DOI:10.12154/j.qbzlgz.2022.06.003

基于文本语义与动态网络结构的科研合作网络链路预测研究*

周 静 庄艳阳 (中国人民大学应用统计科学研究中心;中国人民大学统计学院 北京 100872)
周季蕾 (中国人民大学信息学院 北京 100872)

摘 要: **[目的/意义]**近年来科研合作成为学术成果产出的重要途径之一,科研合作网络的链路预测成为提升科研效率、促进学科资源融合的重要方式之一。但是现有研究鲜有将科研合作网络看作动态时变演化网络进行建模,并考虑网络的文本语义属性和动态网络结构。**[方法/过程]**文章提出了融合文本语义信息和动态网络结构信息的科研合作网络动态链路预测模型。首先,文章以统计学者为例,收集了国际统计四大期刊在2011—2020年间发表的所有文章,基于论文合著关系构建了科研合作网络。其次,构建了科研合作网络的文本特征和动态结构特征并对其进行了分析。最后,结合节点的语义特征和动态拓扑结构特征,本文使用动态逻辑回归对学者合作关系进行链路预测。**[结果/结论]**结果表明,科研合作关系的动态演化受到多方面因素的共同影响,例如上一年度是否合作、学者间的研究方向相似度、学者已发表论文的引用情况等。文章的研究结论对增强学者间的联系和提升科研合作效率具有重要借鉴意义。

关键词: 科研合作网络 链路预测 文本语义 动态网络结构

Research on Scientific Research Cooperation Network Link Prediction Based on Text Semantics and Dynamic Network Structure

Zhou Jing Zhuang Yanyang
(Center for Applied Statistics;School of Statistics,Renmin University of China,Beijing,100872)
Zhou Jilei (School of Information,Renmin University of China,,Beijing,100872)

Abstract: **[Purpose/significance]** In recent years, the cooperation among scholars has become one of the important ways to output academic achievements.Co-authorship link prediction is one of the important topics to improve scientific research efficiency and promote the integration of discipline resources. However, little studies have treated the co-authorship network as a dynamic time-varying network and integrated textual semantics and dynamic network information. **[Method/process]** Therefore, we propose a co-authorship link prediction model which incorporatenodes' unstructured semantic attributes and dynamic network features.First, we take statisticians as an example and collect all papers in four top statistical journals from 2011 to 2020, which helps us construct a statisticians' co-authorship network. Second, we extract unstructured semantic attributes and dynamic network feature to enrich nodes representation. Finally, based on these features, we make a link prediction based on a dynamic logical regression model. **[Result/conclusion]** The results have shown that several factors can significantly influence co-authorship, such as whether two researchers cooperated last year, the research similarity between two researches, citation-related factors, etc. The result of this paper is of great significance to enhance the relationship between scholars and the efficiency of scientific research cooperation.

Keywords: scientific collaboration network link prediction text semantics dynamic network structure

*本文系中国人民大学科学研究基金面上项目"统计视角下的深度学习优化算法研究"(项目编号:21XNA027)的研究成果。

1 引言

随着科技发展的全球化以及研究问题的多元化、精细化和复杂化，科研合作成为学术成果产出的重要途径之一。把学者看作网络的节点，学者之间的合作关系看作网络的边，那么学者的科研合作行为就构成了典型的科研合作网络[1]。分析学者的科研合作网络有助于理清学者合作现状，提升学者科研合作效率，促进学科资源融合，是一个重要的研究问题。科研合作网络具有较高的稀疏性，而科研合作网络的关系预测能够在稀疏网络中实现潜在合作者的精准推荐，从而增强不同学者之间的联系，提高网络密度，促进学科发展和知识传播，最终有效推动科研合作效率。这也是本文致力于科研合作关系预测的主要动机之一。

为应对科研合作网络关系预测面临的挑战，本文提出了结合文本语义信息和动态网络结构信息的科研合作网络链路预测模型。由于科研领域学科众多，本文选择统计学科作为重点研究对象，这是因为随着数字时代的到来，统计学成为推动大数据和人工智能发展的重要学科。其次，本文借助文本分析方法和社交网络分析方法挖掘了多种网络属性，包括网络拓扑结构特征、网络节点语义特征（如研究主题特征和研究方向相似度特征）和动态网络结构特征（如上一年的合作情况）。最后，基于包含上述属性的动态时变合作演化网络，使用动态逻辑回归模型进行学者合作关系的动态链路预测。结果表明，学者合作关系的动态演化受到多方面因素的共同作用，例如上一年度是否合作、作者研究方向相似度、已发表论文的引用和被引情况等都能显著影响学者之间合作的可能性。

2 相关研究回顾

科研合作网络分析是复杂网络研究领域的一个重要问题，近年来，有关科研合作网络的研究在各学科领域都取得了迅速发展。早期研究主要从宏观的视角对科研合作网络进行分析。例如，基于科研合作网络的密度、中心度、聚类系数等拓扑结构特征，发现科研合作网络的结构特点，识别科研合作网络中的高影响力作者[2-4]。之后，部分学者从更加微观的视角对科研合作网络进行了更深层的子群分析[5-6]和社区发现分析[1,7]。这种基于微观视角的科研合作网络分析也催生出一批对科研合作关系预测的热点研究。科研合作关系的预测主要依赖复杂网络的链路预测方法。链路预测能够根据已知网络结构预测网络中任意两个节点产生连接的可能性[8]，对挖掘和分析社交网络的演变至关重要。链路预测方法主要分为基于节点相似性的方法[9]、基于最大似然估计的方法[10]和基于概率模型的方法[11]。当前科研合作关系预测的相关研究主要依赖于节点相似性的链路预测方法，即通过科研合作网络节点的相似性指标，度量不同学者之间的相似程度，从而预测双方在未来产生合作的可能性。

然而，现有的科研合作网络链路预测方法存在诸多局限，限制了合作关系预测的效果。第一，以往研究没有充分利用科研合作网络中涉及的语义信息，主要是基于合作网络的拓扑结构进行相似性指标构建，缺乏对合作论文内容的关注。而科研合作关系背后的学术论文涵盖了研究方向、研究领域等语义信息，这些语义信息能够丰富科研合作网络，从而辅助科研合作网络链路预测中节点相似性的测量。第二，以往研究缺乏对历史网络动态信息的考察，主要利用了科研合作网络的静态特征。但是，科研合作网络是一个复杂的时序网络[12]，即网络结构会随时间的推移不断变化。考虑学者合作模式随时间的变化规律将有助于提升合作关系预测的准确性。

从以上相关文献的回顾中可以看出，有关科研合作网络的研究目前已经取得了一些成绩，但也存在一些不足：(1)从研究对象上看，鲜有针对统计学科科研合作网络动态链路预测的研究。与其他学科相比，统计学科是大数据、人工智能技术发展的重要基础，具有应用广泛、学科交叉强的特点，因而面向统计学者科研合作网络的链路预测研究对促进学科融合、推动前沿科学研究具有重要意义。(2)从研究指标上看，链路预测指标的考察仍然不够全面。以往研究主要利用网络拓扑结构属性进行合作网络链路预测，鲜有研究关注科研合作网络具有的语义特征和动态网络特征。(3)从研究方法上看，现有研究主要将科研合作网络看作静态网络进行链路预测建模，鲜有研究对科研合作网络的时序结构进行动态链路预测建模。基于上述问题，本文以统计科研合作网络为例，构建了一种融合节点语义和动态网络结构的统计学者合作网络动态链路预测模型，以期弥补现有研究的理论空白。

3 科研合作数据收集

本文使用的数据爬取自“Web of Science”网站，爬

取范围为统计学领域著名的四大期刊，即 *Journal of American Statistical Association*，*Journal of Royal Statistical Society Series B*，*Annals of Statistics* 和 *Biometrika*。具体来说，在 Web of Science 数据库检索界面中，将时间跨度自定义为2011—2020年，选择按照“出版物名称”检索，分别输入四大期刊的名字并爬取每个期刊的12个字段信息，包括论文标题、作者、出版商、发表日期、关键词、摘要、作者通讯地址、作者所属单位、论文被引次数、引文数及具体参考文献列表。对爬取获得的原始数据，进行数据格式整理、缺失值和重复值的处理以及异常值处理。

首先，数据格式的整理与统一。由于网络爬虫的精细度和灵活性有限，爬取得到的原始数据集存在格式混乱、表示不一等问题。例如，部分字段开头或结尾有数量不等的空格，有部分学者姓名后存在数字编号，不同期刊和时段发表的论文的发表日期表示方式不一等。因此，数据预处理的第一步将对诸如此类的格式问题进行调整和规范，形成统一的数据格式。

其次，缺失值和重复值处理。针对缺失值，首先需要核实缺失原因，若数据缺失的原因是原网页收录信息不全，则剔除相应数据。若因为爬取过程中受网速等客观因素影响造成大量数据的连续缺失，则对相应论文数据进行重新爬取以补全信息。此外，对于个别关键字段随机缺失的情况，可考虑直接删除该条数据。针对重复值，可根据论文标题、发表年份、学者信息等字段进行识别和剔除。最后，异常值的判断与处理。本文对异常值的识别包括两种，第一种异常是数据本身存在明显异常，一般由源网页本身或爬取中的定位错乱等原因造成。第二种异常指的是数据本身脱离了本文的研究范围。例如，文献的发表时间不在本文的研究范围；匿名的论文由于作者信息缺失无法用于学者合作关系的研究；一些非原创论文（诸如评论文章、反驳文章等）也不在本文的讨论范围之类。针对上述异常数据，本文将酌情进行校正或剔除。经过以上数据预处理，本文最终获得2011—2020年发表在国际统计四大期刊上共计3861篇文章，涉及5397位不同的学者。

4 科研合作网络构建与描述性分析

4.1 科研合作网络构建

基于数据预处理后的四大期刊数据集，本小节简述构建学者合作网络的过程。首先，生成学者名单。由于每篇论文的学者有一位或多位，因此需要对学者姓名进行区分和提取，得到与所著论文相对应的学者名单列表。需要注意的是，Web of Science 数据库中学者姓名并不具备唯一标识，即可能出现两位及以上学者同名的情况。此外，由于书写格式的差异，还可能出现同一学者具有多个不同姓名表示方式的情况。数据库中同名学者的识别问题是领域内的一大研究热点和难点。针对这一问题，本文的处理方式是以人工识别和校正为主，根据学者国籍、所在单位等信息进行辅助判断。经处理，本文数据共包含5397位不同的统计学者。

其次，构建学者合作网络。令 $1\leqslant i\leqslant N$ 为合作网络中的第 i 个学者，本文中 N=5379。为了刻画学者间的合作关系，本文用邻接矩阵 $A=(a_{ij})\in R^{N\times N}$ 来表示，如果学者 i 与学者 j 曾经合作写过论文，则有 $a_{ij}=a_{ji}=1$，否则 $a_{ij}=a_{ji}=0$。本文定义自己和自己不存在合作关系，因此，对于 $1\leqslant i\leqslant N$ 有 $a_{ii}=0$。其中在邻接矩阵 A 中存在158个孤立点，也就是说有158位学者从来没有和任何人有过合作。此外，经过计算，该网络共有22772条边，因此网络密度为0.079%，是一个非常稀疏的网络结构。

4.2 科研合作网络特征及动态演化

在学者合作网络中，节点的度为一个学者合作过的其他学者数量，其分布由图1刻画。从图1可以看出，本文构建的学者合作网络是一个无标度网络，节点连边的数量遵循幂律分布，即大部分节点的度很小，而只有小部分节点的度很大。5397位作者中有158位为独立学者，即他们没有合作伙伴。有1042位学者的合作者数量仅为1。合作者最多的是Fan, JQ（Fan, Jianq-

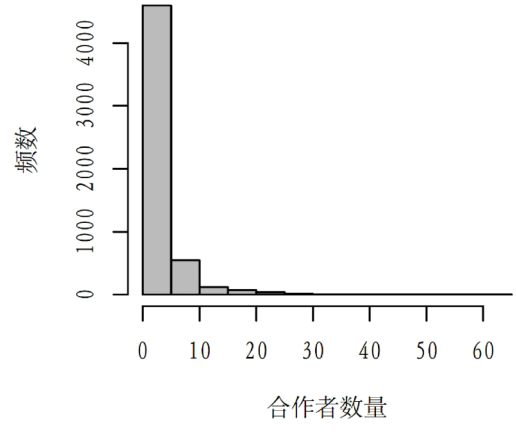
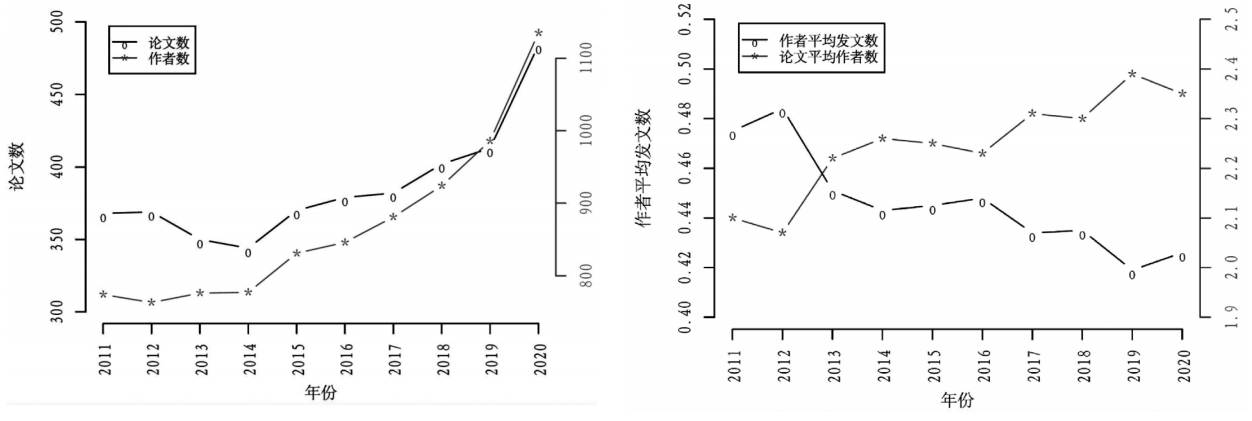


图1 学者的合作者数量分布情况

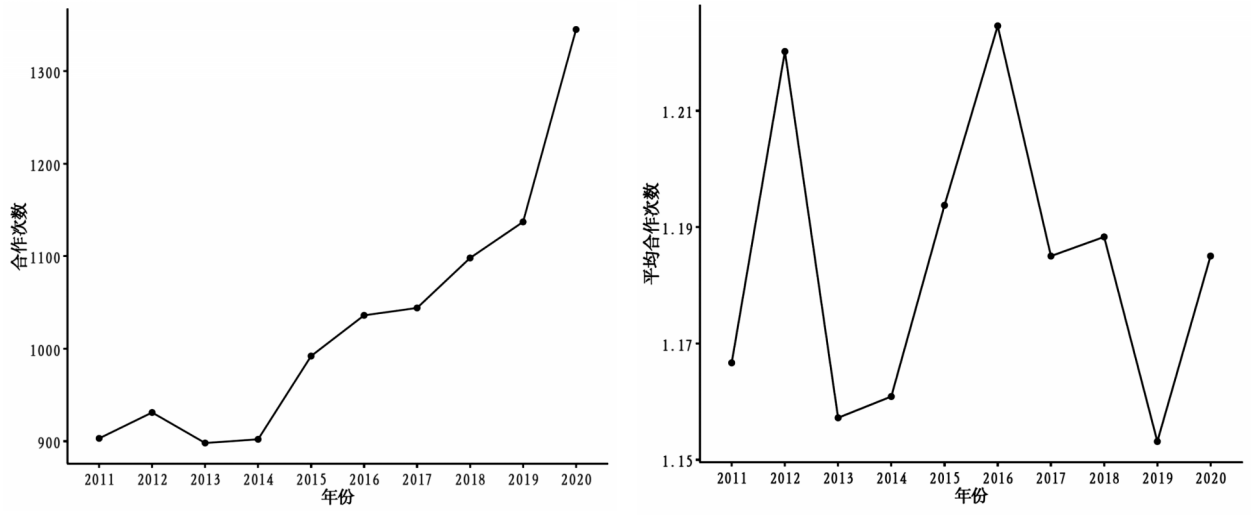
ing),其合作学者数量高达64个。此外,合作者较多的学者还有Zeng, D.L.(Zeng, Donglin)(60个)、Carroll, R. J.(Carroll, Raymond J.)(55 个)、Dunson, D.B. (Dunson, David B.)(55个)等人。上述学者都是国内外较为知名的统计学者,这也说明学术成就和水平较高的学者往往会更经常与他人进行合作。

通过对近10年数据的分析,本文发现统计学者的合作关系随时间的推移呈现出一定的动态变化趋势。首先,考察论文数和学者数的年度变化趋势(见图2(a))。可以看出,近年来,论文发表数和学者数量都呈现出逐年增加的趋势。这表明统计学界的学术创作越来越活跃,同时竞争也越来越激烈,在四大期刊发表论文的难度越来越大。其次,刻画每名学者的平均发表论文数和每篇论文的合作度的年度变化趋势(见图2(b))。可以看到,每篇论文的合作度整体上保持上升趋势,单个学者的平均论文数呈现下降趋势,这也表明统计学者之间的合作在不断加强、合作规模不断扩大,相比于独立研究,合作研究越来越受到学者们的青睐。这一结果与Popescul和Ungar[9]针对2003—2012年的数据分析结果一致,说明在过去近二十年中统计学者的合作模式表现出了相似的变化规律。最后,统计各年份的学者合作次数与每名学者的平均合作次数(见图3)。图3总合作次数随时间的推移呈上升趋势,而单个学者的平均合作次数则呈现上下波动的变化。



(a)　　(b)

图2　论文数、学者数年度变化(左)和学者平均发文数、论文合作度年度变化(右)



(a)　　(b)

图3　总合作次数(左)和学者平均合作次数(右)的年度变化

4.3 基于LDA主题模型的分领域学者合作模式分析

除了合作模式的动态演进变化，学者的研究领域也是合作关系的重要影响因素，而摘要和关键词是论文中心思想和核心内容的集中反映。为此，本小节引入LDA 主题模型[13]对摘要和关键词进行主题提取。主题模型的最优主题数由一致性（Coherence）和主题可解释程度共同确定。一方面，相较于传统的困惑度（Perplexity），一致性被认为是更好的主题模型评价标准，不会随着主题数的增多出现过拟合问题[14]；另一方面，可解释程度高的主题结果有助于凝练研究领域可能涉及的新知识。针对本文的论文摘要和关键词语料，当主题数为6时，LDA 模型具有高一致性（一致性得分为0.415），且主题可解释程度强。因此，本文最终确定了6大研究领域，分别是社交网络分析、生物统计、协方差矩阵估计、变量选择、贝叶斯统计/非参数统计，以及假设检验/时间序列分析。

各主题所对应的前10个代表性主题词如表1所示。表1的结果与前人对统计学者社区发现的探究结果大体一致。例如，Ji和Jin[1]使用2003—2012年统计学四大期刊的数据做社区发现时，其用不同的方法得到的主要分类就包括贝叶斯统计、生物统计、高维数据分析、变量选择、半参和非参数统计；而Gao等[7]在2001—2018年统计学四大期刊的引文网络分析中，划分出了4个重点主题，分别是变量选择、稀疏协方差矩阵估计、函数型数据分析/降维，以及错误发现率（FDR）。因此，我们认为表1的研究领域划分结果符合实际情况，具备较高的可信度。此外，本文的分类结果还出现了“社交网络分析”这一新的研究方向。通过网上查阅检索等方式不难发现这是近几年来快速发展起来的一个新兴领域，这是以往社区发现文献[1,7]未提及的一个研究领域。

表1 基于LDA主题模型的统计研究学科领域分类

主题分类	各主题对应的前10个主题词
社交网络分析	network, graph, structure, random, distribution, dependence, graphical, variable, node, community
生物统计	treatment, effect, causal, patient, control,inference, group, material, individual, trial
协差阵估计	estimate, matrix, covariance, regression, convergence, design, covariates, simulation, parameter, likelihood
变量选择	variable, regression, selection, algorithm, linear, space, spatial, feature, dimension, predictor
贝叶斯统计/非参数统计	bayesian, prior, design, distribution, algorithm, posterior, parameter, mixture, inference, likelihood
假设检验/时间序列分析	test, distribution, hypothesis, time, power, null, sample, series, asymptotic, robustness

LDA模型能够输出每篇论文在每个主题上的概率，将对应概率最高的主题作为论文的研究领域标签，可以获得每篇论文所属的研究领域。进一步地，学者的研究领域定义为该学者所有发表论文涉及的研究领域的集合。在此基础上，对上述6个领域构建了6个子合作网络。表2展示了6个子合作网络的基本统计信息。首先，“生物统计”的平均度明显高于其他五个网络，达到3.907。这说明研究方向为生物统计学的学者平均每人拥有将近4个合作者。“贝叶斯统计/非参数统计”这一领域的平均度最低，仅为2.588。其次，“生物统计”“贝叶斯统计/非参数统计”和“社交网络分析”三个领域的聚类系数都比较高，分别为0.836，0.797和0.791，说明这些领域的学者之间联系密切，相互之间产生的合作较多。最后，“社交网络分析”的网络密度最大（0.004），“协方差矩阵估计”的最小（0.001）。

表2 分领域的学者合作网络基本统计指标

研究领域	节点数	边数	平均度	聚类系数	网络密度
社交网络分析	660	896	2.715	0.791	0.004
生物统计	1564	3055	3.907	0.836	0.002
协差阵估计	3044	4630	3.042	0.755	0.001
变量选择	1142	1711	2.996	0.776	0.003
贝叶斯统计/非参数统计	1007	1303	2.588	0.797	0.003
假设检验/时间序列分析	1166	1603	2.750	0.735	0.002

5 基于动态逻辑回归的科研合作网络动态链路预测

5.1 模型设定

以上分析表明统计学者间的合作关系不仅随着时间推移发生变化，同时还受到节点语义信息（如研究主题）的影响。因此，本小节采取Zhou等[15]提出的动态逻辑回归方法对上述构建的统计学者合作网络进行动态链路预测。

为了将动态网络结构信息考虑到模型中，本小节需要将邻接矩阵按年份重新构建。具体地，考虑一个有N个节点的科研合作网络，进一步假设该合作网络在时刻t可以被观察到，其中$t\in\{1,\cdots,T\}$，则在每个时刻（本文中即每年）t都可以观察到一个邻接矩阵$A_t=(a_{ij}^t)$，其中$a_{ij}^t=1(i\neq j)$表示节点i到节点j在时刻t存在

一条边，本文即两个学者有合作，否则 $a_{ij}^t=0$。定义对 $1\leqslant i\leqslant N$，有 $a_{ii}^t=0$。科研合作网络是一个无向网络，因此有 $a_{ij}^t=a_{ji}^t$。为了刻画 A_t 的分布，Zhou 等[15]提出如下模型：

$$P(a_{ij}^t=1|F_{t-1})=P(z_{ij}=1)P(\tilde{a}_{ij}^t=1|F_{t-1})=a_{ij}\frac{\exp(\beta^T X_{ij}^{t-1})}{1+\exp(\beta^T X_{ij}^{t-1})} \quad (1)$$

其中，$F_{t-1}=\sigma\{A_{t-1},A_{t-2},\cdots,A_0\}$ 为历史网络结构信息，$z_{ij}=z_{ji}\in\{0,1\}$ 为二元随机效应，其中 $a_{ij}^t=z_{ij}\tilde{a}_{ij}^t$，$X_{ij}^{t-1}=(X_{ij,1}^{t-1},X_{ij,2}^{t-1},\cdots,X_{ij,p}^{t-1})^{\mathrm{T}}\in\mathrm{R}^p$ 为 p 维解释性变量，$\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}\in\mathrm{R}^p$ 为相应的待估参数。为了获得参数估计结果，Zhou 等[19]提出了条件似然函数的估计方法，将上述模型的估计转化为求解一个标准的逻辑回归模型的形式：

$$P(a_{ij}^t=1|a_{ij}^{t-1}=1)=\frac{\exp(\beta^T X_{ij}^{t-1})}{1+\exp(\beta^T X_{ij}^{t-1})} \quad (2)$$

上述问题转化为求解以 a_{ij}^t 为二元因变量，以 X_{ij}^{t-1} 为解释性变量的逻辑回归模型。本文之所以采取该模型是因为，首先该模型是建立在动态网络结构的框架下，其次可以在模型中灵活的加入协变量信息（例如文本语义信息）。

5.2 解释性变量探索

针对上述模型设定，本文考虑的所有解释性变量 X_{ij}^{t-1} 如表3所示。可以看到，除了传统的合作网络节点相关属性，本文还进一步考虑了文本语义相关的属性，例如研究主题和研究方向相似度。

令当期为 t，上一期为 $t-1$，接下来选取部分解释性变量探究它们与 a_{ij}^t 的关系。首先，统计在第 t 期有合作关系学者群体和在第 t 期无合作关系学者群体在各个领域的平均论文发表数量（见图4）。为方便书写和表达，这里使用编号I至VI表示社交网络分析、生物统计、协方差矩阵估计、变量选择、贝叶斯统计/非参数统计和假设检验/时间序列六个研究主题。可以看出，在不同领域，第 t 期有合作关系学者群体与第 t 期无合作关系学者群体在平均发表论文的数量上具有一定差异。例如，在生物统计领域，具有合作关系的学者群体平均发表论文数比无合作关系的学者群体高出34.30%，而在协方差估计领域，前者则比后者低20.52%。其次，本文还利用文本语义进行了相似度计算，并考察其对学者合作关系的影响。具体来说，首先，提取出每名作者所发表的所有论文对应的摘要和关键词，其次，通过

表3　合作关系可能影响因素汇总

变量维度	变量名			变量类型	取值范围	备注
网络结构	上一年度是否合作			定性变量	是/否	基准组：否
	共同合作者数量			连续变量	0-24	取值为整数
学者	个人信息		国籍是否相同	定性变量	是/否	基准组：否
			单位是否相同			
	学术产出		发表论文数之和	连续变量	0-17	取值为整数
			发表论文数之差			
	研究领域	在不同领域发表的论文数量	领域I：社交网络分析	连续变量	0-4	取值为整数
			领域II：生物统计		0-9	
			领域III：协差阵估计		0-13	
			领域IV：变量选择		0-6	
			领域V：贝叶斯统计/非参数统计		0-6	
			领域VI：假设检验/时间序列		0-5	
	研究方向相似度			连续变量	0-1	余弦相似度
论文	发表期刊	是否有发表于同一期刊		定性变量	是/否	基准组：否
		各期刊发表论文数	AoS	连续变量	0-10	取值为整数
			Biometrika			
			JASS			
			JRSS-B			
	被引情况	最高引数		连续变化	0-896	取值为整数
		平均被引数			0-817	
	引用情况	最高参考文献数			0-211	
		平均参考文献数			0-211	

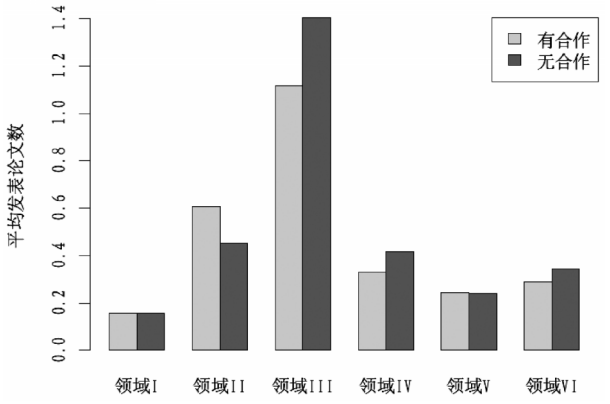
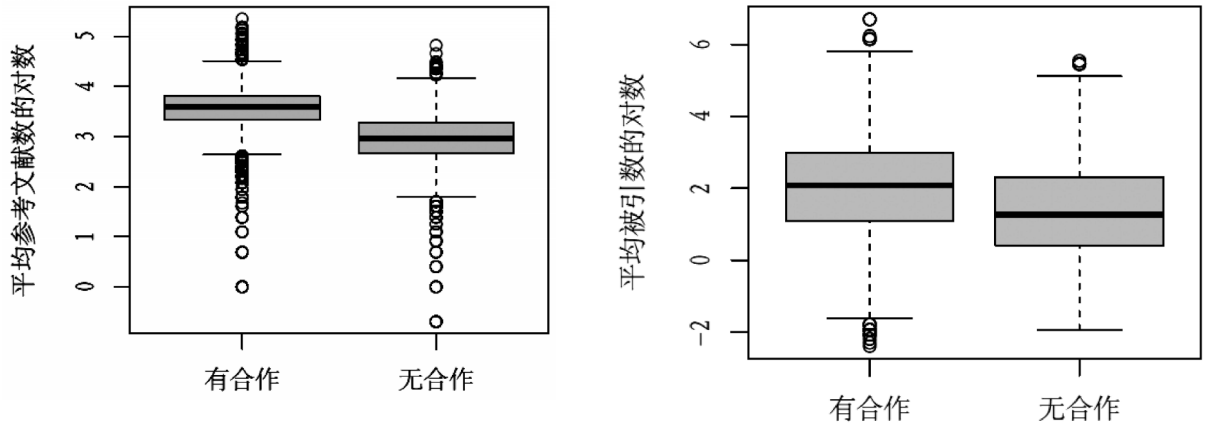


图4　各领域平均发表论文数对比

word2vec模型得到论文摘要和关键词文本信息的词向量表示，最后，利用夹角余弦公式计算每两个作者词向量的相似度，并以此作为学者研究方向相似度的衡量指标。基于上述计算，第 t 期有合作关系的学者群体的平均研究方向相似度高达0.73，第 t 期无合作关系的学者群体的平均研究方向相似度仅为0.13。最后，探索学者所著论文的平均被引次数和所著论文的平均参考文献数量对合作关系的可能影响（见图5）。相较于第



(a) (b)

图5 论文平均被引次数(左)和论文平均参考文献数(右)

t期无合作关系学者群体,第t期有合作关系学者群体发表的论文在平均被引数和平均参考文献数上均较高。

5.3 模型估计结果

本文选择2011—2017年的合作网络数据作为训练集,使用2018—2020年的数据作为测试集,模型估计结果如表4所示。从中可以得到以下结论:首先,网络结构因素,两名学者当年是否合作与他们的往年合作情况和共同合作者数量有关,上一年度发生过合作的、共同合作者数量越多的两名学者之间产生合作的可能性更大。其次,学者因素,相比于国籍,学者所在单位对合作关系产生的影响更为显著,即就职于同一单位的学者之间更有可能组成一个研究团队。发表论文数之差这一变量的估计系数均为正,说明在控制其他因素的影响后,两名学者发生合作的概率与他们的论文产出数量的差异成正比。这意味着,高产学者与低产学者之间的相互合作是一个较为普遍的现象。最后,论文因素,与其他期刊相比,在Biometrika发表论文越多的学者有更大的可能性与他人产生合作,学者所著论文的平均参考文献数量越多,越有可能与其他学者产生合作。

表4 全模型估计结果

变量维度	变量名		估计系数	标准差	P值	备注
网络结构	上一年度是否合作		1.41	0.37	<0.01	基准组:否
	共同合作者数量		6.32	0.25	<0.01	
学者	国籍是否相同		−0.38	0.17	0.03	基准组:否
	单位是否相同		1.78	0.25	<0.01	基准组:否
	发表论文数之和		0.41	0.26	0.11	
	发表论文数之差		0.31	0.07	<0.01	
	各领域发表论文数	领域II	0.15	0.14	0.29	剔除"领域I"以避免产生完全共线性
		领域III	0.23	0.22	0.31	
		领域IV	0.19	0.14	0.16	
		领域V	0.06	0.12	0.62	
		领域VI	0.10	0.11	0.38	
	研究方向相似度		3.80	0.10	<0.01	
论文	是否有发表于同一期刊		−0.25	0.24	0.30	基准组:否
	各期刊发表论文数	Biometrika	0.22	0.70	<0.01	剔除"AoS"以避免产生完全共线性
		JASS	−0.27	0.08	<0.01	
		JRSS-B	−0.10	0.07	0.15	
	平均被引数		0.48	0.17	0.01	
	最高参考文献数		−0.20	0.13	0.13	
	平均参考文献数		0.69	0.14	<0.01	

5.4 模型评价

得到估计系数$\hat{\beta}$后可以依据时刻t的网络特征构造出一个条件似然指标(Conditional Likelihood Index,CLI),其表达式如下:

$$CLI^t(\mathrm{i,j})=\frac{\exp(\hat{\beta}^T X_{ij}^{t-1})}{1+\exp(\hat{\beta}^T X_{ij}^{t-1})} \quad (3)$$

给定一个阈值c,当$CLI^t(\mathrm{i,j})>c$时,则预测$a_{ij}^t=1$,否则认为$a_{ij}^t=0$。选取不同的c会导致不同的预测结果产生。此时通过AUC值可实现对模型性能的综合评价[15]。对测试集2018—2020年的共63328组学者的合作情况进行预测,得到的AUC值为0.896(见下页图6),模型预测精度尚可。

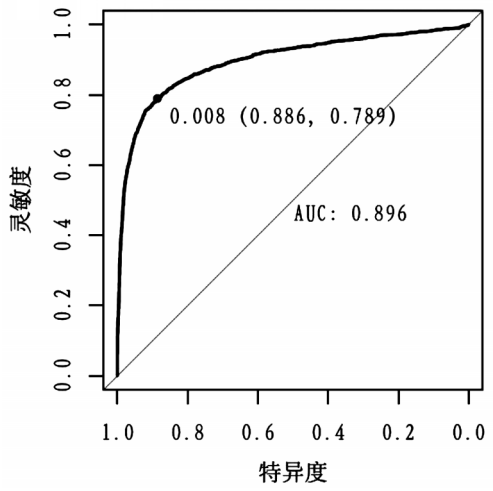


图6　基于全体测试集数据的ROC曲线

为了验证本文实证模型的优势，将本文提出的学者合作网络链路预测模型与经典的基于节点相似性的链路预测方法进行比较，包括共同邻居（Common Neighbors，CN）指标[16]、Salton指标[17]、Sorensen指标[18]和Jaccard指标[19]（计算公式见表5）。令Γ_i^t和Γ_j^t分别表示t时刻节点i和节点j各自的邻居节点的集合。CN指标计算了节点i和节点j的节点相似度，该指标认为，两个节点的共同邻居越多，两节点就越相似，节点之间建立关系的可能性越大。Salton指标又称为余弦相似度指标，是在公共邻居指标的基础上考虑了两个节点的度信息。Sorensen指标是基于Salton指标的改进。Jaccard指标也是基于CN指标的改进，该指标认为两个节点拥有的共同邻居节点占他们所有邻居节点的比例越高，则它们未来发生联系的可能性越大。根据不同方法的AUC值可以看出，本文提出的链路预测方法效果明显优于其他方法，印证了本文方法的优越性。

表5　不同链路预测算法效果对比

铁路预测方法	$a_{ij}^t=1$的计算公式	AUC
本文方法	$\frac{\exp(\hat{\beta}^T X_{ij}^{t-1})}{1+\exp(\hat{\beta}^T X_{ij}^{t-1})}$	0.896
CN	$\lvert\Gamma_i^t \cap \Gamma_j^t\rvert$	0.788
Salton	$\frac{\lvert\Gamma_i^t \cap \Gamma_j^t\rvert}{\sqrt{d_i^t \times d_j^t}}$	0.789
Sorensen	$\frac{\lvert\Gamma_i^t \cap \Gamma_j^t\rvert}{\sqrt{d_i^t + d_j^t}}$	0.764
Jaccard	$\frac{\lvert\Gamma_i^t \cap \Gamma_j^t\rvert}{\Gamma_i^t \cup \Gamma_j^t}$	0.764

6　总结与讨论

如何在科研领域实现学科合作关系的精准预测，是当前科研合作网络分析面临的一个重要问题。为了应对科研合作网络链路预测研究面临的挑战，本文提出了融合文本语义和动态网络结构的学者科研合作网络动态链路预测模型。首先，本文以统计学科为例，构建了一个学者科研合作数据集，丰富和补充了现有的科研社交网络研究资料。其次，借助文本分析法挖掘了网络节点的非结构化语义信息（如研究主题和研究内容相似度）和动态网络结构信息（如上一年的合作情况）。最后，结合节点多方面的拓扑结构特征、语义特征和动态网络结构特征，本文通过动态逻辑回归构建了动链路态预测模型，以分析学者合作关系的动态演化。研究结果表明，融合非结构化语义和动态网络结构的科研合作网络动态链路预测模型不但具有较强的解释性，还能够更好地预测学者合作关系的动态演化。

本文的研究结论对增强学者间的联系和提升科研合作效率具有重要借鉴意义。此外，本文的研究结论对科学社会学和科技政策的制定也具有一定的启示意义。首先，以本文所举的统计学科为例，可以看到，随着科学活动的进步，现代统计学科的研究方向正在发生巨大变化，主要体现在，作者合作模式在不同领域呈现出较大差别。通过构建并对比“社交网络分析”“生物统计”等六个领域的作者合作网络，本文发现统计学者的合作模式和偏好与其研究方向有较大关系，例如，相比其他领域的研究者，生物统计方向的学者与他人产生的合作会更为频繁。因此，在进行合作者推荐等方面的工作时，还应充分考虑到不同领域学者间的合作需求差异，做到有的放矢，从而节省经费、提高资源利用效率，使得科学社会的资源调配更加有效，科学社会分层更加合理。其次，本文的研究结论也可以给科技政策制定者一些新的启示。目前，“合作共赢”已成为当今统计学者的一大共识。根据本文构建的近十年统计学者合作网络的结构演化趋势，各国统计学者之间的合作在不断增强，“合作共赢”在激烈的学术竞争中已经成为一个普遍共识。在此背景下，科研发展更要顺应这一趋势，大力促进研究者之间的相互合作，有关部门应关注到这一庞大需求并有针对性地为学者们

提供相应的平台和机会，进而促进学术产出和知识传播。

本文未来可能的研究方向如下。第一，考虑加权合作网络的链路预测。根据学者合作次数对合作网络的连边赋予不同权重，或能进一步改善预测效果。第二，丰富文本语义信息和动态网络信息相关的指标，例如对研究领域进行进一步细分，考虑“跨领域”学者的研究方向随时间的变化情况等。第三，延长研究时限。在更长的时间范围内展开分时段研究，从而解决部分指标作用的滞后性问题。

参考文献

[1] Ji P, Jin J. Coauthorship and citation networks for statisticians [J]. The Annals of Applied Statistics, 2016, 10(4):1779-1812.

[2] 付 允，牛文元，汪云林，等. 科学学领域作者合作网络分析——以《科研管理》(2004-2008)为例[J]. 科研管理，2009，30(3)：41-46.

[3] 张利华，闫 明. 基于SNA的中国管理科学科研合作网络分析——以《管理评论》(2004-2008)为样本[J]. 管理评论，2010，22(4)：39-46.

[4] Moody J. The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999 [J]. American Sociological Review, 2004, 69(2):213-238.

[5] 陈海珠，李树青，汪圣忠，等. 近代中国农业领域科研合作网络分析[J]. 大学图书馆学报，2019，37(4)：79-87.

[6] 韩童茜，王立梅，许 鑫. 长三角城市群科研合作网络演化研究——基于SCIE和SSCI论文的实证分析[J]. 情报理论与实践，2020，43(10)：151-156.

[7] Gao T, Pan R, Wang S,et al. Community detection for statistical citation network by D-SCORE[J]. Statistics and Its Interface, 2021, 14(3): 279-294.

[8] 吕琳媛，周 涛. 链路预测[M]. 北京：高等教育出版社，2013.

[9] Popescul A, Ungar L H. Statistical relational learning for link prediction [C]. IJCAI Workshop on Learning Statistical Models from Relational Data, 2003.

[10] Claiset A, Moore C, Newman M E.et al. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191):98-101.

[11] Pieter B, Koller A D. Link prediction in relational data[C]. Advances in Neural Information Processing Systems, 2003, 16: 659-666.

[12] 李小龙，张海玲，刘 洋. 基于动态网络分析的中国高绩效科研合作网络共性特征研究[J]. 科技管理研究，2020，40(7)：116-124.

[13] Pepe M S, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve [J]. Biometrics, 2006, 62: 221-229.

[14] Mimno D, Wallach H, Talley E,et al. Optimizing semantic coherence in topic models[C]. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011：262-272.

[15] Zhou J, Huang D, Wang H. A dynamic logistic regression for network link prediction [J]. Science China, 2017 (1): 1-12.

[16] Zhou T, Lu L, Zhang Y. Predicting missing links via local information [J]. The European Physical Journal B, 2009, 71(4), 623-630.

[17] Salton G, McGill M J. Introduction to Modern Information Retrieval [M]. 1986, McGraw-Hill Inc, USA.

[18] Sorensen T A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons [J]. Biol. Skar.,1984, 5: 1-34.

[19] Jaccard P. Etude Comparative De La Distribution FloraleDansUne Portion Des AlpesEt Des Jura [J]. Bull SocVaudoiseSci Nat, 1901, 37: 547-579.

[作者简介]周 静，女，1989年生，中国人民大学应用统计科学研究中心，中国人民大学统计学院副教授。

庄艳阳，女，1996年生，中国人民大学统计学院硕士研究生。

周季蕾，女，1993年生，中国人民大学信息学院讲师(通讯作者)。

收稿日期：2022-03-03