多类型实体演进学术网络:观察、建模和分析

刘佳琪[1)] 傅洛伊[2)] 孔令坤[2)] 甘小莺[2)] 王新兵[2)]

[1)](西北工业大学计算机学院 西安 710129)

[2)](上海交通大学电子信息与电气工程学院 上海 200240)

摘 要 近年来,学术网络经历了快速的发展.该网络结构通常包含多种类型的实体和复杂多样的实体间关系,其中作者、论文和主题是学术网络中最具代表性的三类实体,它们之间存在着类型丰富且随时间演进的交互关系.对学术网络的结构及其演进机制进行研究具有重要意义.然而,大多数相关工作仅考虑了学术网络中单一类型实体内的交互关系,如作者间的合著关系、论文间的引用关系或主题间的关联关系,未能对学术网络中不同类型的实体进行有效整合、给出对多类型实体演进学术网络进行研究分析的统一框架.为了解决上述问题,该文提出了"多实体学术模型",将作者、论文、主题三种类型的实体整合进一个统一的理论框架,并通过研究实体间的连接关系及其演进情况来对学术网络进行刻画.其贡献点主要包括以下几个部分:(1)该文同时考虑了论文、作者、主题对学术网络性质的影响,并对包含 690 万条数据的微软学术网络数据集进行了统计分析,得到以下两个方面的结果:验证了一般社会网络中存在的性质如节点度服从幂律分布、幂律指数随时间收敛、网络稠密化等在多类型实体演进学术网络中同样存在;发现了一些多类型实体演进学术网络中特有的性质,如规模较大的实体往往具有更高的演进速率,幂律指数随时间的波动与收敛及实体间交互演进等,并根据其演进特性提出实体内演进、实体间演进以及交互演进三种演进模式;(2)基于上述观测现象,提出了多实体学术模型:该模型通过构建异构图的方法同时刻画同一类型实体内和不同类型实体间的连接关系,并通过直接演进、间接演进、内部演进等多种策略刻画连接关系的演进模式,具有很强的理论保证.该模型可用于多类型实体演进学术网络的数学刻画,并为学术关系预测、学术影响力传播、推荐算法设计等应用提供理论基础;(3)分别从理论分析和实验验证两个方面证明了模型的有效性:理论方面,通过数学推导证明了多实体学术模型具有节点度随时间呈多项式速率增长、节点度服从幂律分布、网络稠密化等性质;实验方面,根据多实体学术模型生成了相应的仿真网络并证明其具有上述性质.

关键词 学术网络;演进网络;多类型实体;异构连接;交互演进

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.01791

Evolving Multi-Entity Scholarly Networks: Observation, Modeling and Analysis

LIU Jia-Qi[1)] FU Luo-Yi[2)] KONG Ling-Kun[2)] GAN Xiao-Ying[2)] WANG Xin-Bing[2)]

[1)](*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710129)

[2)](*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai* 200240)

Abstract Recent years have witnessed the rapidly growing scholarly information. All the information, when combined together, leads to the formation of the scholarly network that contains three major entities, i.e., paper, author and topic. All the three entities interact with each other as time goes by, which results in an evolving multi-entity scholarly network. As a matter of fact, studying properties of scholarly networks and getting insight of their evolving mechanism have important implications. However, most works focus on single entity of the network, e.g., sub-networks generated by co-authorship, citation or topic relationship; while

收稿日期:2018-11-22;在线发布日期:2019-09-24.本课题得到国家重点研发专项(2018YFB1004705,2018YFB2100302)、国家自然科学基金(61822206,61532012,61602303,61829201,61960206002)、腾讯犀牛鸟(20180116)资助.**刘佳琪**,博士,副教授,主要研究方向为社交网络建模、分析、算法设计.E-mail:jqliu@nwpu.edu.cn.**傅洛伊**,博士,特别副研究员,主要研究方向为物联网、信息传输与优化、多播性能分析.**孔令坤**,博士研究生,主要研究方向为社会计算、数据挖掘.**甘小莺**,博士,副教授,主要研究方向为群智感知、网络经济学.**王新兵**(通信作者),博士,特聘教授,国家杰出青年科学基金入选者,主要研究领域为智能物联网、无线网络.E-mail:xwang8@sjtu.edu.cn.

few of them merge multiple kinds of entities into one single fabric to obtain the understanding of scholarly networks from an overall perspective. To bridge this gap, we are motivated to give the model that incorporates entities of paper, author and topic into one single framework—Multi-entity Scholarly Model (MSM), which amalgamates entities of author, paper and topic into a framework to simulate interactions among different entities, and thus presenting the co-evolution within scholarly networks. Our contributions are listed as below. (1) Our first contribution is to originally explore comprehensive properties in scholarly networks with the concern of multiple entities, i.e., paper, author and topic. Based on scholarly datasets provided by Microsoft, which contain about 6.9 million papers, we use data-mining and other big-data analyzing approaches to observe patterns in the growth of the scholarly network. On one hand, we observe some similar features to those that have already been discovered in many traditional social networks, such as power-law degree distribution, densification, etc. On the other hand, there also exists several unique features in scholarly networks, like faster growth rate of the entity that has a bigger size, varying and converging exponents in power-law distributions with time, and the simultaneous co-evolution of all entities, etc. All these evolving features, with the awareness of multiple entities, can be categorized into three types, i.e., inter-evolution, intra-evolution as well as the co-evolution on the whole. (2) Given empirical observations, our next significant contribution is for the first time establishing a comprehensive modeling of evolving scholarly networks. Combining entities of paper, author and topic in singe fabric, the proposed model captures both the inter-correlation and intra-correlation of the three entities during the evolving process. Particularly, inter-correlation is characterized through tripartite graph whose evolving process follows the mode of preferential attachment and intra-correlation of nodes within each entity is described as intra-degree (which we define) power-law distribution, degree densification, etc. The model can be used in characterizing evolving multi-entity scholarly networks and provides theoretical guarantee for applications such as relation prediction, influence propagation and recommendation. (3) Our third contribution is to validate the effectiveness of MSM through both theoretical analysis and empirical simulation. Based on the constructing methods of random arrival, preferential attachment, edge copying and the assumption of the affiliation relationship inside entities, we successfully obtain the growing rate of nodes' degree, power-law distributions inside or among multiple entities and the densification of the entire network. Further, we also implement simulating approaches to validate that our model can accurately reproduce real scholarly networks.

Keywords scholarly networks; evolving networks; multi-entity; heterogeneous connections; co-evolution

1 引 言

近年来,学术界和工业界的相关研究发现学术信息正在经历大规模的快速增长.在研究人员的不懈努力之下,作者间的合著关系、论文间的引用关系、主题间的关联关系、作者论文间关系、作者主题间关系、论文主题间关系等信息,以及这些信息随时间发生的交互演进模式得到了有效的收集和整理.所有上面提到的信息共同构成了多类型实体演进学术网络(Evolving Multi-entity Scholarly Networks).如图1所示,一篇发表的文章至少包含以下三类实体:文章的标题,称为"论文(paper)",图中以字母P表示;发表该文章的研究人员姓名,称为"作者(author)",图中以字母A表示;以及文章中包含的关键词,称为"主题(topic)",图中以字母K表示.这三种类型的实体之间存在着密切的联系,例如同一篇文章中的作者之间存在着合著关系,作者与论文

之间存在着发表关系，不同论文之间存在着引用关系，论文与所涉及的主题之间存在着包含关系等.另外，上述三种类型的实体及其连接关系会随着时间推移而发生演进，如新的作者、论文、主题个体，以及新的实体间关系会随着时间推移不断产生.这一演进模式导致了网络规模的增长和实体内、实体间关系复杂性的增加.

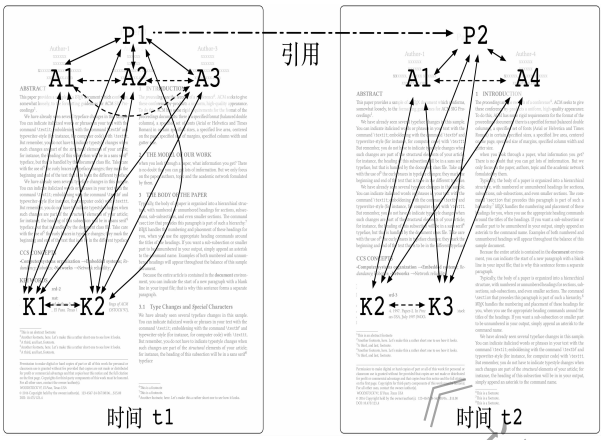


图 1　学术网络中的异构实体间关系

Xia 等人在文献[1]中指出，对学术网络中实体间异构连接关系及其演进过程的研究是非常重要的.例如，对学术网络中作者-论文关系的研究可用于衡量某个特定作者的论著贡献度，并可据此判定该学者的学术成绩.同样地，对主题-论文关系的分析有助于获得每个研究主题的研究热度，并可据此进一步建立主题研究体系.另外，对学术网络的研究分析不仅在学术界研究中具有重要作用，在很多其他实际应用中也是不可缺少的一部分.例如，学术网络研究结果对社会学家理解研究人员间的交互关系，对相关人员整理、制定学术资源分配策略，对学术机构、研究人员、学生的学术生活都有着重要的指导作用.对学术网络中实体间异构连接关系及其演进过程进行研究的重要手段之一是对该过程建立数学模型.建模的意义在于探究学术网络中异构连接关系演进过程的内在机理并利用数学工具对其进行刻画，从而为学术网络的理论分析打下基础，并最终达到改进技术实现方式、提高网络服务质量的目的.然而，目前大多数现有的建模工作[2-6]仅对学术网络中的单类型实体关系进行了讨论，如作者间合著关系、论文间引用关系、主题间关联关系等.这些工作并未将不同类型的实体及实体间关系作为一个整体进行研究，因此也无法取得对于学术网络的全局性理解.

一般来说，对于多类型实体演进学术网络的研究存在以下三个方面的难点：(1) 数据验证方面的困难.目前基于实验数据的相关研究成果非常有限，这是由于大规模、全种类的实际学术网络数据获取过程存在一定困难.尤其是在对多类型实体关系进行研究时，由于需要获得全部实体间交互数据，这一困难变得更加明显；(2) 模型建立方面的困难.相较于现有的单类型实体学术网络模型，对不同类型实体间的交互关系及其演进方式进行数学建模要更加复杂.这是因为多类型实体学术网络中存在着间接演进，如新发表的论文不仅会与某些已有作者和主题建立联系，还有可能使得原本没有关联的作者和主题间产生新的联系；(3) 理论分析方面的困难.虽然已经有不少工作对学术网络进行了数学建模，但其中大部分都缺乏有效的理论支撑，仅具备仿真验证.仿真方法虽然能够在一定程度上为模型性能提供补充验证，但是要从根本上证明模型的有效性必须依赖完备的理论分析，而这一过程往往存在较大困难.

为了解决上述困难，该文提出了多实体学术模型(Multi-entity Scholarly Model，MSM).模型将作者、论文、主题整合进一个统一的理论框架，并通过研究实体间的连接关系及其演进情况来对学术网络进行数学刻画.首先，对一个包含约 690 万个数据条目的大规模数据集进行了统计分析，据此观察得到了多类型实体演进学术网络中实体间的交互演进性质.其次，提出了多实体学术模型，该模型通过构建异构图的方法来对前一步中观察到的性质进行数学建模.最后，分别从理论和实验两方面对所提模型的正确性和有效性进行了验证.工作的贡献点主要包括以下 3 个方面：

(1) 实验观测.首先，通过实验探索了多类型实体演进学术网络中作者、论文和主题之间的交互演进关系.实验所用数据集来自于微软公司开源平台上提供的微软学术网络数据集.该数据集包含 6 个方向各异的研究领域，共 690 万篇文章条目，拥有庞大且丰富的数据资源.实验得到以下两个方面的结果：一方面，验证了一般网络中存在的性质，如节点度服从幂律分布、幂律指数随时间收敛、网络稠密化等在多类型实体演进学术网络中同样存在；另一方面，发现了一些多类型实体演进学术网络中的特有性质，如作者、论文、主题间存在交互演进，规模较大的实体往往具有更高的演进速率等.

(2) 数学建模.其次，提出了多实体学术模型来对多类型实体演进学术网络进行数学建模.该模型

给出了一个能够同时刻画作者、论文、主题三种类型的实体及其演进关系的统一框架.框架的主体是一个包含了三种节点类型的异构图.图中同时包含不同类型实体间的连接关系和同一类型实体内的连接关系,并且这些关系会随着时间推移发生演进.演进过程以优先连接准则(Preferential Attachment)为基础,同时包含直接演进、间接演进、内部演进等多种不同策略.该模型可以有效刻画多类型实体演进学术网络,并为学术网络中的内容分析类工作与算法设计类工作的开展提供理论基础.

(3)分析验证.最后,分别从理论和实验两方面对所提出模型的有效性进行了验证.理论方面,根据模型中的演进算法对模型生成的网络进行了相应的理论分析并计算得到了网络中节点度的增长速率表达式、不同类型实体间节点的度分布、网络密度随时间的变化情况等结果,为模型提供了理论保证.实验方面,通过计算机仿真的方法获得了模型生成网络并对其性质进行了验证,结果证实多实体学术模型可以准确地刻画多类型实体演进学术网络中的各个性质,进一步证明了模型的有效性.

本文首先在第2节中列出相关参考文献,接下来在第3节中给出基于微软学术网络数据集的观测结果,然后在第4节中介绍多实体学术模型,并在第5节中对其进行理论分析;第6节为实验仿真内容;最后,第7节对文章内容进行总结.

2 相关工作

目前已存在一些多类型实体演进学术网络的相关研究工作.下面分别从演进网络和学术网络两个方面对这些工作进行阐述.

2.1 演进网络

一直以来,演进网络都是一个重要的研究方向[7-10],其相关性质已经得到了比较充分的研究.研究成果表明演进网络具有如下性质:节点度服从幂律分布(Power-law Degree Distribution)、稠密化(Densification)、最大连通分支(Giant Component)等.幂律分布指的是网络中节点度服从幂律分布,这一性质已经在众多社交网络如因特网[11]、万维网[12]、学术网络[13]中被观测验证;稠密化指的是网络中边的数目与节点数目之比随着时间推移而增大[14];最大连通分支指的是网络中存在的规模最大的一组互相连接的节点[15],通过计算最大连通分支规模与全网节点规模之比可以衡量网络结构的紧密程度.上述性质已经被证明广泛存在于多种不同的演进网络中.在此基础上,第3节将验证这些性质同样存在于演进学术网络中.

针对上述性质,研究者们已经提出了多个数学模型[9,11,16-18]来对其进行刻画.例如,Kumar等人提出了复制模型(Copying Model)[19],该模型的基本思想来自于实际网络中观测到的一个现象——新网页的生成过程通常为复制某个旧的网页并对其内部连接进行部分替换.Leskovec等人[20]运用克罗内克积(Kronecker Product)矩阵操作来生成演进网络,具体演进过程为迭代地对两个图做克罗内克积来构造自相似图.Leskovec等人提出了森林火灾模型(Forest Fire Model)[14].在该模型中,每个新产生的节点均匀随机地选择一个已有节点并选择其未被访问的邻居进行连接,递归地重复此步骤并最终得到整个演进网络.除了上述模型之外,Chakrabarti等人提出的随机图模型(R-MAT Model)[21],Barabasi等人提出的优先连接模型(PA Model)[22-23]等也可用于演进网络的刻画.

然而,上述相关工作研究的均为社交网络,提出的模型也仅仅刻画了社交网络的性质,并未考虑不同类型实体间的交互关系,因此不能直接用来刻画多类型实体演进学术网络.

2.2 学术网络

目前已有研究工作[2,6,24-25]意识到了学术网络中多类型实体和关系的存在及其对网络分析的重要影响.Mo等人[26]对学术网络中的"主题-主题关系"、"作者-主题关系"和"作者-作者关系"进行了研究.Tang等人[27]提出ArnetMiner模型来对学术网络进行挖掘处理,该模型同时刻画了主题与论文、作者、发表刊物间的关系,然而刻画中使用三个独立模型,未能揭示出各种连接关系的交互过程.Yan等人[28]尝试使用一个统一框架刻画论文-论文关系、作者-作者关系和主题-主题关系,然而,该工作仅研究了同类型实体间的连接关系,未考虑跨类型实体间的交互关系.另外,Yang等人[2]提出了共同主题模型来解决学术网络中的排序和预测问题.Lin[5]和Wang等人[24]研究了主题节点在学术网络中的演进情况.Liu等人[6]研究了学术网络的异构结构并刻画了其中的作者间合著关系.然而,上述这些工作均仅研究了学术网络中的部分连接关系,未能从一个全局角度对其进行刻画.

另外，由于学术网络是一个典型的多类型异质网络，可以使用知识图谱和异质信息网络结构来对其进行建模.(1)知识图谱.该结构假设网络中存在多种不同类型的实体和关系，每一个“(主语实体，关系，宾语实体)”三元组表示一个事件或事实.例如，在学术网络中，“(作者 1，发表，论文 1)”表示事件“作者 1 发表了论文 1”，“(作者 1，研究，主题 1)”表示事件“作者 1 研究主题 1”.利用各个实体和关系的固有联系，可以将其映射至低维向量空间，并借此完成信息抓取[29]、问题回答[30]、语义分析[31]等任务；(2)异质信息网络.该结构是定义在多类型对象和多类型关系上的一个有向图，对象之间存在复合关系，由两者间的元路径表示.例如，在学术网络中，“作者 1-论文 1-作者 2”和“作者 1-论文 2-主题 1-论文 3-作者 2”两条元路径分别代表作者 1 和作者 2 合著了一篇论文和以及作者 1 和作者 2 在同一个研究主题上发表了论文.采用数据挖掘手段对该结构中的信息进行整合、处理，可以完成如聚类[32]、分类[33]、推荐[34]等多项任务.上述两种模式的优势在于结构灵活，能够融合丰富的语义和整合大量信息.然而，它们在刻画真实学术网络方面也存在一些问题.一方面，灵活的结构为网络分析带来一定困难，目前的分析方法大多为机器学习类技术，如随机游走模型、主题模型、矩阵模型等，这些方法对数据依赖较重且不具备理论保证性；另一方面，上述结构均未对知识/信息的演进问题进行考虑，无法刻画学术网络中作者、论文、主题及其关系随时间的变化以及互相影响情况.

3　现象观测

基于真实数据集的统计分析结果对研究网络性质具有重要作用.下面在微软学术网络数据集的基础上，分别从结构性质和演进性质两个角度对多类型实体演进学术网络的性质进行总结讨论.

3.1　数据集介绍

本文采用微软学术网络数据集①(Microsoft Academic Graph)作为研究对象.该数据集由微软公司官方发布，包含多种不同类型的实体及其交互关系信息，如论文标题、作者、研究领域、时间、引用情况和所发表的会议或期刊等数据.数据集收录了 1.26 亿篇文章，发表年份从 1800 年到 2016 年不等.完整的微软学术网络数据集规模庞大，为了避免处理过程过于复杂，本文从中提取出若干个子数据集进行分析研究.为了保证观测结果的普适性，提取数据挖掘、网络研究、文学、金融、人工智能和机器学习六个不同研究领域的子数据集作为研究对象，其中人工智能与机器学习领域是近几年计算机学科的热门研究方向，对其展开分析能够为计算机学科的发展提供重要参考意义.上述六个子数据集均为大规模数据集，其详细统计特性见表 1.

表 1　数据集的统计特性

数据集	论文条目数量	作者条目数量	主题条目数量
数据挖掘	1042279	1703828	403
网络研究	1093537	1391869	774
文学	679350	939544	446
金融	1949028	2716094	968
人工智能	1076665	1202845	956
机器学习	1158514	1212139	953

后续实验过程中对论文、作者、主题实体的产生时间定义如下：

论文：该论文的发表日期；

作者：该作者首篇论文的发表日期；

主题：该主题相关的首篇论文的发表日期.

实验对数据集中的两类关系展开研究：(1)实体内关系.即同类型实体间的关系，具体为作者间的合著关系、论文间的引用关系以及主题间的关联关系(共同论文关系)；(2)实体间关系.即不同类型实体间的关系，具体为作者和论文间的发表关系、作者和主题间的研究关系以及论文和主题间的关联关系.对上述关系的研究将从两个角度进行：

结构性质.用于描述网络中节点间的连接关系，通过节点度分布进行数学刻画，具体包括两部分：节点在实体内关系中的度，称为实体内度，以及节点在实体间关系中的度，称为实体间度.其数学定义如下：

定义 1.　实体内度.对于任意一个类型为 i 的节点 v，其实体内度 d_v^{ii} 定义为节点 v 的类型为 i 的邻居数量.

定义 2.　实体间度.对于任意一个类型为 i 的节点 v，该节点与类型为 j，$j\neq i$，的节点之间的实体间度 d_v^{ij} 定义为节点 v 的类型为 j 的邻居数量.

学术网络中包含作者、论文和主题三种节点类型，因此有 $i\in\{a,p,t\}$.作者节点的实体内度

① https://academic.microsoft.com/

d^{aa} 指的是与该作者建立合著关系的其他作者节点的个数，论文节点的实体内度 d^{pp} 指的是该论文引用其他论文或被其他论文引用的次数，主题的实体内度 d^{tt} 指的是与该主题相关的其他主题节点的个数. 另外，论文-主题关系中，主题的实体间度 d^{tp} 指的是与该主题相关的论文数目，论文的实体间度 d^{pt} 指的是该论文内容涉及到的主题数目. 同样地，在作者-论文关系和作者-主题关系中也存在类似含义.

演进性质. 用于描述网络结构随时间发生的规律性变化，通过节点度及其分布的变化情况进行数学刻画，具体包括三部分：实体内度的演进、实体间度的演进和交互演进.

3.2 结构性质

在对结构性质进行讨论之前，首先给出幂律分布的数学表达式如下：对于一个随机变量 $x\in \boldsymbol{Z}^{+}$，若它服从下列条件则称其服从幂律分布

$$\boldsymbol{P}\{x=k\}=\eta k^{-\varphi},$$

其中 φ 为幂律指数，η 为常数系数.

3.2.1 实体内度

图 2 以引用网络为例展示了各个子数据集的实体内度分布，结果显示其服从幂律分布. 具体的分布参数如表 2 所示. 由于引用量可以在一定程度上反映论文的学术价值，该结果实质上表明学术网络中大多数论文价值有限，仅有少数论文意义重大. 因此，如何甄别并挑选有价值的论文是学术网络研究的重要内容之一，同时也是相关人员整理、制定学术资源分配策略时的重要参考内容. 另外，文中虽未直接给出作者合著网络和主题关联网络中的实体内度分布图，但结果表明其同样服从幂律分布. 与引用网络中结论类似，这个结果体现了研究人员交互频次和研究主题学科交叉的巨大差异.

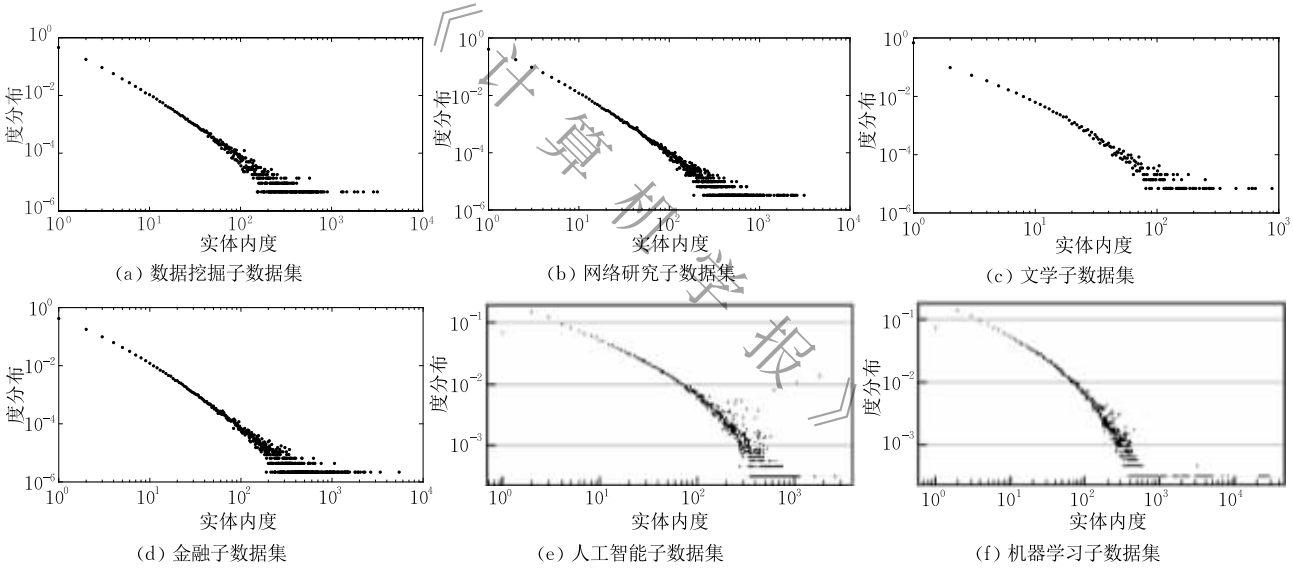


图 2 引用网络中的实体内度分布

表 2 引用网络中的实体内度分布参数

参数值	数据挖掘	网络研究	文学	金融	人工智能	机器学习
φ	2.15	2.27	2.51	2.28	2.51	1.35
η	1.41	3.47	1.99	2.51	20.27	0.05

3.2.2 实体间度

图 3 以金融子数据集为例展示了学术网络实体间度分布，结果显示其服从幂律分布. d^{tp} 和 d^{ta} 的分布表明大多数研究主题热度一般，而少数研究主题具有很高的热度，d^{ap} 的分布表明作者间的论文发表量存在巨大差异，d^{at} 的分布反映了作者跨主题研究能力的差异，并且 d^{pt} 以及 d^{pa} 的分布也反映了论文间性质的差异. 对上述结果加以分析和利用可以更好地理解学术网络，对学术机构、研究人员、学生群体日常研究中的选题、学术交流等活动具有重要指导意义. 另外，很多实际学术网络中，作者、论文、主题这三种类型的实体在数量上存在巨大差异. 例如，一篇论文通常仅会涉及有限的几个主题，而一个主题包含的论文数目通常为数千甚至数万个. 因此，从图 3 中可以观察到，六种类型的实体间度虽然均服从幂律分布，但是他们各自的幂律指数各不相同，详见表 3. 相较于其他两种类型的实体，主题节点实体间度的幂律指数 φ 较小. 这一结果意味着一个主题节点更有可能与大规模的作者和论文相关联.

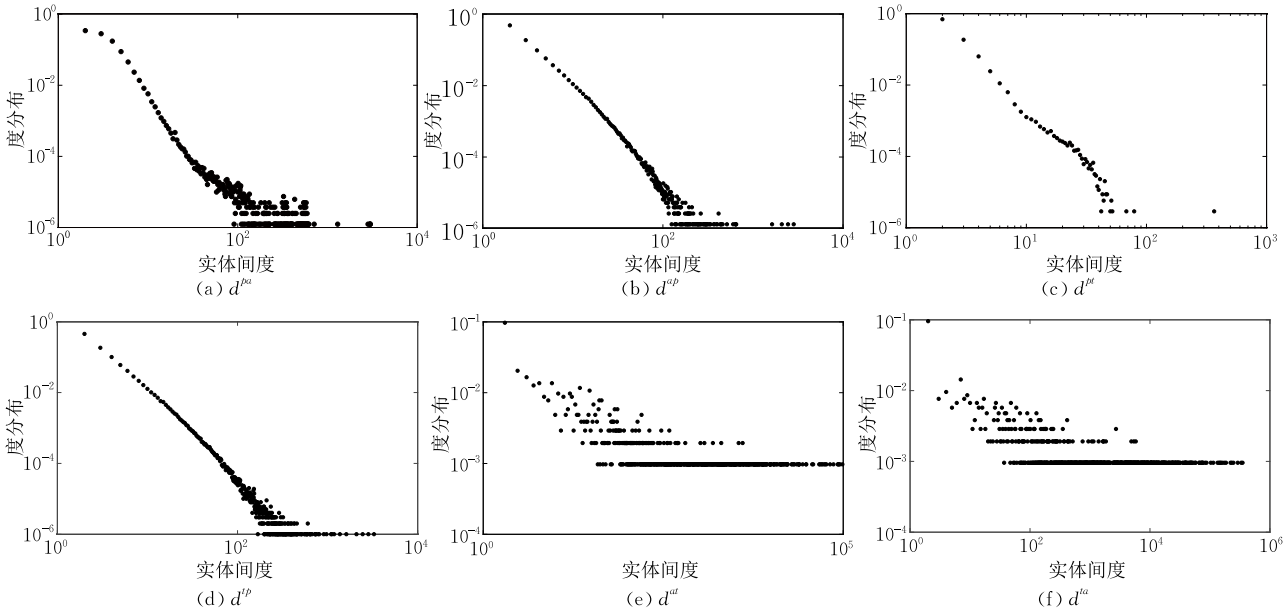


图 3　金融子数据集中的实体间度分布

表 3　金融子数据集中的实体间度分布参数

参数值	d^{pa}	d^{pt}	d^{ap}	d^{at}	d^{tp}	d^{ta}
φ	1.67	3.11	2.38	2.32	0.18	0.13
η	1.70	0.34	0.52	0.32	2.32	2.48

3.3　演进性质

上述性质为学术网络展现出的静态结构性质，该性质同样存在于大多数社交网络中. 下面对学术网络的演进性质，尤其是作者、论文、主题间存在的交互演进等特有性质进行讨论.

3.3.1　实体内度

下面以作者间的合著网络为例对学术网络中实体内度及其分布的演进过程进行讨论.

实体内度的演进. 实验统计了六个子数据集自 1970 年至 2016 年间合著网络总实体内度(合著网络中所有节点的实体内度之和)以及平均实体内度(合著网络中所有节点的实体内度的平均值)的演进情况. 总实体内度的演进情况如图 4(a)所示，结果显示 1970 年至 2000 年间，所有子数据集的总节点度均维持在一个比较小的数值上，除了金融子数据集外其余子数据集中总节点度规模均不超过 10^7 量级. 而从 2000 年开始，总节点度迎来了快速的增长，在短短 16 年间增长至 2×10^7，金融子数据集中的总节点度甚至达到 5×10^7，约为 2000 年数值的 3.5 倍. 这一快速增长得益于新世纪期间科学技术的快速发展. 以此为基础，学者之间拥有了更多的合作机会，进而构建起更多、更紧密的合著关系. 上述结论对于社会学家理解研究人员间的交互关系具有重要指导作用. 另外，平均实体内度的演进情况如图 4(b)所示，与总节点度的增长趋势基本一致. 从图中可以看出，总节点度和平均节点度均随时间呈现增长趋势. 第 4 节中提出的数学模型可以有效刻画上述规律，同时第 5 节将给出该结论相应的理论证明.

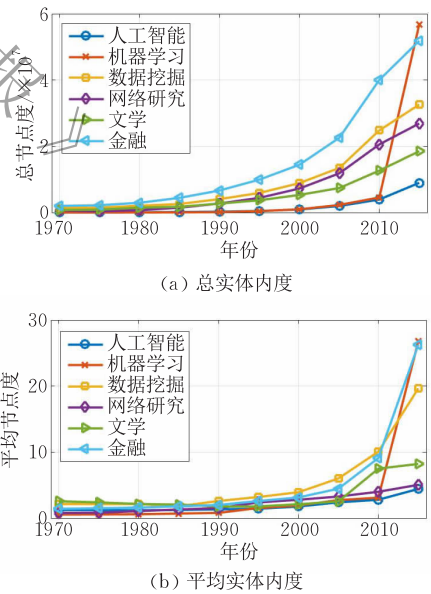


图 4　合著网络中实体内度的演进

实体内度分布的演进. 3.2.1 节中已得到实体内度服从幂律分布，因此，实体内度分布的演进具体体现为幂律指数 φ 的演进. 实验统计了六个子数据集自 1940 年至 2016 年间合著网络实体内度分布幂律指数的演进情况，如图 5 所示. 结果显示幂律指

数 φ 经历了“先波动,后稳定”的演进过程. 在 1990 年之前,六个子数据集的幂律指数 φ 均出现了较大波动. 以数据挖掘子数据集为例,其幂律指数 φ 的取值最低为 1.4,最高为 3.4,波动巨大. 而在 1990 年之后,六个数据集的幂律指数 φ 的取值均稳定在 1.8 左右. 该结果说明学术网络中实体内度分布的演进遵从着前期波动,后期稳定的演进规律. 另外可以看出,由于人工智能领域和机器学习领域近年发展迅猛,其幂律指数仍存在一定波动,并未完全达到稳定状态. “先波动,后稳定”现象产生的主要原因在于两个方面:(1) 前期演进模式尚未稳定. 节点和边的增长速率、各类型节点和边的比例等参数的波动导致实体内度分布存在波动,而后期这些参数趋于稳定,实体内度分布也趋于稳定;(2) 前期节点和边样本过少. 节点和边的产生本身存在随机性,样本数量过少会导致度分布参数的统计结果存在误差,而后期网络规模壮大后不再存在该问题.

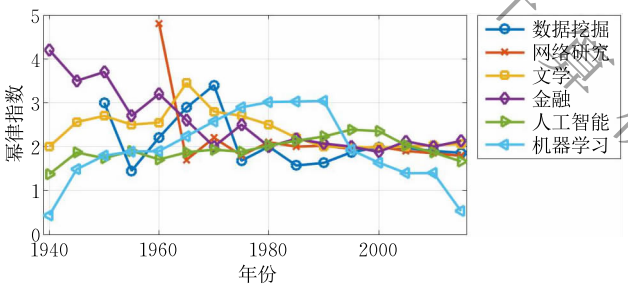
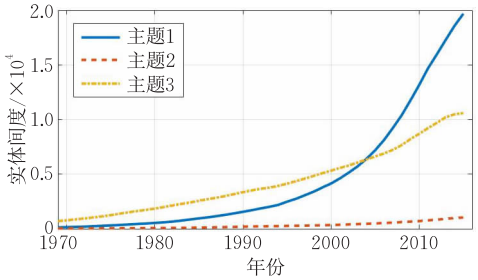


图 5 合著网络中实体内度分布幂律指数 φ 的演进

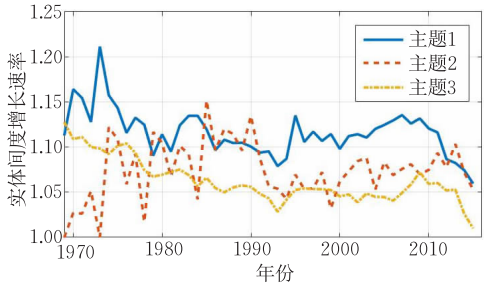
3.3.2 实体间度的演进

下面分别从微观角度和宏观角度,以论文-主题关系网络为例对学术网络中主题节点实体间度的演进过程进行讨论.

微观角度. 为了考察不同主题个体从诞生到壮大的演进过程,实验选取三个具有代表性的主题,观察其实体间度以及实体间度增长速率随时间的变化趋势. 所选取的三个主题分别命名为“主题 1”、“主题 2”和“主题 3”,其实际含义分别为“Principle Component Analysis”、“Simplification”以及“Nerve Action Potential”. 实体间度以及实体间度增长速率的定义与前面章节相同. 如图 6 所示,实验结果表明不同主题个体的实体间度均呈现增长趋势,但是增长速率各有不同且随时间发生变化. 在所选取的三个主题个体中,主题 1 具有较大的增长速率,且在2004年后成为规模最大的主题. 虽然规模较大,但是主题 1 和主题 2 的增长速率呈现下降趋势,如图 6(b)所示. 主题 3 虽然总体规模较小,但是其增长速率呈现增长趋势,具有较大的发展潜力.



(a) 实体间度



(b) 实体间度增长速率

图 6 三个代表性主题个体的实体间度演进

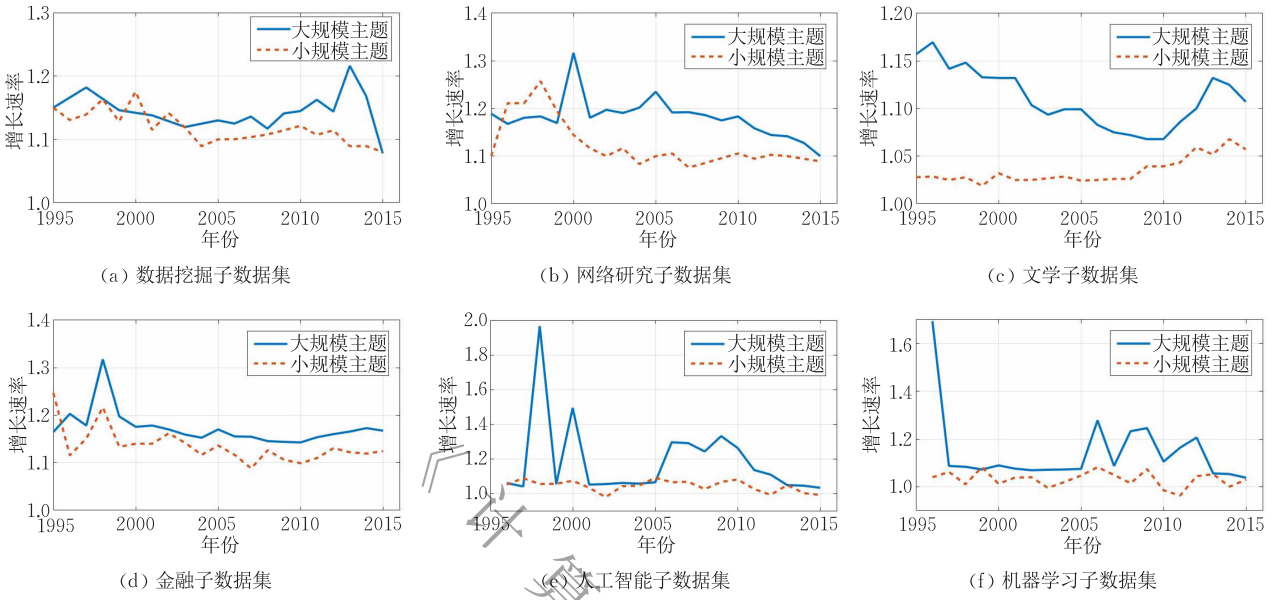
宏观角度. 接下来的实验将观察不同规模主题实体间度增长速率的差异. 实验中主题的规模定义为与该主题相关的发表论文数量. 注意此处仅从数量角度展开观测,主题的质量通常由各种因素联合决定. 实际学术网络中,不同主题之间的规模通常存在巨大差异. 规模较大的主题可能包含超过 5000 篇相关论文,而规模较小的主题可能只包含不足 500 篇相关论文. 那么,不同规模主题的实体间度增长速率是否存在差异?如果存在,导致该差异产生的原因是什么?接下来对上述问题进行研究.

为了后续讨论方便,首先对主题规模在第 t 年的增长速率定义如下:

$$Rate(t)=\frac{d^{tp}(t)}{d^{tp}(t-1)}.$$

接下来从数据集中选出部分主题节点并将它们划分为两组:第一组“大规模主题”包含的是所有规模大于 5000 的主题节点;第二组“小规模主题”包含的是所有规模小于 500 的主题节点. 为了保证对比组间的差异性,规模在 500 至 5000 之间的中等规模主题未被加入对比.

实验过程中，逐年计算这两组主题节点的平均实体间度增长速率并进行比较，结果如图 7 所示. 可以看出“大规模主题”的增长速率整体高于“小规模主题”，符合“rich get richer”规律. 这一现象背后的原因在于，一个较大规模的主题往往更容易吸引到研究者的注意力并进而引发其研究兴趣，被吸引的研究者做出的贡献则会进一步壮大该主题的规模. 因此，大规模主题的增长速率往往可以维持在一个较高的水平. 这个发现对于研究人员的选题过程具有重要的参考价值.



(a) 数据挖掘子数据集　(b) 网络研究子数据集　(c) 文学子数据集

(d) 金融子数据集　(e) 人工智能子数据集　(f) 机器学习子数据集

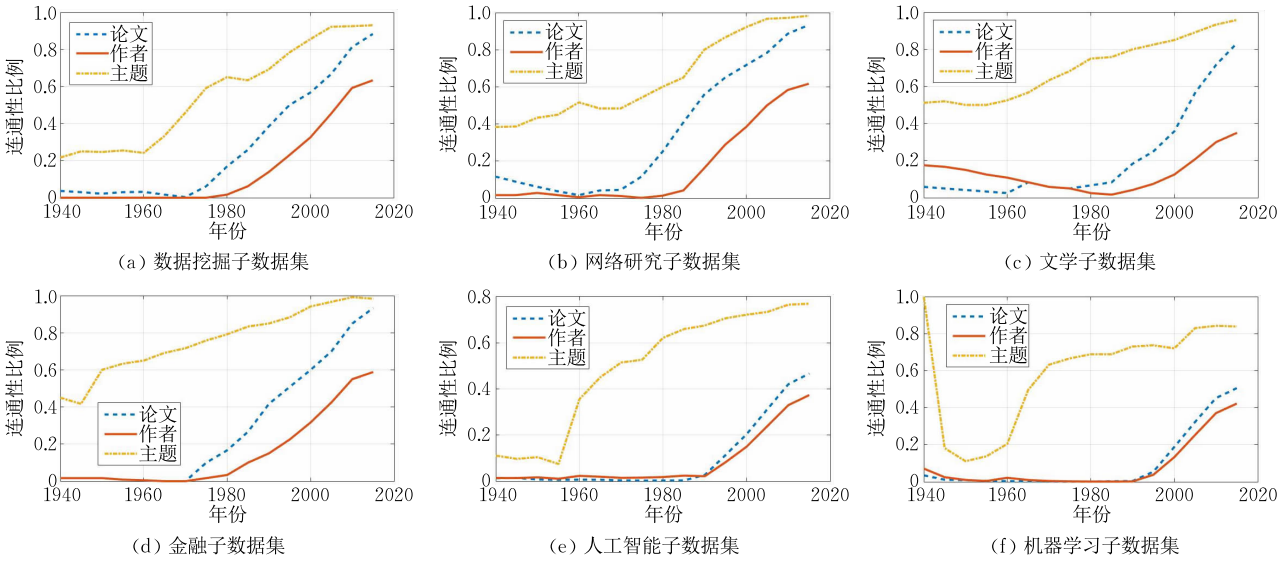
图 7　“大规模主题”和“小规模主题”的实体间度增长速率图

3.3.3　实体间的交互演进

除了实体内度和实体间度的演进，另一个值得关注的地方在于多类型实体学术网络中所有实体间的交互演进. 接下来，本文对多类型实体演进学术网络的一般性演进模式进行总结与刻画.

为了解决上述问题，首先需要了解实际学术网络中作者、论文和主题之间是如何相互促进、共同增强的. 考虑下面的情景——对于某个具有较高发展前景的研究领域，毫无疑问会有大量研究人员对该领域产生兴趣并开展相关研究，这个过程会激发出大量新的研究论文以及大量该领域下的研究主题方向，而新的研究成果又会反过来壮大该研究领域，吸引更多的科研人员参与其中并拓展出更多的新研究主题方向. 另外，上述过程还会使得该领域内论文引用量增加，带来更强的作者合著关系和更紧密的研究主题关联关系. 最终，学术网络的整体结构会变得更加紧密. 因此，对于学术网络中实体间交互演进的研究是非常必要的，有利于研究人员明确学术网络演进的内在机理，进而为相关人员整理、制定学术资源分配策略提供参考.

网络结构的紧密程度可以通过两个指标来进行衡量：连通性比例和网络直径. (1) 连通性比例定义为网络中最大连通子图所包含节点的数量与全网节点数量之比，用于衡量网络的连通程度，该指标数值越大则网络中互相连通的节点数量越大，网络越紧密；(2) 网络直径定义为网络中任意两个节点间距离的最大值，用于衡量网络的小世界特性，该指标数值越小则网络中节点间连通路径的长度越小，网络越紧密. 联合两个指标可以对学术网络中用户关系的紧密程度进行刻画. 基于此，本部分实验内容逐年计算六个数据集中连通性比例随时间的变化情况，并以数据挖掘子数据集为例计算了其网络直径随时间的变化情况，结果分别如图 8 和图 9 所示. 图 8 显示各个数据集的连通性比例随时间增大，同时图 9 显示数据挖掘子数据集的网络直径随时间减小. 以上结果说明多类型实体演进学术网络中每个类型的实体都与其他类型的实体紧密联系，并且拥有相似的增长模式. 这个结果体现了学术网络中所有实体间的交互演进.



(a) 数据挖掘子数据集 (b) 网络研究子数据集 (c) 文学子数据集

(d) 金融子数据集 (e) 人工智能子数据集 (f) 机器学习子数据集

图 8 各个子数据集中连通性比例的演进情况

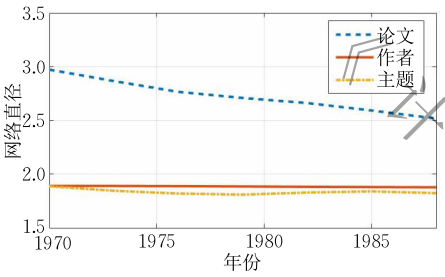


图 9 数据挖掘子数据集中网络直径的演进情况

表 4 对本节所观测到的学术网络共性性质进行了总结. 上述特征性质对于多类型实体演进学术网络中异构关系相互作用的考察是完备的. 接下来,本文将以复现上述共性性质为目标,对多类型实体演进学术网络进行数学建模,并将在后续章节中,分别从理论和实验两个角度对所提出模型的学术网络共性性质刻画能力进行验证.

表 4 学术网络共性性质总结

结构性质		演进性质	
实体内度	实体内度服从幂律分布(性质 1)	实体内度随时间增长(性质 3) 分布幂律指数"先波动,后稳定"(性质 4)	连通性比例随时间增大(性质 7) 网络直径随时间减小(性质 8)
实体间度	实体间度服从幂律分布(性质 2)	实体间度随时间增长(性质 5) 大规模实体的实体间度增速较大(性质 6)	

4 多类型实体演进学术网络的建模

上个章节在若干大规模真实数据集的基础上对学术网络的结构性质和演进性质进行了详尽的观测. 为了从数学角度刻画以上性质,本文提出了多实体学术模型. 该模型从微观角度对网络演进进行刻画,并能够从宏观角度复现所观测到的性质. 在介绍模型的具体内容之前,首先需要了解多类型实体演进学术网络的一般结构,如图 10 所示.

图 10 中包含 3 个节点集合:作者节点集合,论文节点集合和主题节点集合,分别用虚线,点划线和实线的大圆圈表示. 每个节点集合内包含若干节点,用实线小圆圈进行表示. 同一类型实体内的节点关系由对应的节点集合内的实线表示. 不同类型实体间的节点关系由跨节点集合实线表示. 另外,上述网络结构是随着时间推移发生变化的. 例如,当有新的作者节点加入网络时,该节点会与一部分已有节点建立联系甚至影响某些已存在的连接关系,导致网络结构发生变化.

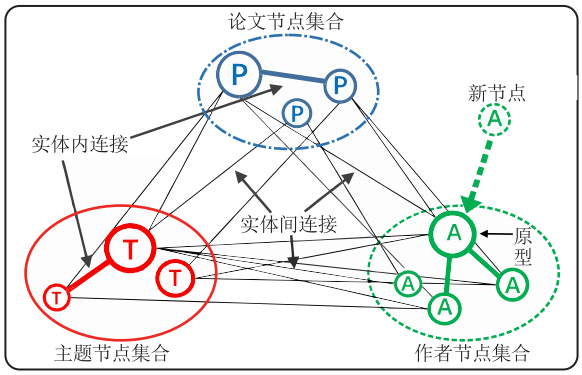


图 10 多实体学术模型结构图

4.1　多实体学术模型

多实体学术模型将多类型实体演进学术网络建模为一个异构图 $G(V_a,V_p,V_t)$，其中 V_a，V_p 和 V_t 分别表示作者节点集合，论文节点集合和主题节点集合. 图中以节点(node)表示作者、论文和主题个体，以边(edge)来刻画同一类型实体内的节点关系以及不同类型实体间的节点关系. $G(V_a,V_p,V_t)$ 包含了数个节点集合和边集合，具体如下：

(1) 3 个节点集合. 同一个类型的所有节点表示为一个节点集合. 图 $G(V_a,V_p,V_t)$ 包含了三个节点集合，分别为作者节点集合 V_a，论文节点集合 V_p 和主题节点集合 V_t. 另外，这三个集合中的节点分别用符号 v_a，v_p 和 v_t 来进行表示.

(2) 3 个实体内关系边集合. 同一类型节点间的连接关系组成的边集合称为实体内关系边集合. 图 $G(V_a,V_p,V_t)$ 包含了三个实体内关系边集合，分别为，作者-作者关系集合 E_{aa}，论文-论文关系集合 E_{pp} 和主题-主题关系集合 E_{tt}. 如果存在一条边 $(v_a,u_a)\in E_{aa}$，则意味着作者 v_a 和作者 u_a 之间存在着合著关系. 论文-论文关系集合和主题-主题关系集合中的边也遵从同样的定义.

(3) 3 个实体间关系边集合. 不同类型节点间的连接关系组成的边集合称为实体间关系边集合. 图 $G(V_a,V_p,V_t)$ 包含了三个实体间关系边集合，分别为作者-论文关系集合 E_{ap}(或 E_{pa})，论文-主题关系集合 E_{pt}(或 E_{tp})和作者-主题关系集合 E_{at}(或 E_{ta}). 若存在一条边 $(v_a,u_p)\in E_{ap}$，则意味着作者 v_a 发表了论文 u_p. 作者-主题关系集合和论文-主题关系集合中的边也遵从同样的定义.

为了后续讨论方便，表 5 中给出了之后的证明和讨论过程中需要用到的数学符号.

表 5　符号和定义

符号	定义
V_a	作者节点集合
V_p	论文节点集合
V_t	主题节点集合
E_{ii}	类型为 i 的节点间的边集合
E_{ij}	类型为 i 的节点与类型为 j 的节点间的边集合
α_i	一个新节点加入集合 V_i 的概率
β_{ij}	当两个节点 $v_i,u_i\in V_i$ 在集合 V_j 中拥有一个共同邻居时 v_i 和 u_i 间产生一条边的概率
c_{ij}	新节点 $v_i\in V_i$ 与集合 V_j 中节点产生的连接数目
$G(V_a,V_p,V_t)$	多类型实体演进学术网络图
T	总时间间隔

4.2　演进过程

算法 1.　图 $G(V_a,V_p,V_t)$ 的演进过程.

给定总时间间隔 T，新节点产生概率 $\alpha_i\in(0,1)$，参数 $\beta_{ij}\in(0,1)$ 和整数 $c_{ij}>0$，其中 $i\neq j\in\{a,p,t\}$.

初始化：给定一个初始图 $G(V_a,V_p,V_t)$，其中任意一个作者节点 v_a 在集合 V_p 中有至少 c_{ap} 个邻居节点，在集合 V_t 中有至少 c_{at} 个邻居节点. 论文节点和主题节点遵从同样的要求.

FOR $1\leq t\leq T$ DO

1. 节点到达. 一个新的作者节点 v_a 以概率 α_a 到达网络. 同样地，一个新的论文节点 v_p 和一个新的主题节点 v_t 分别以概率 α_p 和概率 α_t 到达网络. 为了论述方便，接下来以作者节点到达为例给出剩余的步骤.
2. 原型选择. 新节点选择作者集合中一个已有节点作为其原型. 节点 u_a 被选为作者-主题原型的概率正比于其 t 时刻的实体间度 $d_u^{at}(t)$，被选为作者-论文原型的概率正比于其实体间度 $d_u^{ap}(t)$.
3. 关系复制. 记被选择的作者-主题原型为 u_a，新节点 v_a 随机并且均匀地复制节点 u_a 的 c_{at} 条作者-主题关系 $(u_a,u_t^1),\cdots,(u_a,u_t^{c_{at}})$，即边 $(v_a,u_t^1),\cdots,(v_a,u_t^{c_{at}})$ 被加入网络. 遵从同样的过程，c_{ap} 条作者-论文边 $(v_a,u_p^1),\cdots,(v_a,u_p^{c_{ap}})$ 也被加入网络.
4. 间接演进. 主题节点 $u_t^1,\cdots,u_t^{c_{at}}\in V_t$ 与论文节点 $u_p^1,\cdots,u_p^{c_{ap}}\in V_p$ 之间产生新的边.
5. 内部演进. 对于任意两个作者节点 v_a 和 u_a，若两者之间产生了一个新的共同主题邻居，则 v_a 和 u_a 之间以概率 β_{at} 产生一条新的边. 同理，若两者之间存在新的共同论文邻居则以概率 β_{ap} 产生一条新的边.

多实体学术模型的演进过程如算法 1 所示，该过程可简单描述如下.

初始化：给定初始图和参数 α_i，β_{ij}，c_{ij}，其中 $i\neq j\in\{p,a,t\}$. 后续理论推导部分将会证明，初始图的选取不会影响到最终演进网络的统计特性. 参数值的选取将会对网络的演进速度和图中各个部分的网络密度产生影响，这些影响将会在接下来的部分进行详细讨论. 在初始化完成后，进入迭代的演进过程，具体包括 5 个步骤：

步骤 1. 节点到达. 作者节点、论文节点和主题节点分别以概率 α_a、α_p 和 α_t 到达网络. 显然，类型 i 的节点的到达概率 α_i 越大，则演进后期网络中该类型的节点数目越大. 为了方便起见，在接下来的步骤中以新加入的节点为作者节点这一情况为例进行阐述. 当新加入的节点为论文节点或者主题节点时，接下来的步骤遵循一个对称的过程.

步骤 2. 原型选择. 新节点选择与其类型相同的节点作为原型，其中每个节点被选为原型的概率正比于其实体间度. 注意作者-论文关系原型和作者-主题关系原型的选择过程相互独立.

步骤 3. 关系复制. 被选择的原型的若干连接关

系被新节点复制，复制过程遵循随机并且均匀复制的原则.此步骤与步骤 2 在多种网络生成模型中均有应用，称为优先选择(Preferential Attachment).这种方法可以生成出服从幂律分布的网络结构.

步骤 4. 间接演进.在该步骤中，与新节点相连的其他类型节点之间也会产生新的连接关系.这种演进模式的产生是由于新节点的加入激发了已有节点间的潜在关系.例如，一篇新的论文产生，意味着该论文的作者开始从事论文所涉及的主题方向的研究，进而与该主题节点建立了联系.该步骤与步骤 3 共同决定了网络的演进速率以及新节点的壮大速率，参数 c_{ij} 越大，则该过程速率越大.

步骤 5. 内部演进.若两个相同类型的节点拥有其他类型的共同邻居，则两者之间以一定概率增加一个新的连接.显然，若两个作者之间拥有合著论文，则他们之间很有可能存在着连接关系.作者间的合著网络正是基于这个现象生成的.

该算法通过概率参数 α_i、β_{ij}，以及原型选择步骤和关系复制步骤对网络演进过程中的随机性进行刻画.为了更好地理解上述演进过程，此处以一个新的作者加入学术网络为例来对该过程进行解释.通常情况下，一个新加入的研究人员会选择一位较为有影响力的研究人员作为自己的参考对象，并且往往后者所做主题会对前者的选题过程产生较为显著的影响.同样地，新加入的研究人员也会更倾向于研读较为有影响力的研究人员所发表的学术著作并将其作为自己文章的参考资料.当新加入网络的节点是论文或者主题时，也可以得到相似的结论.

4.3 多实体学术模型的应用

多实体学术模型可用于多类型实体演进学术网络中各类关系间互相影响及其随时间演进情况的数学刻画.良好的数学建模一方面能够帮助科研人员更加全面地理解和研究学术网络，另一方面可以为学术网络中的其他应用提供理论保证.下面通过两个例子对该模型的应用前景进行具体说明.例子 1，多实体学术模型应用于学术关系预测问题：预测类问题的解决思路通常是归纳、总结网络的形成模式并据此对未来关系的形成结果作出预测.多实体学术模型综合考察各类不同关系间的相互作用及演进情况，能够为该问题提供更加全面精准的网络形成模式刻画，进而为学术关系预测提供更加全方位的理论支撑.例子 2，多实体学术模型应用于学术影响力传播问题：传播类问题中传播途径的有效挖掘十分重要，充足多样的传播路径能够极大地优化最终传播结果.多实体学术模型刻画了实体间的多样化传播路径，如“论文-作者-论文”路径及“作者-主题-作者”路径，能够更全面地刻画潜在传播路径并为传播算法设计工作提供重要辅助.

除了前面提到的多类型实体演进学术网络，多实体演进模型在很多其他异构演进社交网络中也有着广泛的应用场景.例如，多实体学术模型可用于演进兴趣社交网络的刻画.在这类网络中，用户通常拥有不同的兴趣，并且可以根据自己的兴趣选择加入一些小组或者社团，与此同时，小组和社团的内容范围也可能涵盖数个不同的兴趣内容.上述场景中可认为存在用户、兴趣、社团三类节点，三类节点间存在交互关系且随时间演进.因此，通过将多实体学术模型中的作者类型映射为用户类型，论文类型映射为兴趣类型，主题类型映射为社团类型，多实体演进模型可以很好地刻画演进兴趣社交网络中用户、兴趣、社团的交互与演进.除了上述应用场景，该模型还可以应用于一些其他场景，如包含产品、客户、经理三种类型实体的销售网络，包含不同类型角色的通讯网络等.

多实体学术模型关注的学术网络共性问题与第 3 节中基于真实数据集的观测内容一致，并能够有效刻画表 4 中所列举的全部性质.接下来，本文将分别从理论和实验角度对模型的性质刻画能力进行验证.其中，性质 1 到性质 6 的相关理论证明将在第 5 节中给出，并在第 6 节中进行实验验证.性质 7 和性质 8 的形成机理较为复杂，无法给出直接理论验证，因此将在第 6 节中对其进行实验验证.

5 理论分析

本节对性质 1 到性质 6 进行理论证明.相关理论分析结果如下.

5.1 实体间度的增长

假设节点 v_i 在 t_0 时刻到达节点集合 V_i，并且其初始实体间度为 $d_v^{ij}(t_0)=[d_v^{ij_1}(t_0),d_v^{ij_2}(t_0)]$，则该节点在 $t>t_0$ 时刻的实体间度可以表示为

$$d_v^{ij}(t)=\left(\frac{t}{t_0}\right)^{\lambda_i} d_v^{ij}(t_0),$$

其中 $\lambda_i\in(0,1)$ 为一个常数，称为“类型 i 节点的增长指数”，其具体取值由演进算法中的预设参数值决定.另外，该结果的证明过程在附录 A 中给出.

上述结果具有以下两个方面的含义：

(1) 实体间度 $d_v^{ij}(t)$ 会随着时间 t 以多项式形

式速率增长，其中增长指数为 $\lambda_i \in (0,1)$. 这一结论与 3.3.2 节中的实验观测结果一致，进而说明多实体学术模型可以很好地刻画学术网络中实体间度的增长速率.

(2) $d_v^{ij}(t)$的两个部分，即该类型节点与其他两种类型节点产生的连接关系中的两类实体间度是同一数量级的. 例如，对于一个主题节点 v_t，存在 $d_v^{tp}(t)=\Theta(d_v^{ta}(t))$①. 这一结果意味着对于某个特定主题，与该主题相关的论文数量和主题数量往往具有相同的增长速率. 该结论的合理性可以从下面两个方面进行说明：(a) 直观解释. 网络稳定后，用户的平均论文发表数量通常是保持不变的；(b) 数据验证. 其他文献[35]通过对实际学术网络数据进行统计佐证了该结论.

上述结论中，第(1)个结论揭示了节点实体间度的增长速率，与第 3 节中观测结果一致. 第(2)个结论说明了两类不同实体间度的演进过程存在相似性. 其中，第(2)个结论进一步从理论上证明了某个在某一类型的实体间子网络中具有较大影响力的节点(具有较大的度)，在另外一个对应的实体间子网络中也具有较大的影响力. 该结论与实际观察结果一致. 例如，一个参与多个研究主题的作者往往具有较大数量的发表论文，反之亦然.

上述结果证明了多实体学术模型构建的网络中实体间度随时间增长，且由表达式可知，较早加入的节点具有较大的实体间度，同时其实体间度的增长速率较大，该结论与性质 5 和性质 6 一致.

5.2　实体度的分布

接下来将从理论方面证明多实体学术模型中的节点度服从幂律分布. 与之前的分析和讨论过程相同，证明过程将从实体间度分布和实体内度分布两个方面展开.

5.2.1　实体间度分布

对于图 $G(V_a,V_p,V_t)$中的任意节点 $v_i \in V_i$，当 $t\to\infty$，该节点的实体间度分布满足

$$P\{d_v^{ij_1}(t)=x\}\propto x^{-\frac{1}{\lambda_i}-1},$$

$$P\{d_v^{ij_2}(t)=x\}\propto x^{-\frac{1}{\lambda_i}-1}.$$

这个结果说明节点的实体间度 $d_v^{ij}(t)$服从于指数为 $-\frac{1}{\lambda_i}-1$ 的幂律分布，证实了多实体学术模型在刻画网络中节点度分布方面的有效性. 这一结论在定理 1 中给出了详细的理论证明，并且在第 6 节中给出了相应的实验验证.

定理 1.　对于 t 时刻的图 $G(V_a,V_p,V_t)$，当 $t\to\infty$，节点 $v_p\in V_p$的实体间度 $d_v^{pa}(t)$和 $d_v^{pt}(t)$均服从幂律分布

$$\boldsymbol{P}\{d_v^{pa}(t)=x\}\propto x^{-\frac{1}{\lambda_p}-1},$$

$$\boldsymbol{P}\{d_v^{pt}(t)=x\}\propto x^{-\frac{1}{\lambda_p}-1},$$

其中 λ_p为类型 p 节点的增长指数，其数值可以根据附录 A 中等式(2)计算得到. 这一结果对于节点 $v_a\in V_a$和节点 $v_t\in V_t$同样成立，它们的参数与上式中的表达形式对称.

证明.　该定理的证明过程可以分为两个部分，即 $d_v^{pa}(t)$的度分布的证明和 $d_v^{pt}(t)$的度分布的证明. 接下来分别对这两个部分进行讨论.

第一部分：$d_v^{pa}(t)$的度分布. 根据附录 A 中等式(1)可知，$d_v^{pa}(t)$的累积分布函数可以按下面的方式计算得到

$$\begin{aligned}\boldsymbol{P}\{d_v^{pa}(t)<x\}&=\boldsymbol{P}\left\{d_v^{pa}(t_0)\left(\frac{t}{t_0}\right)^{\lambda_p}<x\right\}\\&=\boldsymbol{P}\left\{t_0>\left(\frac{d_v^{pa}(t_0)}{x}\right)^{\frac{1}{\lambda_p}}t\right\},\end{aligned}$$

这里 t_0表示的是节点加入网络的时刻. 根据算法 1 中的节点到达规则可知，所有节点是在 0 时刻至 t 时刻间被均匀加入网络的，因此 t_0服从 0 至 t 之间的均匀分布，则

$$\boldsymbol{P}\{t_0>x\}=\frac{t-x}{t},$$

因此，前面的累计分布表达式可以进一步计算为

$$\boldsymbol{P}\{d_v^{pa}(t)<x\}=1-d_v^{pa}(t_0)^{\frac{1}{\lambda_p}}x^{-\frac{1}{\lambda_p}}.$$

接下来，$d_v^{pa}(t)$的概率密度分布函数可以由上式中的累积分布函数得到

$$\boldsymbol{P}\{d_v^{pa}(t)=x\}=\frac{\partial\boldsymbol{P}\{d_v^{pa}(t)<x\}}{\partial x}$$

计算可得

$$\boldsymbol{P}\{d_v^{pa}(t)=x\}=\frac{x^{-\frac{1}{\lambda_p}-1}}{\sum_{x=1}^{|V_a|}x^{-\frac{1}{\lambda_p}-1}},$$

其中$\sum_{x=1}^{|V_a|}x^{-\frac{1}{\lambda_p}-1}$为一个常数形式的归一化系数. 因

① 标准渐进符号定义如下：对于两个非负函数 $f(\cdot)$和 $g(\cdot)$，$f(n)=o(g(n))$表示 $\lim_{n\to\infty}f(n)/g(n)=0$，$f(n)=O(g(n))$表示 $\lim_{n\to\infty}f(n)/g(n)<\infty$，$f(n)=\omega(g(n))$表示 $\lim_{n\to\infty}f(n)/g(n)=\infty$，$f(n)=\Omega(g(n))$表示 $\lim_{n\to\infty}f(n)/g(n)>0$，$f(n)=\Theta(g(n))$表示 $f(n)=O(g(n))$并且 $f(n)=\Omega(g(n))$.

此,可以得到

$$\boldsymbol{P}\{d_v^{pa}(t)=x\}\propto x^{-\frac{1}{\lambda_p}-1}.$$

第二部分:$d_v^{pt}(t)$的度分布.利用第一部分中的方法可以得到

$$\boldsymbol{P}\{d_v^{pt}(t)=x\}\propto x^{-\frac{1}{\lambda_p}-1}.$$

综合第一部分和第二部分中得到的结论,可以证明 $d_v^{pa}(t)$和 $d_v^{pt}(t)$均服从幂律分布.

最后,利用相同的方法可以证明其他的 $d_v^{ij}(t)$,$i\neq j\in\{p,a,t\}$,也服从幂律分布. 证毕.

5.2.2 实体内度分布

对于图 $G(V_a,V_p,V_t)$中的任意节点 $v_i\in V_i$,当 $t\to\infty$,该节点的实体内度分布满足

$$\boldsymbol{P}\{d_v^{ii}(t)=x\}\propto x^{-\omega_i},$$

其中 ω_i是该幂律分布的常数幂律指数.

上述结果证明了模型在刻画网络中实体内度分布方面的有效性.这一结论在定理 2 中给出了详细的理论证明,并且在第 6 节中给出了相应的实验验证.另外,多实体学术模型可以很好地解释 3.3.1 节中图 5 呈现的"先波动、后稳定"现象.一方面,改变 α_i、β_{ij}、c_{ij}等参数会导致节点增长速度、节点类型占比等网络演进参数的改变,造成演进前期网络实体内度分布的不稳定;另一方面,注意该结论的成立条件为 $t\to\infty$,即网络演进成熟,此时节点数量充足,网络中节点度呈现稳定的幂律分布;而当 t 较小时,即网络演进初期,此时节点样本过少,网络中节点度分布并不稳定,与观察现象一致.

定理 2. 对于 t 时刻的图 $G(V_a,V_p,V_t)$,当 $t\to\infty$,节点 $v_p\in V_p$的实体内度服从幂律分布

$$\boldsymbol{P}\{d_v^{pp}(t)=x\}\propto x^{-\omega_p},$$

其中 ω_p为一个常数.这一结果对于节点 $v_a\in V_a$和节点 $v_t\in V_t$同样成立,它们的参数与上式中的表达形式对称.

证明. 该定理的证明利用了 Lattanzi 和 Sivakumar 的论文[36]中的结果.在该论文提出的模型中,模型的二分网络结构与多实体学术模型中图 $G(V_a,V_p,V_t)$的二部子图结构一致,因此该论文的部分结果在本模型中同样成立.在该论文的定理 4 和定理 8 中,作者证明了网络中节点的实体内度服从幂律分布.

因此,利用相似的方法可以得到

$$\boldsymbol{P}\{d_v^{pp}(t)=x\}\propto x^{-\omega_p}.$$

另外,对于节点 $v_a\in V_a$和节点 $v_t\in V_t$也可以得到一致的结果. 证毕.

上述结果证明了多实体学术模型构建的网络中实体间度和实体内度均服从幂律分布,该结论与性质 1 和性质 2 一致.另外,由前文分析结果可知,该模型能够对性质 4 进行有效刻画.

5.3 实体内度的增长

接下来研究网络中实体内度随时间的增长.如果 $\max\{\lambda_j,\lambda_k\}\geqslant\frac{1}{2}$成立,则图 $G(V_a,V_p,V_t)$的平均实体内度随时间增长,即节点集合 V_i中的节点间边的数目与点的数目之比随时间增长.

这一结论在定理 3 中给出了详细的理论证明,并且在第 6 节中给出了相应的实验验证.

定理 3. 对于 t 时刻的图 $G(V_a,V_p,V_t)$,当 $t\to\infty$,节点集合 V_p的中节点的平均实体内度,即节点间边的数目与点的数目之比,满足

$$\frac{|E_{pp}(t)|}{|V_p(t)|}=\begin{cases}\Theta(1), & 0<\max\{\lambda_a,\lambda_t\}<\frac{1}{2}\\ \Theta(\log t), & \max\{\lambda_a,\lambda_t\}=\frac{1}{2}\\ \Theta(t^{2-\frac{1}{\lambda_j}}), & \frac{1}{2}<\max\{\lambda_a,\lambda_t\}<1\end{cases}.$$

这一结果对于节点集合 V_a和节点集合 V_t同样成立,它们的参数与上式中的表达形式对称.

证明. 根据算法 1 可知,集合 V_p内的节点的实体内连接实际上是由它们在集合 V_a和集合 V_t中的共同邻居生成的,并且两者的生成过程互相独立.

当一个节点 $v_a\in V_a$的实体间度满足 $d_v^{ap}(t)=x$,即该节点在集合 V_p中具有 x 个邻居节点时,集合 V_p中所有节点 v_a的邻居会以概率 β_{ap}互相连接,并因此生成平均数量为 $\beta_{ap}\binom{x}{2}$的作者-作者连接关系.在集合 V_a中,实体间度为 x 的节点的平均数量为 $|V_a|\boldsymbol{P}\{d_v^{ap}(t)=x\}$.因此,集合 E_{pp}中由集合 V_a中度为 x 的节点生成的作者-作者连接关系的平均数量为

$$\text{Contribution}(x)=\beta_{ap}\binom{x}{2}|V_a|\boldsymbol{P}\{d_v^{ap}(t)=x\}.$$

在网络演进的过程中,每个时隙均有常数个新节点被加入网络,因此可以得到在 t 时刻下 $|V_p|=\Theta(t)$.结合定理 1 中的结果可以得到

$$\begin{aligned}|E_{pp}(t)|&=\sum_{x=1}^{|V_p|}\text{Contribution}(x)\\&=\sum_{x=1}^{|V_p|}\beta_{ap}\binom{x}{2}|V_a|\boldsymbol{P}\{d_v^{ap}(t)=x\}\end{aligned}$$

$$=\Theta\left(\sum_{x=1}^{|V_a|}x^2x^{-\frac{1}{\lambda_a}-1}t\right)$$

$$=\Theta\left(\sum_{x=1}^{t}x^{-\frac{1}{\lambda_a}+1}t\right),$$

其中求和上限取$|V_p|$是因为集合V_a中的节点在集合V_p中的邻居数目上限为$|V_p|$. 利用p级数求和公式，即

$$\lim_{n\to\infty}\sum_{x=1}^{n}\frac{1}{x^p}=\begin{cases}\Theta(1), & p>1\\ \Theta(\log n), & p=1\\ \Theta(n^{1-p}), & 0\le p<1\end{cases},$$

可以得到

$$|E_{pp}(t)|=\begin{cases}\Theta(t), & 0<\lambda_a<\dfrac{1}{2}\\ \Theta(t\log t), & \lambda_a=\dfrac{1}{2}\\ \Theta(t^{3-\frac{1}{\lambda_j}}), & \dfrac{1}{2}<\lambda_a<1\end{cases}.$$

结合前面得到的结论$|V_p|=\Theta(t)$，可以计算得到最终结果.

另外，对于节点集合V_a和节点集合V_t也可以得到一致的结果. 证毕.

上述结果证明了多实体学术模型构建的网络中平均实体内度随时间增长，该结论与性质 3 一致.

本节从“微观-宏观”联合角度对多实体学术模型的理论性质进行了分析. 具体来说，5.1 节中实体间度增长的相关结果揭示了某个特点节点规模随时间的变化，属于微观角度. 5.2 节和 5.3 节中讨论了学术网络中的实体度分布以及平均实体内度增长情况，属于宏观角度. 需要说明的是，微观与宏观两个角度从来都不是独立存在的，而是紧密相关的. 5.2 节与 5.3 节给出的结论虽然属于宏观范畴，但这种宏观实际上是每个节点和边的微观演进叠加后所得到的宏观结果. 因此，对这些宏观性质的分析，实质上体现了微观-宏观的内在联系，打破了二者的界限.

6　仿真实验

除了前面给出的理论分析，接下来进一步利用实验仿真的方式对多实体学术模型进行分析，证明其对多类型实体演进学术网络数学刻画的有效性. 需要说明的是，多实体学术模型旨在对不同学术网络数据集的共性性质进行复现并探究其内在机理，而非对某个特定数据集进行精准拟合. 因此，本节的仿真实验着重考察其复现第 3 节中观察到的结构性质和演进性质的能力. 本节重点考察多实体学术模型刻画交互演进性质(性质 7 和性质 8)的能力，为其提供有效验证. 另外，为了更加全面的对模型进行考察，本节同时对部分已得到理论证明的性质(性质 1，性质 2，性质 3 和性质 6)展开进一步的实验验证.

参数设置：参数设置如表 6 所示. 根据算法 1 中描述的网络演进过程可知，网络中类型i的实体中节点数目与预设概率α_i的值成正比，并且在t个时隙后该类型节点的平均数目为$\alpha_i t$. 在仿真过程中，为了模拟学术网络中作者节点和论文节点数目较大，主题节点数目较小的现象，三个概率参数设置为$\alpha_a=0.5862$，$\alpha_p=0.4136$和$\alpha_t=0.0002$. 在这样的参数设置下，经过 200 万个时隙后，论文节点、作者节点、主题节点的平均数目分别可以达到 827 636，1 171 977 和 386. 除此之外，另外两个参数的值设置为$c_{ij}=2$和$\beta_{ij}=0.2$.

表 6　参数设置

参数	值	参数	值
α_a	0.5862	c_{ij}	2.0
α_p	0.4136	β_{ij}	0.2
α_t	0.0002	/	/

接下来，对多实体学术模型生成的仿真网络性质进行分析，结果如下：

(1)结构性质. 结构性质的实验验证结果如图 11 和表 7 所示. 从实验结果中可以看出，多实体学术模型构建的网络中实体内度和实体间度均服从幂律分布，与性质 1 和性质 2 一致.

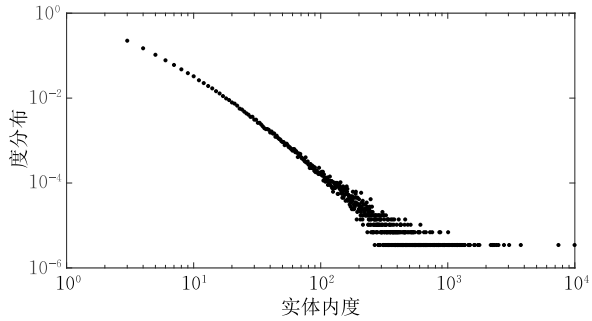


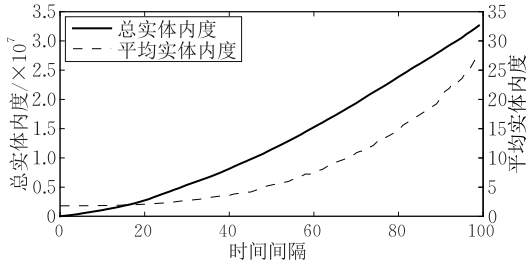
图 11　论文的实体内度分布

表 7　仿真网络中实体间度分布参数

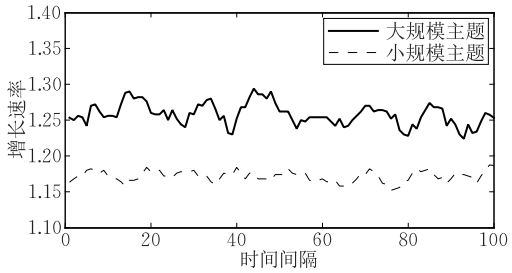
参数值	d^{pa}	d^{pt}	d^{ap}	d^{at}	d^{tp}	d^{ta}
φ	2.19	3.61	2.30	3.82	0.23	0.26
η	0.28	2.01	0.09	0.65	2.18	2.08

(2)演进性质. 验证将从下面三个方面展开.

①实体内度的演进. 如图 12(a)所示，网络中所有作者节点的总实体内度和平均实体内度均随时间

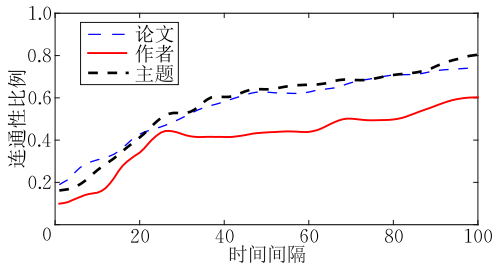


(a) 实体内度的增长

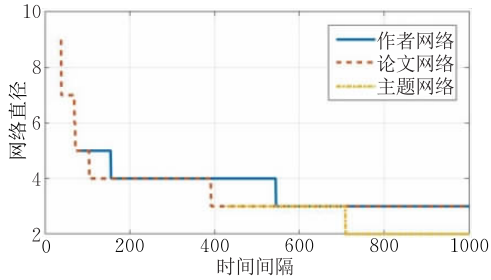


(b) 不同规模主题的演进

图 12 实体内度和实体间度的演进



(a) 连通性比例



(b) 网络直径

图 13 实体间的交互演进

推移呈现增长趋势. 这一结果与 3.3.1 节中的观测结果一致,进一步证明了多实体学术模型能够对表 4 中的性质 3 进行有效刻画.

② 实体间度的演进. 如图 12(b)所示,大规模主题与小规模主题相比具有较大的增长速率. 其中大规模主题的平均增长速率为 1.26,而小规模主题的平均增长速率为 1.17. 这一结果与 3.3.2 节中的观测结果一致,进一步证明了多实体学术模型能够对表 4 中的性质 6 进行有效刻画.

③ 实体间的交互演进. 各个不同类型实体的演进趋势是一致的. 如图 13(a)所示,作者节点、论文节点和主题节点的连通性比例 C_g/C_G 随着时间以相似的趋势增长. 如图 13(b)所示,作者网络、论文网络和主题网络的网络直径也随着时间以相似的趋势下降. 这一结果与 3.3.3 节中的观测结果一致,证明了多实体学术模型能够对表 4 中的交互演进性质(性质 7 和性质 8)进行有效刻画.

根据仿真结果可以总结得到,多实体学术模型可以有效地刻画演进学术网络中论文节点、作者节点以及主题节点之间的交互演进,进一步验证了该模型对真实多类型实体演进学术网络的有效刻画.

多实体学术模型的主要功能在于从数学角度对演进学术网络的共性性质进行刻画与复现. 由该模型建模得到的网络性质与第 3 节中各个真实数据集的观测性质具有“数值存异,趋势一致”的特点. 例如,图 13(a)和图 8 分别显示了建模结果与真实数据集观测结果连通性比例随时间的演进情况. 建模结果中主题网络的连通性比例自初始的 0.2 增长至最终的 0.8,增长过程连续平缓;观测结果中六个真实数据集的连通性比例初始值为 0.2 至 0.5 不等,最终值为 0.8 至 1 不等,增长过程起伏较大. 然而,虽然数值结果上略有差异,建模结果与真实数据集观测结果中主题网络的连通性比例均随时间增大,总体趋势一致. 除了连通性比例性质,其他学术网络性质也满足该特点. 如前所述,本文的主要贡献在于对演进学术网络进行数学建模并刻画其共性性质. 这种建模方式虽然在刻画某个特定数据集时稍显欠缺,但是能够帮助研究人员弱化数据集个体差异,关注学术网络区别于其他类型网络的本质性质,对学术网络的深入研究具有重要意义.

7 总 结

该文对多类型实体演进学术网络进行了研究. 首先,在六个拥有百万级条目的数据集的基础上对真实学术网络进行了实验统计,观测了网络的结构性质和演进性质. 接下来,根据得到的观测结果,提出了多实体学术模型来刻画论文、作者、主题三种类型实体的内部连接关系和相互连接关系,以及连接关系的演进过程. 最后,通过理论分析和实验验证证明了多实体学术模型可以准确地刻画多类型实体演进学术网络的性质.

该工作是对多类型实体演进学术网络中性质观察、建模和分析的一次尝试和探索. 除了本文所涉及的内容,该问题还存很多潜在的研究方向. 例如,在真实网络中,节点的数量通常会非常庞大并且伴有

非均匀的演进特性，如何有效建模并刻画这一特征也是未来的重要研究方向之一.

参 考 文 献

[1] Xia F, Wang W, Bekele T M, Liu H. Big scholarly data: A survey. IEEE Transactions on Big Data, 2017, PP(99): 1

[2] Yang Z, Hong L, Davison B D. Academic network analysis: A joint topic modeling approach//Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara Falls, Canada, 2013: 324-333

[3] Kajikawa Y, Ohno J, Takeda Y, et al. Creating an academic landscape of sustainability science: An analysis of the citation network. Sustainability Science, 2007, 2(2): 221

[4] Li J, Xia F, Wang W, et al. Acrec: A co-authorship based random walk model for academic collaboration recommendation //Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea, 2014: 1209-1214

[5] Lin Y-S. Topic evolution of innovation academic researches. Journal of Small Business Strategy, 2016, 26(1): 25

[6] Liu J, Li Y, Ruan Z, et al. A new method to construct co-author networks. Physica A: Statistical Mechanics and its Applications, 2015, 419: 29-39

[7] Atzmueller M, Ernst A, Krebs F, et al. On the evolution of social groups during coffee breaks//Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea, 2014: 631-636

[8] Wu Y, Pitipornvivat N, Zhao J, et al. egoSlider: Visual analysis of egocentric network evolution. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 260-269

[9] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA, 2005: 177-187

[10] Xu Ke, Zhang Sai, Chen Hao, Li Hai-Tao. Measurement and analysis of online social networks. Chinese Journal of Computers, 2014, 37(1): 165-188(in Chinese)
(徐恪，张赛，陈昊，李海涛. 在线社会网络的测量与分析. 计算机学报，2014, 37(1): 165-188)

[11] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. ACM SIGCOMM Computer Communication Review, 1999, 29(4): 251-262

[12] Kleinberg J, Kumar R, Raghavan P, et al. The Web as a graph: Measurements, models, and methods. Computing and Combinatorics, 1999: 1-17

[13] Newman M E. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(2): 404-409

[14] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 2

[15] Molloy M, Reed B. The size of the giant component of a random graph with a given degree sequence. Combinatorics, Probability and Computing, 1998, 7(3): 295-305

[16] Zheleva E, Sharara H, Getoor L. Co-evolution of social and affiliation networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 1007-1016

[17] Watts D J, Strogatz S H. Collective dynamics of small world networks. Nature, 1998, 393(6684): 440-442

[18] Kleinberg J. The small-world phenomenon: An algorithmic perspective//Proceedings of the 32nd Annual ACM Symposium on Theory of Computing. Portland, USA, 2000: 163-170

[19] Kumar R, Raghavan P, Rajagopalan S, et al. Stochastic models for the Web graph//Proceedings of the 41st Annual Symposium on Foundations of Computer Science. Redondo Beach, USA, 2000: 57-65

[20] Leskovec J, Chakrabarti D, Kleinberg J, et al. Kronecker graphs: An approach to modeling networks. Journal of Machine Learning Research, 2010, 11(Feb): 985-1042

[21] Chakrabarti D, Faloutsos C. Graph mining: Laws, generators, and algorithms. ACM Computing Surveys, 2006, 38(1): 2

[22] Barabasi A-L, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509-512

[23] Barabasi A-L, Albert R, Jeong H. Mean-field theory for scale-free random networks. Physical A: Statistical Mechanics and Its Applications, 1999, 272(1): 173-187

[24] Wang X, Zhai C, Roth D. Understanding evolution of research themes: A probabilistic generative model for citations //Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 1115-1123

[25] Liu J, Yao Y, Fu X, et al. Evolving k-graph: Modeling hybrid interactions in networks//Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing. Chennai, India, 2017: 29

[26] Mo G Y, Hayat Z, Wellman B. How far can scholarly networks go? Examining the relationships between distance, disciplines, motivations, and clusters. Communication and Information Technologies Annual, 2015: 107-133

[27] Tang J, Zhang J, Yao L, et al. ArnetMiner: Extraction and mining of academic social networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 990-998

[28] Yan E, Ding Y. A framework of studying scholarly networks //Proceedings of the 17th International Conference on Science and Technology Indicators. Montreal, Canada, 2012: 917-926

[29] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping

relations//Proceedings of the Meeting of the Association for Computational Linguistics: Human Language Technologies Association for Computational Linguistics. Portland, Oregon, 2011: 541-550

[30] Lukovnikov D, Fischer A, Lehmann J. Neural network-based question answering over knowledge graphs on word and character level//Proceedings of the International Conference on World Wide Web. Perth, Australia, 2017: 1211-1220

[31] Berant J, Chou A, Frostig R, Liang P. Semantic parsing on freebase from question-answer pairs//Proceedings of the Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1533-1544

[32] Sun Y, Norick B, Han J, et al. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2012: 1348-1356

[33] Kong X, Yu P S, Ding Y, Wild D J. Meta path-based collective classification in heterogeneous information networks//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, USA, 2012: 1567-1571

[34] Shi C, Zhang Z, Luo P, et al. Semantic path based personalized recommendation on weighted heterogeneous information networks//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York, USA, 2015: 453-462

[35] Borner K, Maru J T, Goldstone R L. The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences, 2004, 101(Supplement 1): 5266-5273

[36] Lattanzi S, Sivakumar D. Affiliation networks//Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda, USA, 2009: 427-434

附录 A.

定理 4. 对于 t 时刻的图 $G(V_a, V_p, V_t)$，若某一个节点 $v_p \in V_p$ 在 t_0 时刻加入网络，且加入初始时刻的实体间度为 $d_v^{pj}(t_0)=[d_v^{pa}(t_0), d_v^{pt}(t_0)]$，则该节点在 t 时刻的实体间度满足

$$d_v^{pj}(t)=\left(\frac{t}{t_0}\right)^{\lambda_p} d_v^{pj}(t_0).$$

这一结果对于节点 $v_a \in V_a$ 和节点 $v_t \in V_t$ 同样成立，它们的增长指数与上式中的表达形式对称.

证明. 在每一个时间间隙 $t-1$，节点 $v_p \in V_p$ 的实体间度 $d_v^{pa}(t)$ 会在以下两种情况下发生增长.

(1) 情况 1. 一个新的节点加入集合 V_a 并连接到节点 v_p，这个情况下节点 v_p 的实体间度会增加一，即 $d_v^{pa}(t)=d_v^{pa}(t-1)+1$.

(2) 情况 2. 一个新的节点加入集合 V_t 并连接到节点 v_p，则节点 v_p 会和新加入的节点在集合 V_a 中的 c_{ta} 个邻居产生连接关系，进而导致节点 v_p 的实体间度增加，即 $d_v^{pa}(t)=d_v^{pa}(t-1)+c_{ta}$.

由网络的演进过程可知，每个时隙中有新的节点加入集合 V_a 的概率为 α_a，有新的节点加入集合 V_t 的概率为 α_t. 而在关系复制步骤，每一个已存在节点被选作为新节点的原型的概率与其实体间度成正比，原型选定后再从原型的边中随机均匀地复制若干个. 实际上，这个阶段中每一条已存在的边会被等概率的选取到并复制. 因此，情况 1 发生的概率为 $\alpha_a c_{ap}\dfrac{d_v^{pa}(t-1)}{|E^{pa}(t-1)|}$，其中 $|E^{pa}(t-1)|$ 表示关系集合在 $t-1$ 时刻时的平均总边数. $|E^{pa}(t-1)|$ 可由如下关系式计算得到

$$|E^{pa}(t-1)|=(\alpha_p c_{pa}+\alpha_a c_{ap}+\alpha_t c_{tp} c_{ta})(t-1),$$

$|E^{pt}(t-1)|$ 和 $|E^{at}(t-1)|$ 也可以利用同样的方法计算得到. 同理，情况 2 发生的概率为 $\alpha_t c_{tp}\dfrac{d_v^{pt}(t-1)}{|E^{pt}(t-1)|}$. 综合以上两种情况，可得

$$d_v^{pa}(t)-d_v^{pa}(t-1)=\alpha_a c_{ap}\frac{d_v^{pa}(t-1)}{|E^{pa}(t-1)|}+\alpha_t c_{tp} c_{ta}\frac{d_v^{pt}(t-1)}{|E^{pt}(t-1)|},$$

同样地

$$d_v^{pt}(t)-d_v^{pt}(t-1)=\alpha_t c_{ta}\frac{d_v^{pt}(t-1)}{|E^{pt}(t-1)|}+\alpha_a c_{ap} c_{at}\frac{d_v^{pa}(t-1)}{|E^{pa}(t-1)|}.$$

结合给定的初始条件

$$d_v^{pj}(t_0)=[d_v^{pa}(t_0), d_v^{pt}(t_0)],$$

可以得到

$$\begin{aligned} d_v^{pj}(t)&=\left(\frac{t}{t_0}\right)^{\lambda_p} d_v^{pj}(t_0)\\ &=\left[\left(\frac{t}{t_0}\right)^{\lambda_p} d_v^{pa}(t_0), \left(\frac{t}{t_0}\right)^{\lambda_p} d_v^{pt}(t_0)\right] \end{aligned} \tag{1}$$

其中

$$\lambda_p=\frac{\sqrt{\Delta}+\alpha_a c_{ap}|E^{pt}|+\alpha_t c_{tp}|E^{pa}|}{2|E^{pt}||E^{pa}|} \tag{2}$$

$$|E^{pt}|=\frac{|E^{pt}(t-1)|}{t-1},$$

$$|E^{pa}|=\frac{|E^{pa}(t-1)|}{t-1},$$

$$\Delta=(\alpha_a c_{ap}|E^{pt}|-\alpha_t c_{tp}|E^{pa}|)^2+4\alpha_a\alpha_t c_{ap} c_{tp} c_{at}|E^{pt}||E^{pa}|.$$

通过同样的方法，可以得到集合 V_a 和集合 V_t 中节点的实体间度的增长公式. 证毕.

LIU Jia-Qi, Ph.D., associate professor. Her research interests include modeling, analysis and algorithm design in social networks.

FU Luo-Yi, Ph.D., special associate researcher. Her research interests include Internet of Things, information cascading optimization and multicast analysis.

KONG Ling-Kun, Ph.D. candidate. His research interests include social computing and data mining.

GAN Xiao-Ying, Ph.D., associate professor. Her research interests include crowd sensing and network economics.

WANG Xin-Bing, Ph.D., distinguished professor. His research interests include intelligent internet of things and wireless network.

Background

Scholarly networks contain massive scholarly information that can be mainly categorized into three entities, i.e., paper, author and topic, which exhibit a co-evolution over time. Although scholarly networks have attracted much attention over the past years, most works focus on single entity of the network, e.g. sub-networks generated by citation, co-authorship or topic relationship; while few of them incorporate different entities into an entirety to provide a systematic understanding of scholarly networks at scale. We bridge this gap by proposing a multi-entity scholarly model (MSM) with strong theoretical guarantees, which amalgamates entities of paper, author and topic into one single framework to simulate interactions among different entities, and thus presenting the co-evolution within scholarly networks.

First of all, using real scholarly datasets—Microsoft Academic Graph with 6.9 million publications, we observe properties that belong exclusively to scholarly networks, such as varying and converging exponents of power-law distributions with time, degree densification in each of the three aforementioned entities. We also observe interesting evolving patterns like simultaneous co-evolution of all the three entities, faster growth rate of entities with larger size, etc. Based on our observations, we propose the MSM that jointly captures both intra and inter correlations among different entities during the evolving process. Through both theoretical analysis and empirical simulation, we further characterize the MSM as capable of reproducing evolving patterns of real multi-entity scholarly networks.

There remain some future directions which wait for exploration. On one hand, the structure of our MSM can be mapped into counterparts to model other kinds of multi-entity social networks. On the other hand, since the number of entities in social networks might be extremely large or even also evolving, there is room for our current model to improve to accommodate more complicated multi-entity social networks.

This work was supported by the National Key R&D Program of China (Nos. 2018YFB1004705, 2018YFB2100302), the National Natural Science Foundation of China (Nos. 61822206, 61532012, 61602303, 61829201, 61960206002), and the CCF Tencent RAGR (No. 20180116).