中央财经大学

硕 士 学 位 论 文

基于统计研究领域的异质网络社区发现

Community Detection Based on Heterogeneous Network of Statistical Researches

作者姓名：王思雨

分类号_____ 密级_____

U D C _____ 编号_____

中央财经大学

硕 士 学 位 论 文

学位论文题目：基于统计研究领域的异质网络社区发现

姓　　名　王思雨

学　　号　2017210761

学　　院　统计与数学学院

学位类别：□学术硕士 ☑专业硕士 □同等学力

学科专业　应用统计

指导教师　潘蕊　副教授

第二导师

提交论文日期：　2019 年 5 月 25 日

独 创 性 声 明

本人郑重声明：所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中央财经大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名： 王思雨

2019 年 5 月 20 日

学位论文版权使用授权书

本学位论文作者完全了解中央财经大学有关保留、使用学位论文的规定。特授权中央财经大学可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校按规定向国家有关部门或机构送交论文和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名： 王思雨　　**导师签名：** 潘莎

2019年 5月 20日　　2019年 5月20日

摘要

异质网络（Heterogeneous Networks）有别于同质网络，指的是将复杂系统所包含的网络信息抽象成不同类型节点与不同类型的链接关系。目前绝大部分社区发现方法的对象都是针对同质网络的，并且假定网络结构的节点与连边关系都是相同类型的。这种假设为社区发现算法的研究提供了很大的便利。与同质网络的社区发现相比，基于异质网络的社区发现可以挖掘出更多的社区信息和更准确的社区结构。本文所研究的对象正是基于统计研究领域的学术论文数据构建的统计领域的多维异质网络，数据来源于 Web of Science 论文数据库网站。

首先，本文介绍了包含三种异质节点的统计研究领域网络的搭建方法。然后，基于拓扑关系与节点属性，较为详细地讨论每层网络中节点之间的相似性。其次，为了发现其中相对紧密稳定且具有现实意义的社区结构，本文采用 LDA 主题模型和标签传递算法等社区发现方法对统计研究领域的三维网络进行了多维结构、多模信息、语义信息以及链接关系等特征的提取分析和建模表示。

通过实证分析，本文所提出的方法发现了结构相对紧密，具有较高的模块度，包含三种异质节点且拥有明确实际意义的多维科研社区。其中包括古典统计，经济金融统计，生物医药统计和理化统计的 4 个期刊社区，4 个期刊社区下的 20 个论文社区以及在 20 个论文社区下的 92 个合作者社区。发现的科研社区具有较为明确的主题意义，所以可以为科研工作者推荐同一领域的研究刊物，研究成果和研究合作者，并可以帮助研究者结合自身的情况把握未来的研究方向和进程。本文所提出的方法和传统同质社区发现方法对比，能够充分利用统计科研网络中的异构信息和节点属性信息，能够发现重叠社区并很好地处理多维复杂关系，从而提高算法的准确性。此外，本文提出的方法可以通过并行处理使该社区发现操作的时间复杂度下降。

关键词：异质网络；社区发现；重叠社区；LDA 主题模型；标签传递算法

Abstract

Heterogeneous network is different from homogeneous network, which transforms the network information extracted from the complex systems into the different types of nodes and links. At present, the majority of community detection methods are worked on homogeneous networks and such hypothesis that the nodes and the links of a certain network structure are all the identical type is made. Such hypothesis provides much convenience for the study on the community detection method. Compared with the community detection in homogeneous networks, the community detection based on heterogeneous networks can find out more association information and more accurate community structure. The study object of this paper is just based on the heterogeneous network which is constructed by using the academic paper data in the statistics research field. The data is collected from the database website of 'Web of Science'.

Firstly, the method of network construction, containing three kinds of heterogeneous nodes in statistics study field, is introduced in this paper. What's more, on the basis of the topological relation and attributes of nodes, elaborate discussion for the similarity among the nodes in every layer is done. Secondly, through the community detection methods, such as LDA topic model and label propagation algorithm, multi-dimensional structure, multimode information, semantic information and link connection in three-dimensional networks will be extracted and utilized to seek out the community structures which are relatively close and have the practical significance.

In the empirical analysis section, by using the method presented by this paper, multi-dimensional scientific research communities which have relatively compact structure, higher modularity and explicit practical significance are found out, including 4 journal communities, 20 paper communities under those 4 journal communities, and 92 author communities under those 20 paper communities. The scientific research communities detected can provide science researchers with the journals, research achievements and research collaborators in the same field and can enable them to master the research direction and progress in future. Compared with traditional methods of homogeneous community detection, such as Louvain, the method proposed by the paper can make the most of heterogeneous information and node attribute information contained in the statistical science research network, find out the overlapping communities and cope with complex relationship of multi-dimensional structure well to improve the accuracy of algorithm. Moreover, such method can reduce the complexity of operation time for community detection with the parallel processing.

Keywords: heterogeneous network; community detection; overlapping community; LDA topic model; label propagation algorithm

目录

第一章　绪论

1.1　研究背景及意义

随着信息时代的来临，人们之间的合作与交流越来越密切，于是本文的社会被编织成了一个又一个的复杂网络。除了已被本文熟知的互联网、车联网和各种各样的社交网络以外，还有一个正在不断壮大的网络——科研合作网络。一方面，由于教育和科学技术水平的不断提高，越来越多的学者和组织机构加入了科学研究中，这使得承载着人类智慧结晶的科学论文在近几年里呈现出爆炸式的增长。另一方面，由于身处在信息化的现代，通讯手段的发达使得学者们之间的合作更加频繁，并逐渐突破了地理限制。在这样的科研背景下，人类新的知识和新的研究领域不断的涌现，研究者之间的论文合作网络逐渐变得错综复杂起来并趋于海量化。在如此庞大的论文合作网络之中，如何正确把握某一研究领域的研究现状和发展趋势，找到有前途的研究课题，甚至找到适合自己的杂志期刊以及志同道合的合作者成为了学者们共同关注的焦点。

本篇论文的主要目的就是通过构建 2014 年至今的关于统计与概率领域的论文科研合作的多维异质网络，来实现统计领域网络上的主题社区发现，进而指导和启发学者在统计与概率领域中发现更加有前景的研究主题，更加核心的作者、研究团队以及更加适合自己的杂志期刊。本文正是基于此类研究目的，在多维异质网络视角下，提出了一种结合节点属性信息的社区发现算法，并希望借助网络与网络之间连边关系，节点与节点之间的连边关系以及节点们的辅助信息，实现对论文主题社区更加准确、快速的理解和提取。为科研人员了解该领域的现状、获取更多有价值的论文以及制定该领域的研究方向提供更多的参考信息。

1.2　主要研究内容

1、基于 LDA 主题模型的杂志层社区发现。这一步主要针对的对象为杂志层的社区发现。LDA 主题发现模型是一种可以根据文本信息来提取主题并根据所提取的主题来计算每一个实体属于每个主题的概率。每一个杂志节点都包含 2014 年以来在该杂志上发表论文的摘要文本信息。利用每一个杂志节点所包含的所有摘要

文本信息来建立统计学领域的文献语料库以及字典，进而建立 LDA 主题模型。根据每个杂志节点对每个主题的隶属度来确定每本杂志的所属研究领域和研究主题，在本文之中也就是主题社区。

2、基于标签传播模型的论文层社区发现。根据层间连边的完全传递效应，先将论文层的节点划分为其所属杂志主题的研究领域子网络。在这些基本的研究领域子网络中对子网络论文节点实现并行计算来进行社区发现。每一个论文子网络中的节点都包含作者关键词，附加关键词，发表时间，摘要以及出版社等信息。综合这些节点属性，本文定义了论文节点之间的相似度，进而确定论文节点之间的传播概率。最后利用该传播概率矩阵借助标签传递算法实现每一个论文子网络的社区发现。并且最后集中子网络的研究结果来得到论文层的社区发现。

3、基于标签传递模型的作者社区发现。根据论文层和作者层的层间连边的完全传递效应将作者合作层划分为类似于论文社区的作者子网络。在这些合作者子网络中，利用并行计算进行合作者主题社区发现。每一个作者子网络中的作者节点都包含关键词，摘要，所在大学以及地址等属性信息。综合这些节点属性，定义作者节点之间的节点相似度，进而求得节点之间的传播概率。最后利用该传播概率矩阵和标签传递算法进行合作者的主题社区发现。

1.3 主要研究方法

1、统计科研论文数据的获取。利用 python 上的 beautifulsoup 和 selenium 库在 web of science 网站的 JCR 统计与概率研究领域排行模块上动态爬取 2014 年以来 27 个知名统计学杂志上的论文文献数据用以构建论文主题异质网络。

2、数据结构的创建与清洗。将每一篇论文进行特征提取和拆分，从而抽象成杂志，论文和作者等数据网络节点并加以储存。对于每个节点的属性数据结构，为了方便调用和建模，本文采用 json 树状结构加以储存。并且对每个节点属性的非结构化文本数据进行正则化清洗和整理，结构数据进行了格式上的统一，使得数据格式保持一致，方便建模的使用。

3、异质网络的搭建。统计领域的异质多维网络主要分为三层次，包括杂志层，论文层以及作者层。在三个网络层次内分别存在着杂志节点，论文节点以及作者节点。一方面，层内有着相同类型的边链接。论文层的层内连边为相互引用的关系。作者层内之间的连边则是论文的合作关系，并且该链接包含权重，权重是作者之间的合作论文数量。另一方面，杂志与文献，文献与作者之间具有层间连边。杂志与文献之间的连边代表着从属关系，也就是说文献与其发表的杂志之间存在着层间连边。文献与作者之间的连边代表着创作关系，也就是说文献与其作者存在着层间连边。很自然的本文可以看出，每种连边的类型都不同。于是，按照这样的理念，设计构成了一个三维统计领域的科研网络。

4、设计层与层之间连边的联系效应。一方面，由于针对统计科研文献网络的社区发现主要是发掘论文主题社区，所以一个具体的异质社区尽管包含的节点是异质的，但是节点的主题都应当是相同的。也就是按照某几个主题建刊的杂志，所包含的论文应该也是属于这几个主题的文章，而一篇文章的合作者们也应该是属于该篇文章的主题社区。另一方面，在三维异质网络中的上层网络对下层网络具有一对的多映射关系。所以主题社区的现实关系与三维网络的层间连边很相似。所以，本文有理由相信，在统计科研异质网络中的上层社区发现后的社区，可以通过层与层之间的链路，将社区信息完全传递给下层节点。也就是说下层节点由于两边之间的主题从属关系，会首先被上层的网络的社区发现结果圈定初始的范围，并划分为相应子网络。然后，会结合该层的拓扑结构和节点的属性信息进行进一步的分类和社区发现。

5、模型评估。结合了统计领域的现实研究状况，分析每一个多维社区的实际意义，来对社区发现的结果进行评估和验证。例如，对发现的每一个杂志社区的主题进行检验，来判别是否为相似的杂志，从而判断 LDA 主题模型的效果。除此之外，还进行了多维社区关于推荐功能的实际测试来判定是否对一个合作者选择研究方向，投稿杂志和研究合作者有帮助和指导作用。

1.4 论文创新点

1、在科研合作网，社交网络等许多现实的网络中，节点往往不止一种实体。但目前的大多数社区发现算法只关注单模静态网络，忽略了现实生活中网络之间也

会存在联系，也就是说网络的网络这一概念，因此会忽略掉很多信息，导致社区划分不准确。本文通过构建多维异质网络，进行社区发现，实现了异质网络的社区发现。发现的社区具有较为紧密的结构和丰富的现实意义，可以为统计科研研究者提供更丰富的信息。

2、在单层网络中，多数现有方法只能基于网络中的节点拓扑结构来发现社区，然而，现实网络数据除拓扑结构信息外，往往包含关于节点或边的辅助信息，仅利用拓扑结构信息会忽略很多现实意义，使得社区发现能力的下降。比如引用大量著名文章的论文，在拓扑结构上可能和学术经典距离比较近，只考虑拓扑结构，可能会被分到统一个社区。但实际上考虑了论文内容后，论文可能进入不了大牛论文社区。本文对每一层的网络节点信息进行了考虑，使得社区划分更接近实际情况。

3、社区发现效率的提高。首先在多维网络中，通过层间链接，上层会将下层的网络划分为众多的子网络，也就是说对下层的社区发现有着指导作用，会使下层网络在不用进行全局社区发现只进行局部社区发现的情况下，就有着很好的划分效果。其次，进行局部的社区发现，就可以在服务器上进行并行操作，提升运算效率。虽然会丧失一些全局的准确性，但因此会极大地提升大型网络社区发现的效率。最后，利用本文方法发现的社区，由于层间的连边关系，会具有上层的社区标签。社区的多个标签，会使社区更富有实际意义，有利于对社区的相关活动进行指导，使社区更加立体化，更加精细化。

1.5 论文结构和安排

第一章：绪论。介绍有关于多维异质网络的研究背景，指出构建统计领域的多维异质网络并进行社区发现的目的及意义。并且对于论文所使用的研究方法做了简单的介绍。

第二章：文献综述。一方面，在总结大量有关于社区发现和异质网络的国内外文献的同时，对社区发现的模式进行系统的梳理与比较，总结社区发现的历史研究框架与历程，并且对比了几种不同的社区发现方法在论文科研网络中的应用；另一方面，总结近年来在新兴的异质网络中应用较为广泛的社区发现方法，特别是基于科研论文数据的社区发现。

第三章：相关理论基础。在文本挖掘算法和自然语言处理方面，研究了 LDA 主题模型和 word2vec 等文本分析方法。在构造层内节点相似度方面，主要介绍

本文所利用的拓扑相似度构造方法以及属性相似度的构造方法。在社区发现算法中，本文主要介绍了所使用的标签转移算法及其适用性。

第四章：实证分析。依据所获得的 2014 年以来的论文数据构造统计科研领域的异质网络。并计算了各层内节点与节点之间的相似度以及转移概率矩阵。其次，基于论文摘要的应用 LDA 主题模型实现对杂志主题的抽取以及对杂志的分类。然后，对于与相关主题杂志有层间链接的论文子网络分别实施标签传递社区发现算法来发现“主题—论文社区”。最后，对于与相关论文有层间链接的合作者子网络实施标签传递社区发现算法来发现“主题—论文—作者”社区。

第五章：研究结论和展望。总结论文分析结果，得出一定结论，对比传统社区发现算法与本文社区发现方法。对实证和算法中存在不足的地方进行剖析，并指出可能存在改善的地方。对未来的进一步研究进行展望。对统计科研异质网络的探索给出一定的意见。

第二章　文献综述

社区发现作为复杂网络研究领域内的重要研究课题，对于认识复杂网络的结构以及其功能具有非常重要的指导作用。比如，社交网络中具有相同从属关系或者相同兴趣爱好的个人会自发组成社区以更好地交流和分享经验。又如，蛋白质的网络结构中有相互作用的氨基酸同样会构成相似的生物社区来负责相关联的物质的生产过程。因此，社区发现有助于网络研究者更加深入地认识社区本身的结构与功能，有助于提升本文对复杂网络整体的动力学行为特征理解，同时也为本文研究复杂网络的结构提供了一个相对中观视角[1]。

2.1 异质网络定义

在现实世界的复杂网络结构里，如物联网，社交网络，车联网以及电子商务等，既包含相同类型节点通过链接关系所形成的网络结构，同时也包含不同类型的节点通过链接关系形成的网络结构，此时抽象出来的复杂网络被称为异质网络。

对于一个包含众多信息的网络，如果属于网络结构中的节点类型数量 A＞1或者连边类型数量R＞1，这时的信息网络被称为异质网络。所以，也就是说，异质信息网络就是指网络结构中所含有的节点信息或是连边信息不是相同类型的。很明显，包含不同类型节点和连边的异质信息网络会含有更加丰富的网络结构信息。相较于同质网络来说，利用异质网络进行研究往往能够得到更为准确、更为丰富的信息。

通过上述对于异质网络的定义和解释，异质信息网络根据节点和连边类型的不同自然可以被分为多模网络(multimode network)和多维网络(multi-dimensional network)两种类型[2]。其中，多模网络更侧重于指网络中的节点是不同类型的，每一模便是指代一种节点类型。特别地，单模网络便是指前面所讨论过同质网络，它是多模网络中最简单基本的一种网络形式。

2.2 社区发现算法概述

复杂网络的社区发现在国外的研究起源比较早，最早可追溯到图论和拓扑学等[3]。国内针对社区发现的研究开始的比较晚，但研究成果和研究热度呈逐年上升趋势。首先，关于社区的定义方面，国内外学者提出了许多关于社区的定义，

但却始终未能够达成统一的数学模型，不过，这也在一定程度上为社区发现带来了方法和假设的灵活性[3]。Newman 等[4]认为社区内的节点一定是连接紧密的、而社区间节点之间的连接关系则相对松散。并且根据这种理念制定了社区发现的 Q 模块度。并将其作为判断社区发现算法的效果度量。Fortunato S 等人[5]表明社区可以按照某几种属性对网络结构进行的分类，并且由相应的节点和连边组成。陈清华等人[6]则考虑社区是一种由网络结构中的节点构成的内聚子图。该子图的内部节点间存在着大量的链接，而不同子图之间的节点链接相对较少。

在社区发现算法领域，目前国内的外学者们所提出的社区发现算法大多集中在了有关于静态网络的社区发现算法上。其中，最基本的是基于网络图的拓扑结构以及图分割理论的社区发现算法。其中，最具代表性的也是最经典的算法是 Newman-Girvan 算法[6]。这是一种基于节点边介数的网络分裂思想，通过寻找位于社区之间边介数较大的边并将其移除来使得社区结构清晰。为改善 GN 算法对于大型网络的计算效率不高的问题，Blondel V 等人[7]提出了 Louvain 方法。这是一种基于贪婪算法思想的一种节点凝聚算法，其算法思想是不断融合能使模块度 Q 增长最大的社区，直到网络中所有的节点都被融合为一个社区。Zhang P 等人[8]又提出了 edge-clustering 方法，其基本思想是不断地删除原网络中含有较小聚类系数的连边，同时改变邻接矩阵并重新计算所有边的聚类系数，直至网络中所有的连边都被删除，网络变成一堆节点。

与图分割不同，部分研究者基于图聚类理论来进行社区发现。例如 Kernighan-Lin 算法[9]，基于 Laplace 图特征值的谱平分法等[10]以及 Luxburg U[11]于 2007 年提出的 Spectral Clustering 方法。再有就是，Luca Donetti 等人[12]通结合聚类和谱平分方法，提出了一种基于谱平分方法的聚类算法。

除了以上的几种经典方法之外，应用比较广泛的是流分析方法。流分析适用于有向有权网络。基本思路是发现在网络中的某种流动（物质、能量、信息）所形成的社区结构。其中，比较经典的是 Pons P 和 Latapy M 于 2005 年提出的随机游走算法(Walk Trap)[13]，其基本思想是用两点分别到第三点的流距离之差来衡量两点之间的相似性，从而为划分社区服务。并且该方法能将相似性强的节点们分配进入同一社区。2007 年，Raghavan U 等人[14]提出了标签传递算法，其根本思想是对每个节点赋予一个初始标签来标志着其所在社区。经过每次迭代，每个节点标签根据其大多数的邻接节点的标签进行投票修改，当算法收敛后把具有相同标签的节点划分为同一个社区。为了适应各种实际情况下的社区发现，LP 算

法通过改进标签传递方式以及转移矩阵的计算方式的不同，发展出许多衍生算法，例如 S-LP 算法，H-LP 算法以及 BM-LP 算法等[15]。

然而，现实复杂网络的数据除了网络图的拓扑信息之外，还往往包括了关于节点或连边的辅助信息，所以仅仅通过网络的拓扑结构信息会造成社区发现的精度降低，效率变低，社区实际意义差。以被最广泛应用的 Newman-Girvan 方法为例。该方法对网络中的噪音非常敏感，且容易得到多个不同的社区发现结果[16]。所以，一些基于辅助信息并结合网络图基本结构的方法被提出，如结合拓扑结构和节点的辅助信息的 Biogeography Based Optimization (BBO)算法[17]和 ACM 算法[18]。孙怡帆等人[19]还曾利用节点辅助信息改进的 Louvain 方法用来进行微博有向网络的社区发现。还有一些学者利用流分析和节点辅助信息提出一些社区发现算法，如通过辅助信息改进 LP 算法得到的 MUM 算法[20]及其衍生的用于检测重叠社区的 OCD-MUM 算法[21]。闫光辉等人[22]还曾利用基于节点信息 LDA 的 LP 算法进行微博主题网络社区发现。也有一些学者通过图聚类和节点的辅助信息结合，探究一些社区发现算法。如 Huang X 等人[23]在 2015 年通过多值属性图中的子空间聚类问题，提出的 SCMAG 方法。不过，以上包含节点辅助信息的模型大多对节点信息的类型有一定的要求，比如，要求一些节点的信息是结构化的，因此不一定适合所有含辅助信息的网络数据。近年来，由 Tang Anh 等人提出的 SAC 方法由于其对辅助信息要求宽松的特点受到了越来越多研究者的关注[24]。SAC 方法对节点辅助信息类型限制相对其他方法较宽松，适合对包含连续型、离散型和文本型等多种变量信息的复杂网络数据进行社区发现任务。

至此，以上社区发现研究主要针对的对象皆为单维网络，即网络中的节点都为同质的。而当研究对象转变为多模多维网络时，节点的类型便会不尽相同，每一种类型的节点实体和连边也会多种多样，因此社区发现的算法，技术和概念会有很大的变化。王金龙等人[25]通过综合数据挖掘以及最优化理论，在基于学术文献的异质网络中应用最大化分割算法，划分出有着相近主题的科研社区。而 Deng C 等人[26]在系统地分析了在多维网络中挖掘隐藏社区的问题后，提出了一种发现多维网络中的多维社区结构的方法。Zhang H 等人[27]则提出了一种在动态多维网络(multilayer networks)中实现社区挖掘的静态视角。这是一种通过使用时间的信息对多模网络进行结构分析，并利用静态视角加以讨论，最后在脑电波多层网络中进行实证验证其有效性。

2.3 异质网络社区发现

异质网络中蕴含了丰富的文本信息，其中包括了与社区结构有关的大量的语义信息，越来越多的研究者倾向于采用文本主题模型的手段对其中的文本语义信息实现集成建模，用以提升社区发现结果的准确性和提高社区发现的现实意义。这些文本主题模型主要包括 LSA 主题模型(Latent Semantic Analysis)[5]、PLSA 主题模型(Probabilistic Latent Semantic Analysis)[6]以及 LDA(Latent Dirichlet Allocation)[7]模型等。其中 LDA 是近年来应用最广泛的主题发现模型，是 PLSA 模型的广义泛化。

近几年来，在异质网络社区发现的研究过程中，有关于排序的聚类算法逐渐发展，排序与聚类之间的研究手段相互促进、相互补充，相互增强，可以得到比较好的社区发现研究结果。在此类型方法中，RankClus[20]与 NetClus[21]算法是比较早提出的相对经典的方法。RankClus 算法提出了一种有关于异质网络排序并结合聚类的组合方法。该组合方法所基于的背景为“作者—会议”的双类型(bi-typed)异质网络的社区发现问题。根据作者和会议类别实现排序，并按照目标对象来确定聚类对象的向量，随后迭代调整每个研究对象的分类，最终获得较为准确的作者以及会议类型的划分。NetClus 算法则针对的是更为一般的异质信息网络结构——星型的网络信息结构。与 RankClus 的思想相同，NetClus 也是一个基于排序和聚类的迭代方法，也就是说，通过排序的结果来提升聚类的效果。但与 RankClus 明显不同的是，NetClus 能够对具有星型网络结构的任意数量社区类型对象进行处理。并且产生的聚类结果也并不是针对单个类型的网络对象的集合，而是拥有着相似网络拓扑结构的输入网络的子网络聚合。

又因为异质信息网络结构拥有着多模且多维度的特异性，所以研究者可以将异质的信息网络结构进行数据重构，并将其转化为相对简单的网络结构模型，进而实现异质网络中的社区发现。Liu 等人[27,28]提出了基于链接分析以及重构的方法，对多维异质信息网络实现重构。并将多维网络中的连边或超连边作为一类节点，从而把多维异质网络转换为二分图进行研究(bipartite graph)。

文献[30]创新性地提出了 GenClus 算法，从而解决了异质网络中常见的信息的缺失性与不完整性的现实问题，并且可以依据使用者和研究者的不同需要定义不同类型关系的重要性。

2.4 本章小结

从以上文献综述中不难看出，由于异质信息网络的固有特性和目前信息化时代爆炸式的信息增长量，异质网络的多维社区发现受到研究人员的越来越多的关注。目前已有比较多的研究方法，但是该领域还仍然处于一个探索阶段，具有比较大的发展前景。同时，对于该领域的深入研究也面临着挑战。第一，大量的复杂网络研究只是把社区发现的重心放在了网络的拓扑结构信息方面，忽略了网络节点属性等方面所提供的丰富信息；第二，绝大部分的研究只是停留在静态视角，然而真实网络是在时时刻刻变化着的，很多研究成果显然都不能追踪网络的动态变化过程；第三，现有的社区发现的研究成果大部分都是关于单模网络，忽略了网络之间也有相互连接的现实问题，并且缺乏网络与网络之间的相互制约和相互联系，忽视了网络之间的信息传递，因此忽略了一大部分信息；第四，对于多维网络的研究成果，很多都是想办法将多维网络进行压缩变换成单维网络，再利用普通和常用的静态社区发现算法进行探究。这样做显然能降低研究问题的复杂性，并且只要符合相对的场景也能获得比较好的研究成果，但这样也会造成算法忽略很多必要信息；第五，目前大多数对异质网络的研究只考虑到了层与层之间的对应关系，即考虑多层网络为二部图，忽略了层内连边与层间连边之间的相互联系；第六，大多数的静态社区发现算法的时空复杂度过高，因此无法在大型网络里很好地运行；第七，多数异质网络的社区发现算法的应用场景还是比较特殊，存在着一定的特异性。从上面的讨论中不难看出，在社区发现甚至异质网络的社区发现仍有很多的问题需要探究和解决。

第三章 相关理论

3.1 文本预处理

由于统计科研论文网络所包含的主要特征是摘要文本，其次才是一些结构数据，所以势必要对论文的摘要文本进行文本特征提取处理。计算机不能够像人脑一样阅读和理解文本的内容，因此要使计算机能够处理文本内容，则需要首先对文本进行预处理和特征提取，转换为计算机程序能够高效且快速利用的结构化数据。文本预处理就是从文本中提取并且过滤掉无用词汇，从而构成代表文本特征的关键字或者关键词汇。由于论文摘要所使用的语言皆为英文。且英文单词是独立的，所以分词直接采用空格分词。

3.1.1 对文本单词的预处理

首先，对于英文文本进行处理时，经常会出现具有缩略语的情况，比如‘won’t’或‘can’t’。对于这种问题的处理方式则采用正则表达式将这些缩略语进行补全。

其次，英文论文摘要文本中常常会有一些希腊字母，俄语字母，特殊符号等非英文字符以及数字、换行符、标点符号等等，这些字符都会严重干扰到论文的摘要分析，所以将这些非英文字符实施替换空格操作。因为空格这样不会影响英文的分词。

而且，英文单词会根据不同环境进行词形的变化，如动词的现在进行时和完成时以及形容词的比较级和最高级等。本文利用 nltk 自然语言处理模块中的词性标注功能对摘要中的每个单词进行标注后，根据其词性对单词进行词形还原。

众所周知，英文的在很多情况下会出现字母大写以及单词变化形式（如单复数以及时态变化）的状况，这会使在进行摘要文本分析时不能很好地区别和归类相同意思的单词。因此对于每个单词又进行了转化小写和词源变化，使每一个单词都变成词源形式。

在进行完上述文本处理后，将每一篇论文的摘要利用空格来切分进而转化为单词列表。至此已将每一篇论文的摘要都转化为了相应的具有实际意义的单词列表，这样处理能够使计算机更好的提取和识别每篇文章的特点。并且经过处理后的摘要的单词列表可以构成相应的语料库和字典。

3.1.2 过滤停词

一篇摘要文本的中心内容主要通过名词、动词和形容词等实词来展现。而虚词以及在各种摘要文本中经常出现的高频词汇对一篇文章的内容提示并无实际意义，这些无实际意义的字或是词在自然语言处理中被称为停用词。通常情况下上将停用词大致分为如下两类：

(1)论文摘要常用词。就是因为这类词随处可见，在所有主题类型中都会频繁的出现，使得这些词只含有与其它词相同的特征，因此没有区分类别的能力。比如“method”和“statistics”这两个词几乎在每篇摘要中均会出现，这样的词无法保证给出代表一篇文章的特征。

(2)一些虚词，包括语气助词、副词、介词或连接词等，通常本身没有实际意义，并且与主题信息没有丝毫关联,如常见的介词“of”、“in”等,指示代词:“that、those”，指代词:“he、she、it、we” 词汇等等。适当地降低摘要文本处理过程中的虚词出现的频率，可以有效地提高关键词出现密度，因此更能突出实词的具体意义和归类信息。

同时，为了节约词典的存储空间，提升每一篇摘要样本的特征，需要将停用词从文本词汇中剔除，令其不再参与 LDA 主题模型的搭建，进而减少噪音。去停用词的做法是依据停用词表和文本关键词中频繁的实词进行匹配。如果有词存在于表中，则表明该词为停用词或是无意义的词，应从文本词汇集中删除；如果词不在表中，则保留。

3.2 LDA 主题模型

主题，是一个摘要所蕴含的整篇文章的中心思想，也是论文内容的主体与核心[31]。同样的一个摘要文本既可以代表一个主题，也可以包含多个主题。主题表示着一个概念或者某个方面。从词的角度出发，就是一组词上的概率分布表示着某一特定的主题。LDA（Latent Dirichlet Allocation）主题模型是在概率隐性语义索引（Probabilistic Latent Semantic Indexing, PLSI）的基础上扩展而得到的三层贝叶斯的概率模型。LDA 文本主题模型的结构分为三层，从上至下分别为文章、主题、词项。首先，它的基本思想是将摘要文档看作为众多隐性主题集合的概率分布，同时又将每个主题看成相应词汇集合的概率分布。经过上述处理过程，摘要文档便可以看作为有关于主题的概率分布，而主题便是词汇项的概

率分布。有概率公式如下：

$$p(w_n|M_m) = \sum_{k\in K} p(w_n|K_k)p(K_k|M_m) \quad (3\text{-}1)$$

它代表词汇项w_n出现在文章M_m中的概率为主题K_k中w_n出现的概率乘以文章M_m中主题K_k出现的概率，其中$k \in K$。其中，N为所有词汇项，M为所有摘要文章，K为所有文章的所有主题。

将上述公式利用矩阵表示，如下图 3-1 所示：

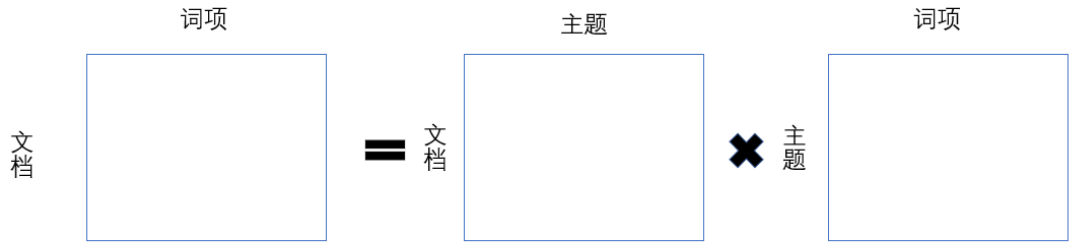

图 3-1：LDA 基本思想的矩阵表示

其中每一个“论文摘要-词汇项”矩阵表示每篇摘要文章关于词汇项的概率分布情况，“论文摘要-主题”矩阵则表示每篇摘要文章中有关于主题的概率分布情况。“主题-词项”矩阵则表示着每个主题针对各个词汇项的概率分布情况。

对于已知的摘要文章，“文章-词汇项”概率矩阵是可以通过统计文章词频得到。LDA 主题模型就是依据“文章-词汇项”矩阵，从而学习得出“论文-主题”概率矩阵以及“主题-词会项”概率矩阵。

有关于 LDA 主题模型的三层贝叶斯概率模型的设计如下图 3-2 所示：

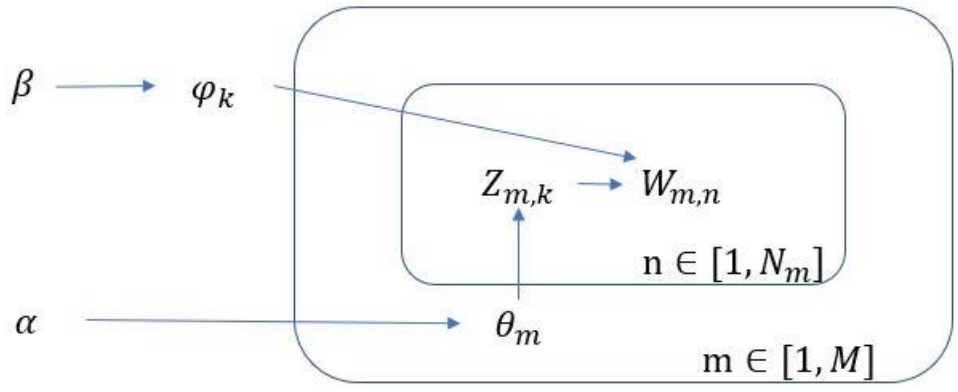

图 3-2：贝叶斯概率模型设计示意图

在图 3-2 中，α代表“论文-主题”分布的贝叶斯先验参数。β则表示“主题-词汇项”分布的贝叶斯先验参数，θ_m则表示第m篇论文摘要中的主题分布，φ_k

表示的是第 k 个论文主题中存在的词汇项分布，$Z_{m,k}$代表第m篇文章的第 k 个主题，最后$W_{m,n}$表示的是第m篇文档的第 n 个词。

LDA 主题模型的模型建立流程，主要是从指定的训练摘要文档之中估计出控制参数α与β。α是每个摘要文档的关于主题多项式分布中的Dirichlet 先验参数，β则表示每个主题中的词汇项多项式分布中的 Dirichlet 先验参数。参数 α 与 β 的估计，可用 EM参数估计与 Gibbs 抽样[27]进行估计得到，本文在这里不再赘述。

3.3 标签传递算法

Raghavan 等人[14]于 2007 年提出了标签传递算法。该算法的基本思想是：在初始状态下先为每个网络节点给定唯一的标签，然后将网络节点标签更新为其相邻节点中出现最频繁的标签，直至算法收敛。也就是说，网络中所有节点的标签与其邻居节点中出现最为频繁的标签保持一致为止。

该算法的运行过程中利用了信息流在网络节点中的传播机制，将标签快速地传播给连接较为紧密的相邻节点，使得联系较为紧密的节点具有相同或相似的标签，也就是分类，以此达到划分社区的目的。LP 算法通常利用的策略是异步更新标签，并且每次迭代之前都要对网络中的节点实施随机化排序。LP 算法具有接近线性的时间复杂度。其中，给与网络节点标签的时间复杂度为$O(n)$，每次迭代运算的时间复杂度为$O(m)$。所以，找出所有社区的复杂度则是$O(n+m)$，算法思想虽然比较简单，但是迭代次数难以估计。

标签传播方法的算法重点是找到每个节点将自己的标签传递给其他定点的概率矩阵。本文也正是利用异质信息网络中的层内节点的拓扑结构关系和层内节点之间属性的相似性来确定每一个节点将自己的标签传递给层内网络结构中的其他节点的概率。

3.4 节点相似度衡量

本文中所涉及到的多维网络中有论文层网络和作者层网络是需要运用到标签传递算法来进行社区发现的。也就是在这两层中需要计算网络节点之间的相似度，进而找出层内转移概率矩阵。对于相似度的衡量，本文主要采取的是拓扑相似度和属性相似度两种手段的线性加和。接下来，本文将分别讨论两种相似度的计算构造方法。

3.4.1 拓扑相似度

在统计科研论文引用模式中，如果作者对其他研究者的研究内容感兴趣时，无须得到授权便可以直接阅读并引用其他研究者的论文，而被引用的论文是绝对不可能引用后者的论文的。且每一篇论文只能引用其他论文一次。这样，论文引用网络可以被看作是一个单向无权的网络。

Ron 等人[14]认为两个不同网页节点之间的最短网络路径长度越小，两者的链接相似度越高，拓扑相似度越大。在论文引用网络里，如果论文 p 和 q 的引用关系越紧密，则 p 链接 q 途径的中间论文节点便越少。因此，本文有理由认为论文引用网络间的链接相似度也和最短路径成反比，即随着两个论文间的最短路径的增长，他们的关系会逐渐减弱。如果论文 p 和 q 间没有任何相同路径，那么两点间的最短路径长度被认为无限大，即他们两者之间没有任何连接相关性。使用最短路径长度计算 p 和 q 之间的链接相关性的公式如下式所示:

$$S_{pq} = \frac{1}{2^{spl_{pq}}} + \frac{1}{2^{spl_{qp}}}, \tag{3-2}$$

其中，spl_{pq}表示从论文 p 到 q 的最短路径长度，当 p 到 q 之间不存在路径时，S_{pq}趋近 0。不过，上述公式并没有考虑到论文的指向共性问题。根据复杂网络中的小世界效应可以知道，朋友的朋友是朋友。所以，即使 p 与 q 间没有路径，但如果它们存在着共同的引文或是某一篇论文的共同引文，p 和 q 也可能存在拓扑的链接关系;并且 p 与 q 之间很可能仅仅因为相似的主题而产生共同引用和共同被引用。根据 bibliographic coupling 方法[21]的思想: 如果网页c_1，c_2均被网页a_1，$a_2 \cdots a_x$直接或间接引用的情况下，两者之间存在关系的可能性会很大。根据这样的思想，本文引入广义 Jaccard 系数，用以计算论文之间的共引用和共被引用相似性。

1）论文共引用相似性。如果论文 p 和 q 中所引用的论文集之间存在着交集，那么 p 和 q 之间存在论文共引用相似性。该指标度量是一种直接的链接相似性，公式被表示为：

$$co_{pq} = \frac{|O_p \cap O_q|}{|O_p \cup O_q \cup I_p \cup I_q|} \tag{3-3}$$

2）论文共被引用相似性。如果引用论文p和q的论文存在着交集，那么论文 p 和 q 之间存在着共被引用相似性，并且公式被表示为：

$$ci_{pq} = \frac{|I_p \cap I_q|}{|O_p \cup O_q \cup I_p \cup I_q|} \tag{3-4}$$

公式中 O_p表示网络节点p的出度，I_p表示网络节点p的入度。由最短路径长度和论文的引用与被引用共性得出相似度，进行线性加和。那么，论文间链接相关度的公式，表示为

$$link_{pq} = \alpha(co_{pq} + ci_{pq}) + \beta S_{pq} \quad (3\text{-}5)$$

其中，$link_{pq}$是大于 0 且小于 1 的值，并且α=β=1/2。

接下来，在作者合作网络里，很自然地联想到作者之间的合作是双向的。且作者之间可以有多个论文的合作，如果合作的次数越多说明作者之间的相关度比较大，每一次的论文合作都会使作者之间的表示合作次数的连边的权重增大。也就是说两个作者之间的合作论文越多则代表两作者之间的拓扑网络距离也就越短。这样，作者合作网络便是一个无向加权网络。同样本文同样利用最短路径长度计算连接相关度来代表合作者网络中的作者 p 和作者 q 的节点拓扑相关性，如公式所示:

$$S_{pq} = \left(\frac{1}{2^{spl_{pq}}} + \frac{1}{2^{spl_{qp}}}\right)\frac{1}{E_{pq}} \quad (3\text{-}6)$$

公式中，spl_{pq}表示从作者 p 到作者 q 的最短路径长度，E_{pq}表示 p 与 q 之间的边权重，也就是两者合作的论文总数。当作者 p 和 q 之间不存在路径时，S_{pq}趋近 0。与论文网络的计算方式相似，合作者网络也存在着复杂网络中的小世界性，也就是说，有共同合作者的两个合作者有共同的研究课题的可能性比较高。因此在合作者网络中引入广义的 Jaccard 指数，计算作者间的共同合作者数量。

3）作者共同合作者相似度。如果指向论文作者 p 和 q 的合作者存在交集，则作者 p 和 q 存在共同合作者相似度，公式表示为：

$$ci_{pq} = \frac{|O_p \cap O_q|}{|O_p \cup O_q|} \quad (3\text{-}6)$$

公式中，O_p表示作者网络节点 p 的度，O_q表示作者节点 p 的度。将最短路径长度和合作者共同合作者相似度进行线性加和得出计算作者间链接相似度的公式表示为：

$$link_{pq} = \alpha ci_{pq} + \beta S_{pq} \quad (3\text{-}7)$$

其中: $link_{pq}$是大于 0 小于 1 的值，并且设定α=β=1/2。

3.4.2 属性相似度

（1）标签信息

由于论文层和作者层中都存在相关的标签属性信息。比如，论文层的关键词以及作者层的所属大学等。这些属性信息在论文社区和合作者社区的发现中起着至关重要的作用。例如，关键词是作者根据论文的内容对论文的一种自我评价，在一定程度上代表了作者对于论文主题的认定，如“Bayesian”、“Network”和“Lasso”等。社区发现除了发现已知存在且在直观上连接比较紧密的社区以外，更希望能够发现结构上虽然不十分明显紧密的却具有某种实际意义的社区。那么，加入标签属性信息将会为发现一些潜在的并且具有一定实际价值的社区提供一些帮助。用这种方式发现的社区也会对后续的论文，期刊以及目标合作者的推荐等应用起到帮助。

标签属性信息是非结构化文本数据信息。那么，通过分词、归类等文本预处理步骤可以将论文或是作者的标签属性信息总结成为一个一个的独立节点。并可以将它们整理成为一个N × T维的标签矩阵W。在矩阵中，N是该网络节点总数，T是预处理后得到的标签总数。由于每个标签也被视为节点，因此该矩阵中的非零数字等同于网络节点和标签节点之间的连边关系。其中W_{ij}＝1表示网络节点i被打上了第j个标签。那么，通过这种处理方法就得到了一个由网络节点和标签构成的二部网络G_{tag}。为了使表达方式统一，本文在G_{tag}中加入了连边方向，使二部网络成为一个有向网络。具体的做法如下：如果网络节点使用二部网络中的某个标签，则存在着一条由节点到该标签的有向边。那么，在有向网络G_{tag}里网络节点是只有出度的节点，节点的出度等于该节点所使用的标签数量。而标签与网络节点相反，是只有入度的节点，其入度等于使用该标签的网络节点的数量。

基于以上构造的“节点-标签”矩阵来度量节点之间相似性，所采用的方法通常是统计两网络节点所共有的标签数量[19]。但从某种角度上来讲，社区发现的最终目的就是要最大限度地拉大不同社区之间的距离，因此需要考虑到标签对节点的相对特征度。本文认为属性标签给与节点的特征度是随其使用的节点的增加而递减的，也就是说，被更多网络节点所共同使用的标签对网络节点区分度相对较小。而被一小部分网络节点所利用的标签由于其相对非主流，因此更能表示该部分网络节点的基本特征。本文为了使社区发现符合这种网络机制采用了给每个标签赋予权重$\frac{1}{lgI_k}$的方法，其中，I_k为第k个标签在二部网络G_{tag}中的节点入度，即

为具有第 k 个属性标签的网络节点的数量。类似于共同相似度的概念，如下式所示，本文将网络节点 i 和节点 j 的标签相似度s_{ij}^{tag}定义为:

$$S_{ij}^{tag} = \sum_{k=1}^{T} I(W_{ik} = 1) \times I(W_{ik} = 1) \times \frac{1}{lgI_k} \quad (3\text{-}8)$$

对于本文所研究的问题——异质多维论文合作网络中，有两个层次的网络具有这种标签信息，比如论文网络层和作者网络层。其中，论文网络的节点信息中包含着关键词。因此在构造论文节点相似度的时候，会相应地加入标签相似度。而对于作者合作网络中，作者所属的大学和地址都有可能是标签信息。而且这种标签信息并不是唯一对应的，因为每个作者在发表不同论文时所在的大学地址并不一定是相同的，所对应的大学和地址可能是多个。所以在计算作者节点相似度的时候，会计算标签相似度。

（2）文本信息

社区具有同质性指的是社区的节点对象间通常会拥有相同或相似的性质[15]。根据文献[12]，论文节点会在主题上所体现出极大的相关性。因此本文计算论文及作者的主题相关度，并将文本相似度作为计算节点相似的主要部分。具体做法为：首先为论文和作者节点匹配其摘要。论文文本信息很自然地用其本身的摘要形成它的文档。而对于作者节点，抽取将每个作者从 2014 年到 2018 年之间的全部摘要，组成文本信息文档。然后，再利用 LDA 主题模型提取每篇摘要文档的主题，并将主题结果存储在矩阵之中;最后利用主题特异度公式来计算主题间的特异程度，其值越大则说明主体之间的差异程度越大，值为 0 时说明主题相同。为了计算主题之间的特异度，现定义如下矩阵:

首先，矩阵DT。D代表了论文作者的数量，T则表示从全部摘要中提取出的主题数量。矩阵中的每一个元素DT_{ij}代表着第i个作者的摘要文档集合中包含着属于主题 j 的单词数目。接下来对于 DT 的每一行进行标准化，使得 $\|DT_i\|$=1。那么，DT_{ij}则代表着作者 i 的摘要文档对主题 j 的归属程度。那么，DT 间接展现出每个论文作者对论文主题喜好。在这里给出对于作者 i 与作者 j 的主题特异度的定义，用dist(i, j)进行表示。如下列公式所示：

$$\mathrm{dist(i,j)} = \sqrt{2 \times D_{js}(i,j)} \quad (3\text{-}9)$$

其中，$D_{js}(i,j)$则是指两概率分布间的 Jensen-Shannon 散度，其计算公式为：

$$D_{js}(i,j) = \frac{1}{2}(D_{KL}(DT_{i.}||M) + D_{KL}(DT_{j.}||M)) \quad (3\text{-}10)$$

这其中，$M=\frac{1}{2}(DT_{i.}+DT_{j.})$，为这两个概率平均的分布值；$D_{KL}(P||Q)=\sum_i P(i)lg\frac{P(i)}{Q(i)}$，是Q到P的KL散度。两个概率分布的相异程度也可以被KL距离衡量。那么，将主题特异度转化为主题相似度的计算公式则为：

$$S_{ij}^{topic}=1-\sqrt{2\times D_{js}(i,j)} \tag{3-11}$$

从公式中不难发现，$topic_{ij}$是大于0而小于1的值。两个作者或是两篇论文之间的主题相似程度越高，说明两个作者或是两篇论文之间所关注话题越相似。

（3）属性相似度

经过上面标签相似度和文本相似度的讨论，可以看出所以网络节点i与网络节点j的属性相似度包含标签相似度，主题相似度。本文对于综合这两种相似度采用的方法是线性加总的方式。如下公式所示：

$$attr_{ij}=\alpha S_{ij}^{tag}+\beta\, S_{ij}^{topic} \tag{3-12}$$

$$\alpha+\beta=1$$

3.4.3 节点之间的相似度

根据节点连接相似度以及属性相似度的计算与综合方法，本文将论文和作者层内节点之间的相似度公式设定为：

$$sim_{pq}=\alpha\, attr_{pq}+\beta\, link_{pq} \tag{3-13}$$

$$\alpha+\beta=1$$

将sim_{pq}视为标签传递算法中所需要的相似度矩阵中的p行q列的值，也就是p节点与q节点的相似程度。经过对行标准化后，相似度矩阵则变为了节点p与q之间的标签传递概率。利用这个标签传递概率矩阵进行标签传递算法。与单纯基于拓扑相似度所得出的传递概率矩阵不同的是，本文方法中的节点不但与相邻有连边的节点有传递概率，并且由于网络节点之间的属性相似，不相邻的节点之间也会存在着标签传递的现象，这样更有助于发现具有相似主题的社区。这种方式更具有实际意义，因为现实中并不是只有相邻关系的两个个体才会存在着社区相似的状况，很多情况下，人们之间即使没有交集但兴趣和爱好也会相同。这也正是需要推荐系统的原因——将平时接触不到的两点进行连接。

3.5 评估方法

为了对算法所发现的社区结果进行客观评价，会对本文的社区发现算法发现的基础社区——作者层的合作者社区计算集团的密度和传导率。首先，密度（density，d）和传导率（conductance，c）两个指标可以从社区的拓扑性质上对社区发现结果进行评价与比较。其中，密度d代表着社区内的边数占总边数的相对比例，用于衡量算法对网络社区的划分效率。因为如果社区在直观上的连边的比例较高说明社区比较紧密；而传导率c则代表着网络中每个社区指向该社区以外的边占该社区总边数的相对比例，用来测量网络中社区与其他社区之间的连通性。密度指标的定义如下：

$$\mathrm{d} = \sum_{l=1}^{K} \frac{M_t}{M} \tag{3-14}$$

其中，M_t表示社区 l 中的边数，K 是算法所发现的社区总数，M 是网络中的总边数。而传导率为如下定义：

$$\mathrm{c} = \sum_{l=1}^{K} \frac{O_l}{O_l + 2l_l} \tag{3-15}$$

其中，O_l代表从社区 l 中的节点指向该社区以外节点的边数总和；l_l表示社区 l 中的总边数。

第四章 统计科研异质网络实证分析

基于统计科研异质网络的社区发现算法在实现的过程中主要包括以下步骤：首先，通过合适的途径获取数量足够具有代表性的论文文献信息；其次，对所获取的论文数据选取合理的数据结构进行拆解，整合并储存；然后，利用 LDA 主题模型进行期刊社区的划分。接下来，在被划分的论文层和作者层的子网络中，对标签传递算法所需要的转移概率矩阵进行构造；最后，通过标签传递模型对已被划分子网络的论文层和作者层的进行社区发现，进而获得自上而下的包含多种类型节点的立体社区结构。

4.1 数据获取和指标选择

4.1.1 数据来源

实证分析部分所采用的文献数据是从“Web Of Science”网站上爬取的 27 个国际知名的统计期刊杂志，从 2014 年到 2018 年的论文文献信息，如图 4-1 展示了爬取界面。其中，所爬取每一篇论文所包含的变量有：文章名 title、期刊名 publisher、doi 号、出版年 published、被引次数 cited、摘要 abstract、作者 authors，出版商 publisher 等。

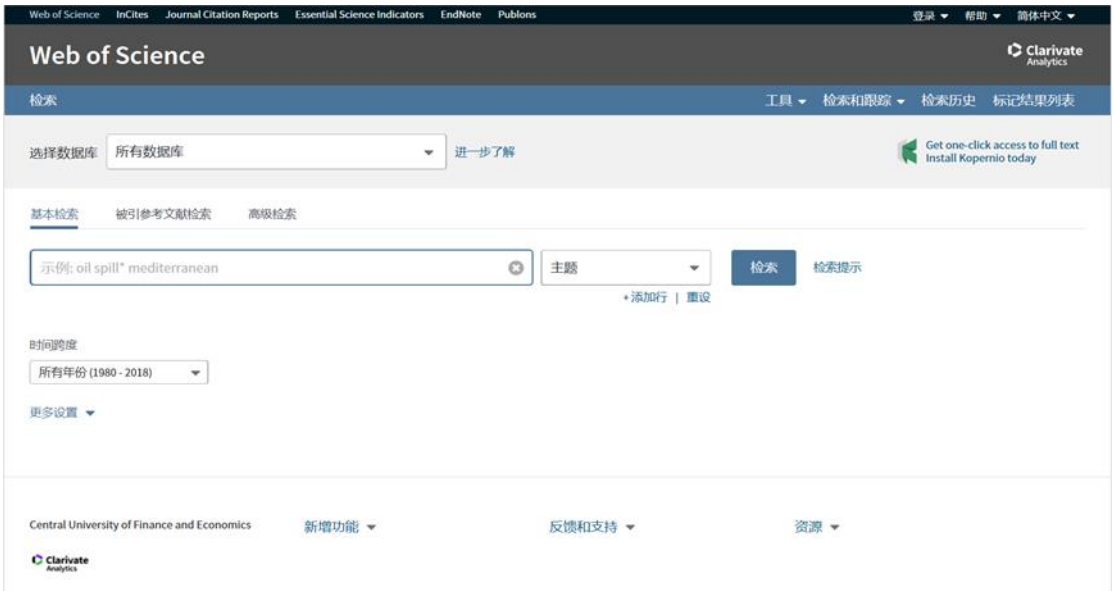

图 4-1 论文信息检索信息页

原始数据的结构为一个信息矩阵，行存储每一篇论文，列为该论文的某一项属性信息。数据选择在 JCR 中排名比较靠前的 27 个刊物杂志，如《统计年鉴杂志》《多元分析杂志》《美国统计协会杂志》《生物计量学杂志》《统计软件杂志》《计算与图形统计杂志》《皇家统计学会杂志》《经济计量学杂志》。因为排名靠前的杂志中的论文被引用的次数已经占据了绝大多数的被引用量。而且，因为其他的刊物属于比较边缘化的，且其他的刊物不是每一年都入选了 JCR，会出现某些年份的断档现象，所以删除了这些刊物。爬取的某一条论文数据的样例如表 4-1 所示：

表 4-1 爬取论文信息示例

Title	Journal	cited	num_cited
Network Vector Autoregression	Annals Of Statistics	no title+Anselin, L.+Spatial econometrics+1999	1
abstract	**keywords**	**university**	**Cite_num**
We consider here a large scale social network with a……	Multivariate time series ordinary least squares social network vector autoregression	Zhu, Xuening@ Peking University Pan, Rui@ Central University of Finance & Economics	35

将爬虫数据进行数据筛选。因为杂志选择的刊登论文一定是符合该刊物的建刊主题。并且国际主流学术杂志的引用周期大概为 5 年左右。也就是说除非是学科领域内的奠基之作以及领域经典，那么引用的周期是一般在 5 年左右。举个例子，拿 2014 年一篇对统计学领域有重要推动作用的经典论文来说。在 2014 年，它会位于引用网络的边缘，换句话说，这篇论文只存在着出度，是一个知识获取者。那么到 2018 年经过 5 年的时间，这段时间足以令其走到引用网络的核心，并对其他论文起着知识贡献的作用。所以，为了保证时效性，本次的研究将范围将制定在最近的一个科研周期内，也就是 2014 到 2018 年的论文。

4.1.2 网络爬虫

Web 爬虫是根据预先设定好的规则和过程自动高效地从符合条件的网页中捕获所需要的信息。互联网中的每个 Web 页面便是爬虫过程中的一个个节点。网络爬虫通过软件程序自动地实现在获取每一个网页节点的信息后，自动跳转到下一个网页节点并按照已编写好的规则获取文章所需要的信息。借助网络爬虫可以众多论文网络信息获取地自动化和高效化，是本文获取所需的统计论文信息的重要手段。

“selenium”是基于 Python 所开发的一个快速,且高层次的动态屏幕抓取以及 web 抓取的框架。它可以抓取网站并从页面中提取结构化和非结构化的数据。利用 selenium 爬虫框架，通过设定一定的规则，并在设定的“Web Of Science”网站上的搜索引擎上抓取到所需要的文本、以及其他属性信息，用来作为论文文本原始信息保存下来。本文编程模拟鼠标以及键盘操作，根据 JCR 排名下的杂志，填入搜索杂志名称，检索相关信息。并在其引擎搜索上键入论文关键词并连入所属网页。最后，利用对爬虫系统设定的正则规则来进行网页中的论文信息抓取。

4.1.3 论文作者姓名的提取与去重

论文的姓名处理过程相对复杂。因为每个期刊的习惯不同，所以采用的作者姓名的缩写方式有着很大的差距。这样会造成作者姓名无法匹配或者出现重名现象，影响网络节点的准确性。本文采取以下组合方式来解决非结构化数据问题：

首先，利用作者唯一的 research ID 进行第一轮匹配，融合部分作者节点。其次，利用正则将每篇论文的作者姓名从字符串中提取出来，但提取的过程中会有一些作者在不同的论文中利用的姓名不同，如“Eckert,Benjamin”这位作者，在他其他的文章中的署名可能是“Eckert,B”，但有时候“Eckert, B”还可能是“Eckert,Bokkin”。那么不一样的名字，就可能会造成重复或遗漏。最后，在数据处理过程中通过一些字符串的包含与否和论文摘要的包含与否，再加上衡量作者姓名字符串的相似程度，将论文的同一作者很好的进行匹配和去重。

利用这种方式，比较好的提取出了作者所对应的每一篇摘要。虽然无法完全的达到作者节点与姓名的完全匹配，但是经过测试，匹配的正确率能够达到 90%以上。那么对于网络社区的实证分析来讲，作者姓名问题所造成的影响可以忽略。

通过以上处理，作者节点会被融合。那么，有些作者会对应多篇论文（在此

案例中也就是多篇摘要）。因为术业有专攻，所以每一篇摘要其实都能反应作者的研究主题和方向。对于此种情况的处理，本文将作者对应的多篇摘要进行拼接，拼接出一个长摘要字符串。

4.2 统计科研异质网络的搭建

4.2.1 异质网络的搭建。

本文将统计研究异质网络分为三层次，包括杂志层，论文层和作者层。三个层次中分别地对应着杂志节点，论文节点和作者节点。其中层间有着相同类型的连边。论文层的连边为相互引用，作者层内的连边为合作关系。杂志与文献，文献与作者之间具有层间连边。按照这样的设计，构成了一个三维统计研究网络，具体网络模式如图 4-2 所示。

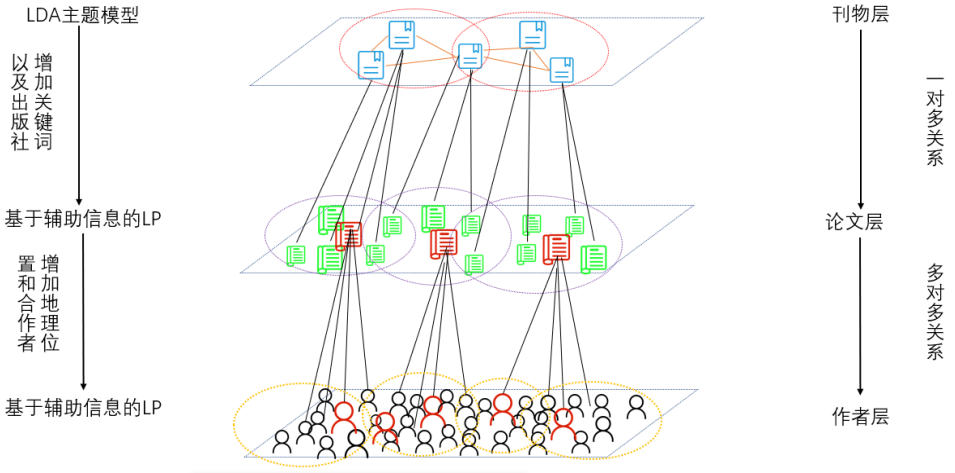

图 4-2：统计科研论文多维网络结构示意图

4.2.2 设计层与层之间连边效应

由于针对科研文献网络的社区发现主要是发掘主题社区，并且在 3 维网络中上层对下层具有一对多映射关系，所以上层网络发现的社区可以通过层与层之间的链路，能将社区信息完全传递给下层。也就是说，下层节点首先会被上层的社区发现的社区结果圈定初始的子网络，然后再结合该层的拓扑结构和节点信息利用社区发现算法进行进一步的分类。

4.3 基于 LDA 主题模型的杂志层社区发现

4.3.1 词向量的转化

对于学术杂志的主题来说，并不是所有的词性都对于 LDA 的主题建模有积极意义。因为介词，连词，动词以及各种各样的形容词和副词在各个主题的应用都很频繁，并且没有特异性，不能代表一个主题的特点。所以本文在进行主题建模时只保留了名词作为文本，删除了其他的词汇。并且对于只出现过一次的名词，也就是词频为 1 的单词，它们反映不出相关主题特征，所以也进行了删除处理。

首先针对杂志的所有文本进行词频分析，来看一下是否有必要进行词向量的转化。因为如果某一词频比较高，会干扰到 LDA 模型的“文章-词汇”Dirichlet 分布先验参数的估计，并且会增加主题模型的困惑度。这样的话，在各个主题中均会出现频率较高的词汇，那么根据词汇所得到的主题的象征意义就会被削弱，导致每个主题之间的分界较为模糊，杂志社区的意义不明确。由于本文提取了论文摘要的所有名词来代表摘要部分，词频统计结果中所有词汇为名词。统计结果显示了词频最高的前二十个词汇如下表 4-2 所示：

表 4-2 词频分布表

序号	词汇	词频（次/篇）	序号	词汇	词频（次/篇）
1	model	0.7068	11	distribution	0.1730
2	data	0.4374	12	estimator	0.1689
3	method	0.3515	13	two	0.1640
4	study	0.2666	14	simulation	0.1527
5	effect	0.2519	15	function	0.1402
6	time	0.2247	16	parameter	0.1392
7	test	0.1911	17	result	0.1377
8	approach	0.1842	18	regression	0.1371
9	treatment	0.1795	19	estimation	0.1239
10	analysis	0.1745	20	trial	0.1224

这里采用统计词频的方式是统计每个词汇在每篇摘要文章中的平均出现次数，平均次数越高说明该词汇在任何统计学论文摘要中都倾向于出现，无法代表一个主题。例如“model”、“data”、“method”、“study”等单词是统计学论文的

基础词汇，无论什么主题都会利用这些词汇，甚至非统计学主题的论文中也会用到这些词汇，所以他们的主题代表性并不强。根据词频分析的结果来看的话，如果单纯依靠摘要文本数据来建立词典然后再建立语料库的方式，LDA 模型的效果并不会太好。

本文选择基本主题较为明显的《ANNALS OF STATISTICS》、《ECONOMETRICA》、《STATISTICS IN MEDICINE》与《JOURNAL OF STATISTICAL SOFTWARE》四本杂志进行词频描述，仍旧只描述名词词频，因为名词词频能很直观的反映主题。在词频描述统计图 4-3 中也可以清楚地看到，尽管四本杂志所涉及的领域有所不同，但每本杂志利用的前 10 高频词汇有很多是相似的，包括“model”,“data”,“result”,“study”等。因为这些单词是统计学研究中不可避免的词汇。不过，如果考虑一些中高频词汇，那么结果便会不同。由图 4-3 本文也可以看出，不同的杂志在用词上有着很鲜明的特征。例如，《ANNALS OF STATISTICS》杂志更偏向于“method”“estimation”“function”等具有较强理论特点的词汇。《ECONOMETRICA》杂志则更倾向于“price”“market”“firm”“preference”等较强经济学倾向的词汇。《STATISTICS IN MEDICINE》更偏向于“treatment”“trial”以及“patient”这样的医学词汇。相比之下,《Journal of Statistical Software》则将注意力集中在了“package”“algorithm”“r”等很明显的计算机词汇上。

根据这样的描述分析结果，本文得出这样的结论：不同主题的杂志会在统计学的基本词汇上添加和自己主题相匹配的特征词汇来构成自己的文章。所以单纯利用词频向量建立 LDA 模型的效果显然是不好的。因此，本文采用将词频向量转化为$tf-idf$向量的方法。也就是说，将语料库中存在的词频向量转化为$tf-idf$向量。这样的处理可以降低在很多论文摘要中都高频率出现的且对主题划分没有实际意义的统计学基础词汇的权重。除此之外，在模型所对应的字典和语料库中加入了一些专业的统计学算法及模型词汇，通过这样的方法使得文本字典更加丰富，分词更加准确，主题更具有实际意义。

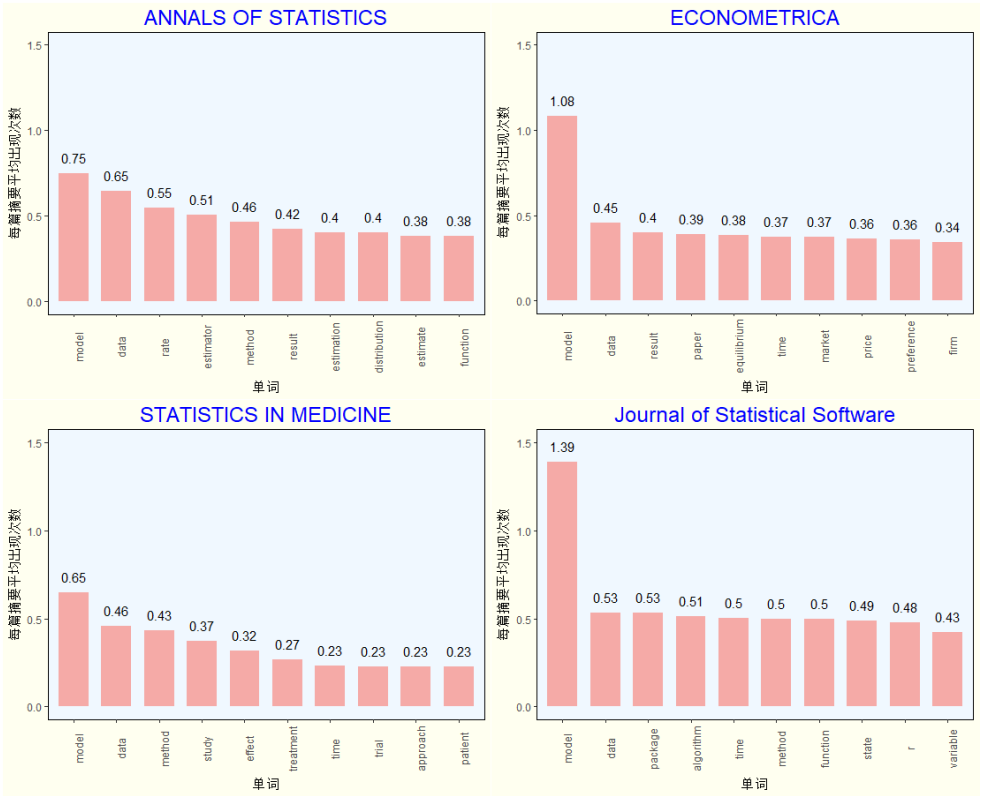

图 4-3 具有代表性的 4 本杂志的词频统计

4.3.2 LDA 主题社区发现结果

按照第二章所介绍的 LDA 模型原理，在 python 上进行了 LDA 主题模型的程序编程。并利用了 2014-2018 年间统计论文的摘要文本信息汇总构建了 LDA 主题模型。经过反复不断地调参尝试，最终构建了具有 4 个主题的论文摘要 LDA 模型。并且确定了每个主题所包含的主题关键词。

首先，本文来看一下 LDA 模型的四个主题关键词的提取结果。如图 4-4 所示为 LDA 第一主题的整体词频分布。在主题分布图中可以看到，几个主题的分布距离是比较大的，且论文主题之间的范围并没有相互的交差。并且每一个主题的边缘主题分布大于 5%，第一主题的边缘主题分布甚至大于 10%，这表明划分为四个主题是比较科学和实际的。而且这四个主题能够很好地涵盖摘要文本的内容特点。接下来便让本文看一下每个主题所包含的最相关词汇以及这些词汇所反映出的主题内涵。

第一主题社区的模型分析结果如下图 4-4 所示。可以看到，与这个主题中最相关的前三十词汇在该主题中的分布要远远超过他们在其他主题的分布。也就是说这些词汇几乎只在第一主题出现，也就是所谓的该主题的特征。这些单词包含

着“patient”“disease”“death”“drug”“hazard”“phase”“health”等一些具有很鲜明的医疗和生物特征的名词，还包含着“probability”“assumption”“covariates”等一些统计学常用词汇。很显然，这种主题的关注重点是在医疗与生物统计上的研究，所以本文可以将这一主题称之为“生物医学统计”，也就是杂志层的生物医学统计社区。

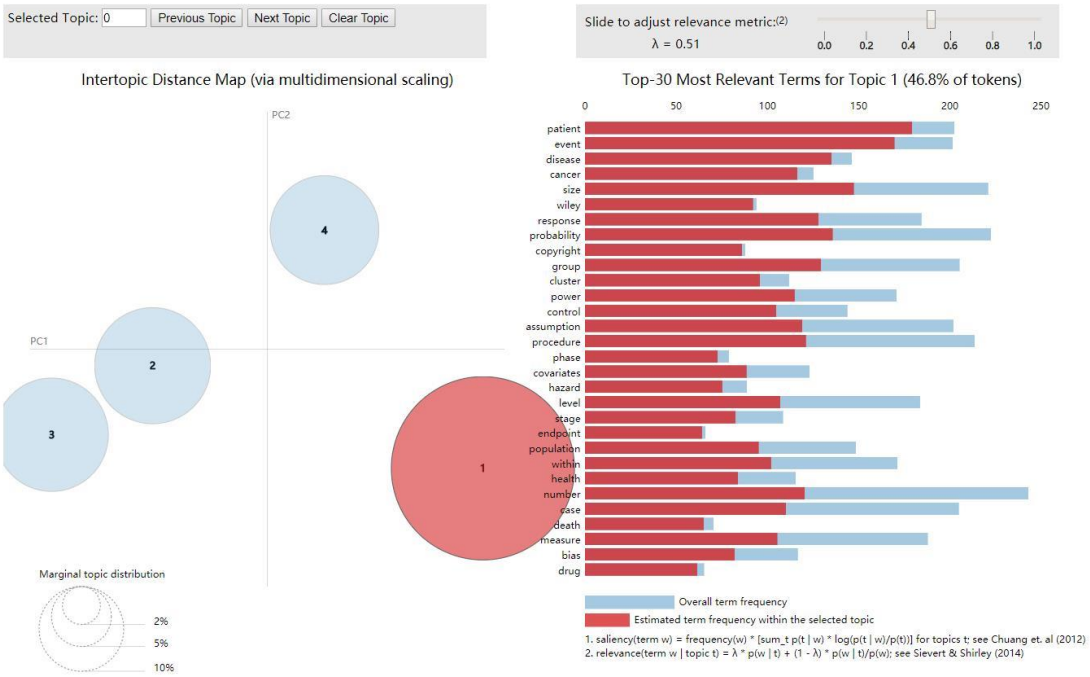

图 4-4 LDA 模型第一主题

接下来是第二个主题的关键词汇。同样可以看到，与这个主题最相关的前三十词汇在第二个主题中的分布频率要远远超过他们在其他主题的分布。如图 4-5 的 LDA 模型结果所示，第二个主题包含着“volatility”,“material”,“garch”,“carlo”“matrix”“density”“factor”“material”等统计学基础词汇。而且很少看到一些专业的词汇。所以本文有理由相信，这个主题所对应的主题是关于统计学的方法论，并且更加偏重于理论研究的。所以，本文将这个主题社区被命名为古典统计。这个主题中比较著名的杂志有《ANNALS OF STATISTICS》，《AMERICAN STATISTICIAN》等。

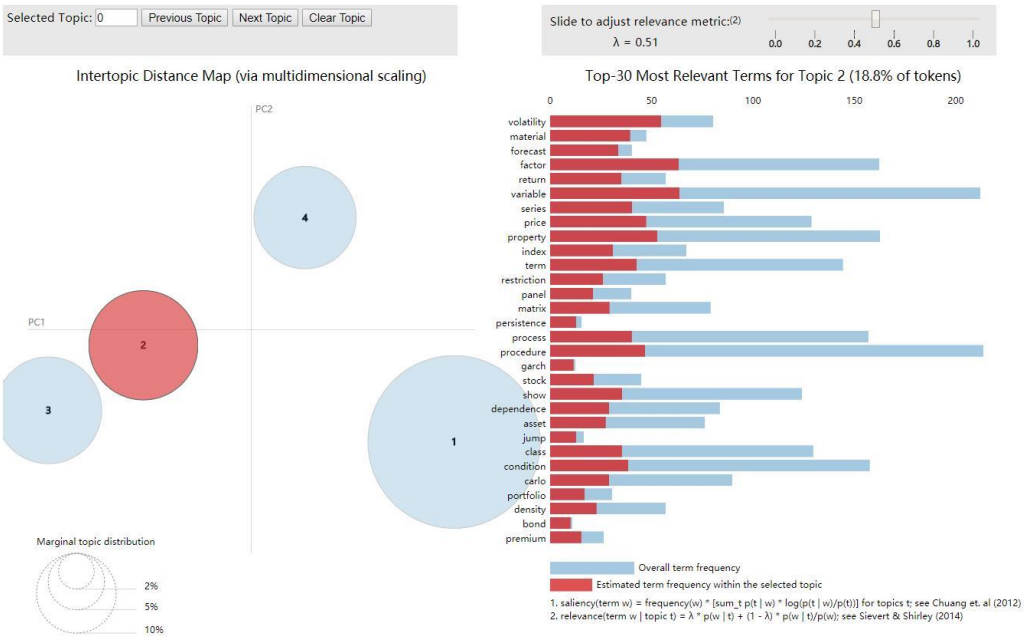

图 4-5 LDA 模型第二主题

根据图 4-6 所示，LDA 的第三个主题充斥了大量的经济与金融学词汇，如“equilibrium（均衡）”,“preference（偏好）”,“market”,“firm”,“price”,“investment”“economy”“insurance”“trade”等。并且这些词汇在该主题中的占比远远高于他们在其他主题中的占比。所以有理由相信这些词汇可以代表这个主题所包含的实际主题意义。不难看出这个主题所展现的全部是有关于经济金融学领域的词汇，所以本文将这个主题命名为“金融经济统计”，所代表的社区也就是“金融经济统计”社区。

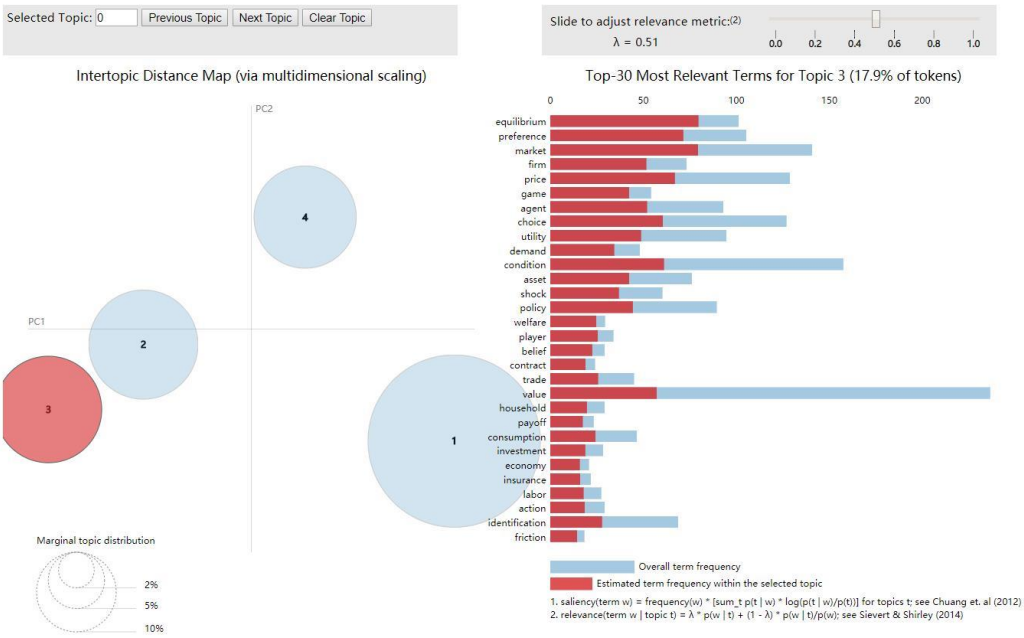

图 4-6 LDA 模型第三主题

根据图 4-7 所显示，LDA 主题模型的第四个主题充斥了大量而又丰富的计算机编程术语，如“package”,“r（r 软件）”,“network”,“tool”,“algorithm”,“array”,“system”,“space”等。并且这些词汇在该主题中的占比远远高于他们在其他主题中的占比。所以有理由相信这些词汇可以代表这个主题所包含的实际主题意义。同时前三十的关键词汇中也不乏一些“statistics”“bayes”等关于基础统计学的词汇。很显然，这个主题主要是关于计算机编程在统计中的应用，也可以称作是统计学方法的计算机实现。所以本文将这个主题命名为“统计软件”，该主题所代表的社区是“统计软件”社区。具有代表性的杂志有《Journal of Statistical Software》，《Journal Of Statistical Computation And Simulation》和《R Journal》等

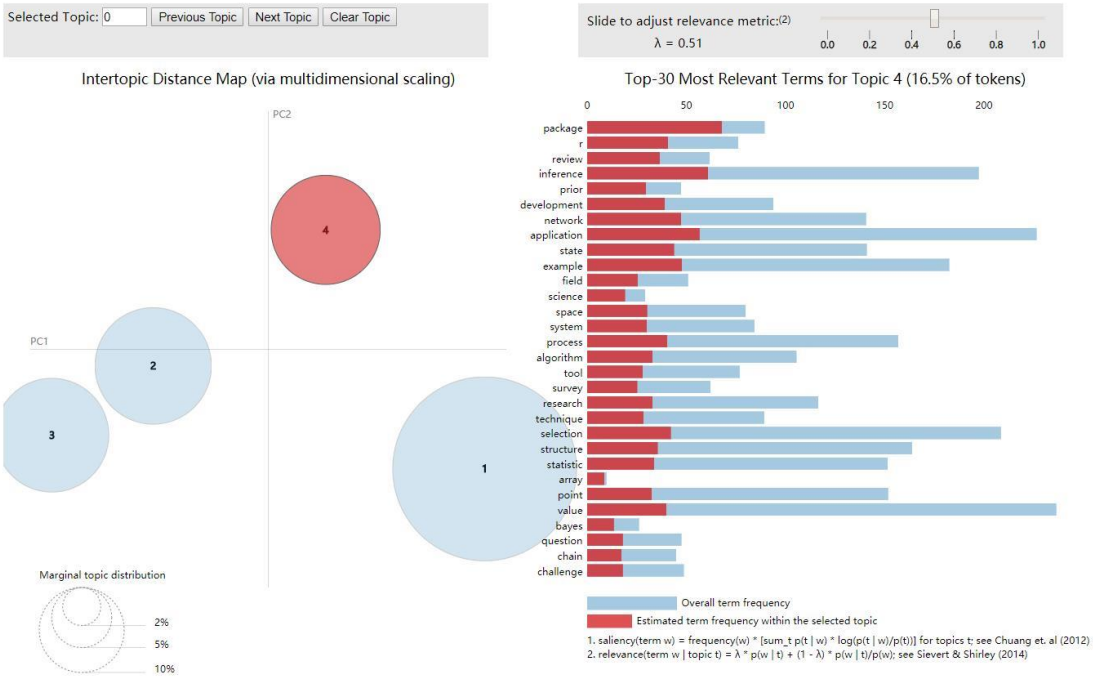

图 4-7 LDA 模型第四主题

到目前为止，本文已经利用 LDA 模型获取了有关于统计学论文摘要的四大主题以及他们所对应的社区，分别是“古典统计”主题，“金融与经济统计”主题、“生物医学”主题以及“统计软件”主题。这说明目前统计学的研究目光大体集中在了这四大研究领域上。如下表 4-2 所示，展现了四个主题中权重排名前 5 的相关词汇：

表 4-2： 四个主题的关键词

主题	古典统计	经济统计	生物医学统计	统计软件
关键词	likelihood	equilibrium	patient	package
	bayes	preference	disease	r
	dimension	market	death	tool
	estimation	firm	drug	algorithm
	distribution	price	health	System

从上表 4-2 中可以看出每个主题都有着比较鲜明的主题特点。他们中所包含词汇可以很好的反映他们所属的主题，在本文中也就是具体的杂志主题社区。除此之外，本文还对用于构建 LDA 模型的每本杂志在每个主题上的归属概率进行了计算。并将每本杂志所属最高概率的主题作为该本杂志的归属主题，从广义上来讲，即为每本杂志所属的主题社区。本文在表 4-3 中列出了在 JCR 中影响因子排名比较靠前的 5 本统计学领域的杂志对于这 4 个主题的归属度。

表 4-3： 影响因子较高的统计学杂志的主题归属度

主题	古典统计	经济统计	生物医学统计	统计软件
JOURNAL OF STATISTICAL SOFTWARE	0.097	0.034	0.110	0.759
AMERICAN STATISTICIAN	0.996	0.001	0.002	0.001
ECONOMETRICA	0.00	0.999	0.00	0.00
MULTIVARIATE BEHAVIORAL RESEARCH	0.990	0.004	0.002	0.004
STATISTICS IN MEDICINE	0.006	0.0	0.889	0.104

从表 4-3 中，基本可以看出每本杂志对于主题的归属比较明确。每本杂志在归属度最大的社区上一般都能达到 0.7 以上。这说明 LDA 主题模型对摘要主题的划分比较鲜明，四大主题的区分度比较高，且能够比较好的区分杂志所属主题和社区，基本符合本文所假定的情况。但对于例如《Stochastic Analysis and Applications》”的杂志，由于这种刊物上的论文既有理论见长的偏向古典统计学论文又有以应用见长的其他论文，所以 LDA 主题模型无法很好地将其进行区分，出现了主题归属度比较模糊的情况。不过本文可以将它划分到两个主题社区中，诸如这样的杂志可以为网络带来交叉重叠社区。那么，对于这样的杂志所连接的下层论文节点也将归属于论文网络的两个子网络——“古典统计”子网络和“生物医学”子网络（或其他子网络）。从而属于该杂志的论文将作为不同子网络的节点进行多次社区发现，进而形成重叠社区。

4.4 基于标签传递的论文层社区发现

首先依据杂志层的主题社区发现结果，并通过杂志层和论文层间的从属连边关系，将杂志层的社区信息传递到论文层。具体的操作是将论文层根据它们所属的杂志进行了四个主题社区的划分，进而形成了 4 个不同的论文层主题子网络——“古典统计”子网络，“生物医药”子网络，“经济统计”子网络以及“统计软件”子网络。其次，利用 3.4 节介绍的网络节点间相似度构造方法进行节点的标签转移概率矩阵构造。然后，验证转移矩阵的稳定性，以防止矩阵是非常返而导致标签转移社区发现算法失效。最后，利用标签传递算法在论文层上实施社区发现，进而确定论文的社区结构。

4.4.1 论文层子网络描述

依据 4.3 节杂志层社区主题社区发现结果，通过杂志层和论文层之间的从属连边的传递效应，本文首先将论文层中的每一个论文节点打上它所属杂志的主题标签，并依据论文节点所属的主题标签来将整体论文层的网络拆分为 4 个子网络，如图 4-8 所示。

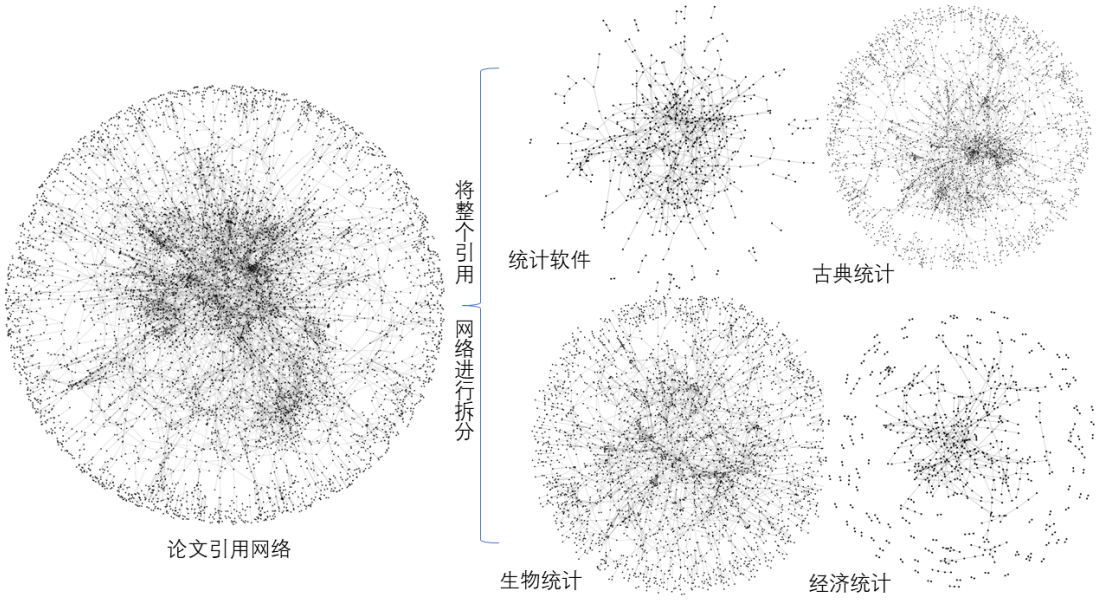

图 4-8 论文层子网络概述

整体的论文层引用网络一共包含了 2014 年—2018 年期间 30 个影响因子最高的统计杂志上刊登的 12485 篇论文。在该网络中论文之间的引用纷繁复杂，网络中

的最大连接组件在网络中央，十分密集紧凑，代表着比较主流的统计研究方向，不难看出主流统计研究中相互之间的引用相对较多。而网络中周围的一些节点以及相应的连边则代表着一些小众的研究，它们与主流的研究联系性不大。将论文层整体网络拆分后得到包含着 824 论文节点的“统计软件”子网络，5326 个论文节点的“古典统计”子网络，5863 个论文节点的“生物医学统计”子网络和 1593 个论文节点的“经济社会统计”子网络。可以在图 4-8 中看到，每一个子网络中均包含一些代表主流研究趋势的大连通组件，其他的小众研究也分布在主流研究的周围并且和主流研究保持着一定的联系。“生物医学”子网络的引用之间比较紧密，主流和小众研究的界限并不明显，整个网络更像是一个整体。相比之下，“经济社会统计”网络则相对稀疏，论文之间的引用比较少，彼此之间形成一个又一个的研究团体。

4.4.2 论文节点相似度构造

首先，论文节点先按照 3.4.1 节所介绍的拓扑相似度构造方法进行拓扑相似度构造。然后，再利用 3.4.2 节介绍的节点属性相似度进行属性相似度的构造。需要注意的是，论文节点能够利用到属性相似度的构造上的有文本属性“摘要”和标签属性“关键词”，“出版社”。

在这之中，标签关键词属性需要进行处理。因为每位作者的运用关键词的习惯不同，所以对同一个关键词的描述也不尽相同，比如对于疾病，有的作者表述成“disasters”，而有的作者标注成“disaster”。又比如对于蒙特卡洛算法，有些作者采用的是 MCMC，有些作者则用其全称“Markov Chain Monte Carlo”所以导致关键词的二部网络过于稀疏，无法很好的代表每一个作者的研究重点词汇。对此问题的处理方法，本文利用 word2vec，n-gram 和字符串相似的手法。具体分为以下三个步骤:

1)首先对比两个单词的字符串相似度，如果大于某一阈值 a，就判断他们是同义词汇的意思，比如“disasters”和“disaster”之间的字符串很相像，所以通过这一步可以检测出很多大小写或是单复数所造成的单词相似。由于这种方法的计算复杂度小所以放在第一个。如果小于某一阈值 b 则将其作为不是同一词汇。如果介于两个阈值 a，b 之间则进行下一步继续判断。

2）利用基于维基百科的语料库将每一个关键词构造 word2vec 向量，对比两

个关键词的相似度，如果相似度高于某一阈值c，则认为两个单词作为同一关键词并进行合并。这种方式可以检测出很过缩写所造成的单词不同，如“MCMC”，“Markov Chain Monte Carlo”。但如果低于某一阈值 d 则将两个单词是为不同。如果介于两个阈值之间则进行第三步。

3）最后通过 n-gram 进行两个相似词汇的检测。首先进行对两个单词的其中一个词汇的前后两个最大概率的单词的选取，然后根据语料库计算这些单词出现后另一个相似单词的后验概率。如果后验概率高于某一阈值 e 则认为量单词相近，可以作为同一单词进行合并。

通过这样的手段虽然不能完全的将所有的语义相似的记单词找出，但不断的调节每一步骤的阈值，达到相对理想的情况，使得而二部网络不至于太稀疏。也减小了计算的复杂度

4.4.3 论文层社区发现结果

利用节点相似度构造后的转移概率矩阵进行标签传递算法的运算。最终在四大子网络中共发现 30 个重叠社区。论文层社区划分的结果概览图，如图 4-9 所示。最大连通子图被标签传递算法进行了社区发现，不同颜色的节点代表所属不同社区的论文，而其他较小的离散连通子图都被当成了一个一个小社区。

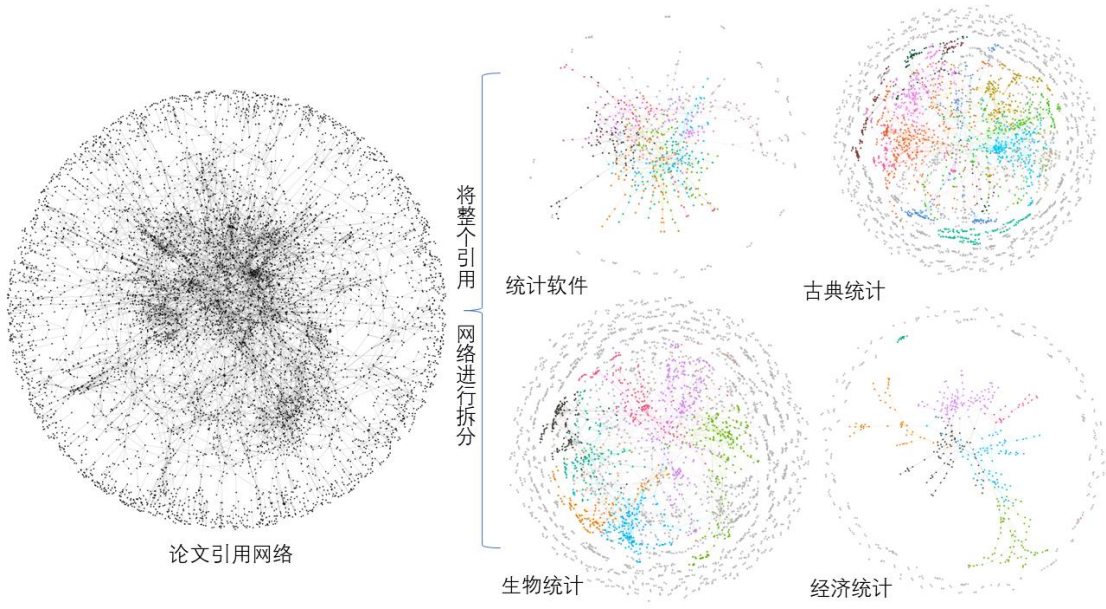

图 4-9 论文层子网络社区发现结果

在此展示相对较小的“生物医疗统计”子网络中的最大连通子图的社区发现结果。生物医疗统计子网络被划分为了 4 个典型社区，并且每一个社区的聚集状态都比较好，社区内的节点间连接紧密，而社区外相对较为稀疏，但是因为利用到了节点的属性相似度的信息，所以有很多看似拓扑连接不紧密的节点，由于他们的属性信息较为相似，他们也被划分在了一个社区。所形成的社区划分网络图如下图 4-10 所示。

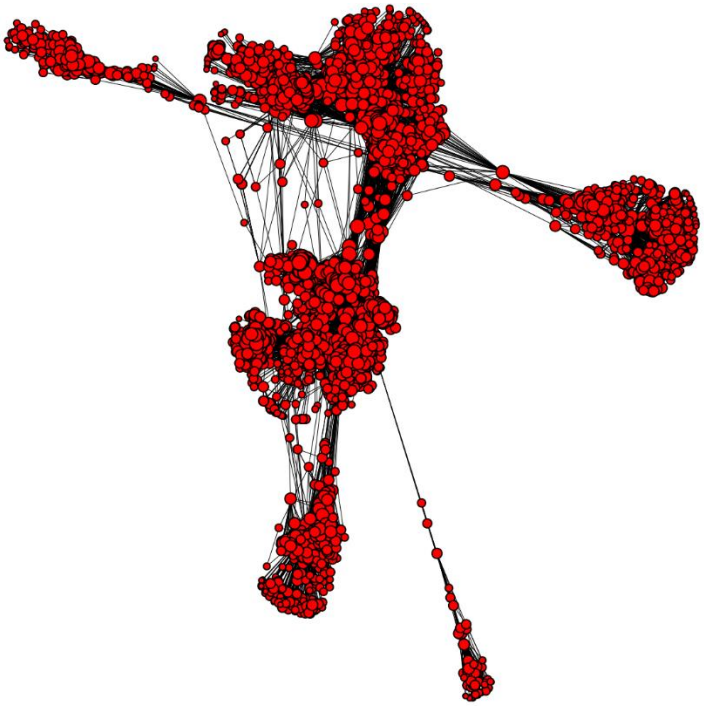

图 4-10 生物医疗统计子网络社区结构

从图中可以看出，除了右下角的小集团以外其他的论文节点都聚集到了四个主要的社区中。接下来对这四个生物医疗社区所包含的摘要分别提取的关键词，并删除高频词，形成关键词的词频统计，如下表 4-4 所示：

表 4-4： 生物医学子网络各社区关键词

医学	疫苗和病菌	病菌的防治	日常生活
clinical	wale	Mortal	drinker
cardio	surveil	Spatial	smoke
fatality	salmonella	infect	strain
leukocyte	vaccine	veget	cuc
trial	aphid	virus	veget

不难从这 4 个社区的高频关键词中得知在这 4 个社区的研究方向。从表 4-4 中可以看出，第一个社区在生物医学的领域中更偏向于医学，包含了很多诸如临床 clinical 和死亡率 fatality 等词汇。第二个社区，通过疫苗和沙门杆菌等词汇，本文可以初步判断研究主要集中在疫苗和病菌的统计分析。第三个社区的研究上，还是集中在医学领域，不过研究的重点集中在了病菌的防治上，因为出现了 virus 和 infect 等词汇。最后一个社区将研究重点放在了人们的日常生活上，更多的统计学词汇集中于普通人的日常作息习惯的研究，出现了关于健康行为的日常单词，如吸烟，精神压力，结肠炎（cuc）等，比较深入普通人的生活。

4.5 基于标签传递的作者合作网络发现

在本文的研究范围内，作者层是多维网络中最大的一个网络层，包含着 21905 个作者节点和 45769 个合作加权连边，作者网络平均度为 4.17，平均加权度为 4.7，也就是说在合作者网络中每个作者在 2014 年—2018 年间大约有 4.17 位合作者，总体网络的概览图如图 4-11 所示。

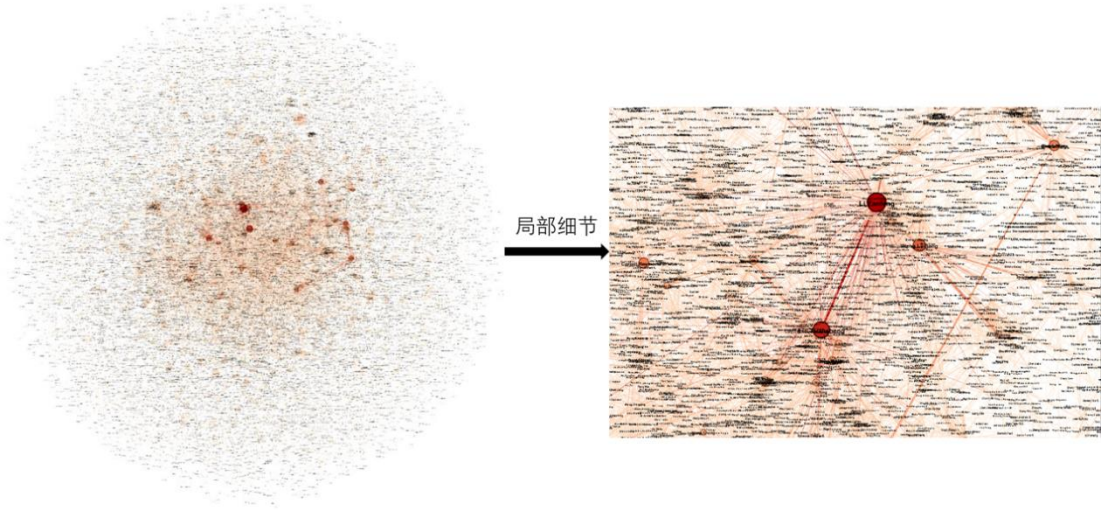

图 4-11 合作者网络概览图与局部细节

首先，对于论文层和合作者层的网络关系，本文有如下考虑：有相互引用关系的论文的作者们，他们研究领域之间的主题也很有可能重合。甚至某些作者在发表论文时会引用自己或者其合作者先前的研究成果。依据论文层的社区发现结果，并通过论文层和作者层之间的归属连边关系，对作者层进行了 30 个论文引用子网络的划分，将作者层节点首先划分为 30 个与其所属论文一致的子网络。由于

存在着论文重叠社区，所以在每一个作者节点可能存在于多个社区。其次，利用3.4 节所介绍的作者层的节点间相似度构造方法进行作者节点的相似度矩阵以及转移概率矩阵构造，最后进行标签传递算法的实施，确定合作者社区。

4.5.1 作者层子网络描述

依据 4.3 节论文层社区主题社区发现结果，并通过论文层和作者层之间的从属连边的传递效应，本文首先将论文层中的每一个作者节点打上它所属论文的主题标签，并依据作者节点所属的论文社区标签将整体的作者层网络拆分为 30 个子网络。如图 4-12 所示，展示的为古典统计作者合作网的作者子网络拆分结果。

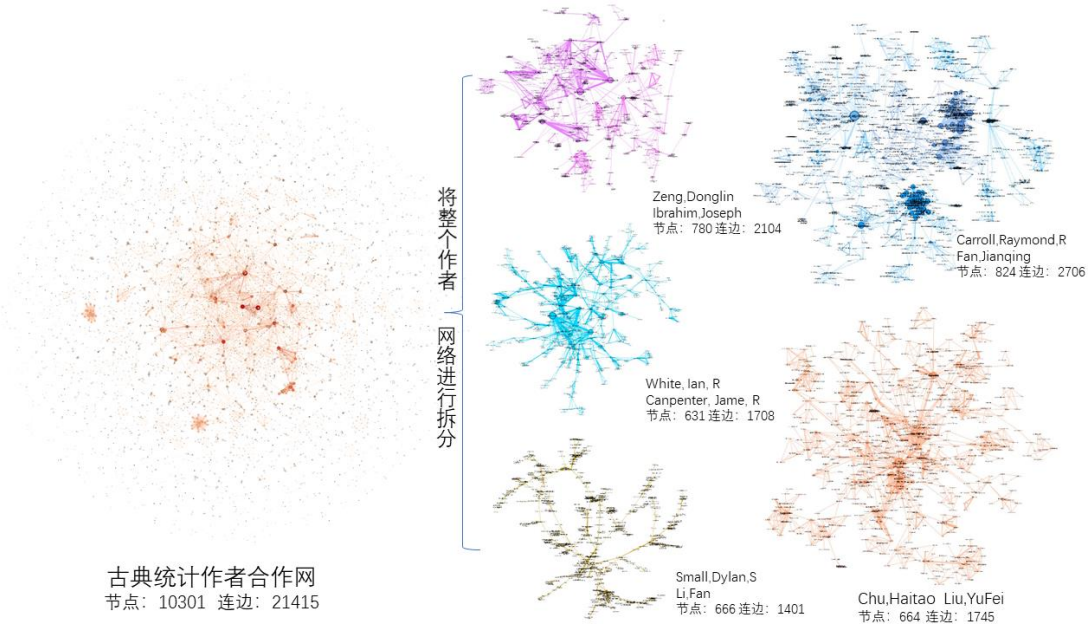

图 4-12 古典统计合作者合作网子网络拆分

在图 4-12 的所有作者子网络中，每个作者的平均度皆为 4 左右，说明在整体网络拆分为子网络的过程中使用的社区发现方法和层间连边对应关系比较好，能够很好的保留原本网络的连接状态和连接关系，不至于过分的切分网络。并且可以看到每一个网络中的中心节点也就是每一个作者子网络的核心作者，这些作者对于整个社区的学术发展起着至关重要的作用。这些核心作者，如 Carroll,Raymond,R 等都具有众多合作者，因此他们所在的社区中的学术交流以他们展开。

4.5.2 作者节点相似度构造

首先，作者节点先按照 3.4.1 节介绍的拓扑相似度构造方法进行节点间的拓扑相似度构造。然后，再利用 3.4.2 节介绍的属性相似度进行节点间的属性相似度的构造。需要注意的是作者节点的信息能够利用到属性相似度的构造上的有文本属性“摘要”，标签属性“关键词”，“大学”和“出版商”等。

由于每位论文作者的出版物大多不只有一篇论文，若单独对每个作者发表的多篇论文分别进行主题分析提取是不现实的。为了准确快速地抽取每个作者感兴趣主题，本文采取把同一作者的所有论文摘要集中到一篇长文档中，于是作者和文档就形成了一一对应的关系，摘要也就成为了作者的节点属性。

最后标签关键词属性同样需要进行合并处理。因为每位作者的运用关键词的习惯不同，所以对同一个名词的描述不尽相同，比如对于疾病，有的作者表述成“disasters”，而有的作者标出成“disaster”，所以导致关键词的二部网络过于稀疏。对此的处理方法类似于论文层，都是利用 word2vec，n-gram 以及字符串相似的手法对每位作者的关键词进行合并和拓展。只不过这里不同的是三个步骤的阈值的确定是要根据在作者网络上的实验。于是，论文的关键词在语义不变的条件下得到了压缩。

4.5.3 论文节点相似度构造

经过标签传递算法的计算，最终在 30 个论文社区四大子网络中共发现 92 个合作者社区，由于篇幅有限，不便展示，对古典统计论文社区中的以 Zeng,Donglin 为核心的合作者社区所确定的作者子网络进行社区发现，结果如图 4-13 所示。子网络的网络结构如左图所示，社区发现结果如右图所示：

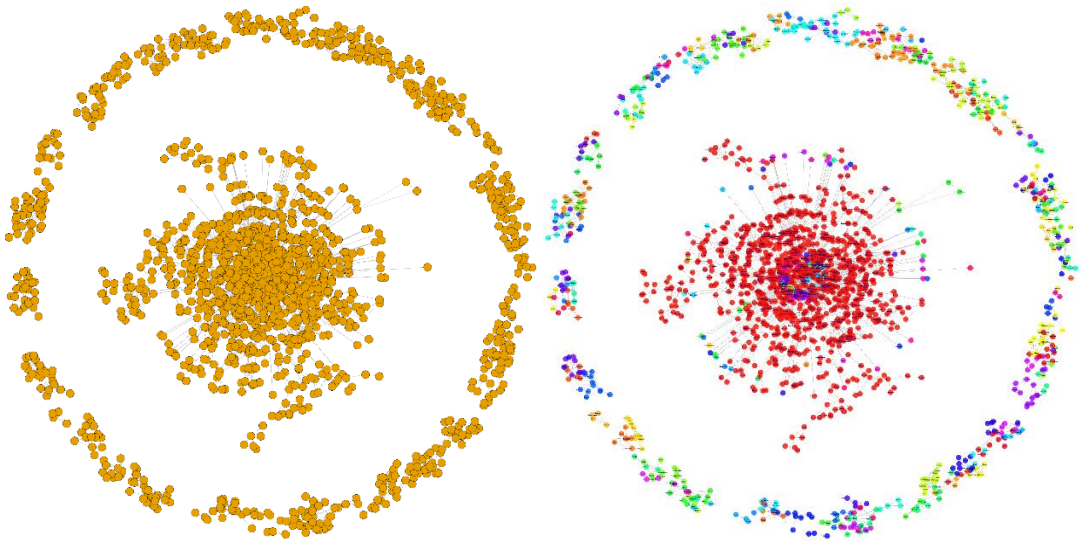

图 4-13 Zeng,Donglin 为核心的合作者社区的子网络图

这里子网络包含 780 个作者节点的情况。使用布局为 kamada—kawai 算法布局，将整个论文合作网络进行展示。每一个节点为一个论文作者，每一条边表明两个作者之间存在合作关系，边的粗细代表着两个合作者的合作的文章数目也就是权重，标志着两个合作者合作的密切程度。可以在上图的网络结构图中发现的是，很多节点都连接在一起而又不与其他的集团相连，他们在复杂网络中称之为一个一个的连通区域。它们的形成部分的原因是由于本文将原本的网络切割成了一个个的子网络，本文将这些不相互连通的组件赋予不同的颜色，成为一个小合作者社区。

可以看到，中间的红色点构成最大连通组件，十分密集，通过肉眼观察大致为一团比较密集的节点聚集，可能是由某几个统计学术大师为中心的合作集团，如著名统计学家 Ibrahim,Joseph 等；而周围散布一圈的多色点为众多的规模很小的多个连通组件，这可能是在统计领域比较小众的研究方向或者是其他国家的统计团体。非常有意思的是在国外统计领域比较出名的教授 Zeng,Donglin，也在这样的一个小的组件中，维系着其他的作者，2014 年以来发表的论文数目达到了 22 篇。

因为作者层应用的网络为加权无向网络，所以利用衡量节点度的指标为加权度分布，即与某一个节点所相连所有边的权重之和。将论文合作网络中的节点的度分布绘制在直方图 4-14 中。可以看出，该子网络的大部分节点度都很小，但也有一小部分节点具有很大的度。在该子网络中，节点度分布大致服从幂律分布，是具有无标度特征的。

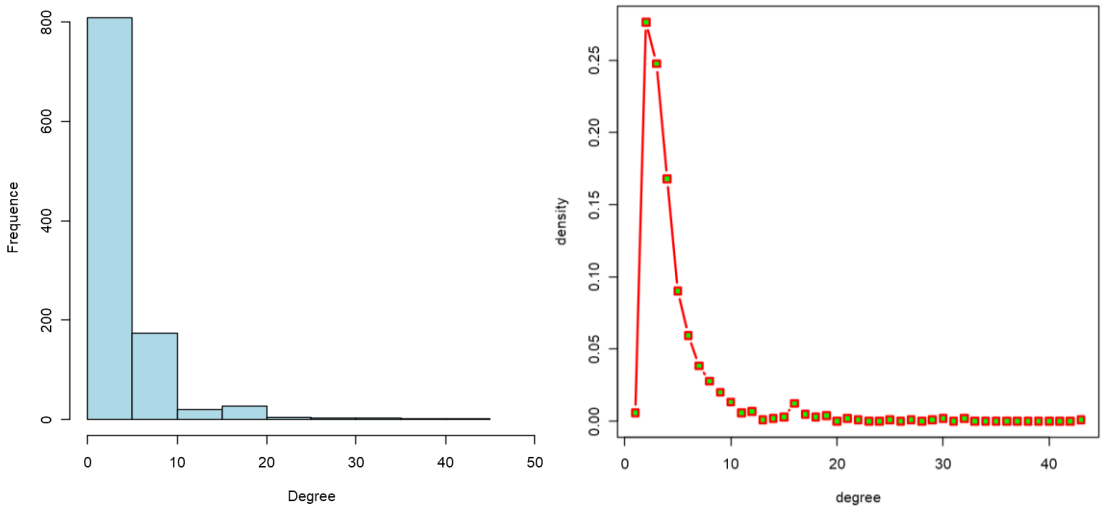

图 4-14 合作者子网络度分布

同时这种度分布也说明了大部分的统计学论文作者的创作能力还是比较有限的，只有少部分统计学大牛才是比较高产的。同时统计一下论文创作者的创作状况。对于该子网络，高产作家的 top5 统计如下表 4-1 所示：

表 4-5： 合作者子网络的高产作家

排名	作者	论文数	合作者数
1	Zeng, Donglin	21	42
2	Ibrahim,Jpseph,G	19	40
3	Lawson,Andrew,B	18	35
4	Dunson, David B	15	34
5	Tilling,Kate	15	35

从表中可以看到，作品数最多的是北卡罗来纳大学 Zeng, Donglin，他的作品数量比较多且合作者数量合作者数目也比较多，是引领社区主题发展趋势的带头人。他的研究领域主要集中在高维数据的处理以及生物信息学上。对该子网络中的最大连通组件实施基于相似度的标签传递算法后，一共得出了 6 社区，社区划分结果如下图 4-15 所示：

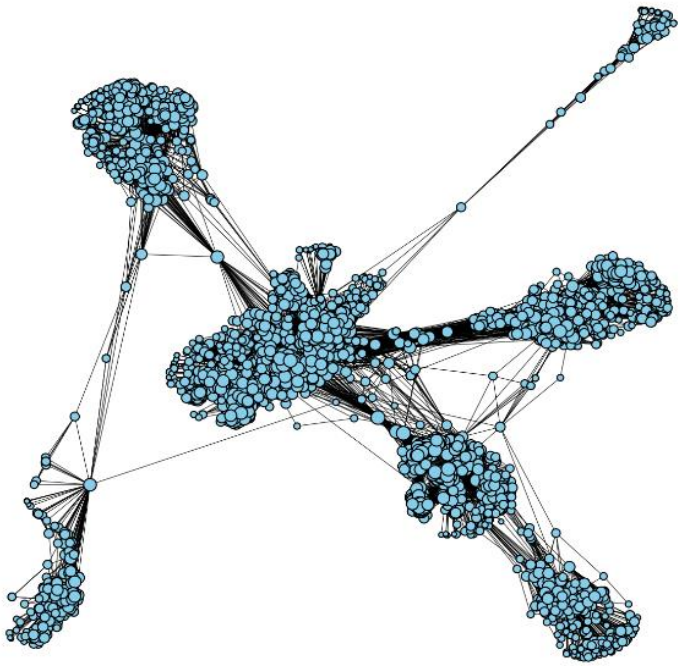

图 4-15 Zeng, Donglin 合作者子网络的社区发现结果

从图 4-11 中可以明显的看出原来的一个巨大的连通组件被聚集为了大概 6 个社区。其中值得注意的是中间的最大社区包含有表 4-5 中的三个高产作者，即包含的“Dunson, David B”，“Zeng, Donglin”和“Ibrahim,Jpseph,G”的社区。查询该社区内作者所在大学，发现大多数作者都来自北卡罗来纳州立大学，北卡罗来纳教堂山分校，而这两所学校又恰恰都位于北卡罗莱纳地区，因此这两所大学又被称为生物统计的黄金区。而这三所大学又是以生物医学统计以及对应的统计理论见长，所以该社区的合作的主要动机应该是地理位置和主题，非常符合实际社区情况。计算这个子网络上的社区密度和传导率，分别为 0.95 和 0.133。在社区实际意义被保证的条件下，本文所得到的社区有较高的密度和传导率，所以本文所发现的社区是比较紧密的，保持着良好的社区拓扑性。至此，统计研究领域的异质网络社区的多维社区被挖掘出来，具体的多维社区的局部示意图如下图 4-16 所示：

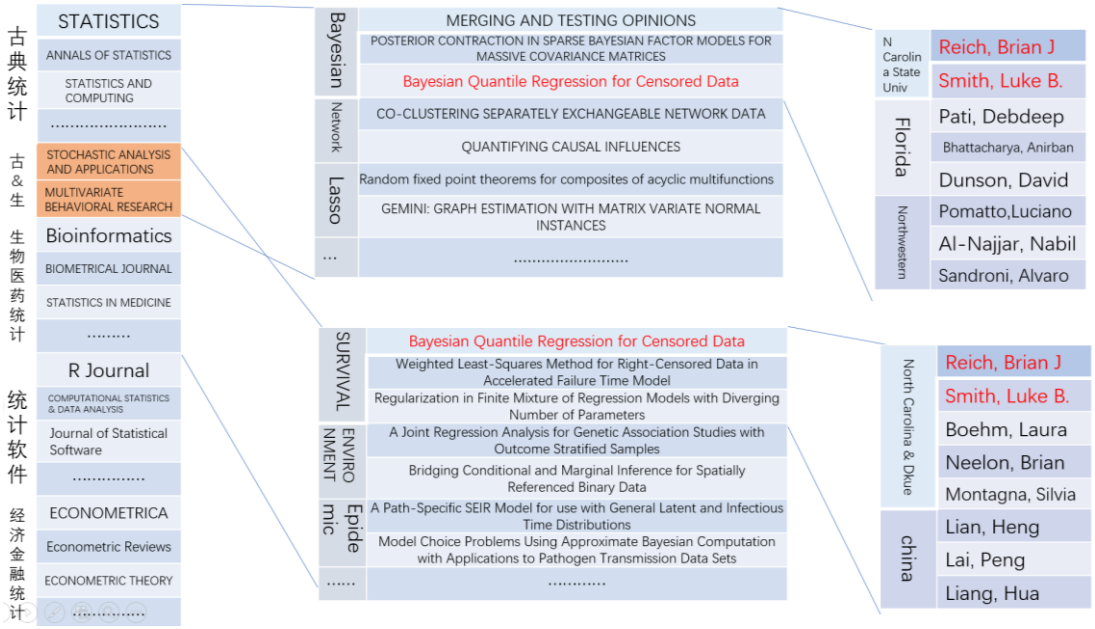

图 4-16 多维社区示意图（红色为交叉社区）

第五章 总结与展望

5.1 论文研究成果

本文在统计研究领域的 2014 年到 2018 年之间的论文数据基础上，构建了基于统计与概率领域的论文科研异质多维网络。本文主要有如下几个成果：

1、首先本文回顾了有关多维异质网络和社区发现的文献，包括基于单模网络的社区发现算法和基于异质网络的社区发现算法。其次，本文选取了异质网络的社区发现方法，对科研论文的异质网络进行了多维社区发现的尝试。

2、利用爬虫收集了大量的论文数据，对统计科研异质网络进行了新的数据结构储存的尝试，并借此搭建好包括杂志层，论文层以及作者层的三层新型异质网络结构，并设计好了层与层之间的连边效应和信息传递状况。

3、利用 LDA 主题模型进行了杂志层的社区发现，顺利的发现了 4 个典型的主题杂志社区并将每本杂志进行了归类，计算出了每本杂志对于每个社区的归属概率，并且通过连边效应将社区信息传递给论文层。论文层基于上层的社区发现信息被划分成了多个子网络。

4、设计了论文层和作者层的节点与节点之间的相似度，其中包括拓扑相似度和属性相似度，比较全面的考虑了层内节点之间的相似状况。充分综合考虑了节点属性的相关信息，使节点的相似度衡量更具有现实意义。

5、通过标签传递算法对计算完节点相关度和概率转移矩阵的论文层和作者层中的子网络进行了社区发现。从而得到了 30 个论文引用社区以及与之对应的 92 个作者合作社区，与传统方法相对比具有运行效率高，计算复杂度小，信息利用率高，社区实际意义强等优点。

5.2 论文研究难点与不足

1、本文所提供的方法并不是一个全局性的社区发现算法，所以只能由上至下的进行社区寻找，这样会损失较多的网络结构信息，不能很好的反映异质节点和异质连边之间的关系。并不能做到下层节点对上层节点的反馈，使得社区无法根据节点的不同进行实时的改变，社区的意义被牢牢地固定在了文本主题上。因此较为依赖 LDA 文本的分析结果。

2、本文所涉及的连边效应的传递问题。本文采取的是上层网络信息完全地传递到下层，并未考虑层内连边对层间连边的影响，孤立地看代了两种连边效应，使得下层社区被上层社区划定了范围，导致社区发现的不完善。

3、再划分子网络是为了追求更高的计算效率切割了子网络之间的连边，使得在进行社区发现时忽略了较多的拓扑信息。

4、因为是一个特定的网络架构，所以此种社区发现算法只能针对本文所提出的异质网络，缺乏算法的有效拓展性。

5、简单的基于拓扑的社区发现算法无法在这样大的网络上进行社区发现，而基于特殊架构的其他异质网络社区发现算法无法再统计研究异质网络上进行社区发现。所以本文社区发现算法无法针对一个数据进行横向对比，也因此无法很好地反映出该算法在时间和计算上的优势。

6、数据上利用了 2014 年到 2018 年之间的统计论文数据，如果能进一步拓展年限，会使得网络的连接更为紧密，会增强社区发现效果。

5.3 本文工作展望

首先，在下一步的工作中，会基于这个统计研究异质网络的架构，尝试着进

行全局的异质网络的发现。在这个过程中会完善连边效应假设。让层间连边的信息完全传递变为部分传递。将层间连边的又上到下的单向传递变为双向传递。将层内连边与层间连边之间的完全无影响的强假设放宽，会增加异质连边的相互影响，这样就会是得算法从一个局部无反馈的网络变为一个全局有反馈的网络。那么网络的异质节点之间可以进行相互监督，社区的发现将不仅仅依赖于主题，还会增加更多信息，当然这需要更高的数学能力。

数据上会进一步增加文献数据，使得网络中的连边逐渐增加，网络中的信息会进一步的增多，那么社区发现的效果和实际意义也会提升。

最后，还会进一步地设计异质网络的结构，使得其异质网络更加普通化，这样能够使其他的异质算法在该统计异质网络上运行，可以增加算法与算法之间的比较。

参考文献

[1] Girvan M, Newman ME. Community structure in social and biological networks[J]. Proc Natl Acad Sci U S A, 2002, 99(12):7821-7826.

[2] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 清华大学出版社, 2006.

[3] Eric D. Kolaczyk, Gábor Csárdi.Statistical Analysis of Network Data with R[M].Springer New York,2014.

[4] Clauset A, Newman M E, Moore C. Finding community structure in very large networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 70(2):066111.

[5] Fortunato S, Hric D. Community detection in networks: A user guide[J]. Physics Reports, 2016, 659:1-44.

[6] Chen Q, Shi D. The modeling of scale-free networks [J]. Physical A Statistical Mechanics & Its Applications, 2004, 335(1):240-248.

[7] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics, 2008, 2008(10):155-168.

[8] Zhang P, Wang J, Li X, et al. Clustering coefficient and community structure of bipartite networks[J]. Physica A Statistical Mechanics & Its Applications, 2008, 387(27):6869-6875.

[9] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 2014, 49(2):291-307.

[10] Pothen, Alex, Simon, Horst D, Liou, Kan-Pu. Partitioning sparse matrices with eigenvectors of graphs[J]. Siam J.matrix Anal.appl, 1990, 11(3):430-452.

[11] Luxburg U V. A tutorial on spectral clustering[J]. Statistics & Computing, 2007, 17(4):395-416.

[12] Donetti L, Muñoz M A. Detecting network communities: a new systematic and efficient algorithm[J]. Journal of Statistical Mechanics Theory & Experiment, 2004, 2004(10):10012.

[13] P. Pons and M. Latapy. Computing Communities in Large Networks Using Random Walks[J]. Computer and Information Sciences. 2005,284-293.

[14] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(3 Pt 2):036106.

[15] Gregory, Steve. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2009, 12(10):2011-2024.

[16] Eaton E, Mansbach R. A spin-glass model for semi-supervised community detection[C]// Twenty-Sixth AAAI Conference on Artificial Intelligence. AAAI Press, 2012:900-906.

[17] Hajeer M, Sanyal S, Dasgupta D, et al. Clustering online social network communities using genetic algorithms[J]. 2013.

[18] Wu, Peng, and L. Pan. "Mining Application-aware Community Organization with Expanded Feature Subspaces from Concerned Attributes in Social Networks." (2017).

[19] 孙怡帆，李赛. 基于相似度的微博社交网络的社区发现方法[J]. 计算机研究与发展, 2014, 51(12):2797-2807.

[20] Lin P, He T, Zhang Y. Microblog searching module based on community detection[C]// International Conference on Intelligent Computing Theories. 2013:112-119.

[21] Kessler, M. (1963). Bibliographic coupling between scientific papers. Journal of American

Documentation, 14, 10-25.
[22] 闫光辉,舒昕,马志程,等. 基于主题和链接分析的微博社区发现算法[J]. 计算机应用研究, 2013, 30(7):1953-1957.
[23] Huang, Xin, H. Cheng, and J. X. Yu. Dense community detection in multi-valued attributed networks. Elsevier Science Inc. 2015.
[24] Tang A, Viennet E. Community detection based on structural and attribute similarities[J]. Achi, 2012:7-12.
[25] 王金龙, 徐从富, 骆国靖. 面向异质关系的社区挖掘[J]. 计算机应用, 2007, 27(12):3016-3018.
[26] Cai D, Shao Z, He X, et al. Community Mining from Multi-relational Networks.[J]. 2012, 3721:445-452.
[27] Liu H, Wang C D, Lai J H, et al. Modularity in complex multilayer networks with multiple aspects: a static perspective[J]. Applied Informatics, 2017, 4(1):7.
[28] Liu Y, Tang J, Han J, et al. Community evolution detection in dynamic heterogeneous information networks[C]// Eighth Workshop on Mining and Learning with Graphs. ACM, 2010:137-146.
[29] 尹沐. 基于文献数据的 Graph OLAP 技术研究[D]. 北京邮电大学, 2013.
[30] Mane K, BaRner K. Mapping topics and topic bursts in PNAS[J]. Proc Natl Acad Sci U S A, 2004, 101(Suppl 1):5287-5290.
[31] Greene D, Doyle D, Cunningham P. Tracking the Evolution of Communities in Dynamic Social Networks[C]// International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2010:176-183.