量也相应会高。从2003年之前的数据来看，即使北京大学的发文量远不如解放军总医院，但其期刊的分布更广，而合作的机构也更广泛，在网络中的表现也更活跃。而在2004年之后，解放军总医院的期刊分布超越北京大学，其度的大小也渐渐逼近北京大学，直至2006年超过它。

中山大学、华中科技大学同济医学院附属同济医院的度的大小则远远不及前面两个机构。但是华中科技大学同济医学院附属同济医院的度与北京大学和解放军总医院的差距渐渐缩小，体现出它也在逐渐扩大自己的合作范围，在合作网络中逐渐活跃起来。

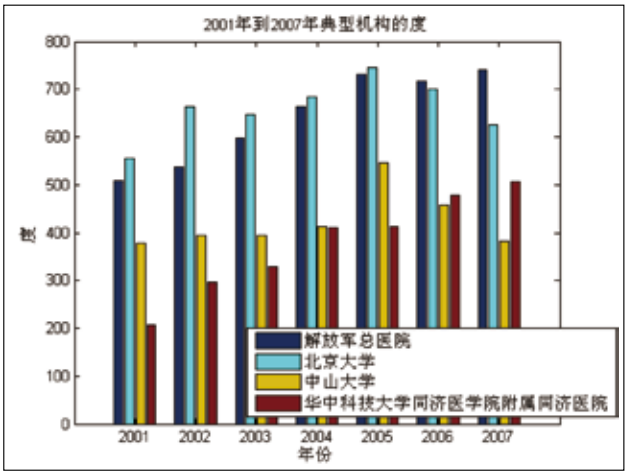

图13 2001年到2007年典型机构的度

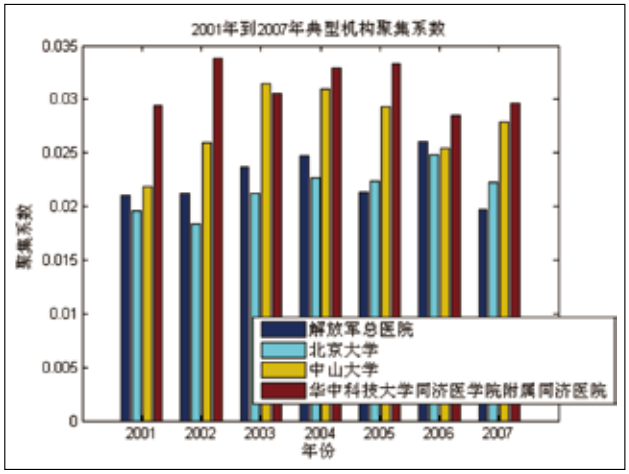

图14 2001年到2007年典型机构聚集系数

图14为典型机构的聚集系数统计。对比机构的度的统计可以看出，度越大的机构聚集系数更小。这说明度越大的机构的合作领域内的其他机构也互相合作的概率更小，也就是说其他机构通过度大的机构而互相联系的概率更大，从而使度大的机构更容易处于中心地位。

综合典型机构的个体分析和网络分析来看，机构的发文量不是衡量机构合作程度的绝对指标，拥有广泛的发文期刊量的机构更容易有更多的合作伙伴，成为网络中度更大的点，从而使自己处于网络的核心地位。

6 总结

本文以1999年到2008年10年间国内的医药文献为研究数据，从静态和动态两个角度，分析了网络中个体和群体的特征和演化规律。综合来看，机构合作网络是一个小世界网络，具有无标度特性，这与作者合作网络一致。同时，机构科研合作网的合作较其他类网络更紧密，但是也呈现出非常明显的局部性特征。这个特征既显示出机构科研合作网的合作交流水平在逐年提升，又显示出机构之间的交流合作只限于局部网络，近一半机构不与或很少与其他机构合作。而通过对医药界机构合作网络的分析发现，医药界机构的科研水平在不断上升，发文量和影响力都在不断扩大。通过典型机构的分析来看，机构的发文量不是决定机构影响力的唯一因素，广泛的发文期刊分布一定程度上反映了机构的合作领域的广泛性，而机构的合作机构数量，也就是网络中节点的度更加明确地体现了机构在群体间的影响力和活跃度。本文所用数据集主要是国内期刊，没有包含国际期刊，研究机构也集中在国内，下一步工作可以进一步扩大数据源，从更广泛的范围分析我国医药领域相关研究机构科研合作的特征。

参考文献

[1] PRICE D J D S, BEAVER D. Collaboration in an invisible college [J]. American Psychologist, 1966(21):1011-18.
[2] NEWMAN M E J. Coauthorship networks and patterns of scientific collaboration [J]. PNAS, 2004(101):5200-5205.
[3] NEWMAN M E J. Who is the best connected scientist? A study of scientific Coauthorship networks [J]. Physical Review E, 2001,64 (1): 016131.
[4] YAN E, DING Y, ZHU Q. Mapping library and information science in China: a coauthorship network analysis [J] .Scientometrics, 2010(83):115–131.
[5] 彭奇志. 基于SCI的科研机构学术成果评估与实证分析[J]. 情报杂志,2008,27(9).
[6] 张鹏,王继民,王建冬. 我国数字图书馆研究论文（2005-2009）的统计分析——社群分析[J]. 数字图书馆论坛,2010(3-4):120-127.
[7] ZHANG J, CHEN C. Collaboration in an Open Data eScience: A Case Study of Sloan Digital Sky Survey [J/OL]. Arxiv preprint arXiv:1001.3663, 2010 [2010-6-13]. http://arxiv.org/abs/1001.3663.
[8] CHEN AL, TANG CJ, TAO HC, YUAN CA, XIE FJ. An improved algorithm based on maximum clique and FP-tree for mining association rules [J]. Journal of Software, 2004,15(8):1198-1207.
[9] 汪小帆,刘亚冰. 复杂网络中的社团结构算法综述[J]. Journal of University of Electronic Science and Technology of China, 2009.
[10] SMALL H G. Co-citation in scientific literature: A new measure for the relationship between publications [J]. JASIS, 1973,24:256-259.
[11] ADLER R J, FELDMAN R E, TAQQU M S. A Practical Guide to Heavy Tails [M]. Boston: Birkhauser, 1998:111-130.
[12] ALBERT R, BARABÁSI A L. Statistical Mechanics of Complex Networks [J]. Rev. Mod. Phys., 1999(74):47-97.

作者简介

温婉婷，硕士研究生，研究方向为数据挖掘、复杂网络。通讯地址：北京邮电大学179信箱 100876。
吴斌，副教授，主要研究领域为数据挖掘、复杂网络及智能信息处理。通讯地址：同上。E-mail: wubin@bupt.edu.cn
王柏，教授，主要研究领域为电信系统软件、分布计算技术、数据挖掘。通讯地址：同上。

The Feature Mining and Evolution Analysis of the Institutional Collaboration Network

Wen Wanting, Wu Bin, Wang Bai / Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, 100876

Abstract: This paper has not only constructed the networks of institutional collaborations by using data of Chinese biomedical literatures, but also studied many statistical properties of the networks through the segment of network in years, such as static parameters, dynamic evolution topology. We found out the static features of the institutional collaboration network. The institutional collaboration network, which shows significant local features, is a “small world” network with the properties of scale free. Moreover, the influence and active state of an institution is reflected not only in the quantity of literatures of a institution, but also in the level of the collaboration with others.

Keywords: Complex network, Collaboration network, Institution

（收稿日期：2010-06-24）