电子科技大学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

博士学位论文

DOCTORAL DISSERTATION

论文题目 社会经济系统的空间结构与动力学研究

学科专业 计算机软件与理论

学　　号 201411060110

作者姓名 高　见

指导教师 周　涛　教　授

分类号 _____ 密级 _____

UDC [注1] _____

学　位　论　文

社会经济系统的空间结构与动力学研究

（题名和副题名）

高　见

（作者姓名）

指导教师　周　涛　　教　授

电子科技大学　　成　都

（姓名、职称、单位名称）

申请学位级别　博士　　学科专业　计算机软件与理论

提交论文日期 2019.04.10　论文答辩日期　2019.05.25

学位授予单位和日期　电子科技大学　2019 年 06 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

Research on the Spatial Structure and Dynamics of Socio-Economic Systems

A Doctoral Dissertation Submitted to

University of Electronic Science and Technology of China

Discipline: **Computer Software and Theory**

Author: **Jian Gao**

Supervisor: **Prof. Tao Zhou**

School: **School of Computer Science & Engineering**

独创性声明

本人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。据我所知，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，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。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谢意。

作者签名：_____ 日期： 年 月 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技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，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日期： 年 月 日

摘 要

社会经济系统是一类重要的复杂系统，涉及到人类经济活动与所处社会环境的复杂相互作用。人类的认识和行为不断发生变化，主观决策过程极大地影响社会经济系统的运行。精准和及时地感知社会经济态势，揭示和理解社会经济发展规律，有重大的理论意义和应用价值。洞察社会经济发展中各方面的状态，并对其发展趋势进行准确的预测，有助于科学地指引社会经济决策。揭示个体的社会经济行为模式，能帮助逐渐实现预测性管理。刻画宏观的社会经济结构，有助于探寻经济发展路径。如何有效地分析社会经济系统的结构与演化规律，是多学科交叉研究领域所关注的重要科学问题，近年来得到了包括计算机科学、网络科学、复杂性科学、统计物理和社会经济学在内的很多相关学科的极大关注。

传统的社会经济研究依靠定性或半定量方法，导致不容易从机制层面认识相关问题。利用传统普查数据计算宏观经济指标，整个过程不但消耗大量资源，而且时间滞后很长。不仅如此，传统分析方法难以洞察经济发展的结构转变，无法刻画经济发展过程中的复杂性，缺乏预测经济发展趋势的能力。近年来，硬件和技术的同步发展推动数据化浪潮，为社会经济研究带来了前所未有的机遇和改变。数据获取方式的进步，提高了大规模社会经济数据的可用性。数据规模和多样性的增加，促进了社会经济分析工具和方法论的变革。逐渐应用的新数据和新方法，提高了社会经济研究的定量化程度，催生了一个新兴的交叉学科研究分支，称为计算社会经济学。本文在计算社会经济学框架下，将分别从微观、中观和宏观层面研究社会经济系统的状态推断和结构建模，进而以理论结合实证的方式探究经济的结构演化和发展策略。特别地，不同层面的研究基于类似的空间网络结构和动力学理论基础。本文主要的研究内容和创新点总结如下：

（1）在微观层面，基于非干预行为数据研究了社会经济预测性管理。通过分析匿名校园卡数据，提出了谨严性指数来刻画个体行为规律程度。发现谨严性与学生成绩显著相关，使用谨严性特征能显著提高排序学习算法对学生成绩的预测效果。基于企业社会化平台数据构建互动网络和社会网络，发现利用员工在网络中的位置能预测其升离职的可能性。特别地，互动网络比社会网络的预测能力强，预测离职比预测升职容易。通过分析大规模在线平台数据，以量化方式揭示了一些社会经济现象，包括团队规模在8人以下能提高员工沟通效率和绩效表现，中国社交圈规模也在邓巴数150人左右，职场中存在身高溢价和性别不平等现象。

（2）在中观层面，基于在线用户评分数据研究了社会经济系统排序。针对信誉排序问题，提出了基于群组的信誉排序GR算法。不依赖传统的产品质量假设，GR算法根据评分群组规模计算用户信誉。真实数据集上的实验结果表明，GR算法对用户的信誉排序比基准算法更准确。进一步，利用迭代寻优过程改进GR算法，提出了迭代信誉排序IGR算法。同时考虑用户数量和信誉计算群组规模，IGR算法的排序准确性和鲁棒性更好。针对产品排序问题，提出了节点相似性CosRA指标，基于此提出的CosRA推荐算法表现更好。进一步，提出了考虑用户信任关系的CosRA+T推荐算法，发现过度依赖信任关系有损推荐效果。

（3）在宏观层面，基于大规模真实数据刻画和分析了社会经济结构。利用企业注册信息数据，刻画了中国区域经济复杂性。发现复杂性ECI指标和Fitness指标对中国区域经济发展的预测能力相当，复杂性与收入不平等性负相关。利用人力和企业数据，分别构建了巴西和中国区域产业空间。发现两者都有“核心-边缘”结构，复杂程度高和低的产业分别占据核心和边缘位置。中国产业空间还有“哑铃型”结构，在时间演化上存在区域竞争。利用微博和简历数据，分别构建了信息和人才流动网络。发现根据网络的结构特征能推断区域经济发展水平。特别地，人才流动网络的预测能力强，结合两个网络的特征能解释大约84%的GDP变化。

（4）在经济发展和结构演化方面，基于空间网络研究了经济演化路径和产业升级策略。利用空间网络模型和传播动力学过程，揭示了网络的空间结构对信息传播的影响。发现空间网络的长边分布能改变靴襻渗流的相变类型，长边分布的幂指数-1为出现相变点不变的双相变的临界值。针对产业空间和地理近邻网络，分别提出了经济发展的相似技术学习途径和近邻区域学习途径。发现两条途径都能促进区域发展新产业，但两者存在替代效应。进一步，以理论分析结合实证数据，研究了发展经济的最优策略。发现缩短距离能提高协同学习效果，引入高铁能提高区域的产业相似性和生产率，两条协同学习途径都存在最优发展策略。

计算社会经济学是一个新兴研究分支，在数据和方法上面临新挑战和新机遇。在未来研究中，值得进一步探索社会经济系统的空间结构与动力学，提高对社会经济态势的感知和对发展规律的理解。长期而言，数据驱动的研究范式必将成为解决社会经济问题的主流方法论，也将深刻地改变社会经济研究的图景。

关键词：复杂网络，社会经济系统，排序算法，经济复杂性，网络结构

ABSTRACT

Socio-economic systems are an important branch of complex systems, which involves the complex interactions between people's economic activities and the social environment in which they live. With the constant change of cognition and behavior, people's subjective decision-making process greatly affects the operation of socio-economic systems. To accurately and timely perceive socioeconomic situation and to reveal and understand the law of socioeconomic development have great theoretical and practical values. Revealing the status of socioeconomic development in many aspects and predicting the development trends with desirable accuracy can greatly help to guide socioeconomic decision-making. Uncovering the socioeconomic behavioral patterns of individuals can contribute to gradually realizing predictive management. Quantifying the macro socioeconomic structure can help to explore the path of economic development. How to effectively analyze the structure and evolution of socio-economic systems is an important scientific issue in the interdisciplinary research field, and it has recently received great attention from many related disciplines including computer science, network science, complexity science, statistical physics and socioeconomics.

Traditional socioeconomic research relies mainly on qualitative or semi-quantitative methods, which makes it difficult to understand relevant issues at the mechanism level. The process that calculates macroeconomic indicators based on traditional census data not only consumes substantial resources, but also follows a long-time delay. Besides, traditional analytical methods have difficulty in tracking the structural transformation of economic development, fail to quantify the complexity of economic development and are lack of predictive power on development trends. The recent simultaneous development of hardware and technology is driving a new wave of big data, which has brought unprecedented opportunities and changes to socioeconomic research. The advances in methods of data acquisition have increased the availability of large-scale socioeconomic data, and the increases in the size and diversity of data have contributed to the transformation of socio-economic analytical tools and methodologies. The application of novel data and methods has gradually increased the level of quantification in socioeconomic research and led to the emergence of a new scientific branch, named Computational Socioeconomics. Under the framework of computational socioeconomics, this dissertation will

investigate the status inference and structural modeling of socio-economic systems from the micro, meso and macro levels, and explore the evolution of economic structure and the optimal strategy for economic development through theoretical and empirical studies. In particular, studies at different levels are based on the similar theoretical basis of network spatial structure and dynamics. The main contents and major contributions of this dissertation are summarized as follows:

(1) At the micro level, the predictive management of socio-economic systems was studied based on unobtrusive behavioral data. By analyzing data recorded by anonymized campus cards, we proposed a novel orderliness measure to quantify the regularity of individual behavior. Orderliness is significantly correlated with student academic performance, and it can largely improve the performance of learning-to-ranking algorithm on predicting student academic performance. Based on the analysis of two employee networks built on data from an enterprise socialization platform, we found that the locations of employees in both networks are predictive to the possibility of their promotion and resignation. In particular, action network has stronger predictive power than social network, and predicting resignation is easier than predicting promotion. Moreover, by analyzing large-scale online platform data, we revealed some socio-economic phenomena in a quantitative way, including keeping team size below 8 can improve employee's communication and performance, the size of Chinese social circle is also around Dunbar's Number 150, and there are height premium and gender inequality in the workplace.

(2) At the meso level, the ranking of socio-economic systems was studied based on online user rating data. To solve the of problem reputation ranking, we proposed a group-based reputation ranking (GR) method. Instead of relying on the traditional assumption of product quality, GR method calculates user reputation based on the size of rating groups. Experiments based on real-world datasets showed that GR method outperforms benchmark methods in the accuracy of ranking users by their reputation. By introducing an iterative process into the GR method, we further proposed an iterative group-based ranking (IGR) method. Considering both the number and the reputation of users when calculating the group size, GR method exhibits better accuracy and robustness in reputation ranking. To solve the problem of object ranking, we proposed a novel vertex similarity measure, named CosRA index, based on which we developed a CosRA-based recommendation algorithm that exhibits better performance. Further, we proposed a trust-based recommendation algorithm, named CosRA+T, and found that relying too much on

trust relations among users is detrimental to recommendation performance.

(3) At the macro level, socio-economic structures were quantified and analyzed based on large-scale real data. Using firm registration information data, we quantified China's regional economic complexity. We found that ECI index and Fitness index exhibit comparable predictive power for China's regional economic development, and economic complexity is negative correlated with income inequality. Using labor and firm data, we built Brazil's and China's regional industry space, respectively. We found that both industry spaces exhibit a "core-periphery" structure, where industries with high and low level of sophistication occupy the core and the periphery of the industry space, respectively. Moreover, China's regional industry space has a "dumbbell" structure, and its time evolution has regional competitions. Based on Weibo and resume data, we built information flow and talent mobility network, respectively. We found that regional economic status can be inferred from the structure of both networks. In particular, talent mobility network exhibits a stronger predictive power, and combining the structures of both networks can explain about 84% of the variance in GDP.

(4) In economic development and structure evolution, the path of economic evolution and the strategy of industrial upgrading were studied based on spatial networks. By leveraging the spatial network model and the spreading process, we revealed the effects of the spatial structure of networks on information diffusion. We found that the distribution of long-range links of spatial networks can change the phase transition of bootstrap percolation, where the exponent -1 of the distribution of long-range links is a critical value for the presence of a double phase transition with two nearly constant critical points. For industry space and geographical adjacent networks, we proposed the inter-industry learning and the inter-regional learning for economic development, respectively. We found that both collective learning channels can increase the probability of development new industries, while they exhibit an alternative effect. Moreover, we explored the optimal strategy for economic development using both theoretical and empirical analyses. We found that reducing geographical distance can enhance the collective learning effects, introducing high-speed rail can increase regional industrial similarity and productivity, and both collective learning channels have optimal strategies for industrial development.

Computational socioeconomics is an emerging research branch, and it faces new challenges and opportunities in both data and methods. In future studies, it is worthwhile to further explore the spatial structure and dynamics of socio-economic systems, and to

improve the perception of socioeconomic situation and the understanding of the law of development. In the long run, data-driven research paradigm will become the mainstream methodology for solving social and economic problems and will profoundly change the landscape of socioeconomic research.

Keywords: complex networks, socio-economic systems, ranking method, economic complexity, network structure

目 录

第一章 绪论

1.1 研究背景与意义

社会经济系统是一类重要的复杂系统[1, 2]，涉及到人类经济活动和所处社会环境的复杂相互作用[3]。不同于非生命的物理系统，社会经济系统更加难以被描述、理解和控制[4, 5]。一方面，社会经济系统的层次和功能结构尚不明确，传统分析方法难以刻画社会经济系统中主体之间的复杂相互作用[6]，对系统各个组成部分的独立分析不能确定系统整体行为[7]。另一方面，社会经济系统的运行状态和发展趋势不易推断，人类的认识和行为不断发生动态变化，普遍存在人类主观决策过程对社会经济系统有很大的影响[8]。如何有效地分析和理解社会经济系统的结构与动力学，既是社会经济学所关注的重要研究问题[9]，近年来也得到了包括计算机科学、网络科学、复杂性科学、统计物理和社会经济学在内的很多相关学科的极大关注。感知社会经济态势和揭示社会经济规律[10, 11]，具有重要的决策支撑价值和社会经济效益。精确感知社会经济发展中若干重要方面的状态，并对其发展趋势做出准确判断，对社会经济决策有指引作用[12]。

社会经济学交叉融合了经济学和社会学[13, 14]，它将经济系统看做社会系统的一部分，从社会文化角度分析经济过程。同时，社会经济学也运用来自社会学的理论和方法，分析经济结构和经济行为。然而，传统社会经济研究存在很多局限性，导致其不容易从机制层面认识相关问题[15]。在数据方面，以问卷访谈等形式获取的小规模数据，容易受到心理防御等因素干扰[16]，难以推广到全体尺度[17]；以普查等形式获取的大规模数据，费时耗力；缺乏社会经济发展实时的、过程性的数据[9]。在工具方面，经济计量分析等传统方法[18]，无法处理相互作用涌现的复杂性，缺乏对社会经济结构的刻画[19, 20]，无法分析社会经济发展的动力学过程[21, 22]，对未来的预测能力也不足。特别地，传统方法对于社会经济态势的感知，大多基于统计数据计算宏观经济指标[23]，不仅耗费大量人力物力，而且时效性很差。例如，统计年鉴数据至少有两年的时间滞后，难以满足实时的社会经济决策。另外，传统GDP等综合指标仅能从单一维度估计经济发展所处的大概阶段，无法体现出经济发展的结构特征和复杂性[24]。GDP总量相同的两个国家，在产品和产业结构上可能差别很大，在未来的经济发展潜力也不同[25]。

随着硬件和技术的同步发展，全世界都在经历数据化浪潮[26]，这为社会经济研究带来了前所未有的机遇和改变[27]。一方面，数据获取方式的进步，提高了大规模社会经济数据的可用性。借助先进数据采集终端和传感设备，能方便地获取

发展过程数据和人类行为数据，包括手机通讯数据[28]、社会媒体数据[29]、卫星遥感数据[30]、网络检索数据[31]等。这些新型数据拥有获取成本低、更新及时、规模大、时空分辨率高和涵盖范围广等优势，弥补了传统社会经济数据的不足。另一方面，数据规模和多样性的增加，促进了社会经济分析工具和方法论的变革。例如，依靠数据挖掘和机器学习等先进技术，如深度学习算法[32]，处理大规模新型数据（如文本内容和图像数据等）；借助网络科学和统计力学等交叉学科的工具和方法[33, 34]，分析社会经济结构和复杂相互作用；利用少量人工标注数据训练机器学习算法，推断全体尺度上难以获得的高价值数据[30]。近年来，新数据和新方法的逐渐应用，提高了社会经济研究的定量化程度，催生了一个全新的交叉学科研究分支，我们称之为计算社会经济学（Computational Socioeconomics）[35]。借助先进工具分析大规模真实社会经济数据，计算社会经济学以量化手段揭示社会经济发展规律，提高了人们对社会经济系统结构、功能和动力学的认识。

计算社会经济学主要关注两方面研究：一是对于社会经济态势的感知和推断，包括个体、群体、区域、国家等层面的状态和发展趋势；二是对社会经济规律的洞察和理解，包括人类移动、城市布局、产业演化和经济发展等方面的规律和机制[15, 35]。对状态的感知有助理解规律，对规律的理解也有助预测状态。作为一个新交叉学科研究分支，计算社会经济学研究在很多方面都面临着挑战。首先，微观层面的非干预行为数据类型多、规模大和结构复杂，不容易从数据中直接抽取行为特征，导致难以揭示人类行为动力学规律，缺乏对个体行为倾向的预测能力[36]。其次，中观层面的社会经济系统存在复杂相互作用，群体行为之间相互关联和影响，导致无法通过单独分析个体状态来确定系统状态，必须从整体上对社会经济状态进行推断[7]。再次，宏观层面的社会经济系统涌现出很多复杂结构，传统方法难以对社会经济结构进行建模，也无法刻画社会经济发展的复杂性，迫切需求新的分析工具[37]。最后，社会经济系统受众多因素影响而不断动态变化[2]，导致难以从机制层面揭示经济结构演化的动力学规律，也不容易提出经济的发展路径和发展策略[38]。尽管面临以上的困难和挑战，计算社会经济学作为一个极具活力的新研究分支，有希望借助新工具和新数据解决或部分解决这些难题。鉴于此，本文将从微观预测性管理、中观系统排序、宏观结构建模、发展路径与策略这四个方面，展开对社会经济系统的空间结构与动力学的研究。

在微观层面，个体行为复杂多样和难以追踪，如何精准地感知和预测个体社会经济状态是一个难题。常用的办法是收集大规模非干预行为数据，建模刻画个体行为特征，揭示人类行为动力学规律[36]。进一步，构建机器学习模型，利用行为特征预测个体状态和行为倾向[39]。以非干预手段收集的行为数据，如移动轨

迹[40]、社交沟通[41]和校园刷卡[42]等数据，有规模大、时空分辨率高和代表性强等优势，对预测性管理研究有重要意义。首先，利用大规模数据揭示人类动力学规律，如在线行为的阵发性[43]、移动行为的可预测性[44]和应急情况下的迁移规律[45]等，为理解人类行为提供深刻洞见，有理论价值。其次，利用行为特征预测个体社会经济状态，如财富状况[46]、学习成绩[47]和离职倾向[48]等，为预测性管理提供决策依据，有应用价值。最后，利用大规模数据揭示社会经济现象，如团队规模效应[49]、男女不平等性[11]和身高溢价效应[50]等，为解决社会经济问题提供指导，有社会经济价值。考虑到真实数据的类型复杂多样，需要针对具体问题选择合适的行为数据和分析工具，以提高对社会经济状态的预测准确性。

在中观层面，社会经济系统中群体行为之间存在复杂的相互作用，个体状态不仅由自身行为特征决定，还受其他个体影响，这给推断社会经济系统状态带来了很大困难。一种有效的方法是对相互作用进行网络建模，利用网络排序算法从整体上推断社会经济系统的状态[51]。特别地，在线评分系统是一种常见的社会经济系统[52]，可以利用“用户-评分”二部分网络建模[53]。通过分析刻画用户评分行为模式的二部分网络结构，不仅能推断用户的信誉水平[54]，还能为用户推荐可能喜欢的产品[55]。基于网络结构分析的这两类应用，本质上是解决社会经济系统的排序问题[56]。将推断社会经济系统整体状态转化为解决社会经济网络的排序问题，主要有两方面的优势。一方面，利用网络排序方法能方便地估计社会经济系统的相对状态。例如，基于评分偏差评价用户相对信誉水平，利用信誉排序检测作弊评分用户，维护评分系统的健康运行[51]。另一方面，利用网络动力学过程和其他辅助信息，容易提高对社会经济的排序效果。例如，借助网络扩散动力学过程和引入用户信任关系，设计新的个性化推荐算法，提高推荐结果的准确性[57]，解决信息过载问题。由于网络结构体现群体行为和相互作用模式，如何通过网络结构分析来提高排序算法对社会经济系统状态的推断效果值得进一步研究。

在宏观层面，经济发展伴随着复杂性的提高和结构的转变[58, 59]，如何刻画经济结构和复杂性是一个用传统方法无法解决的难题。近年来，复杂网络分析方法的广泛应用[60]，为刻画经济结构和复杂性提供了新工具。从网络角度理解经济发展[61]，有重要的理论意义和应用价值。特别地，基于国际贸易数据计算产品之间的接近性，构建产品空间网络[25]，能直观地反映出国家经济发展的整体概貌。实际上，国家所出口产品的多样性和复杂程度不同，在产品空间网络中占据的位置不同，当前的经济结构限制着其未来发展潜力[25]。由此可见，通过对经济结构进行网络建模，能为理解国家经济发展提供更深刻的洞见。通过分析“国家-产品”二部分网络结构，能量化国家经济复杂性[24, 62]。作为一种非货币性指标，经济复

杂性能很好地预测国家经济发展水平[63]。此外，基于大规模数据能构建社会经济网络，如手机通讯网络[64]和在线社交网络[65]，利用网络结构特征能准确地推断区域的社会经济水平[66]。分析经济复杂性和社会经济网络的结构，不依赖于宏观统计数据，对实现精准和及时的社会经济态势感知有重要意义。考虑到不同尺度上数据可用性不同，如区域层面没有国际贸易数据，如何将经济结构建模和复杂性分析拓展到区域层面仍然是一个亟待解决的问题。

在经济发展和结构演化方面，迫切地需要揭示经济发展路径[67]，探寻能快速实现产业升级的道路[38]，研究能最快提高经济水平的发展策略。宏观层面的经济网络建模，为分析经济发展路径提供了基础。产品空间展现出产品之间的接近性，国家更容易发展与已有产品相接近的产品，逐渐占据位于产品空间中心的复杂产品[25]，不断提高经济发展水平。类似地，区域更容易发展与已有产业在技术上相接近的产业，倾向于淘汰与已有产业在技术上关联性小的产业，以此实现产业结构的不断优化升级[67]。研究经济的发展路径[68]和揭示产业发展的路径依赖效应[69]，有助于对区域经济结构调整进行科学地指导。随着区域经济发展水平的不断提高，经济结构会相应的发生转变[58, 59]，这要求区域动态地调整经济发展目标[12]，采用与发展阶段相适应的产业发展策略，以求抓住机遇实现最快的经济发展速度。考虑到经济的结构转变和发展的路径依赖，如何制定行之有效的经济发展策略是一个难题。解决该问题的一种可行方法是利用传播动力学模型[70]，在产品空间网络上模拟经济发展过程，探究产业的选择策略和空间网络的结构对经济发展的影响[12]，以此提出经济发展和产业升级的最佳策略。

综上所述，社会经济系统的结构复杂，传统方法不易精准地感知社会经济态势，也很难深刻地揭示社会经济发展规律。近年来，广泛应用的新数据和新工具，催生了计算社会经济学这一新兴的交叉学科研究分支，使用定量化手段研究社会经济发展中的各种现象。然而，微观层面个体的行为动力学模式复杂，很难精准地感知和预测个体社会经济状态；中观层面群体行为之间彼此关联和影响，不容易直接推断社会经济系统的整体状态；宏观层面缺少刻画社会经济空间结构和分析经济复杂性的方法，对区域经济发展的感知和预测能力不足；整体上对经济发展路径和发展动力学机制缺乏深刻洞见，也不容易提出最优经济发展策略。因此，在计算社会经济学框架下，进一步研究社会经济系统的空间结构与动力学，有非常重要的理论意义和应用价值，为制定科学的社会经济政策提供支撑。

1.2 国内外研究现状

计算社会经济学以定量化手段分析大规模真实数据，致力于精准和及时地感

知社会经济状态，揭示社会经济发展规律，帮助改善社会经济水平。一方面，所基于的大规模社会经济数据，如卫星遥感、手机通讯、社交媒体等，有低获取成本、实时更新和高时空分辨率等优势。另一方面，所使用的交叉学科分析工具，如统计力学、网络分析、机器学习、文本挖掘等，极大地提升了对社会经济发展态势的感知和预测能力。总体而言，计算社会经济学的研究尺度跨越微观、中观和宏观等多个层面，研究内容主要关注三个方面：一是社会经济态势的感知和推断；二是社会经济结构的建模和刻画；三是社会经济发展规律的理解和应用。下面将分别从态势感知、结构刻画和发展规律方面介绍国内外研究进展。

1.2.1 社会经济态势感知研究

社会经济态势主要是指社会经济发展的状态和趋势。这里所说的状态，从微观上讲，包括个体财富情况、职业状态、绩效表现、信誉水平、社会属性、人格特质和情感状况等；从中观和宏观上讲，包括城市的景观感受、设施布局情况、城市发展水平，以及国家和地区的经济水平、财富分布、不平等性、经济多样性等。传统的社会经济态势感知大多依赖普查等手段，不但消耗大量社会资源，而且整个过程有很长的时间滞后。如今，借助现代技术能以低成本实时地获取大规模社会经济数据，利用先进交叉学科工具加以分析，能更精准和及时地感知社会经济态势。下面介绍在感知社会经济态势方面的研究进展。

微观层面的研究利用行为特征推断个体社会经济状态。在电子设备使用方面，Soto等人[71]从手机数据中抽取用户行为特征，构建机器学习模型以高达80%的准确率区分用户的社会经济水平；Gutierrez等人[72]计算科特迪瓦手机用户充值金额的变异系数，以此估计用户的收入水平；Blumenstock等人[73]利用卢旺达手机数据构建预测模型，以此预测用户的家庭财产情况。在移动行为模式方面，Frias-Martinez等人[39]分析了一个拉美国家的手机数据，发现富人有更大的移动范围；Lotero等人[74]分析了哥伦比亚旅行调查数据，发现个体出行早高峰的时间随着财富水平的上升而滞后；Yang等人[75]从手机数据中抽取6种移动行为特征，提出了一种数据融合的方法预测用户社会经济水平；Kassarnig等人[47]分析了学生行为模式与成绩的关系，发现利用课堂出勤率和社会连接能预测成绩。在网络结构方面，Fixman等人[76]分析了墨西哥手机通讯和银行卡数据，发现通讯网络所连接的用户收入水平相近；Luo等人[77]分析了一个拉丁美洲国家的手机通讯网络，发现借助用户在网络中的位置信息能推断其财富状况；Jahani等人[78]利用手机数据构建了自我中心网络，发现网络结构多样性与收入水平非常相关。

利用大规模真实数据能推断个体的职业状态和绩效表现。在利用手机数据方面，Toole等人[80]根据通讯模式的变化检测大规模裁员，如失业个体的通话量

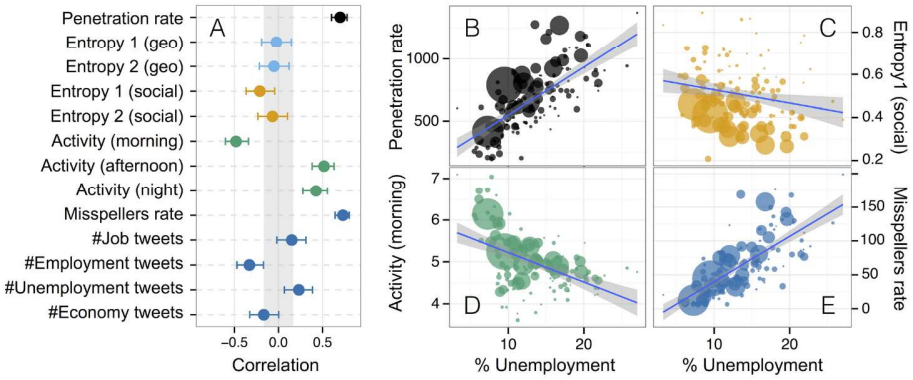

图 1–1 社交媒体用户行为特征与失业率之间的关系[79]

降低51%；Sundsøy等人[81]从手机数据中抽取反映经济和移动行为的特征，训练机器学习模型以73.5%的准确率预测失业个体；Almaatouq等人[82]从手机数据中抽取通话数量和多样性等特征，基于此预测地区失业率。在利用社交数据方面，Feeley等人[83]发现社会网络中与他人沟通多的员工不易离职；Llorente等人[79]从西班牙推文数据中抽取用户行为特征，发现失业率与通讯多样性和推文错频率等特征非常相关（如图1–1所示）；Bokanyi等人[48]分析了美国带有地理标记的推文数据，发现推文节律性能预测失业率。在利用搜索数据方面，Askitas等人[84]利用Google Insights数据预测德国失业率，发现失业相关的搜索关键词与失业率强相关；D'Amuri等人[85]发现使用了GI数据的模型在预测失业率上更准确。在利用非干预行为数据方面，Olguín等人[86]分析了徽章传感器记录的团队成员沟通数据，发现沟通交流能提高绩效；Watanabe等人[87]分析了徽章传感器收集的呼叫中心的员工行为数据，发现团队绩效与休息时员工面对面交流的活跃程度相关。

分析在线平台数据还能推断用户的其他社会经济状态。在信誉评价方面，通常使用在线评分数据，假设产品有唯一质量分数，以用户评分与产品质量的偏差来估计用户信誉。例如，Laureti等人[88]迭代计算用户信誉和产品质量，评分偏离产品质量越小的用户信誉越高；Zhou等人[51]利用评分和产品质量的相似性计算用户信誉；Liao等人[52]以非线性迭代增强高信誉用户在信誉评价中的影响力。在推断人格特质方面，Schwartz等人[89]分析了百万量级的Facebook数据，发现利用语言特征能预测人格特质，如外向性人格的用户喜欢使用“party”等社交用词；Guntuku等人[90]分析了Twitter上的150万张图片推文数据，发现利用图片和点赞能预测用户的大五人格；Segalin等人[91]分析了Facebook上用户资料图片的视觉特征，发现亲和性和外向性人格的用户偏好使用暖色图片。在估计情绪和健康状况方面，Larsen等人[92]构建了一个“We Feel”系统来分析用户在Twitter上情绪表达的变化，将推文情感词分为6大类和25小类；Mohammad等人[93]标注了一个带有情绪强度的推文数据集，发现情感词标签能影响情绪强度，利用单词嵌入和词汇特征能最

好地预测用户情绪强度；Reece等人[94]分析了Instagram上的照片特征，发现抑郁用户的发布频率高、图片中人脸少和倾向于使用滤镜等；Sueki[95]分析了Twitter上与自杀有关的推文数据，发现发布自杀推文与自杀未遂行为显著相关。

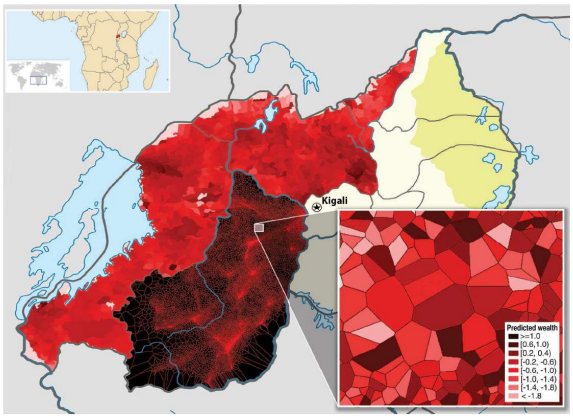

图 1–2 基于手机数据绘制贫困与财富分布地图[46]

宏观层面的社会经济态势感知，一方面聚合分析手机通讯和社交媒体等数据，另一方面使用街景图片和卫星遥感等独到数据。在利用手机数据方面，Eagle等人[66]分析了英国社区之间的手机通讯网络，发现网络结构多样性与社区经济发展指标强相关；Mao等人[96]利用手机通讯网络判断区域相对重要性，提出的通话排名指数能预测区域收入等社会经济指标；Blumenstock等人[46]基于卢旺达手机数据构造上千个特征，利用机器学习模型推断个体财富状况，重建出国家的贫困和财富分布地图（如图1–2所示）。在利用社交媒体数据方面，Liu等人[97]分析了超过2亿微博用户的注册信息，发现用户在线活跃程度能关联区域经济水平；任晓龙等人[65]分析了QQ在线社交网络的用户数据，发现城市的用户数和聊天数与其经济和交通等社会经济发展指标正相关；Holzbauer等人[98]分析了美国社交媒体上用户之间的朋友关系，发现跨州的长程连边数量与社会经济水平强相关。在利用街景图片方面，Salesses等人[99]分析了用户对Google街景图片的评分数据，将城市街景感观与不平等性等社会经济指标联系起来；Naik等人[100]在此基础上提出了Streetscore街景评分算法，给海量街景图片自动打分，从而构建城市感观地图。在利用卫星遥感数据方面，Elvidge等人[101]分析了卫星拍摄的夜间灯光数据得到贫穷指数，绘制出全球范围的贫穷程度地图；Jean等人[30]利用迁移学习方法分析夜间灯光和白天卫星图像特征，使用机器学习算法预测区域内的家庭财产状况。

在利用大规模数据感知社会经济态势方面，虽然已经产生了大量的研究成果，但在解决具体问题的数据和方法等方面，还有待进一步提高。例如，个体层面的非干预行为数据不容易获得，刻画行为规律性的方法还很缺乏；研究大多关注贫困国家和发达国家，针对发展中国家和地区的研究还相对比较少。

1.2.2 社会经济结构刻画研究

社会经济发展从简单统一逐渐向复杂多样过度。传统的社会经济指标，如人均GDP和克强指数[102]，仅能从单一维度估计社会经济发展所处的大致阶段，缺乏对发展过程中所涌现的复杂性和结构改变的刻画[58, 59]。从空间结构的角度分析社会经济系统，有助于更好地感知社会经济状态、理解社会经济运行规律和把握社会经济发展趋势。这里所说的空间结构，从微观上讲，包括个体之间社会网络的社团结构、层次结构和空间网络结构等；从中观和宏观上讲，包括社会经济二部分网络结构、城市空间的基本功能结构，以及构建的产品和产业空间结构等。下面介绍社会经济结构刻画方面的研究进展。

微观层面的研究主要关注社会网络结构分析，包括网络的基本结构参量、社团和层次结构、空间网络结构等。在基本结构参量方面，Barabási[33]在网络科学图书中介绍了很多结构指标，如网络度分布、度相关、节点重要性等。举例而言，Newman[103]提出了计算网络同配系数的方法，用以判断节点度相关性；Kitsak等人[104]提出了网络核数指标，用以度量节点在网络中的重要性。在社团结构方面，Girvan和Newman[105]研究了社会和生物网络的结构，提出了一种社团结构划分算法；Clauset等人[106]提出了一种最大化网络模块度的算法，能计算大规模网络的社团结构。在层次结构方面，Dunbar[107]发现人类只能维持有限规模的社交圈，与大约150人维持亲密关系；Zhou等人[108]发现个人社交关系以分层模式来组织，包含社交规模增加、强度递减的包容圈。在空间网络方面，Kleinberg[109]提出了一种空间网络模型，通过在方格网络上添加长边来构建；Barthélemy[110]综述了空间网络的研究进展，指出很多真实网络都有空间嵌入结构；Lambiotte等人[111]分析了手机通讯网络的地理分散情况，发现社会网络存在空间标度律；Hu等人[112]给出了空间标度律的一种解释，认为空间网络结构与最优信息收集密切相关。

中观层面的研究主要关注社会经济二部分网络的结构特征，以及金融网络的结构稳定性和风险传播。在网络结构方面，Hidalgo等人[24]基于国际贸易数据构建“国家-产品”二部分网络，利用一组线性迭代方程刻画网络结构，提出了经济复杂性ECI指标来预测国家发展潜力；Caldarelli等人[62]利用非线性迭代方程刻画二部分网络结构，提出了经济复杂性Fitness指标；Tacchella等人[113]在此基础上采用非线性迭代同时计算产品复杂性和国家竞争力；Cristelli等人[22]分析了经济复杂性的异质性动力学，发现Fitness指标与收入水平之间有两种关系，即存在高可预测性和低可预测性的两个区域（如图1–3所示）；Hausmann和Hidalgo[114]剖析了“国家-产品”网络结构，发现了国家产品多样性与普遍性负相关的网络结构依据；Bustos等人[115]发现“国家-产品”二部分网络有显著的嵌套结构，基于此能预测

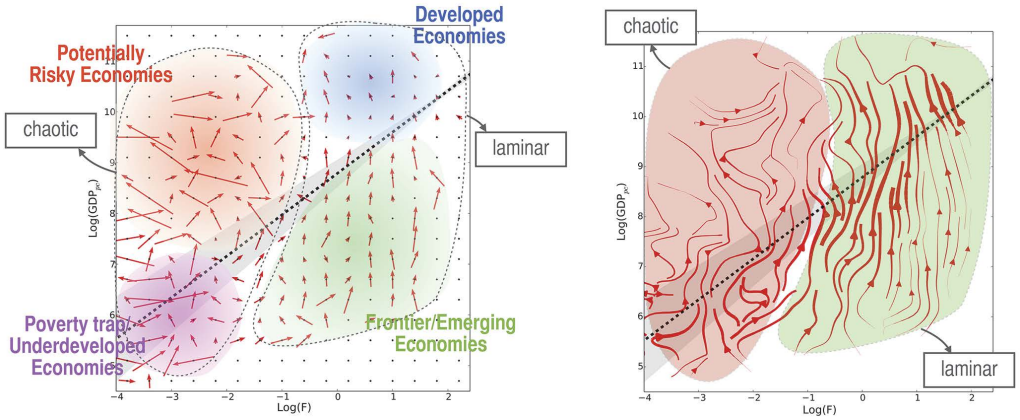

图 1–3 经济复杂性Fitness指标与经济发展的异质动力学相图[22]

产业系统的演化；Stojkoski等人[116]分析了服务业对国家生产结构的影响，发现服务比产品的经济复杂性更高。在金融网络稳定性和风险传播方面，Schweitzer等人[117]分析了金融机构之间通过借贷形成的国际金融网络，发现金融机构在网络中存在很强的相依关系，由此带来的系统风险使金融网络容易受到不稳定因素的影响；Arinaminpathy等人[118]分析了金融系统的稳定性，发现位于金融网络中心的机构反而无法倒闭；Haldane和May[119]提出了一种金融风险传播模型，从理论上分析了金融网络的复杂性和稳定性之间的相互作用关系。

利用大规模社会经济数据还能分析城市空间的功能结构和社会文化结构。在城市功能结构方面，Liu等人[120]基于卫星图像数据分析了城市的时空结构，发现在城市化过程中城市变的更加分散化、多样化和复杂化；Chi等人[121]分析了中国区域层面的手机通讯数据，发现通讯网络有两级分层组织结构，在商业中心和政府驻地都有中心性很强的基站；Yuan等人[122]提出了一种基于分割和主题建模的方法，能基于兴趣点和出租车轨迹数据发现城市的主要功能区；Frias-Martinez等人[123]提出了一种非监督学习算法，根据推文活跃模式对区域聚类，从而推断城市的土地使用类型；Zhi等人[124]提出了一种低秩分解模型，识别行为活动的时空特征，根据签到数据估计城市功能区。在社会文化结构方面，Shelton等人[125]分析了带有地理位置标记的推文数据，发现城市邻里有社交隔离现象；Yip等人[126]基于手机数据分析了香港的人群移动模式，发现富有和贫穷的人群倾向于在各自邻里内移动；Yang等人[127]分析了城市内签到数据，发现本地人访问地点分布广泛，外地人集中到访几个固定地点；Hu等人[128]基于微博数据构建了中国宗教网络，发现宗教存在明显的隔离现象，慈善活动在促进跨宗教沟通方面扮演重要角色。

宏观层面的研究主要关注产品空间、产业空间、技术空间等网络建模和结构分析。在产品空间方面，Hidalgo等人[25]利用共同出口概率估计产品接近性，将其以网络形式表示为“产品空间”（如图1–4所示）。发现产品空间有“核心-边缘”

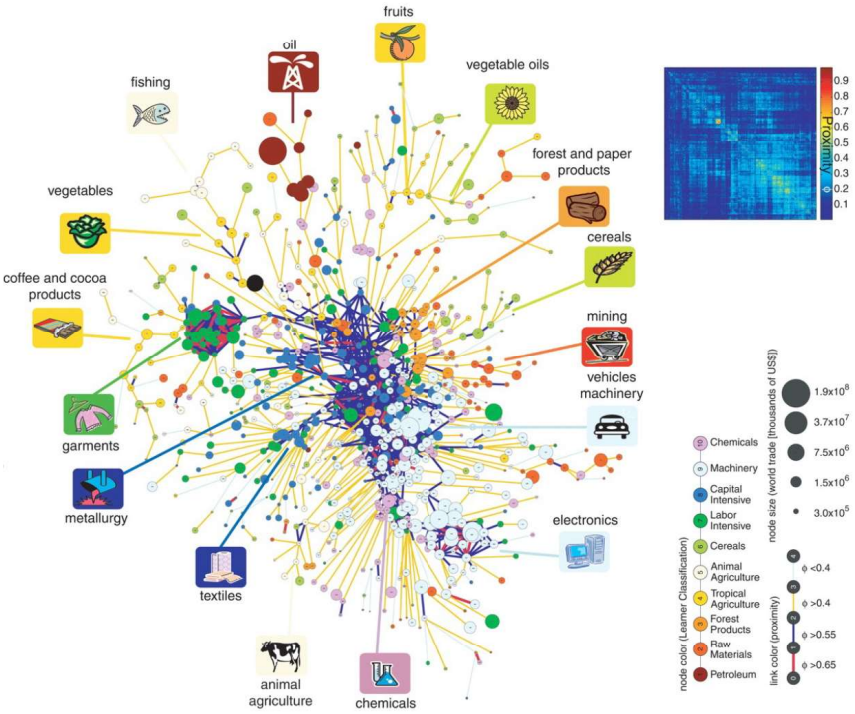

图 1–4 基于国际贸易数据构建的产品空间[25]

结构，富有国家能出口位于中心的复杂产品，而贫穷国家只能出口位于边缘的简单产品。国家容易出口与当前产品接近性高的产品，当前产品结构决定国家未来发展潜力。贺灿飞等人[129]基于贸易数据构建了中国产品空间，发现中国不同地区的产品结构转型方向不同；伍业君等人[130]发现利用产品空间能解释中国出口增长现象，还能识别产品升级的断档风险。在产业空间方面，Guo和He[131]基于中国工业企业数据计算产业接近性，构建了产业空间，发现其“核心-边缘”结构从1999年到2007年发生了重大结构变化，很多副核心融合为一个重要核心。在技术空间方面，Guevara等人[132]基于学术论文数据计算学科领域接近性，构建了学科空间，发现其能预测研究机构在未来进入哪些学科领域；Acemoglu等人[133]基于美国专利数据计算技术接近性，构建了创新空间，发现其能预测未来的创新方向；Alabdulkareem等人[134]基于职场技能数据计算技能接近性，构建了技能空间，发现其极化为高低认知能力的两个社团，这种结构限制了劳动者的职业转换。

在刻画社会经济结构方面，目前的研究大多使用社会网络和国际贸易数据，不容易对区域层面的经济结构进行建模和分析。缺乏对信息服务等没有产品输出的产业的考虑，给刻画区域经济复杂性带来困难。另外，在定量刻画产品空间“核心-边缘”结构的基础上，值得进一步分析空间网络结构对经济发展的影响。

1.2.3 社会经济发展规律研究

深刻地揭示社会经济发展规律，有助于更好地感知社会经济态势。传统研究

受数据和分析工具的局限，大多仅能以定性或半定量的方式解释社会经济现象，缺乏对其背后运行机制的解读。借助交叉学科工具分析大规模真实数据，有助于揭示社会经济系统的运行机制，发现社会经济的统一发展规律，从而制定更科学的经济发展策略。这里所说的发展规律，从微观上讲，包括人类行为的时间和空间规律；从中观上讲，包括城市的标度律和景观布局规律；从宏观上讲，包括发展的学习规律和策略。下面介绍社会经济发展规律方面的研究进展。

微观层面的研究主要关注人类动力学规律[36, 135]，包括时间活跃规律和空间移动规律，以及这些规律在异常情况下的变化。在活跃规律方面，主要分析电子邮件通讯、网页浏览、在线评分等数据，发现行为的时间间隔分布有胖尾性质[136]。例如，Barabási[43]分析了用户回复电子邮件的时间间隔，发现人类行为有长期静默和短期高频爆发等特点；Zhou等人[137]分析了在线电影点播记录数据，发现点播时间间隔分布有幂律尾部；Yang等人[138]分析了两个在线评分数据集，发现用户评分行为有锚定效应。在移动规律方面，主要分析移动距离分布、空间多样性、回转半径等。例如，González等人[139]分析了手机移动轨迹数据，发现个体移动行为有很强的时间和空间规律性；Song等人[44]基于手机数据刻画了个体移动轨迹的真实熵，发现人类移动行为有93%的可预测性；Yan等人[140]分析了瑞士志愿者旅行日记数据，发现群体移动距离近似服从指数截断的幂律分布。在突发情况下，人类的行为规律会发生变化，基于此能判断事件类型和发生地点[141, 142]。例如，Gao等人[143]分析了手机用户的异常行为模式，能区分应急事件和非应急事件（如图1–5所示）；Lu等人[45]基于手机数据分析了海地地震难民的移动轨迹，发现这时人类移动行为仍然有很高的可预测性；Sakaki等人[144]训练机器学习模型对推文进行分类，能很好地推断地震发生的地点；Kryvasheyeu等人[29]分析了飓风期间的推文数据，发现飓风相关推文活跃度能预测灾害强度和带来的经济损失。

中观层面的研究主要关注城市标度律和设施布局。在城市标度律方面，主要包括：人口密度与人类需求指标（如房屋和电力消耗）之间的线性标度律、人口密度与通货指标（如信息和财富）之间的超线性标度律、人口密度与基础设施指标（如道路面积和电缆长度）之间的亚线性标度律等[145]。例如，Louf等人[146]发现美国碳排放总量与人口总量呈现亚线性关系；Alves等人[147]发现巴西城市自杀人数与人口总量呈现超线性关系；Pan等人[148]发现社会连接密度与城市人口密度呈现超线性关系；Louail等人[149]发现西班牙城市活跃中心数与人口总量呈现亚线性关系。在解释标度律方面，Bettencourt[150]提出了一种不依赖基础设施建模的分析框架，能估计标度律的指数；Li等人[151]提出了一个统一模型，能重现超线性和亚线性标度律。在城市设施布局方面，Um等人[152]发现城市商业设

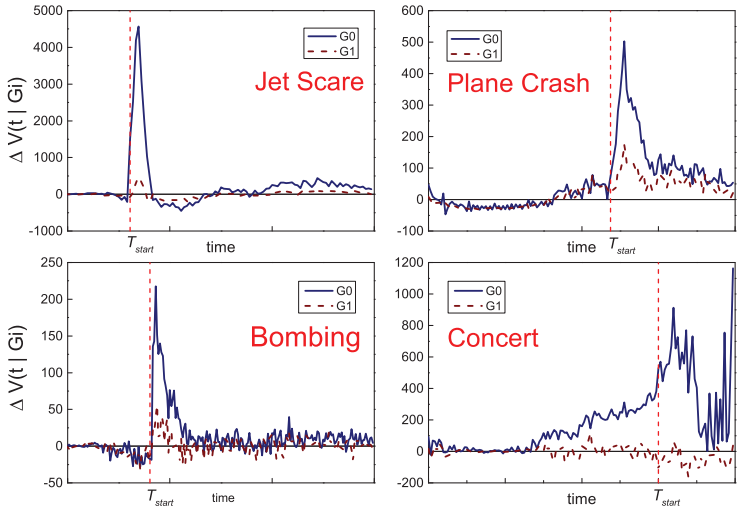

图 1–5 应急和非应急情况下手机通讯量的变化[143]

施密度与人口密度之间存在指数为1的幂律关系，而对于公共服务设施的幂指数为2/3；Hidalgo等人[153]基于Google地图数据计算设施共同出现概率，构建了“设施空间”，基于此提出了一种推荐算法，能为邻里区域推荐缺失设施来改善城市布局；Naik等人[100]基于街景图片数据分析了城市感观变化，发现驱使城市安全性变好的三大因素：好的初始环境、距离城市中心近和受大学教育人口多。

宏观层面的研究主要关注经济发展的学习过程和路径依赖效应，包括相似技术和近邻区域学习途径存在的“相关性准则”[154]，即区域容易发展与已有经济活动密切相关的经济活动。在相似技术方面，主要关注从区域内的相似经济活动中学习。例如，Hidalgo等人[25]分析了国家的产品贸易数据[155]，发现在新产品上取得比较优势的概率受相关产品数量的影响；Boschma等人[156]分析了意大利贸易数据，发现产业多样化发展受益于区域自身的相关产业；Neffke等人[67]分析了瑞典70个区域的企业数据，发现新产业的发展概率与已有相关产业数量正相关；He等人[157]分析了中国制造企业数据，发现区域发展的新产业与已有产业在技术上很接近；Jara-Figueroa等人[158]分析了巴西劳动力市场数据，发现招聘有地域和行业特定知识的员工能提高新企业的存活率。在近邻区域方面，主要关注从周围区域的相同经济活动中学习。例如，Bahar等人[159]发现有很多近邻已经出口某种产品的国家，其出口该产品的概率显著增加；Holmes[160]分析美国沃尔玛商店的地理扩张，发现新店的位置倾向于接近已有很多店的区域；Boschma等人[161]分析了美国新生企业数据，发现区域更容易维持和发展已在周围出现的产业。

宏观层面还特别关注经济发展问题，包括经济结构调整和演化规律，以及经济发展的最佳时机和最优策略。林毅夫[58, 59]提出的“新结构经济学”认为，经济发展的本质是产业和技术不断创新、经济结构不断调整的过程。经济发展阶段是

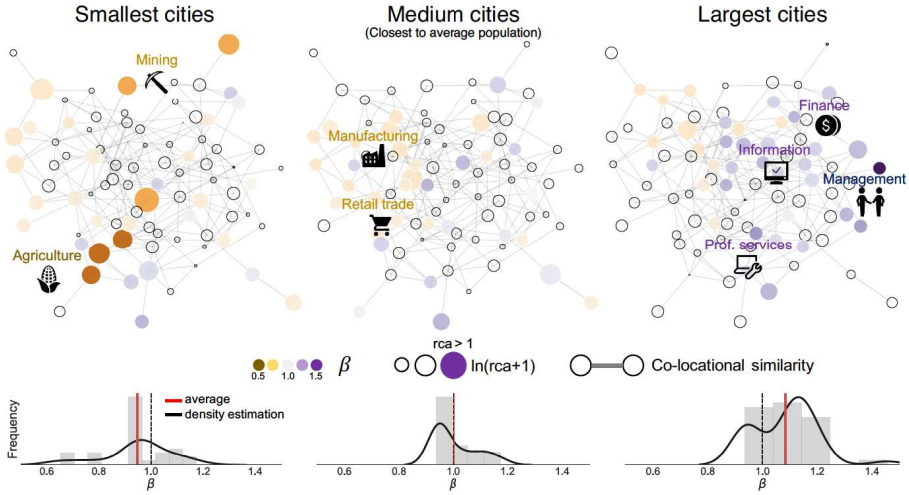

图 1–6 美国不同体量城市的产业空间的结构特征[162]

从低收入农业经济到高收入工业化经济的连续谱，处在不同阶段的国家有不同的素禀赋结构，最优产业结构由禀赋结构决定。当经济发展导致禀赋结构发生变化时，最优产业结构也随之变化，所以产业升级要与比较优势保持一致[163]。在经济结构调整和演化规律方面，Hong等人[162]基于美国就业数据构建了城市产业空间，发现城市的体量决定了其产业结构特征（如图1–6所示）。大城市的产业结构不同于小城市，也不是自己过去产业结构的简单放大。在经济发展策略方面，主要利用贸易数据分析产品发展问题[155]。例如，Alshamsi等人[12]发现出口产品是一个路径依赖过程，需要动态调整目标产品，在合适时机尝试发展位于产品空间中心的复杂产品，能实现最快的经济发展速度；Pinheiro等人[164]发现国家在经济发展的中期阶段倾向于发展不相关的产品，这能为国家在未来带来更快的经济增长；Zhu等人[165]分析了中国产品空间的结构和演化，发现利用外部投入和内部创新能跨越产品空间发展，中国在经济发展过程中一定程度上突破了路径依赖[166]。

在社会经济发展规律研究方面，目前针对中国经济增长和发展路径的讨论还不足，值得进一步基于实证数据分析经济发展的学习路径。另外，在刻画经济结构的基础上，有希望利用信息传播模型探究区域经济和产业发展的最佳策略。

1.3 本文主要创新点

计算社会经济学是一个新兴的交叉学科研究方向，利用先进工具分析大规模真实数据，旨在精准和及时地感知社会经济状态，揭示和理解社会经济运行规律。以复杂网络刻画社会经济系统中的相互作用，分析社会经济空间结构与动力学规律，为揭示社会经济现象提供更深刻洞见。本文从微观、中观和宏观层面展开对社会经济系统空间结构的研究，进一步利用空间网络和动力学模型研究经济结构演化和最优经济发展策略。本文的主要贡献和创新点如下：

一、基于非干预行为数据研究了微观层面的社会经济预测性管理。首次提出了谨严性指数刻画个体行为规律性，发现谨严性与学生成绩显著相关，利用谨严性特征能显著地提升排序学习算法对学生成绩的预测效果。发现互动网络和社会网络的中心性特征能预测员工升离职的可能性，互动网络比社会网络的特征有更强预测能力，预测离职比预测升职容易。分析大规模数据揭示了一些社会经济现象，发现团队规模在8人以下有助于提高沟通和绩效；中国社交圈规模也维持在邓巴数150人左右；求职中存在男女不平等现象和身高溢价效应。

二、基于在线评分数据研究了中观层面的社会经济状态排序问题。提出了一种基于群组聚类的用户信誉排序算法，根据评分聚类形成的群组规模计算用户信誉，不依赖产品有唯一质量的假设，对用户信誉的排序更准确。进一步，提出了基于迭代过程的群组聚类信誉排序算法，群组规模由用户数量和用户信誉共同决定，显著地提高了对作弊评分用户排序的鲁棒性。提出了一种节点相似性CosRA指标，基于此提出的CosRA推荐算法效果更优。进一步，提出了基于信任关系的CosRA+T推荐算法，发现过度依赖信任关系有损推荐效果提升。

三、基于大规模真实数据研究了宏观层面的社会经济结构建模。首次刻画了中国区域经济复杂性，发现经济复杂性ECI和Fitness指标对经济发展的预测能力相当，经济复杂性与收入不平等性负相关。构建了巴西和中国区域产业空间，发现两者都有显著的“核心-边缘”结构，复杂程度高和低的产业分别占据核心和边缘位置。另外，中国产业空间还有“哑铃型”结构，在时间演化上有地区竞争。发现利用信息和人才流动网络的结构特征能推断区域经济水平，人才流动对经济发展的预测能力更强，综合两个网络的特征能最多解释大约84%的GDP变化。

四、基于空间网络研究了经济结构演化规律和最优经济发展策略。揭示了网络的空间结构对靴襻渗流相变类型的影响，发现长边分布的幂指数-1为临界值：当其大于等于-1时，出现一级和二级相变点不变的双相变；当其小于-1时，仅出现相变点随幂指数减小而增大的二级相变。提出了经济发展过程中的相似技术学习途径和近邻区域学习途径，发现两条学习途径都能提高发展新产业的概率，但两者存在替代效应。发现引入高铁能显著地提高区域产业相似性和生产率；两条学习途径都存在最优发展策略，即随机选择产业激活和随机选择区域连接。

1.4 本文研究内容与章节安排

本文将在计算社会经济学框架下系统性地研究社会经济系统的空间结构与动力学，为揭示社会经济现象提供深刻洞见。在介绍计算社会经济学相关知识的基础上，本文将分别从微观、中观和宏观层面研究社会经济系统的空间结构建模和

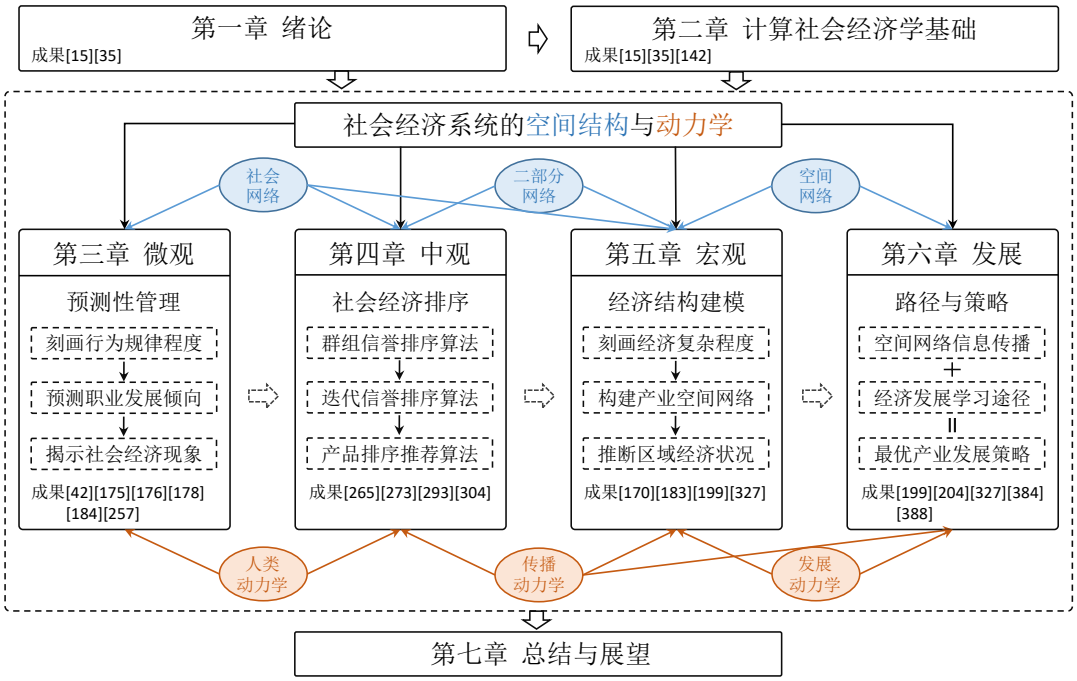

图 1–7 本文各章节内容之间的关系图以及相应研究成果的文献编号

状态推断，进而利用空间网络和动力学模型研究经济结构演化和最优发展策略。概括而言，在微观层面研究社会经济的预测性管理，包括社会行为规律性预测学习成绩、社会网络结构特征预测职业发展和大规模数据揭示社会经济现象；在中观层面研究社会经济系统的排序问题，包括提出一种基于群组聚类的在线信誉排序算法、利用迭代过程改进信誉排序算法和利用复杂网络结构进行排序推荐；在宏观层面研究社会经济空间结构的建模和分析，包括计算区域经济复杂性、建模分析产业空间结构和利用网络结构推断经济水平；进一步，研究经济结构演化规律和发展策略，包括网络的空间结构对信息传播的影响、经济发展中的协同学习效应和基于空间网络的最优经济发展策略。

图1–7展示了本文各章节内容之间的逻辑关系图以及相应研究成果的文献编号。其中，蓝色椭圆体现章节所共同基于的网络空间结构，橙色椭圆体现章节所共同关注的网络动力学。在网络空间结构方面，微观、中观和宏观层面都基于社会网络，中观和宏观层面都基于二部分网络，宏观层面和经济发展与结构演化方面都基于空间网络。在网络动力学方面，微观和中观层面都关注人类动力学，中观、宏观层面和经济发展与结构演化方面都关注传播动力学，宏观层面和经济发展与结构演化方面都关注发展动力学。本文各章的研究内容安排如下：

第一章介绍本文的研究背景与意义、国内外研究现状和本文的创新点与章节安排。首先，介绍计算社会经济学的研究背景和本文研究的理论价值与意义。然后，介绍计算社会经济学的国内外研究现状，包括社会经济态势感知、结构刻画

和发展规律的研究进展。最后，给出本文的主要创新点和各章节内容安排。

第二章将介绍计算社会经济学的基础知识，包括社会经济相关数据和社会经济分析方法，作为本文研究的数据和方法基础。首先，概述来源于社会经济系统的大规模真实数据，依次介绍政府部门统计数据、在线社会媒体数据、非干预行为数据和其他相关数据的收集方式、数据概貌和简单应用案例。然后，概述计算社会经济学常用的交叉学科分析工具和方法，依次介绍传统回归分析方法、复杂网络分析方法和统计机器学习方法，以及各方法中常用的代表模型。

第三章将从微观层面研究社会经济的预测性管理。第3.1节将利用三千万条校园刷卡记录刻画学生行为的规律性，基于时间序列真实熵提出谨严性指数，区分不同规律程度的学生。然后，关联分析谨严性指标和学生成绩，探究行为规律性与学习成绩的关系。最后，结合谨严性和努力程度这两种行为特征，利用排序学习算法预测学生成绩，分析谨严性指标对学生成绩的预测能力。第3.2节将利用社会化平台的非干预行为数据，分别构建互动网络和社会网络。然后，分析两个网络的结构特征，以及员工互动行为模式与绩效的关系。最后，利用网络的结构特征关联分析和预测员工升职和离职的可能性。第3.3节将量化分析在线平台的大规模数据，揭示一些社会经济现象。首先，利用社会网络数据分析团队规模对沟通强度和绩效的影响，探究最佳团队规模。然后，利用手机通讯数据在中国社会和文化背景下分析社交圈规模，验证邓巴数理论。最后，利用匿名求职者简历数据分析职业发展的影响因素，揭示求职中性别和身高的不平等性。

第四章将从中观层面研究社会经济系统的排序问题。第4.1节将针对在线评分系统的用户信誉评价问题，提出一种基于群组聚类的在线信誉排序算法。不再依赖于产品具有唯一质量分数的假设，新算法将用户按照评分相似性进行聚类，根据用户所归属群组的规模计算用户信誉。在三个真实评分数据集上测试，比较新算法和传统算法在应对作弊评分用户攻击方面的表现。第4.2节将引入迭代寻优求解过程，改进得到一种基于迭代过程的群组聚类信誉排序算法。在计算用户的群组规模时，综合考虑用户的数量和信誉水平。在真实评分数据上测试算法性能，分析算法特点和对作弊评分用户的排序效果。第4.3节将针对在线系统的排序问题，提出基于网络结构的推荐算法。首先，提出一种新的节点相似性CosRA指标，基于此提出CosRA推荐算法。然后，将用户信任关系引入CosRA算法框架，改进得到CosRA+T推荐算法，分析信任关系对推荐效果的影响。

第五章将从宏观层面研究社会经济系统的结构建模。第5.1节将利用企业注册信息数据刻画中国区域经济复杂性，分析其对社会经济指标的预测能力。首先根据企业注册地和产业分类构建“省份-产业”二部分网络，然后利用迭代方程基于

网络计算经济复杂性ECI和Fitness指标，最后对比分析两种指标对社会经济指标的预测能力。第5.2节将利用劳动力市场数据和企业注册信息数据分别构建巴西“产业-职业”和中国“省份-产业”二部分网络，基于此利用余弦相似性计算产业接近性构建产业空间。分析产业空间的结构特征和演化规律，以及地区之间的产业竞争。第5.3节将利用信息流动和人才流动推断区域经济发展水平。首先基于社交媒体数据和简历数据分别构建信息流动和人才流动网络，然后关联分析区域经济发展水平与两个网络的结构特征，最后利用两个网络的结构特征预测区域经济发展水平，分析基于两个网络所构造的复合指标的最大预测能力。

第六章将研究经济结构的演化规律和最优产业发展策略。第6.1节将以理论研究网络的空间结构对信息传播的影响。以在方格网络中添加长边的方式构建Kleinberg空间网络模型，分析长边分布对靴襻渗流相变类型的影响。使用数值模拟方法判定相变类型和确定相变点的数值，分析长边分布幂指数存在的临界值。第6.2节将以实证研究经济发展中的两条学习途径和路径依赖效应。首先，分析区域内活跃的相似产业密度对发展新产业的影响，提出产业空间网络上的相似技术学习途径。然后，分析区域周围活跃的邻居区域密度对发展新产业的影响，提出地理近邻网络上的近邻区域学习途径。最后，分析两条学习途径之间的相互作用。第6.3节将以理论结合实证研究经济最优发展策略。首先分析高铁引入对地理近邻学习的影响，然后利用靴襻渗流模型分析两条学习途径的最优发展策略，最后利用国际贸易数据分析三种知识扩散策略对贸易的促进作用。

第七章将总结本文的主要研究内容和展望未来值得研究的方向。首先，总结本文各章节所研究的主要内容、所使用的数据和研究方法、所得到的主要结论，以及研究结果所具有的现实意义。然后，分析计算社会经济学所面临的新挑战，给出未来值得关注的研究方向。最后，针对本文各章节的研究内容，分析现有工作的不足之处，讨论有待解决的问题，提出有希望改进的方向。

第二章 计算社会经济学基础知识

本章将简介计算社会经济学的基础知识，包括主要的研究内容、基于的大规模数据和使用的交叉学科分析方法。在第2.1节中，介绍计算社会经济的研究内容，包括感知社会经济态势和理解社会经济规律。在第2.2节中，介绍来源于社会经济系统的大规模真实数据，包括政府统计数据、社会媒体数据、非干预行为数据和其他相关数据。在第2.3节中，介绍计算社会经济学研究使用的主要分析方法，包括传统回归分析、复杂网络分析和统计机器学习。

2.1 计算社会经济研究内容

近年来，社会经济研究出现了数据和方法的转变。一方面，大规模社会经济数据可用性提高，包括卫星遥感、手机通讯、社交媒体等，有低获取成本、实时更新和高时空分辨率等优势。另一方面，分析工具和计算方法进步，包括机器学习、网络分析、文本挖掘等，提升了感知和预测社会经济状态的能力[15]。新数据和新方法的应用，催生了计算社会经济学这一新研究分支。依靠数据驱动社会经济洞察，计算社会经济学以量化手段分析大规模真实数据，揭示社会经济现象。

图 2–1 计算社会经济学主要使用的数据和研究的内容

图2–1展示了计算社会经济学主要使用的数据和研究的内容。在研究尺度上，计算社会经济学主要涵盖三个层面（国家结构、区域状态、个体属性）和两个应用（应急管理、发展路径）[35]。在研究内容上，计算社会经济学主要关注两部分：一是对于社会经济状态的感知和推断，包括个体、群体、区域、国家等层面的状

态和发展趋势；二是对社会经济规律的洞察和理解，包括人类移动、城市布局、产业演化和国家发展等方面的规律。感知社会经济状态和理解社会经济规律，这两部分研究内容相辅相成，彼此支持和促进。

2.1.1 感知社会经济状态

精准和及时地感知社会经济状态，对于社会经济的决策有重要意义，小到影响个人消费选择，大到关乎国家战略决策[15]。传统方法大多利用普查得到相关数据，然后汇总和计算宏观经济指标，例如GDP，以此大概估计当前经济状况。尽管普查技术逐渐提高，传统方法仍然耗费大量人力、物力和时间等，难以支撑及时的社会经济决策。随着信息技术的进步，先进设备已经能获取社会经济相关的大规模数据，借助现代技术加以分析，能近乎实时地感知、推断和预测社会经济状态。下面，将从微观、中观和宏观层面简介社会经济状态感知的研究内容。

微观层面的研究，关注个体行为特征与状态的关系，利用行为特征推断社会经济状态。这里所说的行为特征，包括设备使用模式、动力学模式、在线平台使用特点和社会网络结构特征等。具体而言，设备使用模式包括：手机通话频率、总量、多样性、充值频率和额度等；动力学模式包括：人类行为的时间和空间行为特征，如移动距离、多样性和行为谨严性等；在线平台使用特点包括：网站搜索关键词、社交网站推文关键词、签到位置和个人信息展示等；社交网络结构特征包括：通讯网络结构多样性、节点重要性和社团结构等。微观层面的社会经济状态包括：财富和收入水平、升职和离职率、绩效表现、人口属性、人格特征、情感和健康状况等。例如，基于手机数据构建社会网络，利用个体在网络中的位置预测其财富[77]；基于社交媒体推文关键词，分析用户的人格特质[89]。

中观层面的研究，关注群体状态推断和城市布局与状态的关系。群体状态包括：在线用户的信誉水平和对商品的偏好等。城市景观布局，一方面包括反映城市面貌和感观的图像数据，如Google街景图片、社交媒体带有位置标记的图片、卫星拍摄夜间光亮数据等，另一方面包括结构化的城市空间数据，如开放地图记录的兴趣点数据、工商记录的企业注册数据、城市规划与统计数据等。利用这些大规模数据，能分析城市的设施布局、功能区域、感观环境、种族和社会经济隔离等。例如，基于签到数据分析用户的时空活动特点，推断城市的主要功能区域[124]；基于用户对Google街景图片的打分训练自动化的评分模型，在安全、绿色、舒适、阶级等不同维度评价城市的感观环境[99]。

宏观层面的研究，关注国家和地区的社会经济结构建模和分析，提出有预测能力的新指标。基于真实数据构建的社会经济网络，包括国际贸易网络、手机和

社交通讯网络、人才和信息流动网络等。一方面，利用网络结构特征关联分析区域的社会经济指标，能实现对指标的预测。例如，利用通讯网络多样性预测区域的复合衰败指数[66]；利用在线用户活跃程度预测区域经济水平和分析区域产业结构[97]。另一方面，分析网络结构特点能设计新指标，更好地预测社会经济状态。例如，基于国际贸易网络计算经济复杂性指标，预测国家未来经济发展潜力[24, 113]。此外，微观和中观层面的数据和方法，也能用于宏观层面的分析。例如，利用手机数据推断个体财富状况，聚合数据重建出国家财富分布地图[46]；结合夜间光亮和白天卫星图像数据，利用机器学习算法预测区域财富分布情况[30]。

2.1.2 理解社会经济规律

传统社会经济研究受数据和方法的限制，大多以定性或半定量的方式研究社会经济现象。对现象背后的社会经济发展规律和社会经济系统的运行机制，都缺乏基于数据分析的定量化理解。近年来，大规模真实社会经济数据逐渐积累，来自统计力学、计算机和复杂网络等学科的分析工具逐渐丰富。新数据和新方法的应用，帮助更好地揭示社会经济运行机制、理解社会经济发展规律，有助于设计出遵循社会经济发展规律的最优经济发展策略，更快地提高社会经济水平[15]。下面，将从微观、中观和宏观层面简介社会经济规律分析的研究内容。

微观层面的研究，关注个体在日常和应急状况下的行为规律，包括时间规律和空间规律。具体而言，行为的时间规律包括：在线行为（如电子邮件通讯、金融交易、网页浏览等）的阵发性[43]、时间间隔分布的胖尾性质[136]、在线评分的锚定效应[138]等。行为的空间规律包括：移动的距离分布、空间多样性、回转半径、节律性等[44, 139]。在应急事件和自然灾害发生时，这些时空行为规律会发生变化，包括社交媒体活跃程度、手机通信频率和时长、移动距离和空间多样性等。例如，分析大规模手机通讯数据，根据通讯总量的时空变化判断应急事件类型[40]；追踪和预测地震时人类迁徙的规模和移动轨迹[45]。

中观层面的研究，关注社交网络结构、城市标度律和景观布局演化规律。社会网络的结构包括：网络度分布异质性、聚类系数、同配性、社团结构、节点重要性、多层重叠性、“邓巴圈”层次结构、空间网络结构等[33]。城市的社会经济标度律包括：人口密度与人类需求指标（如房屋、就业、电力消耗）之间的线性标度律，人口密度与通货指标（如信息、创新、财富）之间的超线性标度律，以及人口密度与基础设施指标（如道路面积、加油站数量、电缆长度）之间的亚线性标度律[145]。城市的景观布局演化规律包括：感观变化规律、商业设施共同出现模式、城市功能区变化和区域拓展等。例如，利用机器学习算法对比分析同一地

点、不同时间的街景图片，发现初始环境好、受大学教育人口多和距离城市中心近，这三个因素能驱使城市的安全性感观变好[100]。

宏观层面的研究，关注经济和社会的结构演化、经济发展的路径依赖和最优发展策略。结构演化规律包括：经济接近性演化、产品和产业空间演化、复杂性演化、语言和姓氏演化等。例如，基于国际贸易数据分析经济复杂性演化和经济发展预测能力[63]；基于中餐食谱数据分析地理接近性对菜系相似性的影响[167]。经济发展的路径依赖包括：相似技术依赖和地理近邻依赖。例如，分析国际贸易和企业注册数据，发现新经济活动的发展依赖于已有经济活动的密度[154]。最优经济发展策略包括：基于产业空间网络优先激活产业的策略、基于地理近邻网络优先连接区域的策略、发展过程中动态调整发展目标的策略等。例如，分析国际贸易和科学论文数据，发现存在最佳时机去发展复杂程度高的产品和研究方向[12]。

2.2 社会经济相关数据

大规模社会经济数据是开展计算社会经济学研究的基础。分析真实数据已经在相关领域成为一种流行趋势，哪怕是相对保守的传统经济学领域，最近五十年来在顶级经济学杂志上发表的实证论文数量占比也超过了70%[168]。在几十年前，还不容易获得第一手科研数据，对于统计年鉴这样重要数据载体，还只能到图书馆借阅。随着信息技术的发展，数据的获取途径逐渐丰富，数据类型、规模和质量都得到大幅提高[15]。计算社会经济学所使用的数据有三个特点：第一，必须是真实数据，用理论模型解释真实数据，用真实数据评价预测效果；第二，尽可能是大规模数据，尽量获取全体尺度数据（如图2–2所示），降低采样偏差带来的风险；第三，一般是及时获取和持续更新的数据，以指导社会经济政策。

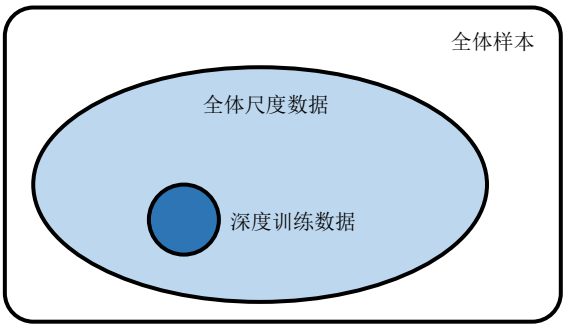

图 2–2 全体数据与全体样本的关系示意图[35]

从来源上讲，社会经济数据包括政府部门收集的数据和新兴私营部门积累的数据[9]。政府部门数据主要分为两类：一类是在政府管理系统中记录的个人和企业相关信息，是有时间序列和截面的高质量面板数据；另一类是科学研究和工程

项目所收集的数据，如人类基因组数据和卫星遥感数据，一般依赖于特殊设备获取。这些数据总量很大，覆盖时间长，在刻画人口变化、收入和支出、经济流动性等方面有独到的优势。私营部门数据是企业在经营活动中收集的数据，如购物网站存储的用户数据。这些数据有更丰富的类型，数据灵活度和分辨率高，方便追踪个体社会经济活动。下面，将介绍研究中常用的政府部门统计数据、在线社会媒体数据、非干预行为数据和其他类型相关数据。

2.2.1 政府部门统计数据

政府部门实行全国经济普查，有助于全面掌握国家的产业发展规模、结构和效益等基本信息。截至2018年，我国已经进行了4次综合性的全国经济普查。国家统计局（http://www.stats.gov.cn）负责编印《中国统计年鉴》，发布最全面和最具权威性的社会经济相关统计数据。统计年鉴实际收录上一年的统计数据，例如《2018年中国统计年鉴》收录2017年全国省份层面的统计数据，内容包括：国民经济核算、人口、就业人员和职工工资、价格指数、城市概况、环境保护、农业、工业和金融业等行业的统计数据。目前，国家统计局开放了国家数据（National Data）网站（http://data.stats.gov.cn），提供《中国统计年鉴》中所有数据的查询和下载，数据可以回溯到二十年前。图2–3展示了国家数据网站页面截图。

图 2–3 国家统计局国家数据网站页面截图[①]

在城市层面，国家统计局城市社会经济调查司编印《中国城市统计年鉴》，收录全国656个建制城市（含地级以上城市和县级市）的城市建设和社会经济统计数据，数据目录与国家层面类似。中国统计出版社还编纂和发布其他分类普查数据，包括《中国农村统计年鉴》、《中国价格统计年鉴》、《中国人才资源统计报告》、《中国劳动统计年鉴》、《中国社会统计年鉴》等。另外，还有一些科研机构

① http://data.stats.gov.cn

主导的调查数据库，例如：中国综合社会调查（CGSS）数据库（中国人民大学），是经济与社会科研数据采集平台；中国家庭追踪调查（CFPS）数据库（北京大学），提供跟踪个体、家庭、社区三个层次的数据。

国际机构也提供很多社会经济宏观统计数据。世界银行（World Bank）发布国家发展指数（WDI, World Development Indicators）相关数据，包括经济数据（收入、增长、贸易、生产率等）、贫困和不平等数据（贫困、消费、收入分布等）、人口数据（人口变动、教育、性别等）、国际连接数据（借贷、贸易、旅游、移民等）等。联合国商品贸易统计数据库（UN Comtrade Database）发布双边国际贸易数据，包括进出口国、产品类别、贸易总额、贸易时间等信息。联合国统计司（UN Statistics Division）发布国家宏观统计数据，包括产业数据、经济数据、环境数据等。另外，国家和机构也提供自己的社会经济统计数据。例如，美国国家统计局（US Census）公布美国社区调查数据、劳工统计局数据、职业技能数据等，由Data USA网站（https://datausa.io）提供数据集成、可视化和开放获取。

2.2.2 在线社会媒体数据

在线社会媒体是人们分享观点和生活的平台，例如国内的微博、微信和QQ等，以及国外的Facebook、Twitter和Instagram等。图2–4展示了一些常见的在线社会媒体平台图标[169]。在线社会媒体数据，包括用户发布的个人简介、推文、照片、视频和地理位置等信息，以及用户之间的点赞、评论、转发、提及、关注等社交关系信息。基于这些数据，能刻画用户行为模式和构建在线社会网络，进而利用行为特征和网络结构特征推断用户和聚合层面的社会经济水平。

图 2–4 常见的在线社会媒体平台图标[169]

新浪微博是中国最大的在线社交媒体平台之一。截至2018年3月，微博的月活跃用户数已经超过4.11亿。用户在注册微博时提供很多信息，包括用户名、昵称、个人简介、教育信息、所在地等。用户可以发布推文，点赞、评论和转发其他用户的推文，通过关注和被关注形成在线社会网络。研究中使用的微博数

据集，包括：Liu等人[97]提供的涵盖大约2亿用户的微博数据，时间跨度从2009年到2012年，包括用户的注册日期、性别、位置信息等；Wang等人[170]提供的涵盖大约4.33亿用户的微博数据，时间截止到2017年初，包括用户的关注关系和位置信息等；Dong等人[171]提供的与2013年雅安地震相关的微博数据集，包括用户的推文、转发、关注关系等；Hu等人[128]提供的中国与宗教相关的微博数据，涵盖6875个宗教用户的个人简介、标签、关注关系等。

Twitter是国际上最常用的在线社交媒体平台之一，其功能与微博类似。截至2018年3月，Twitter的月活跃用户数达到3.36亿。使用Twitter时，用户可以发布长度不超过140个字符的推文，附带图片、地理位置、话题标签等信息。用户可以在Twitter上传头像、填写个人简介、关注用户等，也可以进行评论、转发、点赞和私信等互动。基于Twitter数据的研究工作有很多，例如利用推文数据监控疾病爆发和感知灾害等[142]。研究中使用的Twitter数据集，包括：Leetaru等人[172]提供的超过7000万用户发布的超过15亿条带有地理位置标记的推文数据；Larsen等人[92]提供的带有情感信息的大约27.3亿条推文数据，情感词分为6大类和25小类；Kim等人[173]提供的超过2.87亿条韩国推文数据，带有流感样疾病关键词；Toriumi等人[174]提供的大约3.6亿条在2011年东日本大地震前后发布的推文数据。

企业内部的社会化平台，在形式上类似于微博和Twitter，但仅限于企业内员工彼此沟通交流、开展文化建设和进行工作协作。企业社会化平台一般兼具社交生活和工作协同的功能，不仅记录员工彼此之间的社会互动（如推文的评论、转发、点赞等），还记录员工之间与工作任务相关的沟通交流（如工作任务分配、资料共享、业务汇报、评论回复等）[175]。将数据从社会化平台功能的层面分解开，能构建企业内部的员工沟通网络，包括与工作相关的员工“互动网络”和与生活相关的员工“社会网络”。Yuan等人[176]提供的一家中国企业内部使用的亦群社会化平台数据，包含104位员工在生活和工作中的交互行为记录。

2.2.3 非干预的行为数据

非干预行为数据是用户无意间遗留在社会经济系统，或通过手机、电子徽章、可穿戴设备等现代手段收集的数据。不像通过问卷和访谈等容易引起用户心理防御的方式收集的数据，非干预行为数据很大程度上能反映用户的真实情况。利用先进手段收集的非干预行为数据，一般具有很高的时间和空间分辨率，能以很低的成本获取长时间、大规模的数据，涵盖非常广泛的研究群体，例如数以亿计的手机用户、成百上千万的交通卡和校园卡用户。

手机已经在全球范围内得到普及，即使在非常贫穷的国家，也有相当一部分手机用户。手机几乎随时伴随着用户，忠实地记录着用户的通讯对象、时间、频

率、收发信息等数据。近年来流行的智能手机功能更为强大，其中的很多传感器能记录和提供更多高质量的数据，例如手机GPS坐标、移动轨迹、流量使用、消费记录等。分析这些高时空分辨的手机数据，有助于更好地理解用户行为特征、推断社会经济状态和理解区域经济发展。研究中常用的是手机呼叫记录（CDR, Call Detail Record）数据，例如发展数据（Data for Development）项目开放的四个手机数据集[177]，包括Orange公司的500万匿名用户的CDR数据、部分用户的行为轨迹和通讯网络数据等；Luo等人[77]提供的墨西哥全国1.07亿手机用户的通讯网络数据；Wang等人[178]提供的覆盖中国某城市的超过700万条CDR数据等。

生活中常用到的智能卡片，如公交卡、校园卡、信用卡等，也记录着用户的很多数据。举例而言，公交卡记录着用户的乘车起点、终点、时间、消费金额等信息；信用卡记录着用户的消费时间、地点、金额、类别、信用额度等信息；校园卡记录着学生在校园里的吃饭、购物、签到等信息。智能卡覆盖大量群体，数据有不错的时空分辨率，对理解个体和群体移动规律、推断社会经济状况有帮助。研究中使用的智能卡数据，包括：Dong等人[179]提供的两个国家的信用卡消费记录数据，涵盖欧洲国家10万用户的1000万条记录，以及拉美国家300万用户的6000万条记录；Hashemian等人[180]提供的西班牙银行卡转账记录数据，涵盖450万用户的1.78亿条记录；Cao等人[42]提供的近2万张校园卡的刷卡记录数据，涵盖宿舍洗澡、食堂吃饭、进出图书馆和教学楼打水等近3000万条记录。

可穿戴智能设备，如智能手表、手环、谷歌眼镜、徽章传感器等，独立或借助手机应用程序运行，以非干预形式记录用户的监测数据，包括运动步数、心率、血压、体重等信息。智能手机上安装的其他应用程序，能收集用户的很多数据，例如放音乐、读文章、玩游戏等活动的频率和时间信息等。除此之外，一些功能简单的徽章传感器，也能为科学研究收集用户的会话地点、频率和用户距离等数据。利用这些设备能获得用户极高时空分辨率的数据，帮助推断用户的行为特征、健康状况和社会经济状况。研究中使用的可穿戴设备数据，包括：Aral等人[181]提供的健身数字追踪设备记录的大约110万用户的运动数据，涵盖每天的运动距离、时长、速度和燃烧卡路里等；Olguin等人[86]提供的电子徽章传感器记录的用户交互数据，涵盖用户之间会话的时间比例等信息。

2.2.4 其他类型相关数据

一些从事社会经济管理的政府机构，以及从事专业数据库开发的私营部门，也提供很多经济、社会和金融方面的数据。这些数据结构化、规模大、质量高，一般来自有特定功能的社会经济系统，由专人负责维护和更新。例如，政务系统

中记录的企业注册、纳税、财务等信息；金融公司系统中记录的行业、股票、基金等数据；专业数据服务公司记录的卫星图像、地图、街景等数据。另外，互联网公司也记录着海量业务数据，例如用户注册信息、社会关系、在线检索等。下面将分别介绍信息系统数据、互联网平台数据和地理信息平台数据。

大规模信息系统结构化数据，来自政府政务管理系统和私营部门业务系统。在政府部门方面，工商局掌握所有企业信息数据，包括企业基本信息（如企业名称、类型和行业，以及注册地址、时间和资本等）、股东结构、治理结构和税务情况等。利用这些数据能分析区域产业结构、研判经济走势和预测失信企业[182]等。在私营部门方面，锐思金融研究数据库（http://www.resset.cn）包含股票、行业、指数等一些列近百个子库，提供上市公司和非上市公司的经济和金融数据。这些大规模、结构化的数据库，为计算社会经济学研究提供了数据支持。例如，Jara等人[158]使用了巴西年度社会信息报告（RAIS）数据，涵盖员工基本信息和企业注册信息，图2–5展示了RAIS数据可视化网站（http://legacy.dataviva.info）页面截图；Gao等人[183]使用了中国沪深A股2690家上市公司的注册信息和财务信息数据。

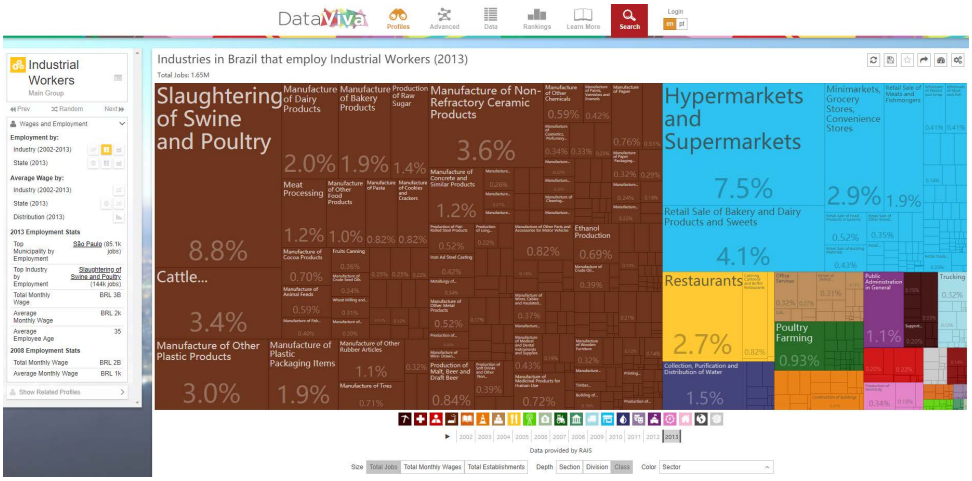

图 2–5 巴西RAIS数据可视化网站（DataViva）页面截图①

互联网平台记录着海量的社会经济相关数据。招聘网站记录求职者个人信息、教育和职业背景、技能特点、求职意向等。例如，Yang等人[184]使用了求职者匿名简历数据，涵盖超过14万大专及以上学历的求职者。购物和娱乐网站，依据平台业务的不同，记录不同类型的数据。例如，淘宝和亚马逊等购物网站，记录产品信息、购物信息、评分信息等。这些数据能用于设计推荐算法、分析用户在线信誉和构建新经济指数等。例如，Lü等人[55]提供了电影评分和购物数据，用于评价推荐算法效果。搜索引擎数据是用户输入的关键词，能用来预测失业率、流行病爆发和经济走势等。例如，Choi等人[185]使用的Google Trends数据涵盖超

① http://legacy.dataviva.info

过32种类型的搜索量，能预测汽车销量、消费信息和失业人数等社会经济指标。

地理信息平台主要记录区域地貌、城市景观、市政建设等信息，包括卫星图像、街景图片和开放地图等数据。卫星图像是拍摄地面的高分辨率图片，捕捉建筑和光亮等信息。夜间光亮数据能用来估计区域经济状况，从而绘制贫困程度地图。例如，Jean等人[30]使用了夜间光亮和白天卫星图像，利用机器学习算法提取图像特征预测家庭财产状况。街景图片数据是利用街景车拍摄的道路场景，能用来量化城市感观。例如，Gebru等人[186]提供了美国近200个城市的大约5000万张街景图片；Naik等人[187]提供了美国5个城市超过100万张街景图片。开放地图数据是基础设施布局、兴趣点和商业设施等信息。例如，Haklay等人[188]提供的开放地图（OpenStreetMap）数据，包括经纬度、道路、关系和标签等信息。

2.3 社会经济分析方法

大规模数据为社会经济学研究提供了基础，数据多样性和规模的增加，也给计算社会经济学在分析方法上带来了改变。首先、传统分析工具天然地不容易处理社交网络、卫星图像和文本内容等新型数据，必须依靠数据挖掘和机器学习等先进分析工具。其次、高时空分辨率的大规模数据能体现更复杂的相互作用，使用传统的分析工具不容易处理相互作用所涌现的复杂性，必须借助复杂系统和网络科学等学科的新工具。再次、基于少量高价值人工标注数据训练机器学习模型，能更加准确地推断全体尺度上难以获得的高价值属性。

传统的社会学和计量经济分析工具，应当与新兴的交叉学科分析工具形成互补，在解决复杂社会经济问题时各展所长。事实上，传统的回归分析方法，能有效地分析控制变量的影响，在解决因果推断等问题上有优势。复杂网络分析方法，能对社会经济系统中主体之间的相互作用进行抽象和建模，方便从网络角度解释社会经济现象，利用网络结构特征进行预测。统计机器学习方法，不仅能对大规模数据进行降维，还能从数据中直接抽取有用特征，利用特征组合做出更准确的预测。下面将依次介绍计算社会经济学中常用这三种分析方法。

2.3.1 传统回归分析方法

回归分析是计量经济学中最为常用的统计分析方法[18]，用以确定两种或两种以上变量间相互依赖的定量关系。进行回归分析时，特别关心的是根据自变量（x）的给定值，考察因变量（y）的总体均值。按照所涉及到变量的个数，回归分析分为一元回归和多元回归。按照因变量和自变量之间呈现的关系类型，回归分析分为线性回归和非线性回归。下面依次介绍简单的线性回归分析方法、解决多

样本问题时使用的合并样本（Pooled Sample）回归分析方法和进行因果推断时使用的双重差分（DID）回归分析方法。

如果回归分析模型中只有一个自变量和一个因变量，并且能用一条直线来表示两者之间的关系，把这样的回归分析称为二变量（Two-Variable）或双变量（Bivariate）线性回归分析。例如，一个简单的线性回归分析模型：

$$y = \beta_0 + \beta_1 x + \mu. \tag{2–1}$$

其中，y被称为因变量，也称为被解释变量、相应变量、预测变量、从属变量等；x被称为自变量，也称为解释变量、控制变量、预测变量、回归量、协变量等；μ称为误差项或者扰动项，代表除了x以外影响y的非观测因素；β_1称为y和x之间关系的斜率参数，也是最为关注的系数；β_0称为截距参数，也就是常数项。进一步，介绍普通最小二乘（OLS）估计方法，对参数β_0和β_1进行估计。总量为n的随机样本记为$\{(x_i, y_i) : i = 1, \ldots, n\}$，对于每个样本$i$得到估计方程为

$$y_i = \beta_0 + \beta_1 x_i + \mu_i. \tag{2–2}$$

进而，基于数据得到参数β_0的估计$\hat{\beta}_0$和参数β_1的估计$\hat{\beta}_1$，得到OLS拟合曲线$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x$。OLS回归分析有些特性，例如样本的OLS残差均值为0，即$\sum_i^n \hat{\mu}_i = 0$；自变量和OLS残差的样本协方差为0，即$\sum_i^n x_i \hat{\mu}_i = 0$；样本均值点$(\bar{x}, \bar{y})$必定落在OLS回归线上，其中$\bar{x} = \sum_{i=1}^n x_i/n$和$\bar{y} = \sum_{i=1}^n y_i/n$。由简单OLS模型推广得到多元线性回归模型，包含$k$个自变量的线性回归方程为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \mu. \tag{2–3}$$

其中，β_0为常数项（截距参数）；其他β_i为对应自变量x_i的回归系数；μ为误差项。在评价自变量x对因变量y的解释能力上，通常使用R^2（称为R-squared或决定系数），定义为$R^2 \equiv \text{SSE}/\text{SST} = 1 - \text{SSR}/\text{SSE}$。其中，$\text{SSE} \equiv \sum_{i-1}^n (\bar{y}_i - \bar{y})^2$；$\text{SST} \equiv \sum_{i-1}^n (y_i - \bar{y})^2$；$\text{SSR} \equiv \sum_{i-1}^n \bar{\mu}_i^2$。取值范围上，$R^2$的数值从0到1。$100 \cdot R^2$表示可解释的差异占总差异的百分比，即$x$可以解释$y$中多少比例的样本差异。如果所有的点都在一条直线上，OLS模型实现对数据的完美拟合，这时$R^2 = 1$。

在回归分析中，一些定性因素会以二值的形式存在，例如个体的性别是否是男性、个体是否拥有电脑、区域之间是否开通高铁等。这些信息通过0-1变量来定义，一般称为哑变量（Dummy Variables）。以性别（$female$）和教育水平（edu）来确定每小时工资（$wage$）为例，所使用的OLS回归方程为

$$wage = \beta_0 + \delta_0 female + \beta_1 edu + \mu. \tag{2–4}$$

其中，$female$为性别哑变量，$female = 1$表示个体为女性的情况（基准组），

$female = 0$表示个体为男性的情况（对照组）。哑变量$female$的回归系数δ_0，表示男性和女性每小时工资的差异（在给定教育水平和误差项相同的情况下）。如果$\delta_0 < 0$，意味着女性比男性的每小时工资少（在其他因素相同的情况下）。然而，这样的回归仅考虑了男女差异，没有控制教育水平对男女的不同影响，即回归线斜率不同。为了控制教育对男女的收益相同，需要引入性别和教育的交叉项$female \cdot edu$，所使用的OLS回归方程为

$$wage = \beta_0 + \delta_0 female + \beta_1 edu + \delta_1 female \cdot edu + \mu. \tag{2–5}$$

其中，回归系数δ_1体现男女在每小时工资（$wage$）相对于教育水平（edu）斜率上的差异。当$\delta_1 = 0$ 时，表示男性和女性的回归曲线斜率相同，即教育带给工资的收益相同。值得注意的是，模型对于常数项的差异（δ_0）没有限制，但默认在所有教育水平上差异都相同。在对比分析男女样本时，为了控制回归的残差相同，一般通过引入性别哑变量的方法，在合并样本上进行回归分析。

因果推断是社会经济分析中的重要问题，也是计量经济分析中的难题。对于实证数据的因果分析，需要找到合适工具变量（IV, Instrumental Variable）来分离出自变量（尤其是政策）对因变量的影响。回归分析中，要求自变量x与残差μ之间是不相关的。如果x和μ之间存在相关性，那么估计得到的β_1不准确，需使用工具变量解决内生性（Endogeneity）问题。一个有效的工具变量z需要满足两个条件：1）工具相关性，即z_i和x_i相关性不为0；2）工具外生性，即z_i和μ_i相关性为0。使用工具变量z估计回归系数β_1时，采用两阶段最小二乘回归（TSLS）。首先，利用x关于z的OLS回归模型分离出与μ不相关的那部分x:

$$x_i = \alpha_0 + \alpha_1 z_i + v_i. \tag{2–6}$$

计算x_i的估计值$\hat{x}_i$，其中$\hat{x}_i = \hat{\alpha_0} + \hat{\alpha_1} z_i$ （$i = 1, \ldots, n$）。然后，将回归分析中的x_i替换为估计值$\hat{x}_i$，即利用OLS模型建立y_i关于$\hat{x}_i$的回归：

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \mu_i. \tag{2–7}$$

由此，可以估计变量x的回归系数β_1。常用的因果推断方法是双重差分DID回归分析[189]，是计算处理组（Treat Group）差分与控制组（Control Group）差分之差。如果处理组没有受到影响，那么趋势应当与控制组一样，即平行趋势（Counterfactual或Parallel）假定。图2–6展示了DID回归分析的工作原理示意图，其中$t = 0$表示政策实施前（Before），而$t = 1$表示政策实施后（After）。处理组趋势线与平行趋势线在$t = 1$时的差值为DID回归系数。更一般的，DID回归模型为

$$y_{i,t} = \beta_0 + \beta_1 (Treat_i \cdot After_t) + \beta_2 Treat_i + \beta_3 After_t + \mu_i. \tag{2–8}$$

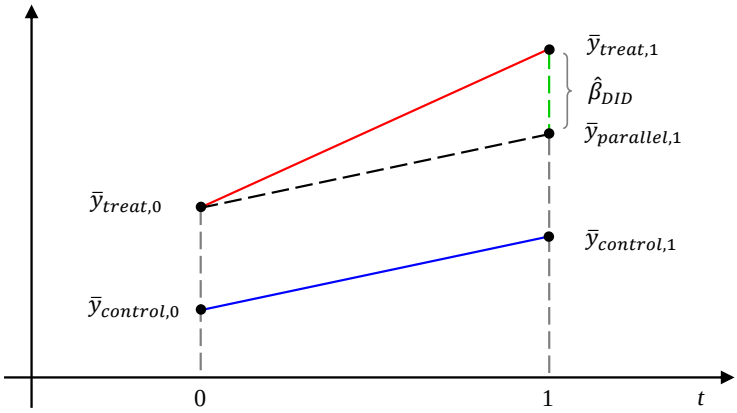

图 2–6 双重差分DID回归分析模型工作原理示意图

其中，$Treat_i$为分组哑变量（$Treat_i=0$表示控制组，$Treat_i=1$表示处理组），表示控制组和处理组之间的固有差异；$After_t$为分期哑变量（$After_t=0$表示政策实施之前，$After_t=1$表示政策实施之后），表示政策实施之后的时间效应；交叉项$Treat_i \cdot After_t$表示处理组在政策实施之后的效应，也就是处理效应；交叉项的回归系数β_1为双重差分DID的数值，表示处理效应的影响大小。如果β_1显著，则表示处理组与控制组之间在政策实施后存在显著差异。

2.3.2 复杂网络分析方法

复杂网络理论在分析社会经济系统中主体的相互作用方面非常有帮助，不同层次的相互作用都能用复杂网络模型刻画。例如，用户之间基于通讯和互动关系形成社会网络，国家与产品之间基于出口关系形成“用户-产品”二部分网络，产业之间基于接近性关系形成产业空间网络等。通过分析复杂网络的结构特征，不仅能解释很多社会经济现象，还能关联分析主体的社会经济水平。利用复杂网络模型，也能构建新的社会经济度量指标，更好地刻画社会经济发展中涌现的复杂性。另外，从网络结构的角度考虑经济发展问题，还能利用传播模型研究产品和产业升级策略，理解经济发展所面临的机遇差异。下面，依次介绍基本的网络结构特征、有代表性的复杂网络模型、社会经济问题的网络建模与分析方法。

（一）网络结构。网络由节点和连边构成，节点数量为N，连边数量为M。已经有很多网络结构特征指标描述复杂网络的拓扑性质，例如度分布指标、度相关指标、节点重要性指标等[33]。度（Degree）是刻画网络中节点属性的最基本概念，对于无向网络G，节点i的度k_i是节点所有连边的总数。对网络中所有节点的度计算平均值，得到网络的平均度（Average Degree），记为$\langle k \rangle$。如果以邻接矩阵$A=(a_{ij})_{N\times N}$表示节点数为N的无向网络，那么节点i的度和平均度为

$$k_i=\sum_{j=1}^{N}a_{ij}=\sum_{j=1}^{N}a_{ji}. \tag{2–9}$$

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{1}{N}\sum_{i,j=1}^{N} a_{ij}. \tag{2–10}$$

对于有向网络，节点度包括入度（In-Degree）和出度（Out-Degree）。节点i的出度k_i^{out}为从节点i指出的连边总数，节点i的入度k_i^{in}为指向节点i的连边总数。节点的出度和入度也能用邻接矩阵表示：$k_i^{out} = \sum_{j=1}^{N} a_{ji}$和$k_i^{in} = \sum_{j=1}^{N} a_{ji}$。网络稀疏性通过网络密度（Density）指标刻画，定义为网络中实际存在的边数与最大可能的边数的比值。无向网络的密度ρ定义为

$$\rho = \frac{M}{\frac{1}{2}N(N-1)}. \tag{2–11}$$

其中，$M = \frac{1}{2}\sum_{i,j=1}^{N} a_{ij}$为网络中的连边数量，$N$为网络中的节点数量。

度分布（Degree Distribution）用来刻画网络的整体性质。对于无向网络，度分布$P(k)$定义为，随机在网络选择的一个节点，它的度为k的概率。对于有向网络，类似地能定义入度分布$P(k^{in})$和出度分布$P(k^{out})$。度相关性（Degree Correlation）用来刻画网络的高阶拓扑特性，通过联合概率分布和相关性等方法能刻画网络的二阶度分布特性。对于度相关的网络，如果节点度是正相关的（大度节点倾向于连接大度节点），则称为同配（Assortative）网络；如果节点度是负相关的（大度节点倾向于连接小度节点），则称为异配（Disassortative）网络。在判断网络是同配还是异配上，可以计算同配系数（Assortative Coefficient）[103]。同配系数取值范围为$r \in [-1, 1]$，其中$r = -1$表示网络是异配网络，$r = 1$表示网络是同配网络。

网络中节点的价值取决于其在网络中所处的位置，处于网络中心的节点一般比处于网络边缘的节点具有更高的价值。在社会网络分析中，通常使用中心性（Centrality）指标刻画节点在网络中所处的位置，即节点重要性指标[190, 191]。常用的中心性指标包括：度中心性、介数中心性、接近中心性、核数指标、PageRank指标、LeaderRank指标等。度中心性（Degree Centrality）是最直接的度量，节点的度越大意味着节点越重要。对于节点数为N的网络，节点所能具有的最大的度为$N-1$。如果节点i的度为k_i，将归一化度中心性指标定义为

$$DC_i = \frac{k_i}{N-1}. \tag{2–12}$$

介数中心性（Betweeness Centrality）[192]利用经过网络中某个节点的最短路径的总数对节点的重要程度进行刻画。具体而言，节点i的介数中心性定义为

$$BC_i = \sum_{s \neq i \neq t} \frac{n_{s,t}^{i}}{g_{s,t}}. \tag{2–13}$$

其中，$g_{s,t}$表示从节点s到节点t的最短路径的总数；$n_{s,t}$表示从节点s到节点t的经过节点i的最短路径的总数。接近中心性（Closeness Centrality）[192]利用节点之间的平均距离来刻画网络中节点的重要程度。具体而言，节点i的接近中心性定义为

$$CC_i = \frac{1}{d_i} = \frac{N}{\sum_{j=1}^{N} d_{i,j}}. \tag{2–14}$$

其中，$d_{i,j}$为节点i到节点j的距离；$d_i = \frac{1}{N}\sum_{j=1}^{N} d_{i,j}$为节点$i$到网络中所有节点距离的平均值。核数指标（k-shell index）[104]通过对网络进行剥壳操作定义节点在网络中的重要性。具体而言，不断重复操作剥去网络中度为1的节点及其相连的边，直到网络中节点度至少为2，这些被去除的节点的核数为$k_s = 1$；接下来，不断重复剥去网络中度为2的节点及其相连的边，直到网络中节点度至少为3，这些被去除的节点的核数为$k_s = 2$；以此类推，得到网络中所有节点的核数指标。PageRank指标[193]是Google对WWW上页面的重要性的判断，也能应用于社会网络节点重要性计算。具体而言，给定网络中每个节点i初始PageRank值（简称为PR值）$PR_i(0)$，满足$\sum_{i=1}^{N} PR_i(0) = 1$；在第$k-1$步时，节点$i$将自己的PR值平均分配给所有连接的节点，将节点$i$的PR值更新为所收到PR值的总和，即

$$PR_i(k) = \sum_{j=1}^{N} a_{ji} \frac{PR_k(k-1)}{k_j^{out}}. \tag{2–15}$$

注意，网络中所有节点的PR值总和保持不变：$\sum_{i=1}^{N} PR_i(k) = 1$。LeaderRank指标[194]是在PageRank算法的基础上，将网络中加入一个全局背景节点（Ground Node）与所有节点进行双向连接，再按照PageRank算法计算网络中节点的PR值。LeaderRank算法通过引入背景节点得到全连通网络，不仅提高了算法的收敛速度，还解决了节点排序的唯一性问题。

（二）常见网络模型。最简单的网络模型是完全规则网络，常见的有三种：1）全局耦合网络：网络中任意两个节点之间都有连边，也就是图论中的完全图；2）最近邻耦合网络：网络中每个节点只与它周围的邻居节点连接，例如每个节点只与周围四个节点相连的方格网络；3）星形耦合网络：网络中$N-1$个彼此不相连的节点都与中心节点连接[60]。完全随机网络与完全规则网络对应，以Erdö和Rényi提出的ER随机图模型[196]为代表。具有固定连边概率的ER随机图可以表示为$G(N,p)$，其中N个节点中随机选取的两个节点之间有一条边的概率为p，网络中的总边数不固定。现实中的网络，通常不是完全规则的，也不是完全随机的。例如，人们通常认识一些周围的邻居，也认识一些远在他乡的朋友。Watts和Strogtz提出了WS小世界网络模型[195]，在规则网络中引入少许的随机性。简单来说，先构造一个环状最近邻耦合网络（含有N个节点），每个节点连接它左

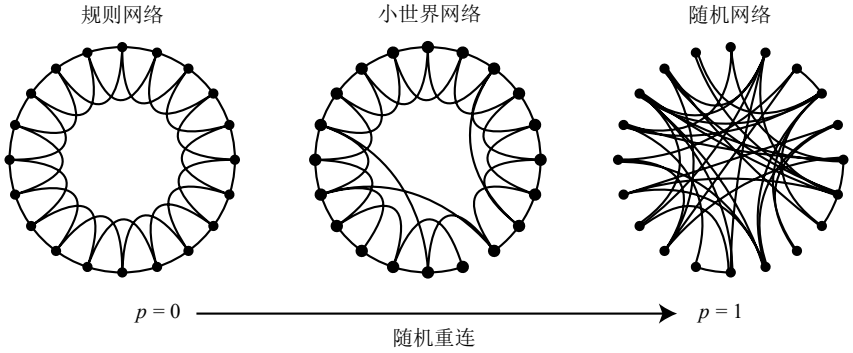

图 2–7 WS小世界网络模型示意图[195]

右相邻的$K/2$个节点（K为偶数）；再以概率p对网络中原有的每条边进行随机地重新连接。如图2–7所示，通过调节参数p的值就可以实现从完全规则网络（$p=0$）到完全随机网络（$p=1$）的过渡[195]。

ER网络和WS网络的度分布能近似的用泊松分布来表示，而有些网络的度分布呈现出幂律形式，节点的度无明显特征长度，这样的网络称为无标度（Scale Free）网络。Barabási和Albert提出了BA无标度网络模型[197]，从一个节点数量为m_0的连通网络开始，每次加入一个新节点，将它与m个已存在的节点相连（$m \leq m_0$），新节点连接到已有节点i上的概率为$\prod_i = k_i / \sum_j k_j$，其中节点$i$的度为$k_i$。BA模型基于网络的增长和优先连接特征，即网络规模不断扩大，新节点倾向于连接大度节点。很多真实网络都有空间结构，Kleinberg提出了Kleinberg空间网络模型[109]。如图2–8所示，Kleinberg模型以二维方格网络为基础，网络中的每个节点（u）除了有4条近程连边连接到周围的4个节点（a,b,c,d）之外，还有1条长程连边连接到1个远程的节点（v）。长程连边并不是随机地添加到方格网络上，节点u连接到节点v的概率与两点之间距离的幂函数$r_{u,v}^{-\alpha}$成正比，即$P(r_{u,v}) \sim r_{u,v}^{-\alpha}$。其中，幂指数$\alpha$控制长程连边的长度分布，也影响网络的空间结构特性。

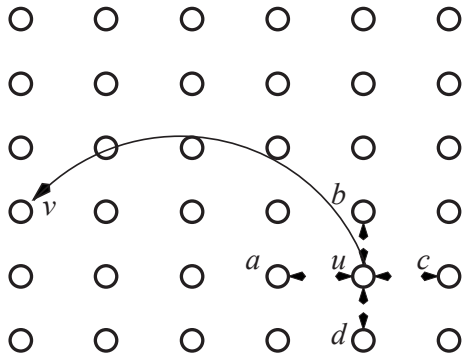

图 2–8 Kleinberg空间网络模型示意图[109]

（三）网络建模与分析。利用复杂网络能分析很多社会经济问题。一种常见的描述社会经济系统（如推荐系统和评分系统）的模型是二部分网络（Bipartite

Network）[198]。网络中包含两类节点，每条边的两个节点属于不同类，同类节点不相连。特别地，用户与产品的关系可以通过“用户-产品”二部分网络$G=\{U,O,E\}$描述。其中，$U=\{U_1,U_2,\ldots,U_m\}$为用户节点集；$O=\{O_1,O_2,\ldots,O_n\}$为产品节点集；$E=\{E_1,E_2,\ldots,E_l\}$为连边集，表示用户对产品的购买、评分、推荐等。二部分网络G能通过邻接矩阵A表示，元素$a_{i\alpha}$为用户i和产品α的连边权重，例如用户i对产品α的评分。类似地，省份与其拥有产业的关系，能通过“省份-产业”二部分网络G和等价的邻接矩阵A刻画，例如元素$a_{i,\alpha}$表示省份i内产业α的企业数量；用户与其评分电影的关系，能通过“用户-电影”二部分网络刻画。

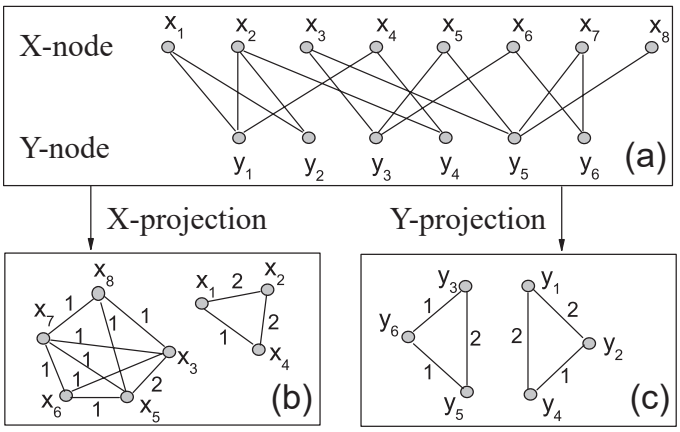

图 2–9 二部分网络中利用节点共同邻居进行投影的示意图[198]

二部分网络能通过网络投影和节点相似性计算构造同类节点组成的网络。如图2–9展示的二部分网络投影示意图[198]，利用节点共同邻居数量将“X-Y”二部分网络投影为仅包含X类节点的网络和仅包含Y类节点的网络。举例而言，X类节点x_1和节点x_2在网络中的共同邻居为y_1和y_2这2个Y类节点。所以，在投影得到的X类节点网络中，节点x_1和节点x_2之间的连边权重为2。更一般地，利用资源分配过程（Resource-Allocation Process）对二部分网络进行投影计算[198]，得到X类节点的投影矩阵为$W=(w_{i,j})_{n\times n}$，矩阵元素$w_{i,j}$定义为

$$w_{i,j}=\frac{1}{k(x_j)}\sum_{l=1}^{m}\frac{a_{i,l}a_{j,l}}{k(y_l)}. \tag{2–16}$$

其中，$a_{i,l}$为X类节点x_i和Y类节点y_l之间的连边权重；$k(x_j)$为X类节点x_j的度；$k(y_l)$为Y类节点y_l的度。类似地，也能利用资源分配过程得到Y类节点的投影矩阵。除了投影计算，也能利用相似性计算构造同类节点组成的网络。以“省份-产业”二部分网络为例，利用余弦相似性计算产业α和产业β的接近性[199]：

$$\phi_{\alpha,\beta}=\frac{\sum_i a_{i,\alpha}a_{i,\beta}}{\sqrt{\sum_i(a_{i,\alpha})^2}\sqrt{\sum_i(a_{i,\beta})^2}}. \tag{2–17}$$

其中，$a_{i,\alpha}$和$a_{i,\beta}$表示省份i在产业α和产业β中的企业数量。另外，还能利用节点共

同出现概率计算节点相似性，构造仅包含同类节点的网络。例如，基于国际贸易中产品的共同出现概率计算产品之间的相似性，从而构造产品空间[25]。

利用网络动力学过程能分析和解释很多社会经济现象。一方面，网络结构变化反映社会经济结构演化。社会经济网络结构随时间和主体的相互作用不断发生变化[200, 201]，例如用户对产品的喜好、企业对员工的技能需求、区域内产业类型的变化等。从网络的视角看，网络中的节点和数量、连边数量和权重等不断改变，导致网络结构不断演化。另一方面，网络上的传播动力学过程能揭示社会经济发展和演化规律。很多社会经济现象都能通过网络上的传播来刻画和解释，例如社会网络上的信息传播、接触过程中的流行病扩散、新产品和社会行为的采纳等。一种有代表性的传播动力学过程是靴襻渗流（Bootstrap Percolation）模型[202, 203]，可以看做是网络上的节点激活过程：（i）所有节点只处于活跃态或非活跃态；（ii）激活节点一直处于活跃态；（iii）每个节点初始时刻以概率p激活；（iv）如果一个非活跃态节点至少有k个邻居处于活跃态，那么该节点将会被激活；（v）不断重复步骤（iv）直到没有新增激活节点。靴襻渗流模型已经被用来研究网络信息传播、国家产品出口和区域产业发展等社会经济问题[204]。

2.3.3 统计机器学习方法

计算社会经济学所依赖的手机通讯、卫星遥感、社交媒体等新型数据，不再是传统社会学和经济学所惯常处理的数据。一方面，处理、分析和理解这些新类型的数据，需要利用数据挖掘和机器学习方法，将统计数据扩展到大规模复杂型数据。另一方面，利用统计机器学习算法从大规模数据中抽取重要特征，能对社会经济状态做出更准确的推断，实现对未来发展趋势的预测。在学习方式上，机器学习算法分为监督式学习、非监督式学习、半监督学习和强化学习等。在算法分类上，机器学习算法包括回归方法、决策树方法、基于核的方法、聚类方法、人工神经网络、深度学习和集成方法等[205]。下面，将简单介绍一些常用的机器学习算法，包括逻辑回归、支持向量回归、排序学习算法和深度学习算法等。

逻辑回归（LR, Logistic Regression）模型[207]是一种用于解决二分类（0-1）问题的机器学习算法，用于估计某种事物发生的可能性，例如用户购买产品的可能性、员工升职的可能性、员工离职的可能性等。逻辑回归以线性回归为理论基础，通过引入Sigmoid函数来考虑非线性因素，从而能处理0/1分类问题。Sigmoid函数也称为逻辑函数（Logistic Function），通过以下公式给出

$$g(z) = \frac{1}{1+\exp^{-z}}. \tag{2–18}$$

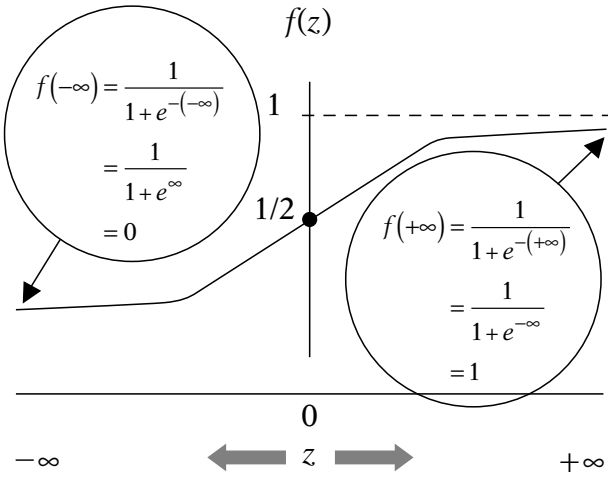

图 2–10 逻辑函数曲线和取值范围[206]

逻辑函数是一条s形曲线，取值范围在$[0,1]$之间。如图2–10所示，逻辑函数在$z=0$处很敏感；当$z\to+\infty$时，函数值趋于1；当$z\to-\infty$时，函数值趋于0。逻辑函数的这一特性，使其能有效地解决0/1分类问题。如果特征数据$x=\{x_1,x_2,\ldots,x_m\}$，对应的分类数据$y=\{y_1,y_2,\ldots,y_m\}$，那么逻辑回归模型如下

$$P(y=1|x;\theta)=g(\theta^T x)=\frac{1}{1+\exp^{-\theta^T x}}. \tag{2–19}$$

其中，θ为需要根据数据估计的参数；$P(y=1|x;\theta)$表示在给定x和θ的条件下$y=1$的概率。使用逻辑回归时，需要选择决策函数和阈值来判断分类。常用的阈值为0.5，相应的决策函数为：如果$P(y=1|x)>0.5$，那么分类$y^*=1$。实际应用中可选不同的分类阈值：如果对召回率（Recall）要求高，则可以选择小阈值；如果对准确率（Precision）要求高，则可以选择大阈值。在逻辑回归中，方程$\theta^T x=0$定义了决策边界，用来识别模型对数据的分类边界位置。

支持向量回归（SVR, Support Vector Regression）模型[208]是一种常用的分类算法，其基本思想是找到一个回归平面，使得训练集中的所有数据点到该平面的距离最近。换句话说，支持向量回归的目的是找到一个函数

$$f(x)=\langle w,x\rangle+b, w\in\mathcal{X}, b\in\mathbb{R}, \tag{2–20}$$

使其在训练数据上对真实目标y尽可能的接近，以尽可能平滑的分割面实现最大的分割度ε。SVR模型的优化函数由以下公式给出

$$\begin{aligned}&\min \quad \tfrac{1}{2}\parallel w\parallel^2\\ &\text{s.t.} \quad \begin{cases} y_i-\langle w,x\rangle-b\le\varepsilon \\ \langle w,x\rangle+b-y_i\le\varepsilon \end{cases}.\end{aligned} \tag{2–21}$$

SVR模型能容忍$f(x)$与y之间最多有ε的偏差，即仅当$f(x)$与y的差别的绝对值超

过ε时才计算损失。为了让SVR模型能处理实际中遇到的非线性问题，通常使用核函数（kernel function）的方法。基本思想是利用映射函数Φ，首先把原始数据映射到一个高维特征空间$\mathcal{F}$，然后在新空间中进行线性回归。常见核函数$K(x_i,x_j)$包括：线性核函数，$K(x_i,x_j)=(x_i\cdot x_j)$；多项式核函数，$K(x_i,x_j)=(x_i\cdot x_j+c)^d$，其中$d>0$；径向基核函数，$K(x_i,x_j)=\exp(-\gamma||x_i-x_j||_2^2)$，其中$\gamma=1/2\sigma^2>0$。

排序学习（L2R, Learning to Rank）模型[209]是一种通过构建排序模型来对列表进行排序的机器学习算法。根据训练数据的不同，排序学习能分为三类：点方式（Point-Wise），不考虑文档之间的关系，将排序问题转化为回归问题或多分类问题；对方式（Pair-Wise），将两个文档的相关度比较关系作为训练数据，训练模型来判断任意两个文档的相关度；列表方式（List-Wise），对给定文档进行人工评分，训练模型来预测和逼近人工评分。RankNet算法[210]是一种有监督的排序学习算法，属于对方式。对于一个表示特征的向量$\mathbf{x}\in\mathbb{R}^p$，RankNet算法通过学习一个评分函数$f:\mathbb{R}^p\to\mathbb{R}$，使得根据评分函数$f$得到的预测排序尽量接近真实情况。将文档$i$比文档$j$排序更高（记做$i\triangleright j$）的预测概率定义为

$$P(i\triangleright j)=\sigma(f(\mathbf{x}_i)-f(\mathbf{x}_j)). \tag{2–22}$$

使用回归方程$f=\mathbf{w}^T\mathbf{x}$预测文档排序，其中$\mathbf{w}$是参数向量。RankNet算法使用交叉熵作为代价函数（Cost Function），其计算公式为

$$\mathcal{L}=-\sum_{(i,j):i\triangleright j}\log\sigma(f(\mathbf{x}_i)-f(\mathbf{x}_j))+\lambda\Omega(f). \tag{2–23}$$

其中，$\Omega(f)=\mathbf{w}^T\mathbf{w}$为正则项。对于评分函数$f$的参数$\mathbf{w}$，代价函数的梯度公式为

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}}=\sum_{(i,j):i\triangleright j}(\sigma(f(\mathbf{x}_i)-f(\mathbf{x}_j))-1)\left(\frac{\partial f(\mathbf{x}_i)}{\partial\mathbf{w}}-\frac{\partial f(\mathbf{x}_j)}{\partial\mathbf{w}}\right)+\lambda\frac{\partial\Omega(f)}{\partial\mathbf{w}}. \tag{2–24}$$

代价函数衡量模型拟合程度，当两个文档不相关时，给予惩罚使他们分开。RankNet算法中代价函数为交叉熵，其形式方便求导，适合梯度下降法的框架。

深度学习（Deep Learning）模型[32]是一种以人工神经网络为基本架构，对数据进行表征学习的算法。在提取数据特征上，深度学习采用非监督或半监督的特征学习和分层提取算法来替代手工，更容易处理大规模内部结构特征复杂的数据。深度学习有很好的迁移学习能力，在一个数据集上训练好的模型，拿到另一个类似的数据集上通过简单的增强就可以使用。深度学习模型大致分为三类[32]：（1）多层感知机模型，代表算法是深度信念网络（DBN, Deep Belief Network）。通过对神经元间的权重的训练，使整个神经网络能够按照最大的概率来生成训练数据。（2）深度神经网络模型，代表算法是卷积神经网络（CNN, Convolutional Neural Network）。使用类似于生物神经网络的权值共享结构，使网

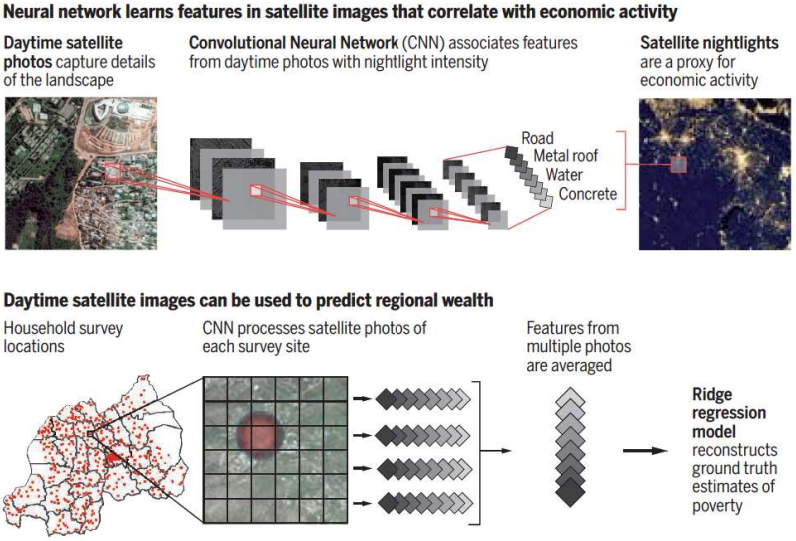

图 2–11 使用CNN模型基于白天和夜间卫星图像预测区域经济水平[211]

络模型的复杂程度降低，能把图像直接输入网络模型。（3）递归神经网络模型，代表算法是递归神经网络（RNN, Recurrent Neural Network）。在时间维度上传递的深度神经网络模型，神经元的输出直接作用于自身，能对序列内容进行建模分析。图2–11展示了利用CNN模型基于夜间光亮数据学习白天卫星图像中的特征，进而使用CNN模型提取白天卫星图像特征来预测区域经济水平[211]。

2.4 本章小结

计算社会经济学是一个新兴的交叉学科研究分支，它用定量化的手段分析社会经济系统中产生的大规模真实数据，研究社会经济发展中的各种现象，特别是与经济发展有关的社会问题，以及与社会过程有关的经济问题。本章从三方面简单介绍了计算社会经济学的基础知识。第2.1节介绍了计算社会经济学的研究内容，包括感知社会经济态势和理解社会经济规律。第2.2节介绍了社会经济的相关数据，包括政府部门和私营部门积累的大规模数据。第2.3节介绍了计算社会经济研究的常用分析方法，包括分析新型数据的交叉学科工具。

计算社会经济学的研究内容，包括感知社会经济态势和理解社会经济规律。感知社会经济发展态势的研究，微观层面关注个体的行为特征与社会经济状态之间的关系；中观层面关注社会经济系统状态、城市景观布局和功能区域状况等；宏观层面关注国家和区域的社会经济水平推断、结构刻画和发展趋势预测。理解社会经济发展规律的研究，微观层面关注个体在日常生活中的时间和空间行为规律，以及在应对突发情况时行为规律的变化情况；中观层面关注群体行为规律、城市社会经济标度律、景观布局演化规律；宏观层面关注国家和区域的经济发展学习过程和路径依赖，以及最佳的经济发展策略。

社会经济相关的大规模数据，包括政府部门统计数据、在线社会媒体数据、非干预行为数据和其他类型数据等。政府部门统计数据来源于国家大规模普查、宏观统计和问卷调查等，以统计年鉴的形式发布。国际机构也提供宏观社会经济统计数据，例如世界银行发布国家发展指数。在线社会媒体数据的来源包括微信和微博等国内平台、Facebook和Twitter等国外平台、企业内部社会化平台等。非干预行为数据是依靠现代手段在非干预状态下收集的用户行为数据，例如利用手机收集人类移动行为数据、利用信用卡收集消费记录数据等。其他类型的数据来源于行政管理平台和互联网业务平台，包括企业信息数据、经济和金融数据、在线购物和评分数据，以及卫星遥感图像数据等。

计算社会经济学研究常用的分析方法，包括传统的回归分析方法、复杂网络分析方法和统计机器学习方法等。传统回归分析是计量经济学和社会学领域常用的方法，包括线性回归分析、二阶段回归分析和双重差分回归分析等。复杂网络分析从网络结构的角度理解社会经济问题，一方面关注网络节点中心性指标等结构特征，另一方面关注简单的网络模型（如规则网络、随机网络和空间网络）和解决具体社会经济问题的网络建模方法（如二部分网络）。统计机器学习方法能从复杂数据中提取重要特征，提高对社会经济状态的预测准确性，常用模型包括逻辑回归、支持向量回归、排序学习和深度学习等。

第三章 微观层面的社会经济预测性管理研究

在数据驱动的背景下，社会经济研究面临方法论的变革，逐渐从定性分析过渡到半定量分析，甚至最终走向定量分析。借助现代手段收集和分析大规模非干预行为数据，以量化方式理解个体行为，有助于逐步实现预测性管理。本章将从三个方面介绍微观层面的社会经济预测性管理研究。首先，介绍利用校园刷卡数据刻画个体行为规律性，分析规律性和学生成绩的关联性，利用机器学习算法预测学生成绩。然后，介绍企业内部社会网络的结构特征，关联分析员工互动模式和绩效表现，进而利用网络结构特征预测员工升离职可能性。最后，介绍企业社会网络群组规模对员工互动和绩效的影响，基于手机通讯数据分析个体的社交圈规模，以及利用求职者简历数据揭示职场中性别和身高等方面的不平等性。

3.1 社会行为规律性预测学习成绩

基于大规模非干预行为数据，分析学生行为特征和提前预测学生成绩，对教育管理意义非凡。一方面，学生行为规律性被认为与学习成绩密切相关。不同于西方所推崇的自由氛围，东亚的教育文化特别强调纪律性和规律性。已有实证研究发现，课堂纪律性与学生成绩显著相关[212]，但仍然缺乏利用大规模数据验证行为规律性与学习成绩之间关系的工作。另一方面，分析非干预行为数据有助于及早发现行为异常的学生。如果有心观察，网瘾学生平时的行为就有别于其他学生[213]。量化分析学生行为数据，能揭示网络沉迷对学习的影响，精准识别网瘾学生。这有助于教育管理者提早采取干预措施，实现对学生的预测性管理。

传统社会经济行为研究主要依靠问卷和访谈获取数据，不但数据的样本规模非常有限，也容易受到心理防御和其他因素的影响。近年来信息技术的快速发展，已经有了更多途径获取高质量数据，用以分析学生的行为、成绩和心理状况，包括手机数据[47, 214]，社交媒体数据[215, 216]和GPS轨迹数据[217]等。特别地，大多数高校都给学生配发一张校园卡，不仅用于学生身份识别，还能为校园消费付款。校园卡使用频率非常高，精确地记录着学生在何时何地进行了何种活动，数据记录一般规模大、时空分辨率高，在刻画学生行为规律性上有独到优势。

本节研究中使用匿名校园卡数据[42]，涵盖近两万名大学生（N=18,960）的四种与学习和生活相关的刷卡记录，时间跨度从2009年9月到2015年7月。具体而言，行为数据包括：3.15万条寝室洗澡记录、19.02万条食堂吃饭记录、3.41万条图书馆进出记录和2.28万条教学楼打水记录。另外，研究中还收集了学生每一学期的

学习成绩。为了保护个人隐私，在数据收集时，对原始数据进行脱敏处理；在数据分析时，移除行为日期，将时间信息粗粒化；仅考虑行为类型，将位置信息粗劣化。同时降低行为数据的时间和空间分辨率，最大程度地降低个体被重新识别的风险[218]。

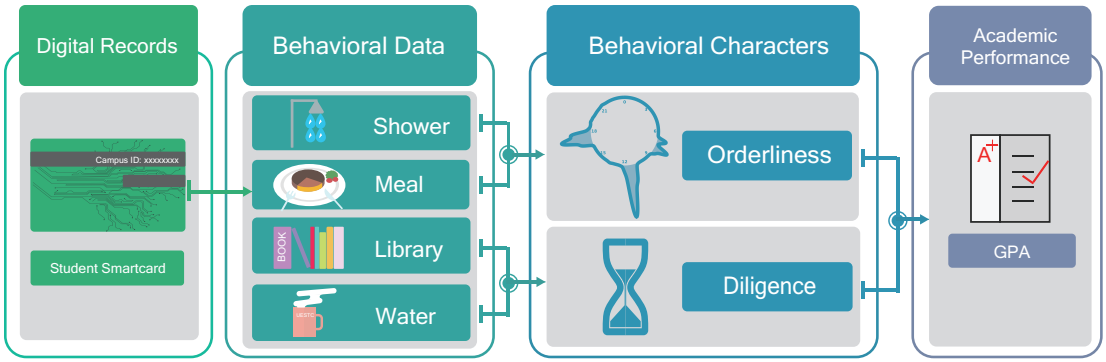

图 3–1 利用非干预行为数据预测学生成绩的分析框架

图3–1展示了研究使用的整体方法论。首先，基于刷卡记录构建两类高阶行为特征：谨严性（Orderliness）和努力程度（Diligence）。然后，关联分析行为特征和学生成绩。最后，利用行为特征训练机器学习模型，预测学生未来的成绩。

3.1.1 利用真实熵刻画校园生活规律性

使用寝室洗澡和食堂吃饭记录来刻画学生校园生活的规律性，这两类是学生的高频非干预行为数据，与学生成绩没有直接关系，也涵盖足够广泛的学生群体。这里所说的行为规律性有两层含义：一是发生时间，同类行为发生的时间要尽可能的接近，例如：集中在8点钟吃饭要比在7点到9点之间吃饭更有规律；二是发生顺序，不同类行为的发生顺序要保持规律，例如：吃饭顺序是早饭→午饭→晚饭→早饭→午饭要比早饭→晚饭→午饭→早饭→午饭更有规律。有了这些认识之后，进一步，通过建模定量刻画学生行为的规律程度。

以寝室洗澡数据为例刻画学生行为规律性，相同方法适用于食堂吃饭数据。假设某学生有n条洗澡刷卡记录，将不同日期的行为记录聚合到一天内考虑，得到时间戳序列$\{t_1, t_2, \ldots, t_n\}$。其中，$t_i \in [00:01, 24:00]$是行为发生的时刻。在数据聚合过程中，将所有行为按照发生时间排序。如果第i个记录发生在第j个记录之前，那么时间序列中保证$i < j$。然后，将一天的24小时等分为48个区间，每30分钟一个区间。其中，0:01-0:30为第1个区间，0:31-1:00为第2个区间，以此类推。按照区间划分，将时间戳序列$\{t_1, t_2, \ldots, t_n\}$转换为相应的离散序列$\{t'_1, t'_2, \ldots, t'_n\}$, 其中$t'_i \in \{1, 2, \ldots, 48\}$。举例而言，如果某学生5次连续的洗澡刷卡记录时间为{21:05, 21:33, 21:13, 21:48, 21:40}，那么对应的离散序列为$\mathcal{E} = \{43, 44, 43, 44, 44\}$。

对于任意的离散序列$\mathcal{E}$，使用真实熵（Actual Entropy）[219, 220]刻画序列的有序性。具体而言，离散序列的真实熵$S_{\mathcal{E}}$的计算公式如下

$$S_{\mathcal{E}} = \left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_i \right)^{-1} \ln n. \tag{3–1}$$

其中，Λ_i表示从$t'_i \in \mathcal{E}$开始的、未重复出现的最短序列长度。如果找不到这样的序列，那么设置$\Lambda_i = n - i + 2$作为序列长度[220]。举例而言，对于离散序列$\mathcal{E} = \{43, 44, 43, 44, 44\}$，计算得到序列长度分别为$\Lambda_1 = 1$、$\Lambda_2 = 1$、$\Lambda_3 = 3$、$\Lambda_4 = 2$和$\Lambda_5 = 2$。进一步，根据公式（3–1）计算得到序列$\mathcal{E}$的真实熵为$S_{\mathcal{E}} = 0.894$。真实熵的大小体现出离散序列的有序程度，真实熵越小，表示离散序列的有序程度越高。值得注意的是，信息熵和一般的多样性指标，无法同时刻画离散序列的时间和顺序特征，无法用于刻画离散序列的有序程度[42]。

在计算离散序列真实熵的基础上，提出行为谨严性（Orderliness）指标，反映行为的规律程度。具体而言，将序列$\mathcal{E}$的有序性$O_{\mathcal{E}}$（即行为谨严性）定义为真实熵的相反数，即$O_{\mathcal{E}} = -S_{\mathcal{E}}$。这样一来，真实熵越小，有序性越高，表示行为越有规律。为了方便跨数据集比较，利用Z-score[221]对有序性进行归一化：

$$O'_{\mathcal{E}} = \frac{O_{\mathcal{E}} - \mu_O}{\sigma_O} = \frac{\mu_S - S_{\mathcal{E}}}{\sigma_S}. \tag{3–2}$$

其中，μ_O和σ_O分别是有序性O的平均值和标准差，μ_S和σ_S分别是真实熵S的平均值和标准差，$O'_{\mathcal{E}}$是序列$\mathcal{E}$所对应学生的归一化有序性。根据学生每学期寝室洗澡和食堂吃饭刷卡记录数据，计算到学生行为的两个谨严性指标，分别是Orderliness (Shower)和Orderliness (Meal)。

学生的努力程度也是一类重要的行为特征，一般被认为与学习成绩密切相关。由于无法得到记录学生实际努力程度的数据，所以基于图书馆进出和教学楼打水记录数据来粗略估计学生的努力程度。一般而言，学生去图书馆的目的是借阅图书或上自习，出现在教学楼大部分情况是上课。如果学生在图书馆进出闸机刷卡或者在教学楼打水刷卡记录的总次数非常多，那么其努力程度可能很大。基于这些考虑，将学生每学期图书馆进出和教学楼打水刷卡记录总次数归一化，计算得到两个努力程度指标，分别是Diligence (Library)和Diligence (Water)。

如果谨严性和努力程度这两类行为特征定义合理，应当能区分不同行为模式的学生。为了验证两类行为特征计算方法的有效性，图3–2展示了真实熵与累计刷卡次数的概率分布。从图3–2(a)和(b)中看到，真实熵（寝室洗澡）和真实熵（食堂吃饭）的概率分布很广，能区分不同规律程度的学生。举例而言，一个规律程度很高的学生（真实熵排序位于5%），集中在晚上9点钟左右洗澡，集中在三餐时间附近去食堂吃饭（如图3–2(c)和(d)中深颜色所示）。对比而言，一个规律程度

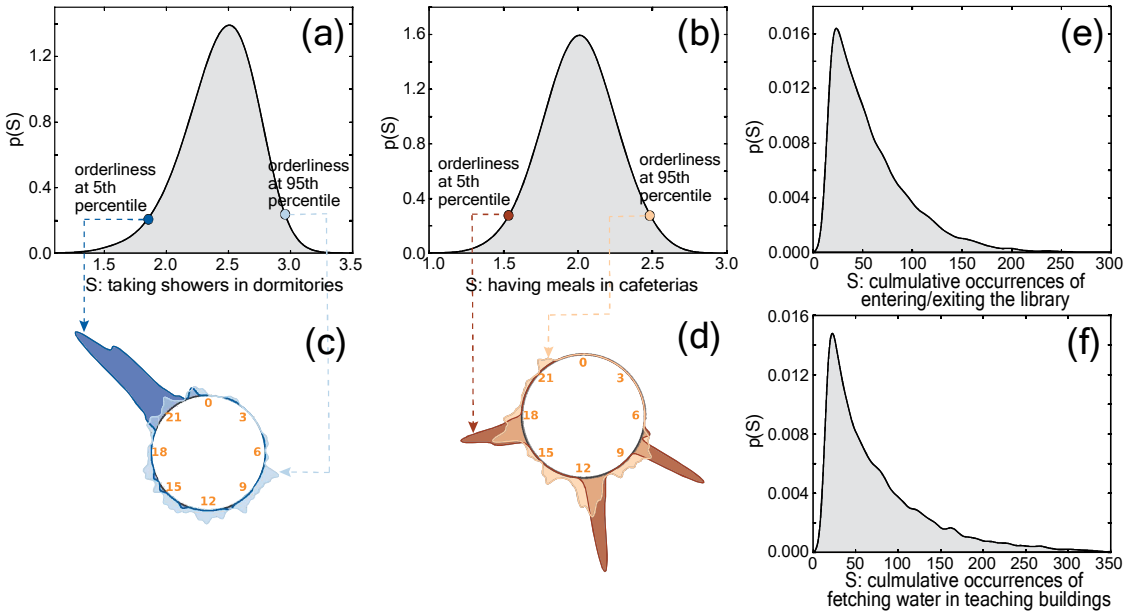

图 3–2 真实熵分布和刷卡总次数分布

很低的学生（真实熵排序位于95%），每天洗澡时间不固定，吃饭时间也不集中在三餐时间附近（如图3–2(c)和(d)中浅颜色所示）。从图3–2(e)和(f)中看到，图书馆进出和教学楼打水的刷卡次数分布也很广，能区分不同努力程度的学生。

3.1.2 关联分析行为规律性与学习成绩

学生行为的规律程度（谨严性），通过寝室洗澡和食堂吃饭数据刻画，这两类数据与学习没有直接联系。普遍认为，生活有规律的学生，自我约束力强，成绩也应该很好。接下来，实证分析行为特征与学习成绩的关系。首先，对于任意一个学生i，将学习成绩以Z-score方式归一化，$G_i' = (G_i - \mu)/\sigma$。其中，$G_i$是学生$i$的成绩（GPA），$\mu$和$\sigma$分别是所有学生成绩的平均值和标准差。然后，关联分析谨严性和努力程度特征与学习成绩，图3–3给出了两者的关系图。从图3–3(a)和(b)看到，行为谨严性与学习成绩正相关。学生的谨严性越高，学习成绩越好。类似地，从图3–3(c)和(d)看到，努力程度也与学习成绩正相关。

为了定量刻画行为特征与学习成绩的关联程度，进一步计算两者的相关系数。考虑到两者之间不是线性相关关系，尤其是努力程度与学习成绩之间（如图3–3(c)和(d)所示），所以采用斯皮尔曼排序相关系数[222]来刻画关联性强弱：

$$r_S = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}. \tag{3–3}$$

其中，N是学生总数，$d_i = r(O_i') - r(G_i')$是学生i的谨严性（Orderliness）和学习成绩（GPA）排序的差值。斯皮尔曼排序相关系数取值范围在$[-1,1]$区间内，绝对值越大，表示关联性越强。经过计算，Orderliness (Shower)与GPA的关联性

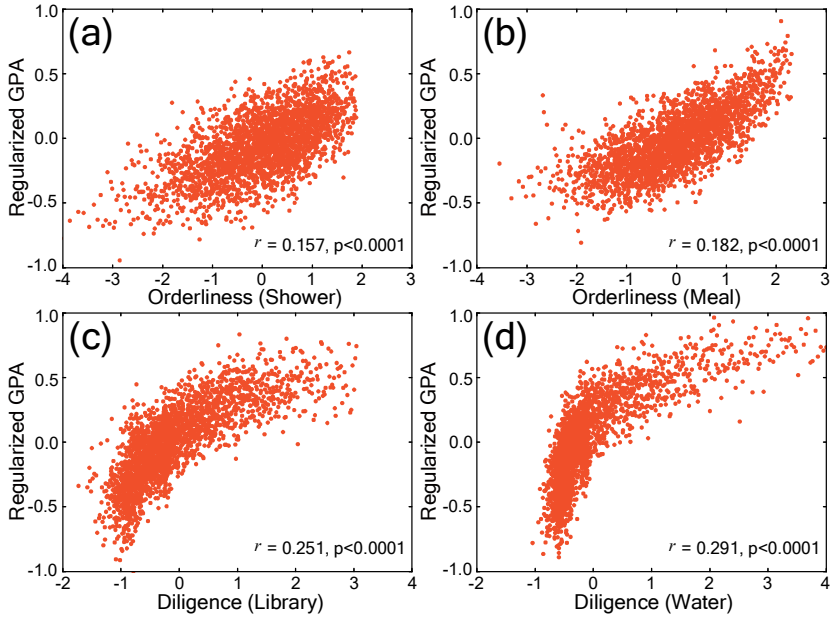

图 3–3 学生行为特征与学习成绩的相关性

为$r = 0.157$，Orderliness (Meal)与GPA的关联性为$r = 0.182$，两者都显著的正相关（$p < 0.0001$）。类似地，Diligence (Library)与GPA的关联性为$r = 0.291$，Diligence (Water)与GPA的关联性为$r = 0.251$，两者也显著的正相关（$p < 0.0001$）。

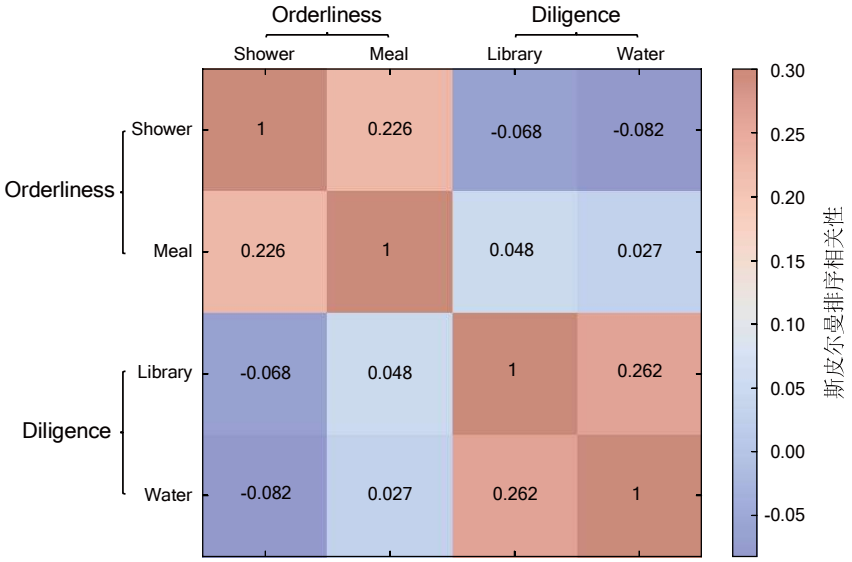

图 3–4 谨严性和努力程度特征之间的关联性

图3–4展示了不同行为特征之间的关联性分析结果。可以看到，相同类型的行为特征之间显著相关，不同类型的行为特征（谨严性与努力程度）之间的关联性不显著。具体而言，两个谨严性行为特征之间的关联性为$r = 0.226$，两个努力程度行为特征之间的关联性为$r = 0.262$，而不同类型的行为特征之间的关联性几乎为0。这些结果说明，谨严性区别于努力程度，对于预测成绩有独立的作用。谨

严性与学生成绩显著正相关的程度与基于问卷得到的五大人格特质的关联程度相当。其中，尽责性与成绩的关联性为0.2左右[223]。更一般地，通过分析大规模的非干预行为数据，教育管理者能进一步挖掘和理解影响学生成绩的关键因素。

3.1.3 基于行为特征建模预测学习成绩

大学的学习和生活环境相对宽松，导致部分学生缺失自我管理，沉迷网络游戏，无心学习，考试挂科，无法完成学业。当教育管理者发现学生成绩不合格、面临退学的时候，实际上已经无法改变局面了。如果有心观察，沉迷网络游戏的学生，在平时就有明显异常的行为表现，比如在宿舍玩游戏、总叫外卖、经常逃课、晚出早归等。通过分析校园卡记录的行为数据，有希望提前发现行为异常的学生，及早采取干预措施。这些愿景依赖于行为特征对学生成绩的预测能力。

已经发现行为特征与学生成绩显著相关，这暗示谨严性和努力程度可以作为两组独立的特征来预测学生成绩。本节研究中采用一种有监督的排序学习RankNet算法[210]，基于行为特征预测学生每学期的成绩排序。具体而言，对于每个学生，给定一个表示行为特征的向量$\mathbf{x} \in \mathbb{R}^p$，RankNet算法尝试学习一个评分函数$f : \mathbb{R}^p \to \mathbb{R}$，使得根据评分函数$f$得到的预测排序尽可能的接近真实排序情况，两者的一致程度通过实际概率和预测概率的交叉熵来测量。根据评分函数，学生i比学生j的GPA排序更高（记做$i \triangleright j$）的预测概率定义为$P(i \triangleright j) = \sigma(f(\mathbf{x}_i) - f(\mathbf{x}_j))$。其中，$\sigma(z) = 1/(1+e^{-z})$是一个Sigmoid函数。

实验中使用一个简单的回归方程$f = \mathbf{w}^T\mathbf{x}$来预测学生的成绩排序，其中$\mathbf{w}$是参数向量。RankNet算法的代价函数（Cost Function）由以下公式给出

$$\mathcal{L} = -\sum_{(i,j):i \triangleright j} \log \sigma(f(\mathbf{x}_i) - f(\mathbf{x}_j)) + \lambda\Omega(f). \tag{3–4}$$

其中，$\Omega(f) = \mathbf{w}^T\mathbf{w}$是正则项。给定所有学生的行为特征向量，以及他们的学习成绩排名，应用梯度下降分析框架来最小化代价函数。对于评分函数f中的$\mathbf{w}$参数，代价函数的梯度通过以下公式计算

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \sum_{(i,j):i \triangleright j} (\sigma(f(\mathbf{x}_i) - f(\mathbf{x}_j)) - 1)\left(\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}} - \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}}\right) + \lambda\frac{\partial \Omega(f)}{\partial \mathbf{w}}. \tag{3–5}$$

成绩排序的预测准确性通过AUC值来评价[224]，其数值等于相对排序能够被准率预测的学生对的比例。AUC值的取值范围从0到1，随机情况是0.5。AUC值超过0.5的程度，反映算法对学生成绩预测能力的大小。

在预测实验中，利用前四学期中任意学期数据所构建的谨严性和努力程度特征训练RankNet模型，预测学生在下一学期的成绩排序。为了方便说明，用O表示

表 3–1 基于行为特征预测成绩排序的准确性AUC值

行为特征	SEM2	SEM3	SEM4	SEM5
O	0.618	0.617	0.611	0.597
D	0.630	0.655	0.663	0.668
O+D	0.668	0.681	0.685	0.683

仅使用谨严性特征，用D表示仅使用努力程度特征，用O+D表示同时使用两类特征。另外，用SEM表示学期，例如SEM3表示利用第二学期的数据训练模型，用以预测第三学期的成绩排序。表3–1展示了不同行为特征组合情况下对成绩排序的预测效果。可以看到，谨严性和努力程度都对学生成绩有预测能力，并且谨严性的引入显著提升了预测准确性。虽然行为谨严性与学习没有直接关系，与努力程度也不显著相关，但它对学习成绩具有独立的预测能力。

3.2 社会网络结构特征预测职业发展

员工是企业最核心的财富，员工的预测性管理对企业发展的重要性不言而喻。虽然经过六十多年的发展，人力资源管理[225]仍旧徘徊在定量学科的大门之外。例如，员工招聘和团队配置依靠经验和直觉做判断，绩效管理和人力测评大多依靠主观评分，员工挽留和关系维系依靠访谈和问卷等形式。实际上，这些人力资源管理措施很大程度上都只是事后补救。当员工没有完成绩效考核、主动提出离职的时候，其实结果很大程度上已经无法改变了。

有预见的科研机构和企业，已经逐渐把大数据和量化分析引入到人力资源管理。例如，谷歌公司推行的“氧气计划”[226]对一万多名员工进行访谈和问卷，建模分析得到成功经理人的八个指标。借助现代技术收集和分析非干预行为数据，能更好地理解员工行为与绩效的关系。例如，利用徽章传感器追踪团队成员的沟通强度[86]，分析发现沟通交流对员工绩效至关重要。成员之间互动频繁的团队绩效表现好，员工在茶歇时间进行无关工作的交流也能提高他们整体的绩效水平[227]。人力资源管理正逐渐从单凭经验转向依靠量化分析，即大数据导航人力资源管理[175, 228]。特别地，借助复杂网络工具分析大规模非干预行为数据，能洞悉员工的真实状态和预判其行为倾向，实现对人力资源的预测性管理。

本节研究中使用的数据，来自一家中国企业内部使用的社会化办公平台。员工在这个平台上，既能与同事进行工作上的交流与合作，如参加不同的工作组、进行工作任务指派和汇报，又能像使用微博一样彼此关注对方动态和分享生活趣事，哪怕两个人没有任何工作往来。数据涵盖104位员工，是匿名化的交互行为记录。将这些数据与企业自身的人力资源数据结合，从三方面开展人力资源的预测

性管理研究。首先，基于行为数据分别构建两个雇员沟通网络，挖掘和分析网络的结构特征。然后，分析员工之间的互动行为模式，关联分析员工的绩效水平。最后，基于两个网络的结构特征，预测员工在未来升职和离职的可能性。

3.2.1 企业在线社会网络结构特征分析

将来自企业社会化平台的匿名化行为数据按照功能分解开。与工作相关的行为数据，如工作内容汇报、工作资料共享、工作任务分配、工作情况评论和回复等，构成员工之间的互动网络（AN, Action Network），节点代表员工，节点之间有向连边的权重为员工之间工作互动的总次数。类似地，与生活相关的行为数据，如生活区的互动、转发、分享等，构成员工之间的社会网络（SN, Social Network），节点代表员工，节点之间有向连边的权重为员工之间生活互动的总次数。如图3–5所示，互动网络（AN）和社会网络（SN）构成了一个雇员双层耦合网络，两层中的员工相互对应，层内连边分别表示工作互动和生活互动关系。

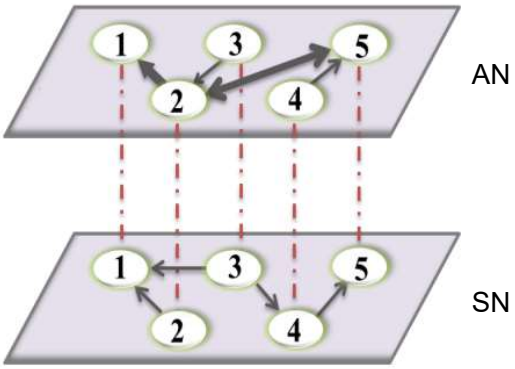

图 3–5 雇员在线互动双层耦合网络示意图

表3–2给出了两个网络的拓扑结构特征参量。具体而言，N为节点数量。$D=|E|/N(N-1)$为有向网络密度，$|E|$是有向连边数量。其他参量均基于无向网络计算，包括：平均度$\langle k\rangle$、平均最短路径$\langle d\rangle$、同配系数r[103]、聚类系数C[195]、度分布异质性H[229]和模块度Q[230]。特别地，异质性H通过度分布的Gini系数刻画，等价于节点度相对均值差异的二分之一，即$H=\sum_{i=1}^{N}\sum_{j=1}^{N}|k_i-k_j|/(2N^2\langle k\rangle)$。可以看到，社会网络比互动网络的密度和平均度大，两个网络都有很高的聚类系数、很短的平均最短路径和很小的模块度。与一般社交网络不同，两个网络都是异配网络（$r<0$）。可能因为互动主要发生在管理层和普通员工之间，前者一般为大度节点，后者通常为小度节点。

图3–6给出了雇员网络的入度（k_i）和出度（k_o）分布。图3–6(a)对应于社会网络，入度表示关注数量，出度表示粉丝数量。可以看到，出度和入度的分布存在明显差异。入度分布的异质性为$H=0.28$，有80%的员工粉丝数小于40人，仅

表 3–2 互动网络AN和社会网络SN的基本结构特征参量

网络	N	D	$\langle k\rangle$	$\langle d\rangle$	r	C	H	Q
AN	97	0.26	35.73	1.64	-0.27	0.76	0.35	0.09
SN	104	0.29	47.04	1.55	-0.41	0.81	0.32	0.08

几个员工的粉丝超过60人。出度分布的异质性为$H=0.59$，有20%的员工关注数量大于70人。这说明，员工在社会网络中更愿意关注其他人，但吸引很多员工的关注并不容易。图3–6(b)对应于互动网络。可以看到，入度和出度的分布基本一致，异质性都接近0.5，表明主动互动和被动互动之间的差异不明显[175]。

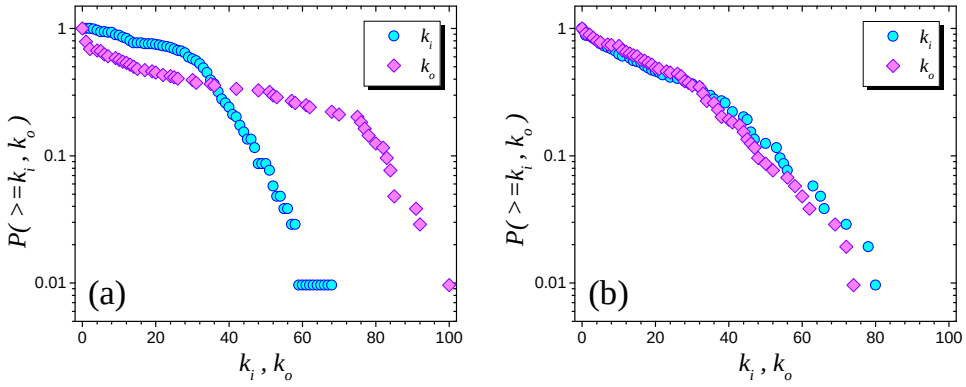

图 3–6 互动网络和社会网络的度分布

进一步，分析雇员双层耦合网络的结构，主要关注网络重叠度和层间度关联性。在重叠度方面，计算得到两个网络有向连边的Jaccard相似性为0.28，暗示两个网络重叠程度不高。进一步，利用随机置乱算法[231]将两个网络中的连边随机化，观察网络重叠度随换边次数增加的改变。具体而言，先计算列归一化的互动网络邻接矩阵A_{AN}和社会网络邻接矩阵A_{SN}差值的F范数

$$Norm_F = \|A_{AN} - A_{SN}\|_F . \tag{3–6}$$

然后，对两个网络中的连边进行随机交叉互换，重新计算$Norm_F$值。结果发现，$Norm_F$逐渐由初始值4.6增加到稳态值5.0左右，总体改变程度不大，佐证了两个网络基本不重叠。在层间度关联方面，发现两层网络中的入度关联性最强，皮尔森相关系数超过0.8，说明粉丝多的员工在互动网络中收到的反馈多。另外，被动行为（如k_i）的层间关联性强于主动行为（如k_o）的层间关联性[175]。

3.2.2 员工间互动行为模式与绩效分析

互动网络和社会网络的结构，能体现员工之间的互动行为模式。已经观察到两个网络重叠程度不高，并且度分布存在明显差异。为了理解互动行为模式，探

究度分布不同的原因，进一步计算有向网络的连边互惠性（ρ）[232]，即网络中节点形成双向连边的趋势。以社会网络为例，双向连边指的是关注其他人的员工，也获得被关注员工的关注。对于有向无权网络，连边互惠性通过以下公式计算：

$$\rho = \frac{\sum_{i \neq j}(a_{ij} - \bar{a}) - (a_{ji} - \bar{a})}{\sum_{i \neq j}(a_{ij} - \bar{a})^2}. \tag{3–7}$$

其中，$\bar{a} = \sum_{i \neq j} a_{ij}/N(N-1)$。如果从节点$i$到节点$j$存在一条连边，那么$a_{ij} = 1$；否则，$a_{ij} = 0$。如果连边互惠性$\rho > 0$，那么网络是互惠网络，网络中容易形成双向连边；如果$\rho < 0$，那么网络不是互惠网络，网络中不容易形成双向连边。计算结果显示，互动网络$\rho(AN) = 0.42$，社会网络$\rho(SN) = 0.19$，说明两个网络都是互惠网络。社会网络的连边互惠性不强，不易形成双向连边，所以出度和入度分布差异大。互动网络的连边互惠性很强，容易形成双向连边，所以出度和入度分布类似。这些结果也表明，工作相关的行为更容易激发员工间的双向互动。

社会网络普遍存在层级结构，社会经济地位高的用户更容易获得别人的关注和互动。以互动网络为例，使用PageRank (PR)指标[193]刻画员工在网络中的社会层级，分析社会层级差异与互动强度的关系。举例而言，当员工1主动向员工2互动时，定义两者的社会层级差异为，员工1的社会层级减去员工2的社会层级：

$$\Delta PR = PR_1 - PR_2. \tag{3–8}$$

图3–7给出了互动网络中员工互动的层级性。从互动次数上看，大多数互动发生层级接近的员工之间，跨越很大层级的互动不多。从互动强度上看，社会层级差别大的员工之间的平均互动强度大。不同于一般社会网络的层级结构，企业社会网络中跨层级互动存在非对称性，员工更倾向于与比自己层级高的员工互动。一方面，普通员工需要经常向领导汇报工作。另一方面，层级差异可能减弱了员工间的互动倾向。总体而言，员工间的互动行为存在一定程度的层级性。

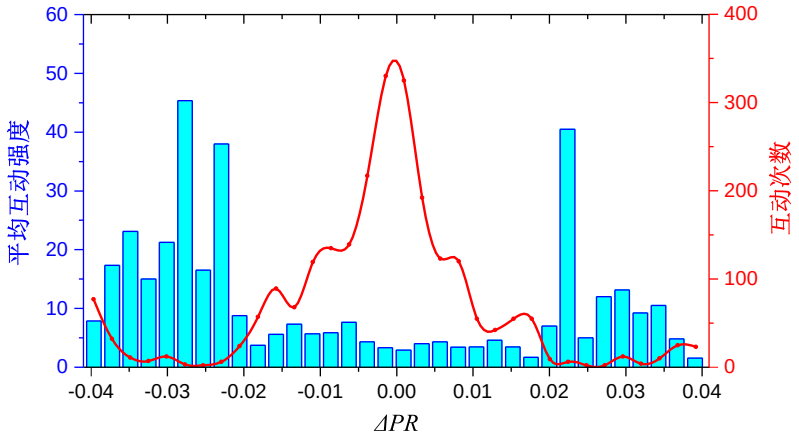

图 3–7 互动网络中员工互动的层级性

除了社会层级性，员工之间的互动模式还与绩效表现有关联。已有研究发现了一些影响员工绩效的网络指标，包括员工的沟通模式[227]、咨询网络中心性[233]、任务阻挠网络中心性[234]等。首先，计算员工在社会网络和互动网络中的12种中心性指标[191]，包括基于有向无权网络的5种指标（入度k_i、出度k_o、节点度k、PageRank指标TPR[193]和LeaderRank指标TLR[194]）、基于有向含权网络的6种指标（入强度s_i、出强度s_o、总强度s、平均强度$\langle s\rangle$、含权网络PageRank指标SPR和含权网络LeaderRank指标SLR）和基于无向无权网络的1种指标（核数指标k_s[104]）。然后，关联分析网络中心性指标和员工半年的平均绩效。

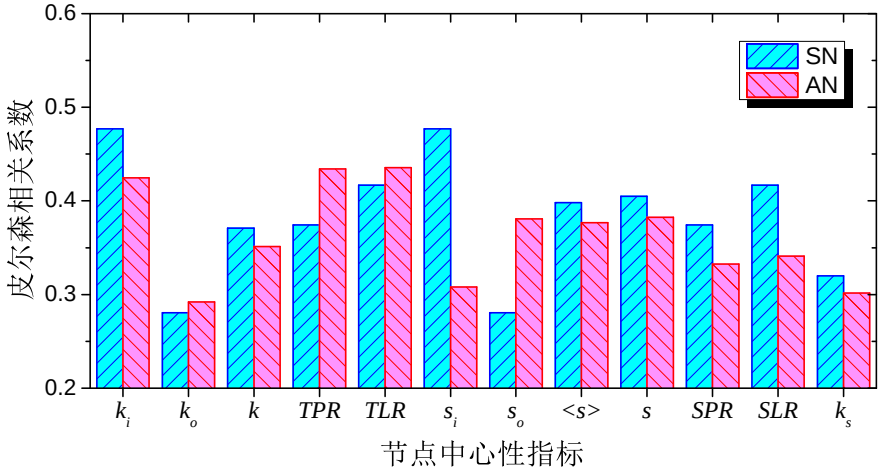

图 3–8 网络节点中心性指标与员工绩效的皮尔森相关系数

图3–8给出了互动网络（AN）和社会网络（SN）中心性指标与员工绩效的皮尔森相关系数。可以看到，网络节点中心性指标与员工绩效呈现不同程度的正相关。关联性最强的是互动网络PageRank指标（AN-TPR）和社会网络入度（SN-k_i），关联性最弱的是社会网络出度（SN-k_o）和出强度（SN-s_o）。总体而言，被动行为与绩效的关联性更强，暗示被动互动更能体现员工在工作中的重要性。另外，社会网络中心性指标与绩效关联性更强，说明社交沟通比工作互动更有助于提高绩效。进一步，分析绩效上的互动层级性，发现绩效相差越大的员工，彼此之间的平均互动强度越小，但没有出现非对称现象[175]。

3.2.3 基于结构特征预测员工升职离职

已有研究发现，绩效表现与员工的离职倾向呈现非线性关系[235]。特别地，处在网络中心位置的员工不容易离职[236]，在朋友网络中与他人更多沟通的员工也不容易离职[83]，类似的网络指标还包括度和介数中心性等[237]。在升职方面，局部社会网络有丰富结构洞的员工，晋升速度可能更快[238]。然而，这些研究大多基于问卷调查和访谈数据，不仅样本的规模有限，还受到员工心理防御等因素的影

响，员工在递交辞呈之前一般不愿意袒露真实想法。近年来，新方法已经能收集大规模非干预行为数据，应用于预测员工离职情况。例如，利用社交媒体数据识别离职个体[79]，利用手机通讯数据检测大规模裁员[80]等。

利用企业社会化平台非干预行为数据，能预测员工升职和离职的可能性。首先，计算互动网络和社会网络的结构特征，关联分析员工升职和离职情况[175]。然后，利用网络结构特征训练机器学习模型，预测员工在未来的升职和离职可能性[176, 228]。员工在互动网络的中心性，反映员工对公司业务的重要程度。员工在社会网络的中心性，反映员工对公司圈子的认同感。一般而言，如果员工积极参与工作交流，对同事生活关心，那么晋升机会大，离职可能性小。相反，如果员工不跟同事讨论工作，也不关心同事生活，可能晋升机会小，也容易离职。

图3–9给出了员工在互动网络（AN）和社会网络（SN）中的节点中心性指标与他们在半年内离职和升职的关联性。其中，图3–9(a)和图3–9(b)分别展示了离职和升职员工在相应节点中心性指标上的排序，离职和升职的员工使用横线标记，左侧蓝线对应于互动网络，右侧红线对应于社会网络。总体而言，半年内出现离职的员工，排序都很靠后（如图3–9(a)所示）；半年内得到晋升的员工，排序都很靠前（如图3–9(b)所示）。这些结果表明，在互动网络和社交网络中越处于中心的员工，越不容易出现离职，也越容易得到晋升[175]。另外，少数排序靠前的员工也出现了离职，暗示员工离职受很多其他因素的影响。

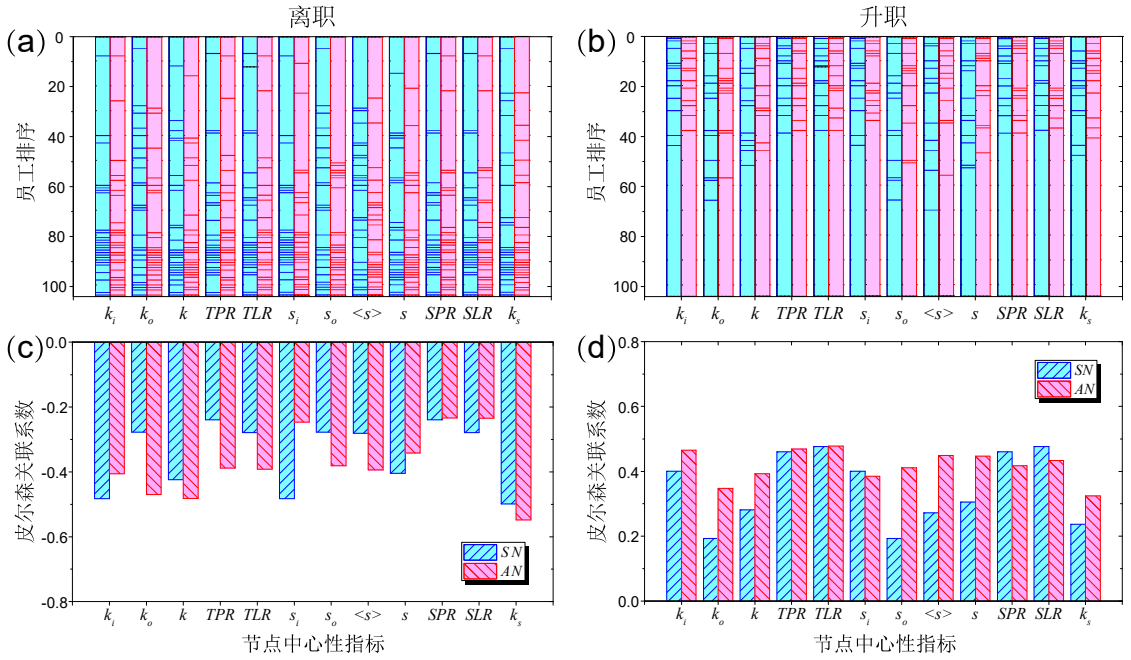

图 3–9 网络节点中心性指标与员工升离职的关联性

进一步，通过计算皮尔森关联系数，量化节点中心性指标与升离职的关联程

度[228]。从图3–9(c)看到，与离职关联性最强的中心性指标包括：互动网络中的度（AN-k）、出度（AN-k_o）和核数（AN-k_s），以及社会网络中的入度（SN-k_i）、入强度（SN-s_i）和核数（SN-k_s）。从图3–9(d)看到，与升职关联性最强的中心性指标包括：互动网络中的入度（AN-k_i），以及社会网络中的影响力指标（SN-TPR、SN-TLR、SN-SPR和SN-SLR）。另外，互动网络中有9个中心性指标与升职的关联性比社会网络强，说明互动网络暗含更多与升职有关的信息。总体而言，将要离职的员工，一般不主动与同事沟通业务、处于网络边缘、很少获得他人反馈；想要未来升职，需要人缘特别好、有高影响力、与同事保持频繁沟通。

员工离职与否（或升职与否）是0-1分类问题，所以能采用Logistic回归模型，基于两个网络的结构特征（节点中心性指标）预测员工离职（或升职）的可能性。以预测离职为例，Logistic回归模型判断员工离职的条件概率为

$$P(1|\vec{x}) = \frac{1}{1+e^{-(b_0+\sum_i^m b_i x_i)}}. \tag{3–9}$$

其中，$\vec{x}=(x_1,\ldots,x_m)$为网络的结构特征向量，$\{b_0,b_1,\ldots,b_m\}$为需要根据数据估计的系数。实验中采用三种结构特征：入度（k_i）、出度（k_o）和核数（k_s），使用AUC指标评价预测效果[224]，AUC超过0.5的程度表示预测准确性的大小。实验中采用“留一交叉验证”（Leave-One-Out Cross-Validation）方法，分别对升职和离职进行预测，计算预测结果准确性的AUC评价指标。

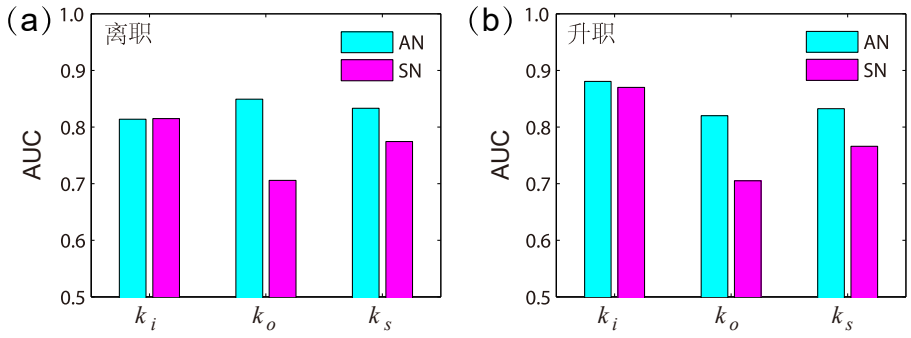

图 3–10 基于网络结构特征预测员工离职和升职的准确性

图3–10给出了基于网络结构特征预测员工离职和升职的AUC数值。可以看到，三种网络中心性指标对离职和升职都有预测能力，互动网络比社会网络对离职和升职的预测能力都更强。位于两个网络的中心位置、连接紧密的员工，容易升职，不容易离职。特别地，互动网络的入度（AN-k_i）对升职预测最有效，互动网络的出度（AN-k_o）对预测离职最有效。另外，社会网络的入度（SN-k_i）也对离职预测非常有效，预测效果最差的指标是社会网络的出度（SN-k_o）。进一步，计算离职和升职预测的准确率（Precision）和召回率（Recall）[239]，以及综合两者的F1指标[240]。最好的情况下，离职预测的F1指标达到0.85左右，升职预测

的F1指标达到0.80左右，这些结果说明预测升职比预测离职更困难[175]。

3.3 在线平台数据揭示社会经济现象

分析真实的大规模社会经济数据，不仅能为预测性管理提供决策支撑，还能揭示和理解很多社会经济现象。在企业管理方面，尤其关注团队规模对成员合作和产出的影响[227]。团队规模太小，缺乏成员的技术和管理互补性；团队规模过大，又影响沟通效率。分析在线平台数据，有希望更好地理解团队规模与产出的关系。更一般地，受脑容量和认知能力的限制，人类能够支撑的社交网络规模是有限的。一个人能维持紧密关系的人数上限是150左右，即邓巴数[107]。极少出现超过150人的大团队一起办公，紧密合作的团队规模要小的多。此外，职场中还可能存很多不平等现象，例如身高、学历、性别等方面的歧视。借助大规模数据分析揭示这些现象，有助于提出消除不平等的建议。

本节研究中使用的数据来自三种在线平台，分别展开三个相关社会经济问题的研究。首先，基于企业社会化平台的互动交流数据，分析团队和群组规模对互动交流模式和绩效水平的影响。然后，基于运营商的手机通讯数据，分析不同文化背景下社交圈子的规模，验证通讯网络中的邓巴数理论。最后，基于在线招聘平台的匿名简历数据，分析职场中的身高溢价现象和预期薪金的影响因素。

3.3.1 社会网络数据分析团队规模效应

团队规模与生产力之间的关系，是经济学、管理学和心理学等学科都非常关注的问题。已有研究发现，团队规模一定程度上影响个体和团队的生产力。例如，团队规模对生产力有非线性的影响[49]：当团队人数从一个增加到两三个时，个人生产力显著下降；当团队人数增加到五六个时，生产力不再显著降低。在完成复杂任务时，团队成员的合作随团队规模的增大而增多，大团队比同等数量个体的生产力更高[241]。当然，团队的组成不是一成不变的。长期存在的大团队有动态更换成员的能力，而小团队需要保持成员的稳定[242]。对科学家来说，如果存在跨国和跨团队的交流合作，更容易产出高影响力的研究成果[243]。

对于不同类型的团队和任务，团队规模的影响也不同。易于收集的社会网络数据，为分析不同场景下团队规模的影响提供了便利。特别地，企业内部使用的社会化平台，记录了员工和团队相关的数据[175]。员工在平台上可以参加不同的工作任务群组，与同事进行工作相关的互动交流，如分配工作任务、上传下载工作文档、讨论工作问题等。利用这些与工作相关的数据，不仅能构建员工之间的

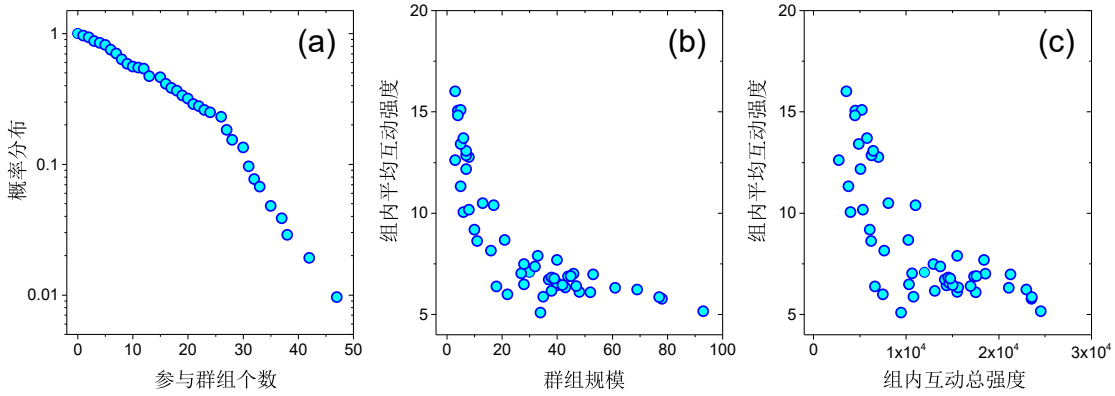

图 3–11 群组规模对组内互动强度的影响

互动网络（AN），还能量化分析群组规模对员工沟通强度和生产力的影响。有助于理解团队的规模效应，预测性地优化配置团队结构。

基于社会化平台上员工的群组参与和互动数据，分析群组规模对员工沟通交流的影响。图3–11给出了群组规模对组内互动强度影响的分析结果。从图3–11(a)看到，员工参与群组个数的累计概率分布呈现双段，分布异质性强。大部分员工的群组个数很少，80%的员工的群组个数少于25个，少数员工的群组个数超过45个。从图3–11(b)看到，随着群组规模的增大，组内平均互动强度降低。如果群组内员工数量维持在8人以下，平均互动强度将保持在较高水平。从图3–11(c)看到，组内平均互动强度与互动总强度呈现负关联。虽然互动总强度增加，但平均互动强度下降。这些结果说明，群组规模的增大，虽然增加了互动总量，但降低了整体的沟通效率，导致平均互动强度下降。所以，把团队规模控制在8人以下，将有利于组内员工保持最频繁的沟通交流。

进一步，分析群组规模对员工平均绩效的影响。增加群组规模会影响团队沟通效率，进而影响个体的生产效率。另外，团队员工的组成结构和社会经济属性，也影响整个团队的绩效表现。图3–12给出了群组规模、组内沟通强度和成员人口属性等特征对组内员工平均绩效的影响。从图3–12(a)看到，当群组规模小于8人时，组内员工平均绩效最高；当群组规模超过15人时，平均绩效迅速降低。如图3–12(b)所示，平均绩效与平均互动强度强关，越积极沟通的团队，平均绩效水平越高。如图3–12(c)所示，平均绩效随互动强度的增大而降低，因为团队规模增加所带来的绩效收益，无法弥补沟通效率降低所带来的绩效损失。

团队的成员配置结构，也会影响团队整体的绩效表现。图3–12(d-f)分别给出了群组内员工的性别比例、平均年龄和平均工龄对平均绩效影响的分析结果。从图3–12(d)看到，群组内男性比例维持在0.6到0.8之间时，平均绩效基本保持稳定。当群组内男性比例增大时，团队的平均绩效会提高。可能的原因是，男性有更大机会进入管理层，而管理层绩效普遍更好。如图3–12(e)和图3–12(f)所示，平均绩

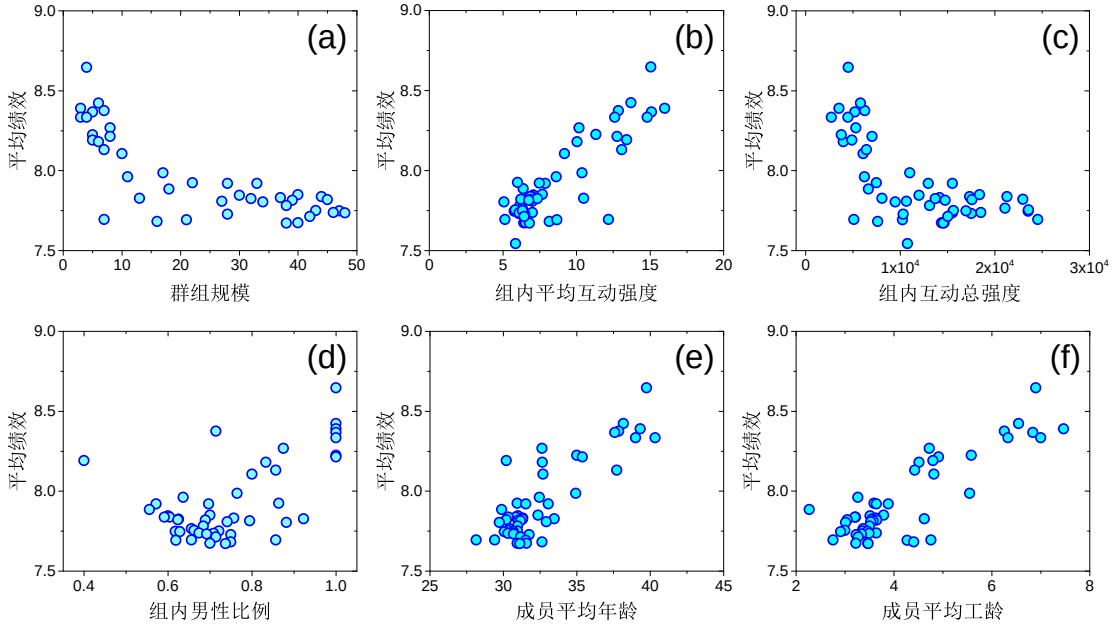

图 3–12 群组规模和社会经济属性对平均绩效的影响

效随组内员工平均年龄和平均工龄的增加而提高，两者都呈现正关联。这说明，更有经验、工作时间更长的员工，绩效表现往往越好。

根据以上群组规模对互动强度和绩效影响的分析结果，总结出一些优化团队结构配置方面的建议[175]。针对提供研究数据的这家企业，优秀团队一般具有如下特点：群组规模维持在10人以下；男性比例维持在0.8-1之间，视工作任务而定；成员平均年龄在32岁左右，新老员工互相搭配；成员平均工龄在4年以上。这些基于在线平台非干预数据的分析结果，将有助于企业更好地进行人力资源管理，以预测性的方式科学地搭配工作团队，尤其关注团队规模和成员组成结构。

3.3.2 手机通讯数据推断社交圈子规模

社交沟通对生活和工作都非常重要。一方面，不论内容是否与工作相关，社交沟通都能提高绩效水平[175, 227]。另一方面，社交沟通存在群组规模效应，随着工作团队和社交圈子的扩大，互动强度和工作绩效会下降[175]。当然，社交沟通的规模不会无限扩大。考虑到大脑容量和社会认知能力的限制[244]，邓巴数理论提出人类只能与大约150人维持亲密的社交关系[107]。换用网络的语言，在自我中心网络（Ego Network）中，个体（ego节点）所保持连接的其他个体（alter节点）平均数量在150左右[245]。特别地，个人社交关系通常以分层模式来组织，包含社交规模增加（5、15、50和150）但社交强度递减的一系列包容圈[108]。

已有研究分析真实数据验证了邓巴数理论。例如，分析在线社交媒体平台数据，发现ego网络中的社交圈[246]；分析Twitter数据，发现邓巴数在100到200之

间[247]；分析Facebook数据，发现在线社会关系有200到300人的上限[248]；分析新奥尔良的Facebook数据，发现在单个城市维系大约65人的社会关系[249]。手机数据也被用来验证邓巴数理论。例如，综合分析手机和问卷数据，发现个体有独特和不变的社会签名[250]。然而，已有研究大多关注网络的整体特征，或无向ego网络的部分特征。事实上，ego网络是有向的，连边方向有不同社会含义。另外，验证邓巴数理论大多基于西方数据，还缺乏在其他文化背景下的讨论。

本节研究中使用超过700万条手机呼叫记录数据（CDR）[178]，覆盖中国一个中心城市，时间从2014年1月到6月。基于CDR数据构建有向含权通讯网络，记为$G(V,E)$。其中，$|V|=N$为节点数量，即手机用户数量；$|E|=L$为连边数量，即社会关系数量。表3–3给出了手机通讯网络的基本统计特征。当用户i主动给用户j打电话时，形成一条有向连边l_{ij}，权重ω_{ij}为打电话的总次数，体现用户i花费多少精力维系与用户j的社会关系。用户j既能回拨电话给用户i，构建一条双向互惠连边[232]，也能直接忽略掉。考虑到通讯方向，任意ego节点i连接的alter节点分为两类：入向节点集（C_i^{in}）和出向节点集（C_i^{out}）。集合规模分别为ego节点i的入度k_i和出度k_o，分别代表ego节点i所能影响和所能维系的社会关系数量。

表 3–3 手机通讯网络的基本统计特征

时间	用户总数	网内用户数	连边总数	时间	用户总数	网内用户数	连边总数
一月	6,520,121	751,643	32,521,180	四月	6,526,250	777,486	32,383,231
二月	6,234,877	742,504	27,600,221	五月	6,561,107	787,614	34,119,390
三月	6,481,767	783,751	32,720,452	六月	6,531,076	787,156	33,461,297

从ego节点的角度出发，定义网络的五种结构特征，分析结构特征与ego社交圈子规模的关系。第一种是入度（k^{in}），刻画ego节点对alter节点的吸引力。第二种是连边总权重（W），刻画ego节点维系社会关系的总花费。ego节点i的连边总权重为$W_i=\sum_{j\in C_i^{out}} w_{ij}$。第三种是平均连边权重（$\overline{w}$），刻画ego节点与alter节点的平均感情亲密度[250]。ego节点i的平均连边权重定义为

$$\overline{w_i}=\frac{W_i}{k_i^{out}}. \tag{3–10}$$

平均连边权重$\overline{w_i}$越大，表示ego与alter之间情感越亲密。第四种是吸引力平衡性（η），刻画ego节点主动维系社会关系的程度。ego节点i的吸引力平衡性定义为

$$\eta_i=\frac{k_i^{in}}{k_i^{out}}. \tag{3–11}$$

吸引力平衡性$\eta=1$，表示社会关系的吸引力达到平衡，即维系和被维系的alter节点数量相等。η数值越大，表示ego节点吸引力越强。第五种是联结平衡性（θ_i），

刻画ego网络中强弱连边相对程度。ego节点i的联结平衡性定义为

$$\theta_i = \frac{|C_i^{in} \cap C_i^{out}|}{|C_i^{in} \cup C_i^{out}|}. \tag{3–12}$$

事实上，θ_i是ego节点i的双向互惠型alter节点的比例，即入向节点集C_i^{in}和出向节点集C_i^{out}的Jaccard距离。举例而言，$\theta = 1$表示ego节点只有互惠型alter节点；$\theta = 0$表示ego节点只有单向社会关系，没有任何互惠型alter节点。

以一月份CDR数据为例构建ego通讯网络，以ego网络联结平衡性θ特征为例计算社交圈子规模临界值。具体而言，计算ego节点的联结平衡性θ与网络规模k^{out}的皮尔森相关系数ρ，分析相关性随ego网络规模增大的变化。图3–13给出了根据ego网络联结平衡性计算得到的社交圈规模临界值。图3–13(a)中颜色越深，表示有同样联结平衡性的ego节点的数量$\log(n)$越多。可以看到，大部分ego节点都有很好的联结平衡性。随着ego网络规模k^{out}的增大，ego节点的联结平衡性逐渐集中在$\theta = 0.4$附近。图3–13(b)具体展示了θ随k^{out}增大的变化。其中，将k^{out}以10为间距划分区间，计算区间内θ的平均值和标准差。可以看到，随着k^{out}的增加，θ开始保持稳定；当k^{out}超过一定值以后，θ显著地下降。这说明，当ego网络超过一定规模时，ego节点的强弱社会连接失去平衡，联结平衡性被打破。

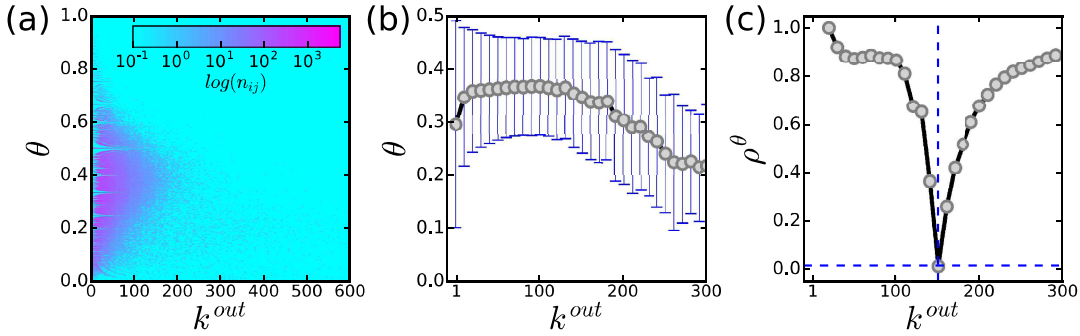

图 3–13 根据ego网络联结平衡性特征计算社交圈规模的临界值

为了确定ego网络规模的临界值，图3–13(c)展示出θ和k^{out}之间的皮尔森相关系数ρ^{θ}。在开始阶段，ρ^{θ}随k^{out}的增大而保持稳定；当k^{out}接近临界值时，ρ^{θ}急剧下降；当k^{out}大约在150时，ρ^{θ}达到最小值；超过临界值以后，ρ^{θ}随着k^{out}的增大迅速增加，然后逐渐保持稳定。利用该方法确定ego网络社交圈规模的临界值为$k_c^{out} = 151$，即ρ^{θ}达到最小值时对应的k^{out}。利用同样的方法，可以判断其他四种网络结构特征所体现的社交圈规模临界值（计算细节在文献[178]中给出）。

为验证结果的鲁棒性，图3–14给出了基于全部六个月数据计算得到社交圈规模的临界值k_c^{out}，涵盖所有五种网络结构特征。以不同颜色柱状图区分同一种网络结构特征在不同月份的计算结果，在柱状图之后展示六个月结果的平均值和标准差。可以看到，社交圈规模的临界值在时间上保持稳定，根据不同网络结构特

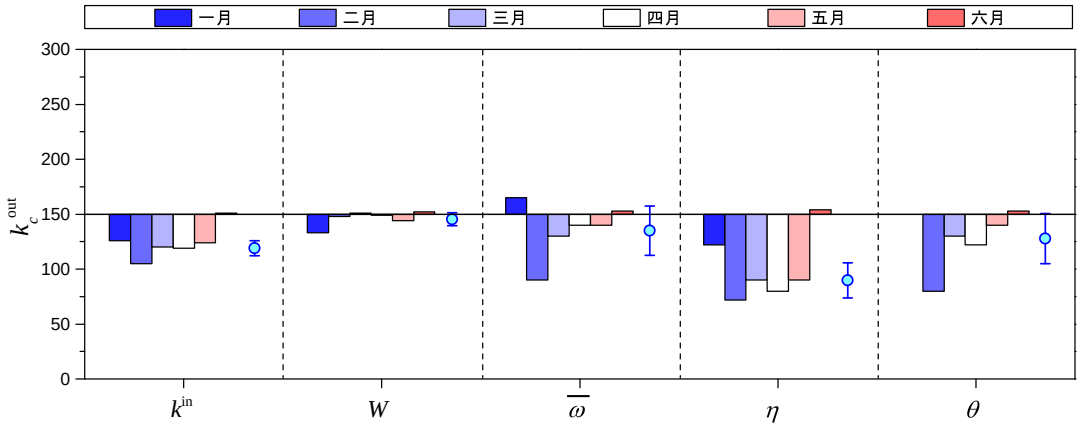

图 3–14 基于手机通讯网络计算得到社交圈规模的临界值

征计算得到的结果比较一致，集中在150附近，很接近邓巴数。一种直观的社会学解释是，随着ego网络规模的增大，ego节点需要花费更多的认知资源来维系社会关系，但ego节点所拥有的认知资源总量是有限的。当ego网络规模超过一定数值时，ego节点没有足够的资源继续维持与所有alter节点的平均感情亲密度和互惠型连边。这些结果说明，社交圈规模对社会网络的组织结构有重要影响。

3.3.3 求职简历数据揭示职场不平等性

数据资源的丰富和分析方法的进步，为揭示很多社会经济现象提供了方便。分析社会化平台数据已经发现，男性有更大机会进入管理层，增大团队内的男性比例有助于提高绩效[175]。这些结果反映出职场不平等现象，企业在招聘过程中可能在求职者身高、性别、年龄、学历、籍贯等方面存在歧视。虽然企业不会将歧视信息写入招聘需求，但职场不平等可能已经现实存在。分析大规模在线平台数据将有机会找到证据，有助于推动不平等性等问题的解决。

身高是一个人最显著的外观特征。身高不仅影响自我评价、认知能力和人格特质[251]，还影响个人职业发展[252]，例如一些职业对从业者身高有要求。很多研究发现，身高有溢价效应（Height Premium）[253, 254]，即高个子在很多方面有好处；身高也存在歧视现象，即矮个子在职场中被区别对待[255]。特别地，身高能影响求职和职场收入。例如，身高影响雇主的招聘意向[256]；身高和薪资存在非线性关系[254]；身高每增加10厘米，大约多赚10%薪资[50]。然而，这些研究使用的数据大多依靠规模有限的问卷和普查获得，也主要针对欧美国家。另外，以往研究一般使用真实薪资数据，无法区分是求职者遭遇职场歧视，还是自身预期薪资低。

本节研究中使用来自两大在线招聘网站的求职者简历数据，涵盖142,190位匿名求职者[184]。求职者提供的个人基本信息包括：身高、性别、年龄、工龄、最高学位、毕业院校、籍贯地、居住地、期望工作地等。为了统计方便，将最高学位

按四类进行数值化：博士学位为4，硕士学位为3，学士学位为2，其他学位为1。类似地，将毕业院校也数值化：“985工程”院校为4，“211工程”院校为3，其他高校为2，剩下的院校为1。另外，简历数据中还有求职者的预期薪资。基于简历数据分析职场中的身高溢价和不平等性，有两方面的优势：一是数据来自东方国家，有利于跨文化比较；二是使用求职者预期薪金，排除某些工作因素的影响。

表 3–4 求职者简历数据的基本统计信息

性别	样本规模	身高	年龄	工龄	学校	学位	预期薪资
男	78,413	173.39	29.67	5.48	2.31	1.81	8039
女	62,651	162.04	27.64	3.8	2.21	1.77	5017

为了保证统计有效性，首先对数据进行筛选：男性身高范围限定[160, 185]厘米，女性身高范围限定[150, 175]厘米；预期薪资范围限定[1000, 50,000]元/月。表3–4给出简历数据基本统计信息。平均而言，男性身高、教育水平和预期薪资更高，毕业院校更好。进一步，图3–15(a)给出男女身高分布。可以看到，身高概率分布能通过正态分布拟合（Ajd. $R^2 = 0.40$）。男性（$\mu = 173$）比女性（$\mu = 162$）平均身高更高，男性（$\sigma = 4.49$）比女性（$\sigma = 3.68$）身高分布更宽广。

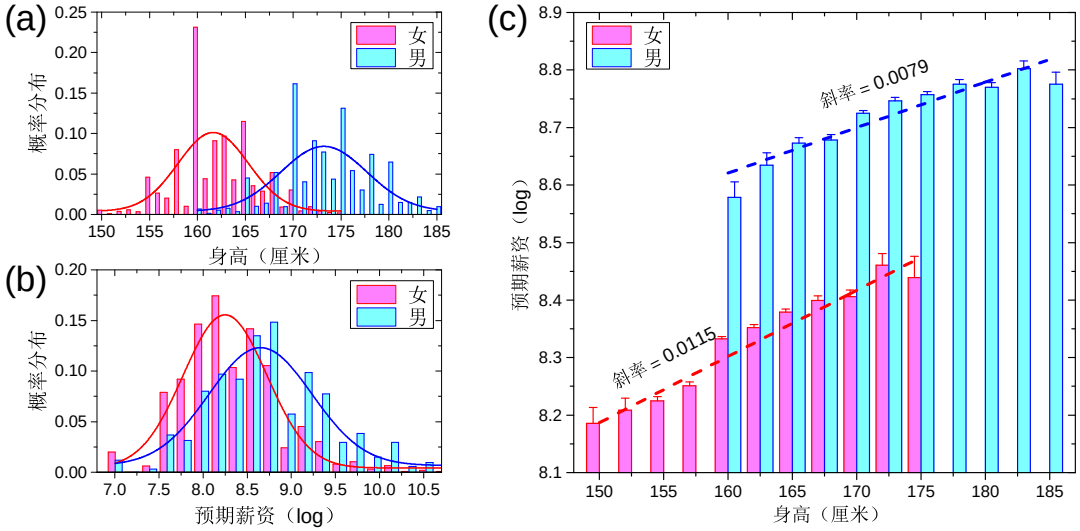

图 3–15 身高对男女求职者预期薪资的影响分析

图3–15(b)展示出男女的预期薪资也符合正态分布（Ajd. $R^2 = 0.85$），其中预期薪资进行了自然对数运算（本文中统一使用log符号表示）。可以看到，男性（$\mu = 8.65$）比女性（$\mu = 8.25$）平均预期薪资更高，男性（$\sigma = 0.58$）比女性（$\sigma = 0.47$）预期薪资分布更宽广。图3–15(c)给出了平均身高对平均预期薪资影响的分析结果。不论男女，平均预期薪资都与平均身高非常相关（男性：Ajd. $R^2 = 0.87$；女性：Ajd. $R^2 = 0.95$）。身高越高的求职者，预期薪资越高，即存在

身高溢价效应。注意到，女性拟合直线的斜率（0.0115）显著大于男性拟合的直线斜率（0.0079），这说明身高对于女性预期薪资的影响程度更大，即女性的身高溢价效应更明显。

进一步，使用回归方法验证身高（Height）对预期薪资（Salary）的影响，控制年龄（Age）、工龄（Seniority）、最高学位（Degree）和毕业院校（School）等因素。另外，也考虑籍贯地（N）、居住地（L）和期望地（T）的经济水平（$GDPpc$）。学位和院校为分类型数据，分别使用虚拟变量D^{Degree}和D^{School}。分析身高对男女影响的差异（斜率和截距）时，将女性（F）作为虚拟变量，基于男女合并样本进行最小二乘回归，控制两个样本的残差相等[18]。所使用的回归方程为

$$\begin{aligned}\log(Salary) &= \beta_0 + \delta_0 F + \beta_1 Height + \delta_1 F \cdot Height \\ &+ \beta_2 Seniority + \delta_2 F \cdot Seniority + \beta_3 Age + \delta_3 F \cdot Age \\ &+ \beta_4 \log(GDPpc^T) + \delta_4 F \cdot \log(GDPpc^T) + \beta_5 \log(GDPpc^L) \\ &+ \delta_5 F \cdot \log(GDPpc^L) + \beta_6 \log(GDPpc^N) + \delta_6 F \cdot \log(GDPpc^N) \\ &+ \beta_7 D^{Degree} + \delta_7 F \cdot D^{Degree} + \beta_8 D^{School} + \delta_8 F \cdot D^{School} + \varepsilon.\end{aligned} \tag{3–13}$$

其中，$F=0$为男性对照组。最值得关注的两个回归系数是β_0（基于男女样本回归的截距差异）和δ_0（基于男女样本回归的斜率差异，相对于身高而言）。

表3–5给出了基于男女合并样本研究身高影响预期薪资的最小二乘回归分析结果（详细结果在文献[184]中给出）。第（1）列仅考虑身高时，男性预期薪资显

表 3–5 基于男女合并样本研究身高影响预期薪资的回归分析结果

变量	(1)	(2)	(3)	(4)	(5)
F	-1.2431^{***}	-0.9988^{***}	-0.1514	0.0195	0.5347^{***}
$Height$	0.0072^{***}	0.0047^{***}	0.0116^{***}	0.0092^{***}	0.0071^{***}
$F \cdot Height$	0.0057^{***}	0.0038^{***}	0.0020^{***}	0.0003	0.0008
$Seniority$			-0.0026^{***}	0.0205^{***}	0.0206^{***}
$F \cdot Seniority$			-0.0023^{**}	-0.0014	-0.0005
Age			0.0711^{***}	0.0481^{***}	0.0482^{***}
$F \cdot Age$			-0.0104^{***}	-0.0083^{***}	-0.0088^{***}
$\log(GDPpc^T)$					0.1925^{***}
$F \cdot \log(GDPpc^T)$					0.0249^{**}
D^{Degree}	NO	YES	NO	YES	YES
D^{School}	NO	YES	NO	YES	YES
Obs.	141,064	141,064	141,064	141,064	141,064
Adj. R^2	0.0946	0.1884	0.2754	0.3303	0.3778

统计显著性水平：$^{*}p<0.1$；$^{**}p<0.05$；$^{***}p<0.01$

著地高于女性（截距差异$\delta_0 = -1.2431$；显著性$p < 0.01$），身高对女性预期薪资的影响显著高于男性（斜率差异$\delta_1 = 0.0057$；显著性$p < 0.01$）。第（5）列控制所有社会经济变量时，男女样本回归的截距差异$\delta_0 = 0.5347$变为正数，且保持显著（$p < 0.01$），说明女性事实上比男性期望更多薪资。另外，男女样本回归的斜率差异$\delta_1 = 0.0008$变的不显著，说明身高对男女预期薪资的影响没有显著差异。分析结果表明，身高在预期薪资上有溢价效应，对男女的影响都一样。

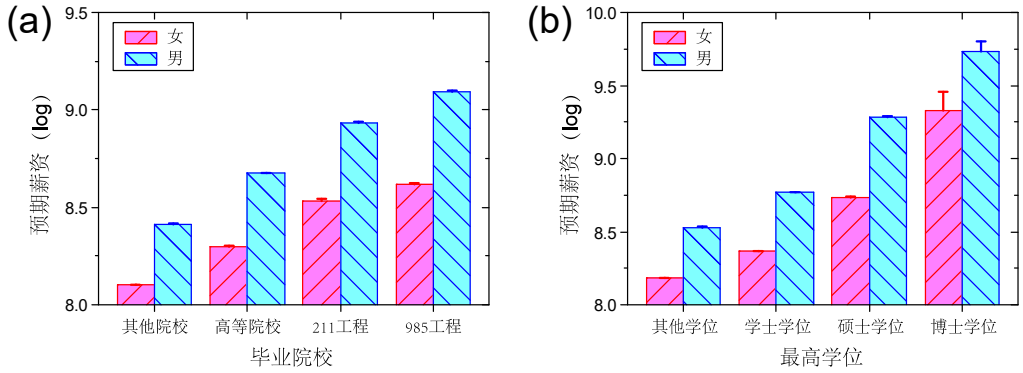

图 3–16 教育背景对求职者预期薪资的影响

其他社会经济属性也会影响预期薪资[257]。例如，图3–16展示了教育程度的影响：毕业院校越好、最高学位越高的求职者，预期薪资越高；毕业于普通高校的男性与毕业于985院校的女性预期薪资相近；女性比男性在预期薪资上相差大约一个学位，即女博士与男硕士相当，女硕士与男本科相当。在工作经验方面，预期薪资先随年龄和工龄的增长而增大，随后基本保持稳定；男女之间存在较大差异，女性要多工作大约5年才能达到男性的预期薪资水平。在地理经济方面，预期薪资随期望工作地人均收入水平的增加而增大，经济水平对男性预期薪资的影响大于女性[257]。这些分析结果对消除职场不平等性有借鉴意义。通过增加工作经验和教育经历，能一定程度上弥补身高和性别等先天条件的不足，尤其女性求职者更应自信地争取职场平等待遇。

3.4 本章小结

本章采用计算社会经济学研究范式，借助交叉学科工具挖掘大规模社会经济数据，推动很多传统学科逐渐向定量学科转变。本章从微观层面研究了社会经济预测性管理问题，探究了个体行为表现与社会经济状态的关系。通过分析非干预行为数据和在线平台数据，能对个体行为进行预测性管理。特别地，个体社会行为有规律性，利用行为规律性等行为特征能预测个体状态；个体互动交流形成社会网络，网络结构特征与个体状态和行为倾向密切相关；群组规模影响社会经济

产出，个体所能维持的社会网络规模有限，职场中存在性别和身高等方面的不平等性。挖掘大规模社会经济数据，有助于解决一些微观层面的社会经济问题。

东亚教育文化特别强调课堂纪律性，普遍认为生活有规律的学生成绩好。然而，这些习以为常的经验至今缺乏科学检验，因为传统方法不易获得大规模行为数据，也缺乏对行为规律性的刻画方法。本章第3.1节以三千万条校园刷卡记录预测学生成绩为例，研究了个体行为规律性的刻画及其对个体状态的预测能力。首先，根据洗澡和吃饭的刷卡记录，采用时间序列真实熵量化个体行为规律程度，首次提出谨严性指标；根据打水和图书馆进出的记录估计用功情况，提出努力程度指标。然后，关联分析两类行为特征与学习成绩，发现谨严性与学生成绩之间显著相关，规律性越强的学生成绩越好；努力程度与成绩也显著相关，但与谨严性不相关。最后，采用排序学习算法基于谨严性和努力程度特征预测学生成绩，发现两类行为特征都对学生成绩有预测能力，谨严性的引入能提高对成绩的预测准确性。研究结果帮助揭示影响学生成绩的因素，对学生的预测性管理和个性化教育有重要意义。一方面，基于行为数据刻画的谨严性指数与学生成绩显著相关，支持了一直强调的生活规律性。另一方面，分析大规模非干预行为数据，能及早发现行为表现和心理状况异常的学生，有助于教育管理者及早采取干预措施。

针对员工进行预测性管理，对企业发展十分重要。然而，传统升离职管理主要依靠问卷、访谈和经验判断等形式，很大程度上只是事后补救。本章第3.2节将量化分析在线平台数据引入人力资源管理，研究了网络结构特征对职业发展的预测性。首先，基于企业社会化平台数据，构建互动网络和社会网络。分析网络结构特征发现，两个网络基本不重叠，都有很高的聚类系数、很短的平均最短路径；社会网络的出度和入度分布差异显著。进一步，分析互动行为模式，发现互动网络的连边互惠性更高；互动存在社会经济地位和绩效方面的层级性，层级相差越大的员工之间的互动强度越小；社会网络的中心性指标与绩效关联性更强。最后，利用两个网络的结构特征预测员工升离职可能性，发现处在网络中心位置的员工，升职可能性大，离职可能性小；互动网络的度和出度、社会网络的入度和核数等指标与离职最相关；预测升职比预测离职更困难，互动网络对两者的预测能力都更强。分析结果有助于人力资源管理逐步转向依靠量化分析数据的预测性管理。分析在线平台大规模数据，以非干预形式洞悉员工的真实状态，有助于提前预测和规划员工职业发展，实现基于数据分析的智能人力资源管理。

利用定量化手段分析大规模社会经济数据，为揭示社会经济现象和解决社会经济问题提供了新途径。本章第3.3节针对三种常见的社会经济系统，研究了在线平台数据对社会经济现象的揭示，为预测性管理提供了依据。首先，基于企业社

会化平台数据，分析了团队规模和团队结构对沟通和绩效的影响。发现互动强度随团队规模的增大而降低，将团队规模控制在8人以下有利于最大化沟通强度和平均绩效；优秀的工作团队有规模在10人以下、男性比例稍高、新老员工搭配等特点。然后，基于手机通讯数据构建ego网络的五种结构特征，分析和验证了邓巴数理论。发现人类所能维持的社交圈规模在150人左右；当ego网络规模超过该临界值时，ego节点无法继续维持与所有alter节点的亲密联系。最后，基于匿名求职者简历数据，分析了职场中身高和性别等方面的不平等性。发现身高有溢价效应，高个子的平均预期薪资高；身高溢价效应对男女都显著，但不存在性别差异；在预期薪资上，女性比男性相差大约一个学位。基于大规模在线平台数据揭示的这些社会经济现象，为解决一些社会经济问题、消除职场不平等性提供了新思路。

第四章 中观层面的社会经济系统排序研究

社会经济系统中的主体之间存在复杂的相互作用，导致推断系统状态时不仅要考虑个体行为特征，还要关注个体相互作用和相对关系。复杂网络能对主体相互作用进行建模，在不容易直接推断社会经济系统状态时，可以借助基于网络结构的排序方法解决问题。本章将从三个方面介绍中观层面的社会经济系统排序研究。首先，介绍一种基于群组聚类的在线用户信誉排序算法，将用户按照评分的相似性进行聚类，根据归属群组规模计算用户信誉和排序。然后，介绍一种基于迭代过程的群组聚类信誉排序算法，将迭代寻优过程引入信誉排序算法框架，提高用户信誉排序的准确性和鲁棒性。最后，介绍两种基于网络结构的个性化推荐算法，分析网络节点相似性和信任关系对产品排序效果的影响。

4.1 基于群组聚类的在线系统信誉排序算法

随着互联网和信息经济的发展，在线电子商务平台逐渐被大众广泛使用。在面对数以亿计的商品和服务时，消费者也面临严重的信息过载问题[258]，不容易准确地判断所有产品的质量。为了帮助消费者做判断，在线平台通常引入评分系统，用户能对产品进行评分[259]。产品收到的评分一定程度上反映其质量情况，也进一步影响消费者的购买决定。然而，很多评分实际上并不可信，比如用户受心理等因素影响给出偏差评分[138]、故意进行作弊评分等[260]。这些偏差和作弊评分，不但影响其他用户对产品的选择，也不利于推荐系统发挥作用[55, 201]。设计有效的方法识别不可信评分和用户，对维护评分系统的健康运行非常重要。

在解决用户虚假评分问题上，最普遍的做法是建立在线信誉评价系统，根据用户评分行为对其进行信誉评价和排序。用户信誉机制已经成为在线社会经济系统的基石，例如促进社会化电子商务、帮助企业评价求职者、提高推荐算法的效果[261]。在评分系统中，用户与产品之间通过评分行为所形成的相互作用能通过“用户-产品”二部分网络进行描述[198]。信誉评价系统通过分析用户对产品的评分行为，即所形成的二部分网络的结构特征，评价用户的信誉水平。

本节研究中使用几种在线评分数据集，包括MovieLens、Netflix和Amazon等，提出和验证一种新的在线用户信誉排序算法。首先，介绍产品唯一质量假设，即产品有唯一质量分数体现其质量情况，简介基于该假设的一些传统用户信誉排序算法。然后，分析用户在产品评分时所形成群组规模的含义，提出一种基于群组

聚类的在线用户信誉排序算法。最后，介绍排序算法的常用评价指标，基于真实评分数据集测试排序算法效果，分析新算法的性能和优势。

4.1.1 研究背景与传统排序算法

在线评分系统中用户对产品的评分行为模式能反映其可信程度[262]。不顾产品好坏而随机的给产品评分，或每次都故意提高或降低对产品的评分，这些都是典型的作弊评分行为[260]。已有研究一般基于产品唯一质量假设，即每个产品对应于唯一且固定的、最客观的质量评分，能反映产品质量的好坏情况。每个产品的固定质量评分，通过被用户评分的平均值来估计[51]。已有的大部分信誉评价排序算法，根据用户评分与产品固定质量分数的偏差来评价用户的信誉，进而通过信誉排序对作弊用户进行过滤。用户评分相对于产品质量分数的偏差越大，计算得到的信誉分数越低，也越有可能是作弊评分用户。

首先，介绍在线评分系统和信誉评价算法的相关符号。在线评分系统用一个含权“用户-产品”二部分网络$G=\{U,O,E\}$来描述。其中，$U=\{U_1,U_2,\ldots,U_m\}$为用户节点集，包含m个用户；$O=\{O_1,O_2,\ldots,O_n\}$为产品节点集，包含n个产品；$E=\{E_1,E_2,\ldots,E_l\}$为网络连边集，包含l次用户对产品的评分。对于离散评分系统，二部分网络G能用评分矩阵A表示，元素$A_{i\alpha}\in\Omega=\{\omega_1,\omega_2,\ldots,\omega_z\}$为节点$i$和节点$\alpha$之间的连边权重，即用户$i$对产品$\alpha$的评分。信誉评价模型分析二部分网络$G$和评分矩阵$A$，为用户$i$计算信誉分数$R_i$，根据信誉分数对所有用户进行排序。传统基于产品质量的信誉排序算法，假设产品α有唯一质量分数反映其真实质量Q_α。由于缺少基准信息，一般通过产品收到的平均评分来估计其质量分数：

$$\hat{Q}_\alpha=\frac{\sum_{i\in U_\alpha}R_iA_{i\alpha}}{\sum_{i\in U_\alpha}R_i}. \tag{4–1}$$

其中，$A_{i\alpha}$为用户i对产品α的评分，R_i为用户i的信誉，U_α为评分产品α的用户。

然后，介绍几种传统的基于产品质量的信誉排序算法。最直接的是迭代排序（IR）算法[88]，以迭代形式计算用户信誉和产品估计质量。通常情况下，用户i的评分向量A_i与产品估计质量$\hat{Q}$存在差异δ_i，计算公式为

$$\delta_i=\frac{\sum_{\alpha\in O_i}(A_{i\alpha}-\hat{Q}_\alpha)^2}{k_i}. \tag{4–2}$$

其中，O_i为用户i评分的产品集。用户i的信誉分数与评分偏差成反比，定义为

$$IR_i=(\delta_i+\varepsilon)^{-\beta}. \tag{4–3}$$

其中，β为可调参数，ε为微小偏量。用户i的初始信誉值为$IR_i=1/n$，然后将公式（4–1）中的R_i替换为IR_i，不断更新公式（4–1）和公式（4–3）直到用户信誉IR_i和产品估计质量$\hat{Q}_\alpha$收敛。注意，每次迭代后要归一化用户信誉IR_i。

Zhou等人[51]提出了基于相似性的排序（CR）算法，根据用户评分向量和产品估计质量之间的相似性来计算用户信誉，提高了算法应对作弊评分攻击的鲁棒性。首先，利用皮尔森相关系数计算用户i的临时信誉分数：

$$TR_i = \frac{1}{k_i} \sum_{\alpha \in O_i} \left(\frac{A_{i\alpha} - \mu(A_i)}{\sigma(A_i)} \right) \left(\frac{\hat{Q}_\alpha - \mu(\hat{Q}_i)}{\sigma(\hat{Q}_i)} \right). \tag{4–4}$$

其中，$\mu(A_i') = \sum_\alpha A_{i\alpha}'/k_i$和$\sigma(A_i') = \sqrt{\sum_\alpha (A_{i\alpha}' - \mu(A_i'))^2/k_i}$分别为评分向量$A_i'$的均值和标准差。如果$TR_i < 0$，那么信誉分数$CR_i = 0$；否则，信誉分数$CR_i = TR_i$。用户$i$的初始信誉分数为$CR_i = k_i/n$，即用户度$k_i$的平均值。然后，将公式（4–1）中的$R_i$替换为$CR_i$，不断迭代更新公式（4–1）和公式（4–4）直到收敛。

在CR算法框架下，Liao等人[52]提出了基于信誉重新分配的排序（IARR）方法，提升了高信誉用户在信誉评价中的影响力。对于用户i，在迭代过程中以非线性的方式重新分配基于CR方法得到的用户信誉分数：

$$IARR_i = CR_i^\theta \cdot \frac{\sum_j CR_j}{\sum_j CR_j^\theta}. \tag{4–5}$$

其中，θ为可调参数。当$\theta = 1$时，IARR方法退化为CR方法。通过引入两个惩罚参数，Liao等人[52]还改进出一种信誉重分配迭代排序（IARR2）算法。

考虑用户活跃模式对其信誉水平的影响，Liu等人[54]提出了一种改进的迭代信誉排序（IRUA）算法，活跃的用户在信誉评价中更重要。不同于公式（4–1），IRUA方法在估计产品α的质量分数时，考虑评分用户的度：

$$\hat{Q}_\alpha'' = \max_{i \in U_\alpha} \left\{ \frac{k_i}{n} \right\} \cdot \hat{Q}_\alpha. \tag{4–6}$$

同时考虑利用公式（4–4）计算出的TR_i和用户的度k_i，更新用户i的信誉分数：

$$IRUA_i = \begin{cases} \left(\frac{k_i}{k_{\max}} \right)^\theta \cdot TR_i, & TR_i \geq 0 \\ 0, & TR_i < 0 \end{cases} \tag{4–7}$$

其中，$k_{\max}$为所有用户的最大度，θ为可调参数。当$\theta > 0$时，活跃用户的信誉在迭代过程中被增强；当$\theta < 0$时，不活跃用户的信誉在迭代过程中被增强。

4.1.2 基于群组聚类的信誉排序算法

传统的基于产品质量的信誉排序方法，假设产品有唯一分数体现其质量。然而，在线评分系统本质上是社会化平台，用户评分行为受到个人偏好等众多因素影响[264]。对于相同的产品，不同用户可以给出多个合理的评分，都能反映产品的质量情况。这种情况类似于完成难度不同的在线任务，例如在不同场景下数物品个数，用户所形成的群组规模一般不同[263]。对于简单任务，如图4–1(a)所示，

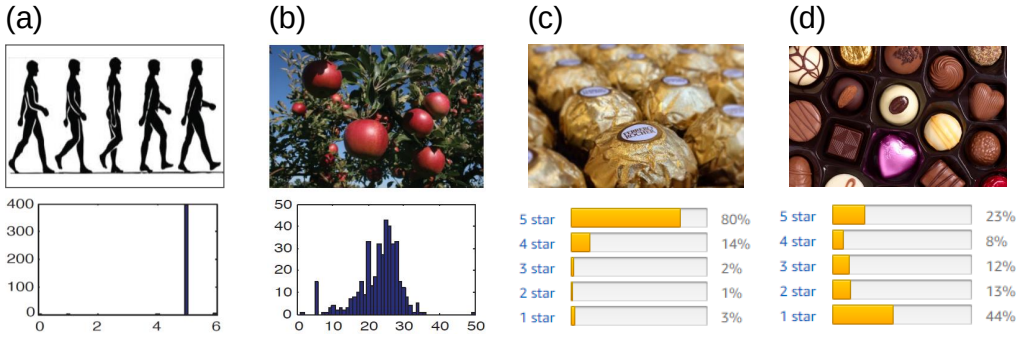

图 4–1 在线评分任务难度与用户所形成群组规模的关系[263]

用户给出答案很一致，形成少数较大规模群组。对于困难任务，如图4–1(b)所示，用户给出答案很多样，形成很多小规模群组。用户所属群组规模，一定程度上能反映用户的可信程度[263]：所属群组规模越大，用户行为越可信。

类比到在线评分系统，尽管产品真实质量无从所知，仍然能通过产品收到评分的分布情况估计产品质量是否容易判断。如果收到的评分很集中，如图4–1(c)所示，说明用户很容易判断产品的情况，那么产品的质量显而易见。如果收到的评分很分散，如图4–1(d)所示，说明用户很难判断产品的情况，那么产品的质量就很难说。对于质量难以把握的产品，一个随机评分是可以接受的，因为用户意见并不统一。对于质量显而易见的产品，一个偏差评分则难以接受，因为用户意见明显偏离大众。这样一来，与大众选择保持一致的用户，将形成大规模的群组，获得高信誉，因为普通用户有从众心理[264]；与大众选择出现背离的用户，只能形成小规模的群组，获得低信誉，因为作弊用户获得很少的支持。

根据以上分析，传统信誉排序方法所依赖的产品唯一质量假设不再适用。本文提出一种基于群组聚类的信誉排序（GR）算法[265]，将用户按照评分进行聚类，再根据所属群组的规模计算用户信誉和排序。具体而言，GR方法首先根据评分对用户进行群组划分。对于任意产品α，将评分数值为ω_s的用户放入群组$\Gamma_{s\alpha}$：

$$\Gamma_{s\alpha} = \{U_i \mid a_{i\alpha} = \omega_s, i = 1, 2, \ldots, m\}. \tag{4–8}$$

考虑到用户对不同产品进行评分，同一个用户i能被划分到k_i个不同的群组，也就是用户评分产品的总数。然后，计算所有群组的大小，得到群组规模矩阵：

$$\Lambda_{s\alpha} = |\Gamma_{s\alpha}|, \tag{4–9}$$

即对产品α评分ω_s的用户总数。将群组规模矩阵Λ列归一化，构建评分回馈矩阵：

$$\Lambda^*_{s\alpha} = \frac{\Lambda_{s\alpha}}{k_\alpha}. \tag{4–10}$$

进而，将原始评分矩阵A按照评分回馈矩阵Λ^*映射到信誉回馈矩阵A'。具体而言，

用户i从评分$a_{i\alpha}$获得的信誉回馈为

$$A'_{i\alpha} = \Lambda^*_{s\alpha}. \tag{4–11}$$

映射过程中，矩阵列标需要满足$a_{i\alpha} = \omega_s$。如果用户i没有对产品α进行过评分，则$A'_{i\alpha}$为空，相应位置以“-”表示，不参与后续运算。

基于信誉回馈矩阵A'计算每个用户的信誉值。如果用户获得信誉回馈的均值小，那么评分行为偏离大众，信誉应该不高；如果获得信誉回馈的方差大，那么评分行为不稳定，信誉也应该不高。基于这两点考虑，将用户i的信誉定义为

$$R_i = \frac{\mu(A'_i)}{\sigma(A'_i)}. \tag{4–12}$$

其中，μ和σ为信誉回馈向量A'的均值和标准差。对于用户i，两者计算公式分别为

$$\mu(A'_i) = \sum_{\alpha} \frac{A'_{i\alpha}}{k_i}. \tag{4–13}$$

$$\sigma(A'_i) = \sqrt{\frac{\sum_{\alpha}(A'_{i\alpha} - \mu(A'_i))^2}{k_i}}. \tag{4–14}$$

事实上，信誉R_i的计算公式与向量A'_i的变异系数[266]互为倒数，刻画用户i获得评分回馈频率分布的分散程度。最终，将所有用户按照信誉值R排序，得到前L个信誉排序最低的用户，更可能存在作弊评分行为。

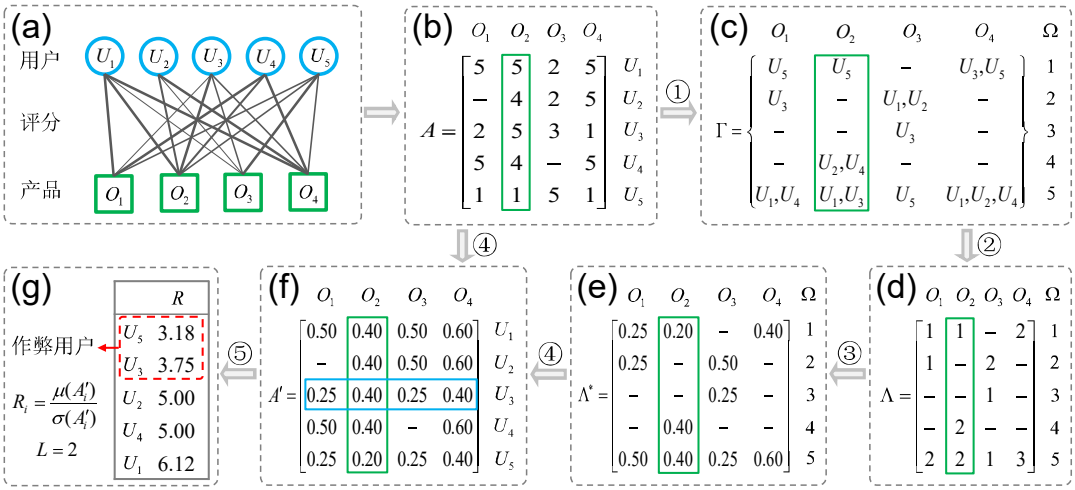

图 4–2 基于群组聚类的在线用户信誉排序算法示意图

图4–2展示了基于群组聚类的在线用户信誉排序算法示意图。其中，箭头旁边的序号表示算法的运行流程。具体而言，图4–2(a)给出了一个“用户-产品”二部分网络G，5个用户U对4个产品O进行了评分。图4–2(b)给出了与二部分网络G对应的评分矩阵A。其中，行对应于用户，列对应于产品，元素为用户对产品的评分。图4–2(c)给出了将用户根据评分聚类得到的群组Γ。以产品O_2为例（以绿

色竖直线框标注），仅用户U_5评分为1，所以$\Gamma_{1,2}=\{U_5\}$；没有用户评分为2和3，所以$\Gamma_{2,2}$和$\Gamma_{3,2}$为空；用户U_2和U_4评分为4，所以$\Gamma_{4,2}=\{U_2,U_4\}$；用户U_1和U_3评分为5，所以$\Gamma_{5,2}=\{U_1,U_3\}$。图4–2(d)给出了群组规模矩阵Λ，按照产品O_2划分得到的群组规模分别为{1,-,-,2,2}。其中，“-”代表空值，不参与后续运算。图4–2(e)给出了列归一化的群组规模，即评分回馈矩阵Λ^*。如图4–2(f)所示，将原始评分矩阵A根据评分回馈矩阵Λ^*映射，得到信誉回馈矩阵A'。以用户U_3对产品O_2的评分$A_{3,2}=5$为例，映射得到的信誉回馈为$A'_{3,2}=0.40$。图4–2(g)给出了所有用户的信誉排序。以用户U_3为例，信誉值$R_3=\mu(A'_3)/\sigma(A'_3)=3.75$。当检测列表长度为$L=2$时，信誉排序最低的2个用户，即用户$U_5$和$U_3$，被认为是作弊评分用户。

4.1.3 算法性能和实验结果分析

为了评价基于群组聚类的用户信誉排序算法的性能，使用三个真实在线评分数据集测试算法对用户信誉的排序效果。其中，MovieLens和Netflix是电影评分数据集，Amazon是商品评分数据集。这三个数据集都采用5分制评分规则：1分代表最差评价，5分代表最好评价。为了保障评分的可靠性，数据集中仅保留评分次数超过20次的用户，以及他们所评分过的产品。表4–1给出了这三个数据集的基本统计信息。其中，二部分网络的稀疏度通过$S=l/(mn)$计算，m为用户总数，n为产品总数。可以看到，MovieLens数据集中用户和产品的平均度都最大，二部分网络最稠密；Amazon数据集平均度最小，网络最稀疏。

表 4–1 真实在线评分数据集的基本统计信息

数据集	用户数	产品数	用户平均度	产品平均度	网络稀疏度
MovieLens	943	1682	106	60	0.063
Netflix	1038	1215	47	40	0.039
Amazon	662	1500	36	15	0.023

由于缺乏作弊评分用户的基准信息，所以模拟两种已知的作弊评分用户：一种是恶意型（Malicious）作弊评分用户，每次都以等概率给最高5分或最低1分；另一种是随机型（Random）作弊评分用户，每次从1分到5分中随机选择评分。具体而言，在真实评分数据集中随机选取d个用户，将他们的原始评分替换为模拟作弊评分。这相当于测试数据集中已知d个作弊评分用户，比例为$p=d/m$。另外，将作弊用户的活跃程度定义为$q=k/n$，其中k为作弊评分评分用户的度。注意，k能作为可调参数。当k小于用户原始的度时，在用户原始评分中随机选择k个替换为作弊评分；否则，随机选取相差的部分进行替换。利用这种方法，基于三个真实数据集构造分别包含恶意型和随机型作弊评分用户的测试数据集。

使用两种指标评价用户信誉排序算法的效果。第一种是召回率（Recall）指标[239]，刻画作弊评分用户多大程度上包含在长度为L检测列表中：

$$R_c(L) = \frac{d'(L)}{d}. \tag{4–15}$$

其中，$d'(L) \leq d$为L长度的列表中能检测到作弊评分用户的数量。召回率指标越高，说明信誉排序算法越准确。由于$R_c(L)$仅关注排序最高的L个用户，所以同时使用一种无关检测列表长度的AUC指标[224]。给定所有用户的排序，AUC指标的含义可以理解为，随机选一个作弊评分用户比随机选一个非作弊用户排序高的概率。计算ACU指标时，每次随机选一个作弊用户和一个非作弊评分用户，比较他们的信誉排序。在N次独立比较中，如果有N'次作弊用户的信誉排序比非作弊用户的信誉排序低，有N''次他们排序相等，那么AUC指标通过以下公式计算：

$$AUC = \frac{N' + 0.5N''}{N}. \tag{4–16}$$

随机情况下，AUC=0.5。所以，AUC超过0.5的程度，表示算法准确性的大小。

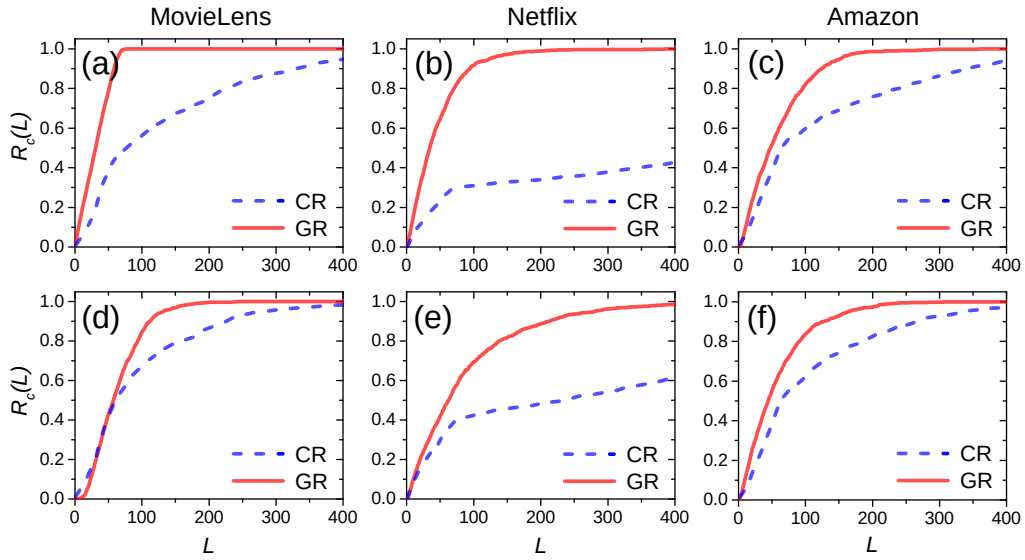

图 4–3 排序算法召回率随检测列表长度的变化

在评价排序算法时，产生包含50个作弊评分用户的三种测试数据集，对每种作弊评分都设置$d = 50$；重复50次实验计算结果的平均值。图4–3给出了CR算法和GR算法得到的召回率$R_c(L)$随列表长度L的变化。其中，图4–3(a-c)针对恶意型作弊评分用户，图4–3(d-f)针对随机型作弊评分用户。可以看到，GR算法在检测两类作弊评分用户上都比CR算法效果更好，尤其当L大于d时。另外，相比于随机型作弊评分用户，GR算法对恶意型作弊评分用户的排序结果更准确，尤其当L小于d时。表4–2给出了CR算法和GR算法对恶意型和随机型作弊评分用户排序的AUC结果。可以看到，GR算法在每个数据集上都比CR算法表现好，说明GR算

法在检查作弊评分用户上更准确。特别地，GR算法更擅长检测恶意型作弊评分用户，CR算法更擅长检随机型作弊评分用户。另外，两种算法都在Netflix数据集上效果最差，暗示该数据集中可能已经包含了很多作弊评分或异常评分用户。

表 4–2 信誉算法对作弊评分用户排序的AUC结果

测试	恶意型	恶意型	随机型	随机型
数据集	CR	GR	CR	GR
MovieLens	0.876	0.994	0.914	0.959
Netflix	0.543	0.977	0.668	0.930
Amazon	0.824	0.941	0.877	0.949

为了全面评价排序算法对作弊评分用户的检测效果，改变人工加入作弊评分用户的比例p和作弊评分的比例q，在不同参数组合下评价算法性能。图4–4给出了GR算法对作弊评分用户的检测效果。其中，图4–4(a-c)针对恶意型作弊评分用户，图4–4(d-f)针对随机型作弊评分用户。图中颜色深浅表示召回率$R_c(L)$的数值大小，检测列表长度设置为$L = d$。考虑到数据集稀疏程度的不同，分别设置参数p和q位于不同范围。可以看到，GR算法在检测恶意型作弊评分用户上总体表现好，尤其针对不活跃的作弊评分用户。当作弊评分用户比例p很小时，GR算法在MovieLens和Netflix数据集上对随机型作弊评分用户的排序准确性低，在Amazon数据集上对两种作弊评分用户的排序准确性都低。另外，随着p的增大，GR算法的排序准确性提高。这些结果说明，恶意型作弊评分用户更容易被检测；MovieLens和Netflix数据集中可能原本就有不少作弊评分用户；GR算法在检测活跃程度不高的作弊评分用户上效果更好。

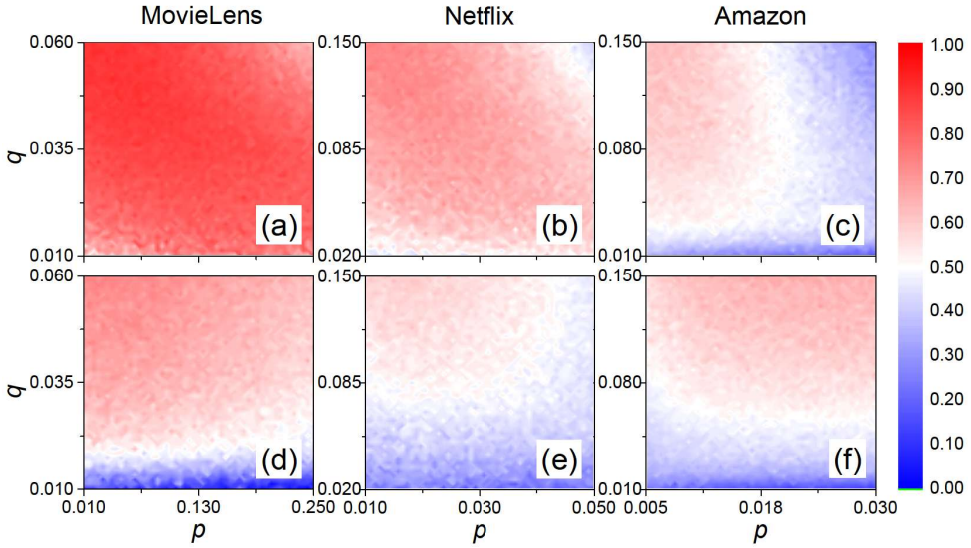

图 4–4 基于群组聚类的信誉排序算法检测作弊评分用户的效果

图4–5给出了GR算法与CR算法在作弊评分用户检测准确性上的对比。其中，图4–5(a-c)是针对恶意型作弊评分用户，图4–5(d-f)是针对随机型作弊评分用户。图中颜色深浅表示两种算法召回率差值的大小，即$\Delta R_c = R_c^{GR} - R_c^{CR}$，检测列表长度设置为$L = d$。可以看到，召回率的差值$\Delta R_c$在大部分区域都大于0，说明GR算法整体上表现更出色。具体而言，GR算法在检测恶意型作弊评分用户上有非常明显的优势，在检测随机型作弊评分用户上优势不突出。另外，当作弊评分用户比例p和他们打分比例q都很小时，两种算法召回率的差值ΔR_c很大，说明GR算法在检测活跃程度较低的作弊评分用户上更有优势。一般而言，不活跃的作弊评分用户不容易被检测出来，这间接体现出GR算法的优势。

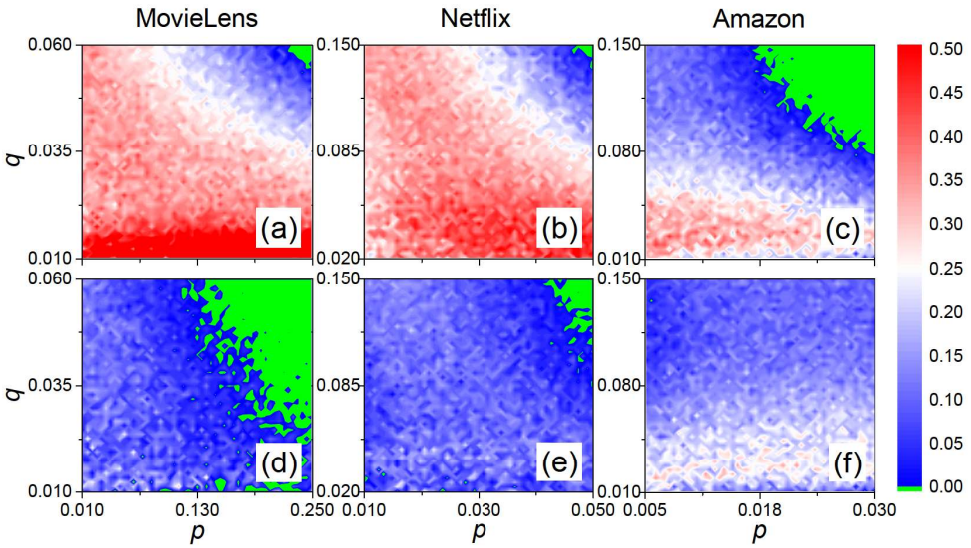

图 4–5 信誉排序算法检测作弊评分用户的效果对比

总结而言，提出的GR算法在检测作弊评分用户上很有优势，也区别于传统信誉排序算法。实验结果显示，GR算法在检测恶意型和随机型作弊评分用户上效果好，准确性和鲁棒性都很高，尤其对于检测不活跃的作弊评分用户。另外，GR算法的效率非常高，时间复杂性为$O(m^2)$，比大部分传统迭代算法都低。最为重要的是，GR算法根据用户评分所形成群组的规模计算用户信誉分数，不依赖产品有唯一质量分数的假设，为评价在线用户信誉提供了全新思路。

4.2 基于迭代过程的群组聚类信誉排序算法

根据评分形成的群组规模来评价用户的可信程度，实际上是考虑用户的评分相似性和从众倾向[265]，这不同于传统信誉排序算法计算用户评分与产品质量之间的偏差或相似性[51, 88]。在基于群组聚类的用户信誉排序算法中，如果用户总是稳定的落入大群组，那么其信誉水平高。然然，算法仅根据评分计算一次用户信

誉分数，用户评分和群组规模是固定的。在已有研究中，很多传统信誉排序算法使用迭代过程改善排序效果，提升算法应对作弊评分用户攻击的鲁棒性。例如，基于相似性的信誉排序算法[51]，以迭代方式更新用户信誉和产品估计质量。

实际上，不同信誉的用户所给出的评分，对产品质量的估计能力不同。高信誉用户给出的评分，更能体现产品质量的实际情况，因为他们的判断更可信。本节研究中将迭代过程引入基于群组聚类的信誉排序算法，同时考虑用户信誉和评分相似性来计算群组规模，利用迭代过程提高算法的排序准确性和鲁棒性。首先，介绍迭代算法提出的动机，阐述传统迭代寻优过程。然后，介绍基于迭代过程的群组聚类信誉排序算法，分析新算法对作弊评分用户的排序效果。最后，介绍不同信誉排序算法的特点，分析用户活跃度和产品流行度对信誉排序的影响。

4.2.1 研究背景与迭代寻优过程

信誉排序算法根据用户的评分行为评价其信誉水平。例如，在基于相似性的排序算法中，用户评分与产品估计质量之间的相似性决定了用户信誉分数[51]。反过来，用户信誉分数又影响用户评分对产品估计质量的计算。用户信誉分数越高，他们的评分在产品质量估计上的作用越大。在反复的迭代过程中，高信誉用户的信誉值逐渐增大，对产品质量估计的影响逐渐提高；低信誉用户的信誉值逐渐减小，对产品质量估计的影响逐渐降低。将迭代寻优过程引入信誉排序算法，有助于提高信誉排序算法的准确性，以及在应对作弊评分用户攻击时的鲁棒性。

迭代寻优过程是排序算法中常用的迭代方法。最初提出和使用迭代寻优过程的目的，是为了提高线性系统中数值求解的准确性[267]。在网络科学研究领域，先后提出了很多基于迭代寻优过程的计算方法。例如，针对超链接网页提出的带有迭代寻优过程的HITS算法[268]，以及针对二部分网络提出的资源分配过程[198, 269]。下面将以这两种算法为例，介绍迭代寻优过程的基本思想和具体算法。

带有迭代寻优过程的HITS算法[268]，是一种基于网络链接的模型，最初被用来识别万维网（WWW）中连到很多不同机构（Authority）的核心页面（Hub）。由于机构页面与核心页面存在互相强化和互相影响的关系，所以需要引入迭代寻优来破解这样的循环。在迭代寻优过程中，同时维护和更新每个页面的权重。具体而言，可以对核心页面和机构页面之间的关系进行如下描述：首先，将核心页面的权重初始化为$x^{(i)}$，将机构页面的权重初始化为$y^{(\alpha)}$。初始化赋值之前，对两个权重都进行归一化处理。然后，更新核心页面i的权重$x^{(i)}$为

$$x^{(i)} \leftarrow \sum_{\alpha \in Y^{(i)}} y^{(\alpha)}. \tag{4–17}$$

其中，$Y^{(i)}$为指向核心页面i的一组机构页面。反过来，可以根据核心页面权重来继续更新机构页面权重。对于机构页面α，其权重$y^{(\alpha)}$被更新为

$$y^{(\alpha)} \leftarrow \sum_{i \in X^{(\alpha)}} x^{(i)}. \tag{4–18}$$

其中，$X^{(\alpha)}$为被机构页面α所指向的一组核心页面。最终，将公式（4–17）和公式（4–18）进行反复的迭代，同时更新核心页面权重$x^{(i)}$和机构页面权重$y^{(\alpha)}$，直到两者的权重都达到稳定值。HITS算法是一个典型的迭代寻优过程，已经被广泛应用于在线社会经济系统排序，例如Google搜索引擎和国际贸易排序[270]。

最原始的资源分配（Resource-Allocation）过程[198, 269]，是另一种常用的迭代寻优过程，等价于网络中从共同邻居出发进行的二步随机游走[271]。以“用户-产品”二部分网络为例，介绍资源分配过程的基本思想。首先，将任意产品α的资源初始化为f_α。然后，将所有产品上已经初始化的资源，按照以下方式进行分配：

$$f'_\alpha = W \cdot f_\alpha. \tag{4–19}$$

其中，f'_α为产品α所最终获得的资源数量；W为资源在二部分网络中的转换矩阵。具体而言，转换矩阵W中的元素$\omega_{\alpha\beta}$通过以下公式计算：

$$\omega_{\alpha\beta} = \sum_{i=1}^{m} \frac{A_{i\alpha} A_{i\beta}}{k_i}. \tag{4–20}$$

其中，$A_{i\alpha}$和$A_{i\beta}$分别为用户i对产品α和产品β的评分；k_i为用户i的度，即所评分产品的个数。实际上，$\omega_{\alpha\beta}$度量了产品α和产品β之间的相似性，也就是来自所有两步路径的贡献之和[198]。资源分配过程能被用来解决很多与网络相关的问题，例如推荐算法的准确性[272]和网络连边的可预测性[271]。

迭代寻优过程已经被用来提高用户信誉排序算法的效果。例如，基于相似性的信誉排序算法[51]以迭代过程不断更新用户信誉和产品估计质量；基于信誉重新分配的排序方法[52]利用迭代过程以非线性的方式重新分配用户信誉和计算产品估计质量。这些传统基于产品质量的信誉排序算法，核心思想是以信誉赋予评分权重，信誉水平不同的用户，其评分对产品质量的估计能力不同。根据类似的思想，自然地将迭代寻优过程引入到基于群组聚类的用户信誉排序算法。信誉水平不同的用户，其评分对群组规模的影响不同：用户信誉水平越高，越能影响群组规模；反过来，用户所属群组规模越大，其信誉水平也越高。

4.2.2 基于迭代的信誉排序算法

基于群组聚类的用户信誉排序GR算法，不再依赖产品有唯一质量分数的假

设，而是受到在线社会平台上用户评分行为有相似性的启发[265]。一方面，质量清晰的产品，大众容易判断，产品收到的评分容易趋于一致。另一方面，高信誉水平的用户，在判断产品质量时有从众心理，容易与其他人的评分保持一致[263]。这样一来，用户评分所形成的群组规模，能反映用户在评分行为上的可信程度。GR算法在计算群组规模时，默认所有用户有相同的贡献，没有考虑用户信誉水平的影响。这意味着，作弊评分用户和高信誉用户，有一样的权重决定群组规模。事实上，根据群组聚类的思想，应当限制信誉水平低的用户对群组规模的影响，增强信誉水平高的用户在计算群组规模时的作用。

根据以上分析，用户群组的规模不仅应当由有相同评分模式的用户数量决定，还应当考虑群组内用户的信誉水平。在计算群组规模时，不同信誉水平的用户贡献能力不同。信誉高的用户，有更大权重决定群组规模；信誉低的用户，对群组规模的影响能力有限。进一步，利用群组规模更新计算用户信誉，总是落入大组的用户信誉水平高，总是偏离大众、落入小组的用户信誉水平低。这样一来，利用迭代寻优过程改进GR算法，反复迭代计算用户信誉分数和用户评分聚类所形成的群组规模，有希望提高群组规模在刻画用户信誉方面的效果。

本节研究中将迭代寻优过程引入GR算法框架，改进得到一种基于迭代过程的群组聚类用户信誉排序（IGR）算法[273]。IGR算法以群组规模为核心刻画用户的不同评分模式，能够更加准确地对用户信誉进行评价和排序。首先，对于离散评分系统，根据评分将所有用户划分成群组。具体而言，将用户i的评分向量A_i映射为产品评分矩阵$B^{(i)}$，矩阵元素$B^{(i)}_{s\alpha}$由以下公式给出：

$$B^{(i)}_{s\alpha}=\begin{cases}1 & \text{if } A_{i\alpha}=\omega_s\\ - & \text{otherwise}\end{cases}. \tag{4–21}$$

其中，$A_{i\alpha}$为用户i对产品α的评分值ω_s；s为不同离散评分的总数；符号“$-$”代表空值，不参与后续运算。然后，根据产品评分矩阵B，计算用户评分所形成的群组Γ。具体而言，对产品α给出相同评分ω_s的用户归属于同一个群组$\Gamma_{s\alpha}$，即

$$\Gamma_{s\alpha}=\{U_i|B^{(i)}_{s\alpha}=1\}. \tag{4–22}$$

其中，$B^{(i)}_{s\alpha}$为用户i的产品评分矩阵中的元素。显然，用户i所归属的群组个数等于用户的度k_i，即所评分产品的总数。同样地，所有给产品α评过分的k_α个用户，归属于s个用户评分组，其中s是不同离散型评分的总数。

然后，计算用户评分所形成群组的规模。与GR算法不同，改进得到的IGR算法在计算用户评分群组规模$\Gamma_{s\alpha}$时，不仅考虑产品评分矩阵$B^{(i)}$，还考虑用户信誉R_i。具体而言，在所有用户初始化相同信誉分数后，例如设置$R_i=1$，同时考

虑用户的群组归属情况和用户的信誉分数，将用户群组规模$\Lambda_{s\alpha}$定义为

$$\Lambda_{s\alpha}=\sum_{i=1}^{m}R_iB_{s\alpha}^{(i)}. \tag{4–23}$$

其中，$B^{(i)}$是用户i的产品评分矩阵，R_i为用户i的信誉分数，m为系统中用户总数。考虑到二部分网络中产品的度分布有异质性，用户评分群组规模的绝对数值不适合比较。所以，将用户评分群组规模矩阵Λ进行列归一化，得到评分回馈矩阵Λ^*，即$\Lambda_{s\alpha}^*=\Lambda_{s\alpha}/k_\alpha$，其中$k_\alpha$为产品$\alpha$的度。然后，将用户评分矩阵$A$根据评分回馈矩阵映射为$\Lambda^*$。具体而言，用户$i$从其评分$A_{i\alpha}$得到的信誉回馈$A'_{i\alpha}$定义为

$$A'_{i\alpha}=\begin{cases}\Lambda_{s\alpha}^* & \text{if } A_{i\alpha}=\omega_s\\ - & \text{otherwise}\end{cases}. \tag{4–24}$$

其中，$A_{i\alpha}$为用户i对产品α的评分，ω_s为相应的评分数值。经过映射以后，信誉回馈矩阵A'相比于原始评分矩阵A更能体现用户的信誉情况。

进一步，引入迭代寻优过程，基于信誉回馈矩阵A'不断计算和更新用户的信誉分数。参考GR算法[265]，根据信誉回馈向量的均值和标准差计算用户信誉分数。具体而言，在迭代过程中将用户i的信誉分数更新为

$$R_i=\frac{\mu(A'_i)}{\sigma(A'_i)}. \tag{4–25}$$

其中，μ和σ分别表示进行均值和标准差运算。用户信誉分数R和用户评分群组规模Λ分别根据公式（4–23）和公式（4–25）迭代更新，直到用户信誉分数的变化小于给定阈值，例如$\Delta=10^{-4}$。具体而言，通过$|R-R'|=\sum_i(R_i-R'_i)^2/m$度量用户信誉的变化程度。其中，$R'$为前一步计算得到的用户信誉，$m$为用户总数。最终，将所有用户根据信誉分数降序排列，排序最低的前L个用户，更有可能是作弊评分用户。注意，当不使用迭代过程时，IGR算法退化为GR算法。

图4–6给出了基于迭代过程的群组聚类用户信誉排序算法示意图。其中，箭头旁边的序号表示算法的运行流程，矩阵中的“-”代表不参与运算的空值。具体而言，图4–6(a)展示了原始评分矩阵A，其行和列分别对应于用户和产品。图4–6(b)展示了用户i的评分矩阵$B^{(i)}$。以用户U_4为例（竖直蓝色线框标注），$B_{5,1}^{(4)}=B_{4,2}^{(4)}=B_{5,4}^{(4)}=1$。图4–6(c)展示了以用户信誉赋权计算得到的群组规模矩阵Λ。以产品O_2为例（横向绿色线框标注），$\Lambda_{4,2}=R_2\times B_{4,2}^{(2)}+R_4\times B_{4,2}^{(4)}=2$。图4–6(d)展示了通过对$\Lambda$列归一化得到的评分回馈矩阵$\Lambda^*$，例如$\Lambda_{4,2}^*=2/(1+2+2)=0.40$。图4–6(e)展示了根据$\Lambda^*$映射评分矩阵$A$得到的信誉回馈矩阵$A'$，例如$A'_{4,2}=0.40$。图4–6(f)展示了所有用户的最终信誉分数向量$R$，例如用户$U_4$的信誉分数$R_4=8.32$。其中，$R'$为前一步计算得到的信誉分数向量，例如初始化

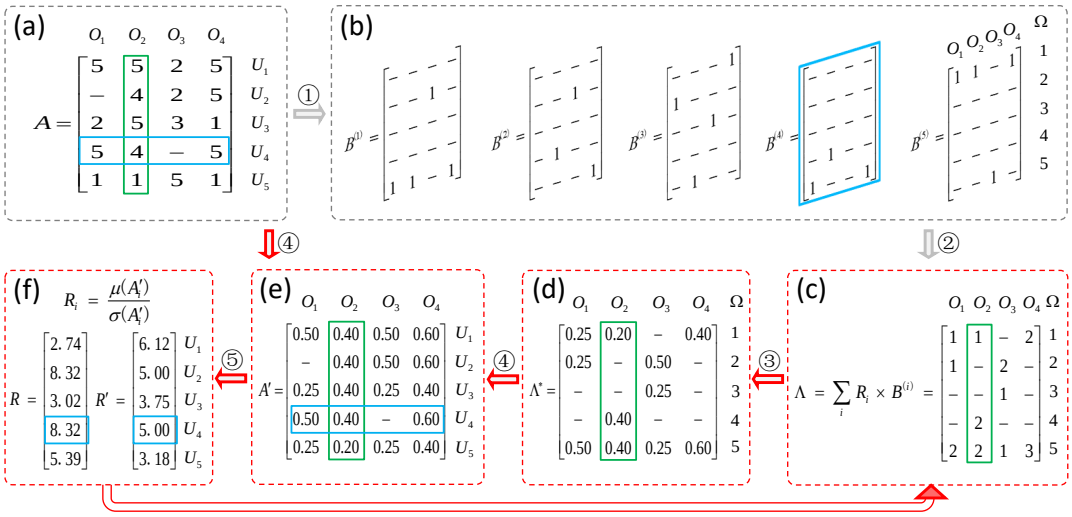

图 4–6 基于迭代过程的群组聚类用户信誉排序算法示意图

时$R'_4 = 5.00$。根据图4–6(c-f)不断迭代更新，计算用户信誉R'和群组规模Λ（红色箭头标注），直到所有用户信誉分数向量R的改变量小于给定的阈值。

4.2.3 算法特点与排序效果分析

使用两个真实在线评分数据集（MovieLens和Netflix），测试和对比信誉排序算法性能。这两个数据集都使用5分制评分：1分表示最差，5分表示最好。具体而言，MovieLens数据集包含943个用户和1,682部电影，共计10万次评分；Netflix数据包含3,000个用户和2,779部电影，共计197,248次评分。比较基于这两个数据集构建的“用户-产品”二部分网络，发现MovieLens网络有更大的用户平均度，更小的产品平均度，更大的网络稀疏度。基于这两个真实评分数据集，构造包含恶意型和随机型作弊评分用户的测试数据集。其中，恶意型作弊用户每次以等概率随机打1分或5分，随机型作弊用户每次从1分到5分随机打分。作弊用户数量为d，占所有用户的比例为$p = d/m$。构造测试数据集的细节，参见本章第4.1.3节。

在评价信誉排序算法效果时，首先分析算法得到的信誉分数对用户的区分能力。IGR算法的对比算法包括：IR算法[88]、CR算法[51]、RR算法[52]和原始的GR算法[265]。图4–7给出了不同信誉排序算法得到的用户信誉分数分布情况。其中，图4–7(a-e)是基于MovieLens数据集计算得到的结果，图4–7(f-j)是基于Netflix数据集计算得到的结果。可以看到，使用IR算法计算得到的用户信誉分数呈现类泊松分布；使用CR算法、GR算法和IGR算法计算得到的用户信誉分数呈现类正态分布；使用RR算法计算得到的用户信誉分数呈现类指数分布，如图4–7(c)和(e)所示，很多用户的信誉分数为0。为了量化用户信誉分数分布的区分程度，计算辛普森（Simpson）多样性指标[274]，记为$1 - D$，数值越大代表信誉分数分布的区

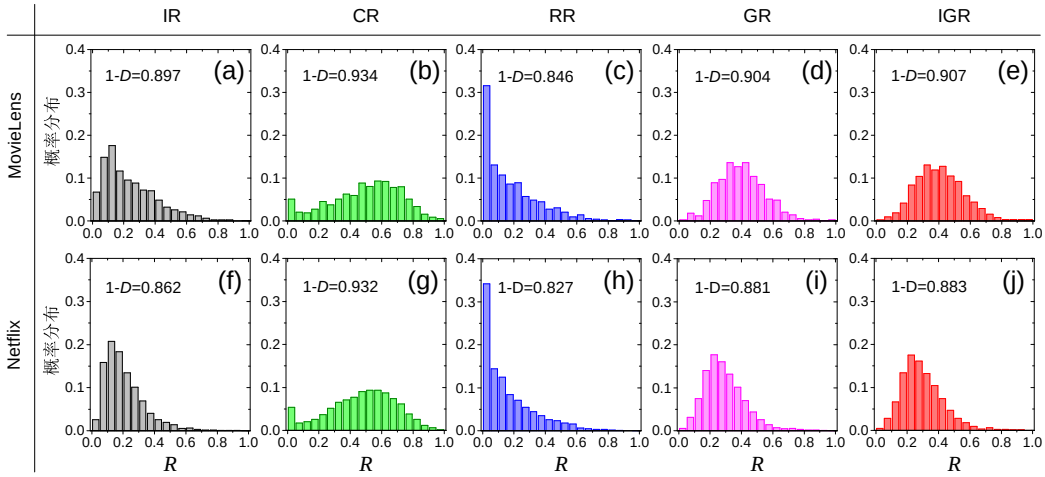

图 4–7 不同排序算法计算得到所有用户的信誉分布情况

分程度越大。计算结果显示，CR算法对应的$1-D$数值最大，GR算法和IGR算法对应的$1-D$数值相近，RR算法对应的$1-D$数值最小。总的来说，使用CR算法、GR算法和IGR算法计算得到的信誉分数对用户的区分能力更强。

进一步，验证信誉排序算法的自恰性。直观上，如果用户评分与产品估计质量的偏差越大，那么用户的信誉应该越低。所以，一个设计合理的信誉排序算法得到的用户信誉分数R，应当与公式（4–2）给出的用户评分偏差δ呈现很强的负关联。两者之间的关联性越强，说明信誉排序算法的自恰性越好。图4–8(a-b)分别展示了不同排序算法在MovieLens和Netflix数据集上得到的用户信誉分数R与用户评分偏差δ之间的关系。可以看到，基于群组聚类的GR算法和IGR算法，为评分偏差小的用户稳定地分配高信誉分数，为评分偏差大的用户稳定地分配低信誉分数；当用户评分偏差大时，其他三种传统的基于产品质量的排序算法（IR算法、CR算法和RR算法）结果波动很大。表4–3第一行给出了R与δ之间的皮尔森相关系数$\rho(\delta, R)$。可以看到，IGR算法对应的相关系数在MovieLens（ML）和Netflix（NT）数据集上分别为-0.817和-0.820，负关联程度都强于GR算法所给出的结果，说明IGR算法有更好的自恰性。另外，IGR算法和GR算法对应的关联性强于IR算法、CR算法和RR算法，说明他们的自恰性相对最好。

在排序算法的特点方面，首先分析用户活跃程度对信誉排序的影响。其中，用户活跃程通过用户度k估计。图4–8(c-d)展示了用户信誉分数R与用户度k之间的关系，表4–3第二行给出了对应的皮尔森关联系数。可以看到，IR算法得到的R与用户度k存在显著的正关联，说明IR算法偏好活跃程度高的用户。其他四种排序算法没有显著的用户度偏好，因为R与k的关联性几乎为0。然后，分析用户评分热门产品的程度对信誉排序的影响。追热程度通过用户评分产品的平均度$\phi = \sum_{\alpha \in O_i} k_\alpha / k_i$来估计。其中，$O_i$为用户$i$评分的产品，$k_i$为用户$i$的度，$k_\alpha$为产

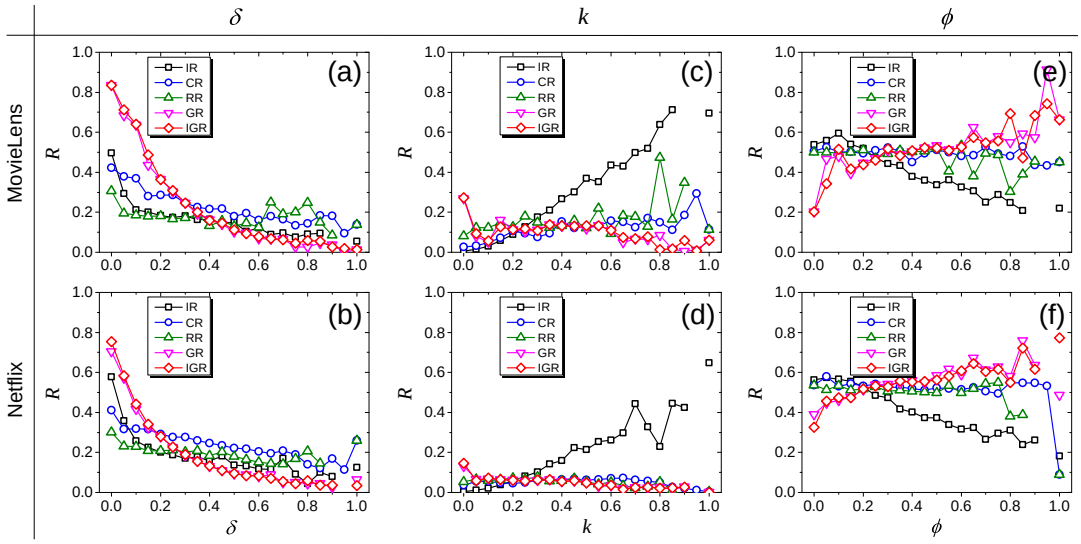

图 4–8 用户信誉与评分偏差、度和趋势度之间的关系

品α的度。图4–8(e-f)展示了用户信誉分数R与追热程度ϕ之间的关系，表4–3第三行给出了对应的皮尔森关联系数。可以看到，IR算法得到的R与追热程度ϕ存在显著的负相关；CR算法和RR算法得到的R与ϕ几乎不相关；GR算法和IGR算法得到的R与ϕ存在微弱的正相关。这些结果说明，IR算法偏好活跃的、不追热门产品的用户；CR算法和RR算法没有用户活跃程度和追热程度的偏好；GR算法和IGR算法不受用户活跃程度影响，但稍微偏好追热门产品的用户。

表 4–3 用户信誉与评分偏差、活跃程度和追热程度之间的皮尔森关联系数

相关	ML	ML	ML	ML	ML	NT	NT	NT	NT	NT
系数	IR	CR	RR	GR	IGR	IR	CR	RR	GR	IGR
$\rho(\delta,R)$	−0.447	−0.454	−0.319	−0.817	−0.820	−0.464	−0.393	−0.281	−0.735	−0.763
$\rho(k,R)$	0.876	0.232	0.172	−0.052	−0.042	0.787	0.054	0.004	−0.095	−0.090
$\rho(\phi,R)$	−0.475	−0.024	−0.029	0.214	0.205	−0.379	−0.043	−0.057	0.237	0.216

最后，利用含有两种（恶意型和随机型）作弊评分用户的测试数据集，分析信誉排序算法对作弊评分用户的检测效果。评价算法排序效果时，使用召回率$R_c(L)$指标[239]和AUC指标[224]。其中，在计算召回率$R_c(L)$时，检测列表长度设置为$L=d$，即人工生成作弊评分用户的数量。$R_c(L)$指标仅关注排序靠前的L个用户，数值越大表示信誉排序越准确。AUC指标关注用户整体排序情况，随机情况下的AUC值为0.5。所以，AUC值超过0.5的程度，表示信誉排序准确性的大小。测试数据集中，作弊评分用户的比例为$p=d/m$，其中m为用户总数量。针对每种参数设定，进行50次独立重复实验，计算50次结果的平均值作为最终结果。

图4–9给出了信誉排序算法应对恶意型作弊评分用户的排序准确性。其中，图4–9(a-b)对应于R_c评价指标。可以看到，当p较大时，IGR算法比GR算法的排序

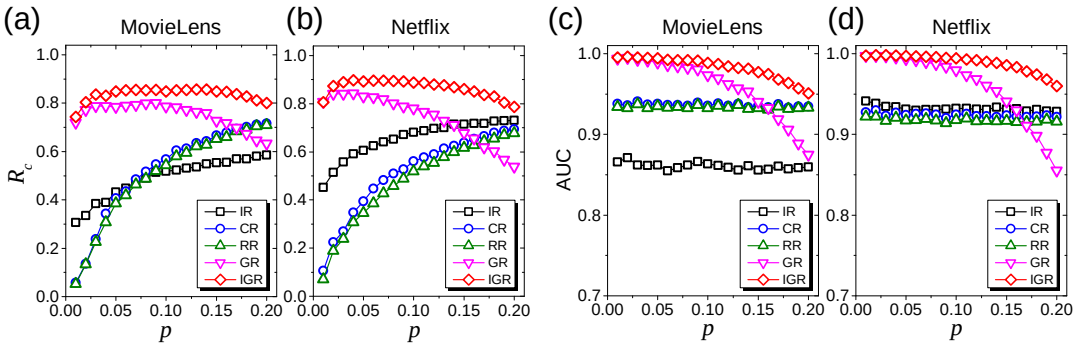

图 4–9 信誉算法应对恶意型作弊评分用户的排序准确性

准确性更高、鲁棒性更强；当p相对较小时，IGR算法和GR算法比其他三种算法的排序准确性更高。CR算法和RR算法有类似的表现，排序准确性R_c随p的增大而增大。IR算法的效果依赖于数据集，但总体表现优于CR算法和RR算法。图4–9(c-d)对应于AUC评价指标。可以看到，IGR算法给出的整体排序更准确，AUC值达到0.95左右；当p较大时，IGR算法比GR算法的鲁棒性更强；CR算法和RR算法的鲁棒性也不错，AUC维持在0.92左右；IR算法的排序准确性依赖于数据集，在Netflix数据集上表现更好。总体而言，在应对恶意型作弊评分用户时，IGR算法和GR算法比传统算法的排序效果好，IGR算法比GR算法的鲁棒性强。

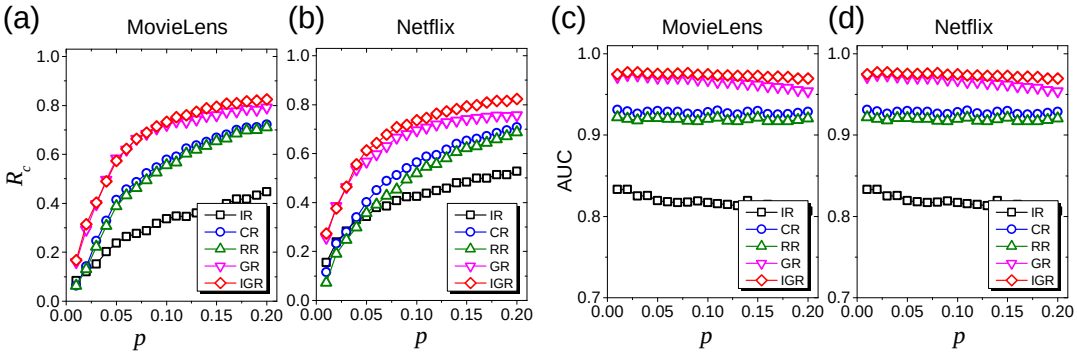

图 4–10 信誉算法应对随机型作弊评分用户的排序准确性

图4–10给出了信誉排序算法应对随机型作弊评分用户的排序准确性。其中，图4–10(a-b)对应于R_c评价指标。可以看到，IGR算法和GR算法比其他算法的排序准确性都更高，尤其当p较大时，IGR算法比GR算法的鲁棒性更强。CR算法和RR算法的效果相当，都比IR算法表现好。召回率R_c随着p的增大而增大。具体而言，当p从0逐渐接近0.05时，R_c迅速增大；当p超过0.05并继续增大时，R_c的增大幅度减小。这些结果暗示，原始数据集中可能含有大约5%的真实随机型作弊评分用户。图4–10(c-d)对应于AUC评价指标。可以看到，GR算法和IGR算法的排序准确性优于其他算法，AUC值在0.96附近；CR算法和RR算法的排序准确性稍逊色，AUC值在0.92附近；IR算法的排序准确性最差。总体而言，在应对随机型作

弊评分用户时，基于群组聚类的两种算法比传统算法的排序效果好。特别地，当随机型作弊用户比例较大时，IGR算法比GR算法表现出更强的鲁棒性。

4.3 基于复杂网络结构特征的推荐排序算法

电子商务平台提供数以亿计的产品和服务，一方面，消费者几乎无可避免地面对信息过载的困扰[258]。另一方面，商家也面临棘手的问题，包括如何服务不同需求的客户、进行个性化推送和提高用户满意度等。推荐系统的广泛应用给用户和商家都带来了好处[55, 259]，很多在线平台已经采用个性化推荐系统[55, 275]，例如Amazon网站推荐图书[276]、AdaptiveInfo网站推荐新闻[277]、Netflix网站推荐电影[278]等。推荐系统的核心是推荐算法，在分析历史购买和产品评分记录的基础上，推荐算法给用户推荐可能喜欢和购买的产品。

在线平台上用户对产品的购买和评分，可以通过“用户-产品”二部分网络描述[53]，网络中每条连边代表购买或评分关系。针对每个用户，个性化推荐算法首先对系统中的所有产品进行排序，然后根据排序结果给用户推荐最合适的产品。如何利用二部分网络的结构特征或借助社会关系等辅助信息，提高推荐算法的排序结果准确性和推荐产品的多样性是一个非常关键的问题[272]。从网络的角度来看，给用户推荐产品相当于建立用户与产品的连边，是预测二部分网络中缺失连边的链路预测问题[279, 280]，本质上也是社会经济系统中的排序问题。

本节研究中使用来自在线社会经济系统中的真实数据集，将相互作用关系抽象成复杂网络模型，从两个方面研究复杂网络结构特征在提高排序效果上的作用。首先，介绍一种新的计算网络中节点之间相似性的方法，基于此提出一种基于节点相似性的个性化推荐算法，对用户可能喜欢的产品进行预测排序。然后，介绍社会信任关系对用户选择产品的影响，提出一种综合考虑社会信任关系和网络节点相似性的个性化推荐算法，借助信任关系进一步提升对产品的排序效果。

4.3.1 利用节点相似性实现个性化排序

针对在线评分系统，已有研究提出了一系列推荐算法[55, 275]。其中，基于用户的协同过滤（UCF）和基于产品的协同过滤（ICF）是最具代表性的两种推荐算法[281]，分别利用相似用户对产品的选择和产品间的相似性来计算测量值[282]。近年来，一些研究将动力学过程引入到推荐系统，提出了很多基于网络上扩散过程的推荐算法[283]，包括热传导（HC）[284]和物质扩散（MD）[285]等。受这些方法的启发，后续研究提出了不同初始条件和不同混合方式的推荐算法。例如，Zhou等人[286]利用度依赖的资源初始化分配策略，提出了一种新的网络推荐算法；Jia等

人[287]利用资源接收节点的影响力对MD算法进行改进，提出了一种基于节点影响力的推荐算法；Zhou等人[272]将HC算法和MD算法相结合，提出了一种混合推荐算法；Liu等人[288]考虑网络连边权重的影响，提出了一种含权的HC算法。

基于协同过滤和基于网络扩散的推荐算法，都依赖于网络中节点相似性的计算[289]。基于协同过滤的算法常用余弦相似性（Cosine）[290]，倾向于推荐热门产品，虽然推荐结果准确，但缺乏多样性。基于网络扩散的推荐算法利用资源分配（RA）指数[269]刻画节点相似性，倾向于为大度节点分配更多资源，导致推荐结果的多样性不足。实际上，余弦相似性和资源分配指数在一定程度上互补，结合两者将有希望提高整体的推荐效果[291]。如何设计一种适用于推荐系统的相似性指标是一个重要问题，节点相似性本身也能刻画网络的很多结构特征[292]。

首先介绍推荐系统的基本符号表示，然后简介余弦相似性和资源分配指数，最后提出一种新的网络节点相似性CosRA指标[293]。推荐系统应用于在线评分系统$G(U,O,E)$，其中$U=\{U_1,U_2,\ldots,U_m\}$为用户集合，$O=\{O_1,O_2,\ldots,O_n\}$为产品集合，$E=\{E_1,E_2,\ldots,E_z\}$为连边集合。自然地，将二部分网络G表示为邻接矩阵A。如果用户U_i购买过产品O_α，则$A_{i\alpha}=1$；否则，$A_{i\alpha}=0$。推荐算法为每个用户提供一个推荐列表O_i^L，包含L个对用户i而言推荐分数最高的产品。

下面介绍Cosine和RA指数的计算方法。产品α和产品β之间的余弦相似性为

$$S_{\alpha\beta}^{Cos}=\frac{1}{\sqrt{k_\alpha k_\beta}}\sum_{i=1}^{m}A_{i\alpha}A_{i\beta}. \tag{4–26}$$

其中，k_α和k_β分别为产品α和产品β度。资源分配过程等同于“用户-产品”二部分网络中从共同邻居出发的二步随机游走[271]。产品α和产品β之间的RA指数为

$$S_{\alpha\beta}^{RA}=\sum_{i=1}^{m}\frac{A_{i\alpha}A_{i\beta}}{k_i}. \tag{4–27}$$

其中，k_i为用户i的度。实际上，资源分配指数是简单版物质扩散过程的转移矩阵中的元素[198]。将Cosine和RA指数相结合，提出一种新的网络节点相似性CosRA指标。具体而言，将产品α和产品β之间的相似性CosRA指标定义为

$$S_{\alpha\beta}^{CosRA}=\frac{1}{\sqrt{k_\alpha k_\beta}}\sum_{i=1}^{m}\frac{A_{i\alpha}A_{i\beta}}{k_i}. \tag{4–28}$$

实际上，$S_{\alpha\beta}^{CosRA}$在测量产品α和产品β之间的相似性时，同时考虑二部分网络中两类节点的度，将网络中所有二步路径的贡献相加。

进一步，提出一种基于相似性CosRA指标的个性化推荐算法[293]。首先，对于任意用户i而言，将产品α的资源初始化为

$$f_\alpha^{(i)}=A_{i\alpha}. \tag{4–29}$$

其中，如果用户i购买过产品α，则$A_{i\alpha}=1$；否则，$A_{i\alpha}=0$。然后，将所有产品上的资源通过转移矩阵进行重分配：

$$f'^{(i)}=S^{CosRA}\cdot f^{(i)}. \tag{4–30}$$

其中，$f^{(i)}$为所有产品的初始资源向量，$f'^{(i)}$为所有产品的最终资源向量。最后，将所有产品按照最终资源$f'^{(i)}$排序，把排序最靠前的、用户没有购买过的L个产品推荐给用户i。图4–11给出了基于网络节点相似性CosRA的推荐算法示意图。

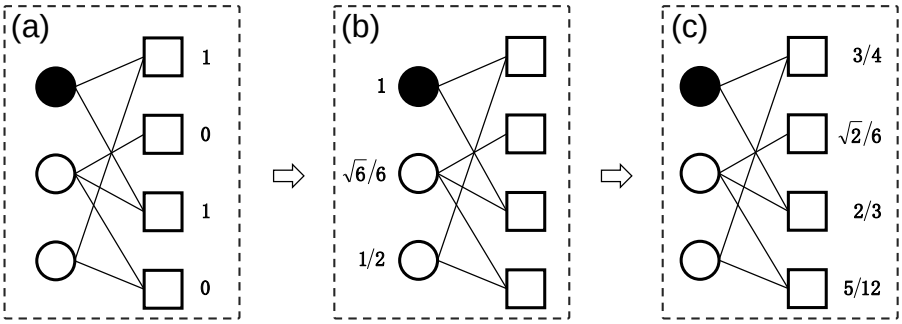

图 4–11 基于网络节点相似性CosRA的推荐算法示意图

在测试推荐算法性能时，使用四个真实在线评分数据集。其中，MovieLens-100K，MovieLens-1M和Netflix为电影评分数据集，采用5分制评分体系。RYM为音乐评数据集分，采用10分制评分体系。构建“用户-产品”二部分网络的连边时，电影评分数据集保留不小于3分的评分，音乐评分数据集保留不小于6分的评分。在分析中，二部分网络看做无权和无向网络，忽略用户的评分数值。表4–4给出了四个数据集的基本统计特征。可以看到，MovieLens-100K数据集对应的二部分网络稀疏度最大，RYM数据集对应的二部分网络稀疏度最小。

表 4–4 四个真实在线评分数据集的基本统计特征

数据集	用户总数	产品总数	评分数量	网络稀疏度
MovieLens-100K	943	1574	82520	5.56×10^{-2}
MovieLens-1M	6039	3628	836478	3.82×10^{-2}
Netflix	10000	5640	701947	1.24×10^{-2}
RYM	33197	5234	609792	3.51×10^{-3}

评价基于节点相似性CosRA的推荐算法时，对比分析基于物质扩散MD的推荐算法[285]和基于热传导HC的推荐算法[269]。采用几种常用的推荐算法评价指标[55, 294]，包括四种准确性指标（AUC，MAP，准确率和召回率），一种多样性指标（内相似度）和一种新颖性指标（流行性）。使用AUC指标[224]评价全局排序准确性，AUC超过0.5（随机情况）的程度表示排序准确性的大小。AUC指标不依赖于推荐列表长度L，其他三种准确性指标受L影响。MAP指标[295]是评价整体排序

准确性的指标，MAP值越大表示排序结果越准确。准确率（P）指标[55]定义为出现在测试集中被推荐的产品数量相对于所有被推荐产品数量的比例，准确率越大代表算法越准确。召回率（R）指标[55]定义为出现在推荐列表中的被推荐产品数量相对于所有测试集中产品数量的比例，召回率越大代表算法越准确。内相似度（I）指标[296]定义为推荐列表中产品之间的余弦相似性，内相似度越低代表算法推荐产品的多样性越好。流行性（N）指标[55]评价算法推荐不流行产品的能力，流行性越低表示算法推荐产品的新颖性越好。

在评价推荐算法效果时，采用十折交叉验证（10-Folder Cross-Validation）策略，将每个数据集划分成评分数量相等的10份，将其中9份作为训练集，剩余1份作为测试集。在10轮计算中每份数据都作1次测试集，将10轮计算结果平均，得到1次实验结果。实验中，设置推荐列表长度为$L=50$，对每个算法在数据集上重复10次实验，计算评价指标的平均值。表4–5给出了算法在四个数据集上的推荐效果。在准确性方面，CosRA算法在所有四个数据集上都表现最好。CosRA算法给出的AUC、MAP、准确率P和召回率R的数值都最大，推荐准确性显著优于HC算法和MD算法。在多样性和新颖性方面，CosRA算法表现稍微优于MD算法，但逊色于HC算法。考虑到不容易平衡推荐算法的所有性能[272]，CosRA算法已经很大程度上提高和平衡了推荐结果的准确率、多样性和新颖性。

表 4–5 推荐算法在四个真实在线评分数据集上的推荐效果

数据集	推荐算法	AUC	MAP	P	R	I	N
ML-100K	HC	0.842	0.037	0.021	0.123	**0.056**	**23**
	MD	0.898	0.325	0.075	0.527	0.355	230
	CosRA	**0.908**	**0.380**	**0.082**	**0.575**	0.335	204
ML-1M	HC	0.881	0.052	0.034	0.162	**0.045**	**198**
	MD	0.885	0.188	0.066	0.297	0.403	1618
	CosRA	**0.895**	**0.223**	**0.074**	**0.350**	0.387	1541
Netflix	HC	0.889	0.002	0.001	0.024	**0.004**	**15**
	MD	0.948	0.207	0.048	0.426	0.368	2369
	CosRA	**0.950**	**0.229**	**0.051**	**0.449**	0.361	2298
RYM	HC	0.933	0.130	0.014	0.361	**0.057**	**214**
	MD	0.941	0.209	0.018	0.471	0.155	1089
	CosRA	**0.952**	**0.292**	**0.019**	**0.482**	0.144	819

进一步，分析CosRA算法的工作机理，即在产品推荐上的特点。图4–12给出了不同算法所推荐产品的度分布。可以看到，MD算法有很大的概率推荐度大的产品，推荐结果有不错的准确性，但缺乏多样性和新颖性；HC算法倾向于推荐小

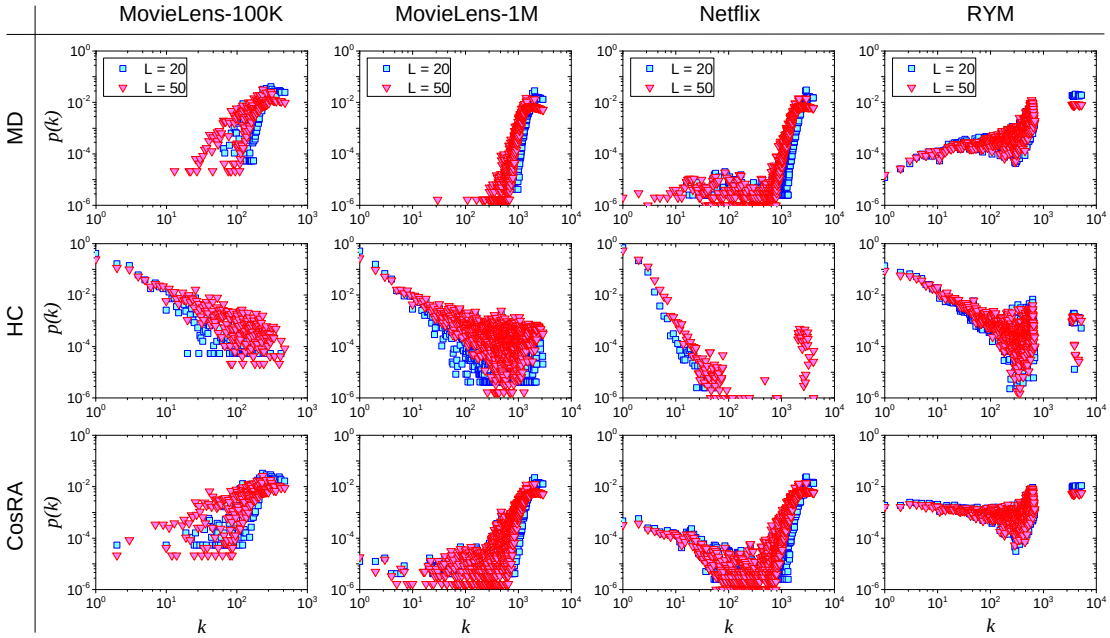

图 4–12 推荐算法在真实评分数据集上所推荐产品的度分布

度产品，推荐结果有很好的多样性和新颖性，但准确性不高。总的来说，MD算法和HC算法都有很强的趋势性，导致最终的推荐效果不理想。提出的CosRA算法在推荐过程中平衡了度大和度小的产品，没有明显的产品度偏好，所以平衡了推荐结果的准确率、多样性和新颖性，算法整体的推荐效果更好。

分析相似性对推荐效果的影响时，引入两个可调参数（η_1和η_2）将CosRA指标泛化为CosRA*指标。具体而言，产品α和产品β之间的相似性定义为

$$S_{\alpha\beta}^{CosRA^*}=\frac{1}{(k_\alpha k_\beta)^{-\eta_2}}\sum_{l=1}^{m}\frac{a_{l\alpha}a_{l\beta}}{(k_l)^{-2\eta_1}}. \tag{4–31}$$

利用泛化的CosRA*指标，构建CosRA*推荐算法。当$\eta_1=\eta_2=-0.5$时，CosRA*算法退化为CosRA算法。图4–13给出了CosRA*算法的推荐效果。当参数η_1和η_2都在0.5附近时，准确性指标（AUC，MAP，P和R）和多样性指标（I）达到最大值，新颖性指标（N）达到最小值。这些结果在四个数据集上保持一致，说明CosRA*算法存在普适的最优参数组合，即$\eta_1=\eta_2=-0.5$（CosRA*算法退化为CosRA算法）。换句话说，原始CosRA算法的推荐效果已经达到最优，调整两个参数不能显著提高CosRA*算法的推荐效果。

总结而言，提出了新的计算网络中节点相似性的CosRA指标，很好的结合了余弦相似性和物质扩散指数的优势。进一步，基于CosRA指标提出了一种新的CosRA推荐算法。在四个真实评分数据集上的测试结果表明，CosRA算法比基准算法表现出更好高的推荐结果准确性、不错的多样性和新颖性。对比不同算法所推荐产品的度分布，发现CosRA算法没有明显的产品度偏好，同时推荐度大和

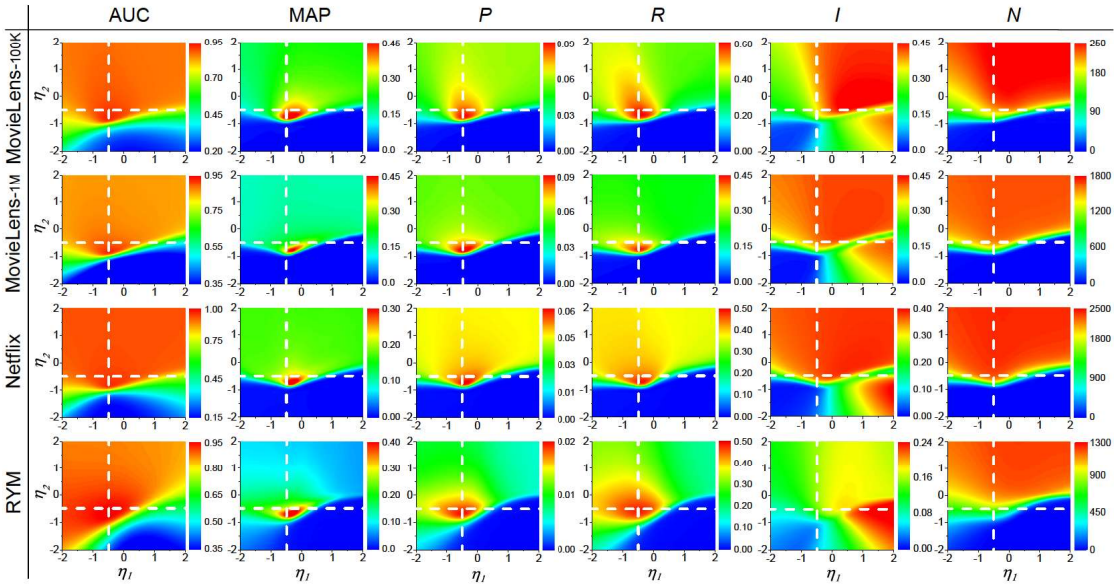

图 4–13 基于节点相似性CosRA的泛化推荐算法的推荐效果

度小的产品，平衡了推荐结果的准确性、多样性和新颖性。最后，将CosRA指标扩展到一般的形式，发现原始的CosRA指标已经在推荐效果上达到最优。这意味着，CosRA指标无需可调参数，在实际应用中更有优势。

4.3.2 借助社会信任关系改善排序效果

传统的个性化推荐算法往往只考虑用户对产品的评分信息，忽略了用户之间存在的社会关系，如信任关系[297]。实际上，用户对产品的选择和对信息的采纳，很多时候也受社会关系的影响[298, 299]，尤其是来自用户信任的朋友的推荐。近年来，一些研究已经将用户之间的信任关系纳入推荐算法的设计框架[261, 300]。例如，Jamali等人[301]结合了信任关系和基于产品的协同过滤，提出一种基于随机游走的推荐算法；Ma等人[302]融合了用户和信任朋友的偏好，提出了一种社会信任集成的推荐算法；Shen等人[303]提出了两个用户信任模型，实现了信任集成的基于用户的协同过滤推荐算法。最近，Wang等人[57]将信任关系引入基于扩散过程的推荐算法框架，分析了信任关系对推荐效果的影响。

受到这些研究的启发，以CosRA算法[293]为基础，将用户信任关系引入资源分配过程[198, 272]，提出一种基于信任关系的CosRA+T推荐算法[304]，分析信任关系对推荐效果的影响。算法设计过程中，不但考虑用户对产品评分的二部分网络的邻接矩阵A，还考虑用户之间的信任关系网络的邻接矩阵$B=(b_{i,j})_{m\times m}$。其中，$b_{ij}=1$，表示存在从节点i到节点j的信任连边；$b_{ij}=0$，表示不存在从节点i到节点j的信任连边。图4–14同时展示了“用户-产品”二部分网络和用户信任网络。

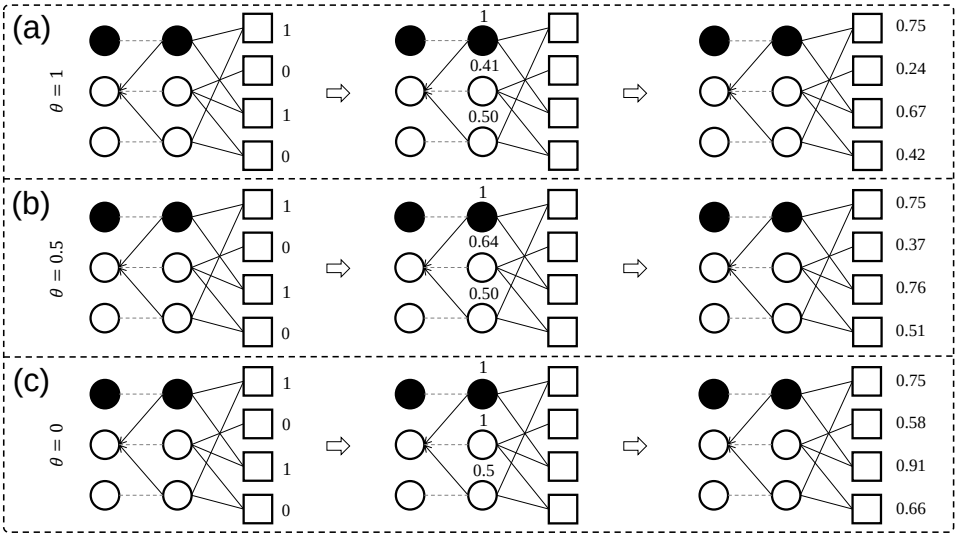

图 4–14 基于信任关系的CosRA+T推荐算法示意图

其中，圆圈和方块分别表示用户和产品；同一行的两个圆圈表示同一个用户，信任关系以跨越不同行的连边体现，表示中间列用户信任左侧列用户。

基于信任关系的CosRA+T推荐算法，分为以下三个步骤：首先，对于给定的目标用户i，为产品α分配初始化资源：

$$f_{\alpha}^{(i)} = A_{i\alpha}. \tag{4–32}$$

其中，如果用户i对产品α进行过评分，则$A_{i\alpha}=1$；否则，$A_{i\alpha}=0$。然后，用户i的所有连接节点，接收从二部分网络中分配过来的、他们评分过产品上的资源。具体而言，对于目标用户i，其邻居用户j收到的资源为

$$f_{j}^{(i)} = \sum_{\alpha=1}^{n} \frac{a_{j\alpha}}{\sqrt{k_j k_\alpha}} f_{\alpha}^{(i)}. \tag{4–33}$$

其中，k_α为产品α的度；k_j为用户j的度；n为产品总数。最后，考虑目标用户i对其他用户的信任关系，将其资源重新分配给产品。如果用户i信任用户j，那么在重新分配之前将用户j的资源总量以参数θ进行标度运算；否则，用户j的资源直接分配给产品。具体而言，对于目标用户i，产品β所最终接收到的资源总量为

$$f'^{(i)}_{\beta} = \sum_{j=1}^{m} \frac{a_{j\beta}}{\sqrt{k_j k_\beta}} (b_{ij} f_{j}^{\theta(i)} + (1-b_{ij}) f_{j}^{(i)}). \tag{4–34}$$

其中，如果用户i信任用户j，则$b_{ij}=1$；否则，$b_{ij}=0$；θ为可调参数；m为用户总数。将用户按资源总量$f'^{(i)}$降序排列，把前L个未被购买的产品推荐给用户i。

图4–14给出了基于信任关系的CosRA+T推荐算法示意图。为了研究信任关系对推荐效果的影响，引入可调标度参数θ。当参数θ从0逐渐增加到1时，CosRA+T算法中信任关系的影响逐渐减弱到消失。具体而言，图4–14(a)展示

表 4–6 在线评分和信任关系数据集的基本统计信息

数据集	用户总数	产品总数	评分总数	稀疏度（S_R）	信任总数	稀疏度（S_T）
Epinions	4,066	7,649	154,122	4.96×10^{-3}	217,071	1.31×10^{-2}
FriendFeed	4,148	5,700	96,942	4.10×10^{-3}	386,804	2.25×10^{-2}

了$\theta=1$的情况，信任关系不影响资源分配；图4–14(b)展示了$\theta=0.5$的情况，受信任用户的资源量被开根号，在重新分配之前增加；图4–14(c)展示了$\theta=0$的情况，在重新分配之前，受信任用户的资源变为1，非受信任用户的资源不变。

使用两个包含用户评分和信任关系的真实数据集测试算法性能。其中，Epinions为用户对产品的评论，FriendFeed为用户对书签的评分，两者都使用5分制评分体系。仅考虑评分不小于3分的情况，基于评分数据构建“用户-产品”二部分网络，即评分网络。另外，基于数据集中的信任关系构建用户之间的信任网络。表4–6给出了两个在线评分和信任关系数据集的基本统计信息。可以看到，Epinions数据集对应的评分网络的稀疏度更大，FriendFeed数据集对应的信任网络的稀疏度更大。为评价推荐算法效果，沿用本章第4.3.1节介绍的准确性指标：AUC、准确率（P）和召回率（R），多样性指标：内相似度（I），以及新颖性指标：流行度（N）。另外，增加两种准确性指标：Ranking Score（RS）和F1指标（F_1）。其中，Ranking Score指标[198]评价相关产品在推荐列表中的相对排序，数值越小表示排序结果越准确；F1指标[239]依赖于推荐列表长度L，评价算法对准确率和召回率的平衡情况，数值越大表示排序结果越准确。

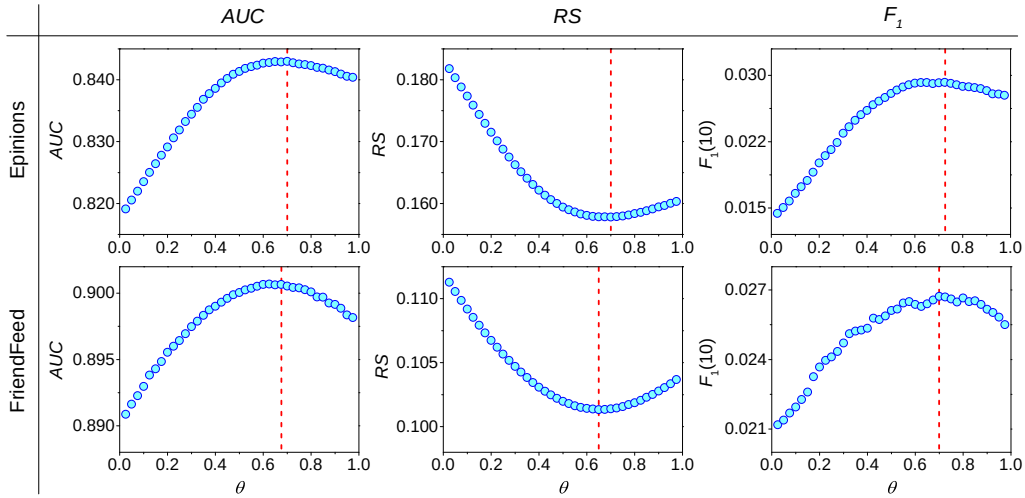

图 4–15 信任关系对CosRA+T算法推荐准确性的影响

图4–15展示了参数θ对CosRA+T算法推荐准确性的影响。在CosRA+T算法中存在可调参数θ，控制用户信任关系的影响程度：θ越大，则信任关系的影响越小。在两个数据集上分别使用了三种准确性指标：AUC，RS和$F_1(L)$。其中，推荐列

表长度设置为$L=10$。可以看到，在不同的准确性指标上，参数θ都存在一个最优值，且两个数据集上结果保持一致。具体而言，在最优值θ^*附近，AUC和$F_1(L)$指标达到最大值，RS指标达到最小值。在Epinions数据集上，最优值θ^*在0.70附近；在FriendFeed数据集上，最优值θ^*在0.65附近。

为了分析最优参数θ^*的普适性，增加考虑两个依赖于推荐列表长度L的准确性指标：准确率$P(L)$和召回率$R(L)$。图4–16展示了两个数据集上最优参数θ^*随L的变化情况。可以看到，最优参数θ^*对L的变化不敏感：Epinions数据集上θ^*在0.70附近，FriendFeed数据集上θ^*在0.65附近。采用10折交叉验证方法进行20轮实验，对每种准确性指标（AUC、RS、P、R和F_1）计算最优参数值的平均值。从图4–16(c)和(f)看到，根据不同指标算出的最优参数的平均值都非常接近，在Epinions数据集上为$\langle\theta^*\rangle\approx 0.70$，在FriendFeed数据集上为$\langle\theta^*\rangle\approx 0.65$。这些分析结果，一方面确认CosRA+T算法中的标度参数θ存在普适的最优值θ^*，另一方面说明引入信任关系能提高推荐效果，但过分依赖信任关系有负面影响。

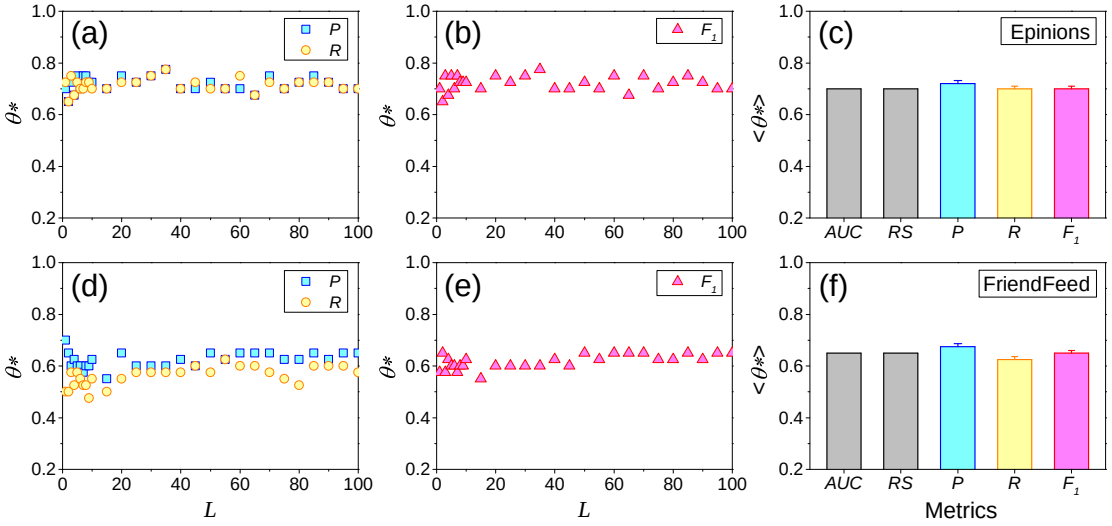

图 4–16 基于信任关系的CosRA+T算法中最优标度参数分析结果

针对不同数据集设置CosRA+T算法的最优参数，与其他三种基准算法（HC算法、MD算法和CosRA算法）对比推荐效果，分析信任关系的作用。推荐表4–7给出了不同算法在两个数据集上的推荐效果。列表长度设置为$L=10$，实验结果为重复20次实验的平均值。在准确性方面，CosRA+T算法在两个数据集上的表现都最好。具体而言，CosRA+T算法极大的优于HC算法和MD算法，对CosRA算法的准确性也有显著的提高。在多样性和新颖性方面，CosRA+T算法优于MD算法，但稍逊色于CosRA算法；效果最好的是HC算法。总体来说，CosRA+T在推荐准确性方面更有优势，也在兼顾推荐结果多样性和新颖性方面表现不错。

综合以上分析结果，发现引入信任关系能显著地提高推荐算法对产品的排序

表 4–7 推荐算法在两个用户评分和信任关系数据集上的推荐效果

数据集	推荐算法	AUC	RS	P	R	F_1	I	N
Epinions	MD	0.8256	0.1735	0.0189	0.0590	0.0286	0.1140	235
	HC	0.7845	0.2161	0.0052	0.0153	0.0077	**0.0245**	**5**
	CosRA	0.8356	0.1641	0.0221	0.0629	0.0327	0.0900	107
	CosRA+T	**0.8382**	**0.1616**	**0.0226**	**0.0651**	**0.0335**	0.0917	101
FriendFeed	MD	0.8925	0.1077	0.0163	0.0683	0.0263	0.1195	73
	HC	0.8833	0.1182	0.0088	0.0370	0.0142	**0.0542**	**11**
	CosRA	0.8978	0.1028	0.0167	0.0633	0.0265	0.0890	35
	CosRA+T	**0.9007**	**0.1000**	**0.0175**	**0.0693**	**0.0280**	0.1008	35

效果，能更好的平衡推荐的准确性、多样性和新颖性。但是过分依赖信任关也会带来负面的影响，导致排序效果反而下降。既考虑网络节点的相似性，又适当考虑用户之间的信任关系，能够最大程度上提高推荐算法的排序效果。

4.4 本章小结

社会经济系统主体之间存在复杂的相互作用，导致推断社会经济系统状态时，不仅要关注个体自身行为特征，还要考虑系统中群体的相对表现。利用复杂网络对相互作用进行建模，在分析网络结构的基础行对群体状态进行排序，是解决问题的一种有效途径。本章从中观层面介绍了社会经济系统中的排序问题研究。针对在线评分系统，将用户按照评分进行聚类分组，根据用户归属群组的规模计算其信誉水平和排序。进一步，将迭代寻优过程引入基于群组聚类的信誉排序，迭代更新用户信誉和群组规模，提高算法对户信誉排序的准确性。最后，考虑网络节点相似性和社会信任关系的作用，提高个性化推荐算法的排序效果。

信誉是社会经济生活的基石，利用排序算法基于在线评分推断用户信誉是一种常用方法。传统信誉排序算法大多基于产品有唯一质量分数的假设，然而产品实际上可以有多个合理评分。本章第4.1节开创性地提出了一种基于群组聚类的在线用户信誉排序（GR）算法，在真实评分数据集上验证了算法效果。首先，将用户按照评分进行聚类分组，对相同产品给出相同评分的用户归入一群组。然后，计算每个群组的规模，列归一化得到评分回馈矩阵。进而，根据评分回馈矩阵，将用户原始评分矩阵映射得到用户信誉回馈矩阵。最终，根据每个用户信誉回馈向量的均值和标准差计算其信誉分数。简单来说，用户信誉由其所归属群组的相对规模来决定，总是稳定地归属大组的用户有高信誉分数。在三个真实评分数据集上测试算法，发现在应对作弊用户攻击方面，GR算法比传统信誉排序算法有更

高的准确性和更强的鲁棒性，尤其在检测不活跃的作弊用户时。提出的GR算法有复杂度低、不依赖产品质量假设等优点，为改进用户信誉排序提供了新思路。

迭代寻优过程能提高线性系统数值求解的准确性，也能用于解决社会经济的排序问题。将迭代寻优与信誉排序算法结合，有助于提高对用户信誉的排序效果。本章第4.2节提出了一种基于迭代过程的群组聚类用户信誉排序（IGR）算法，提高了用户信誉排序的准确性和鲁棒性。改进得到的IGR算法拓展了群组聚类的思想，通过用户信誉加权计算群组规模，不仅关注用户的绝对数量，还考虑用户的信誉水平。高信誉用户有更大的权重决定群组规模，低信誉用户对群组规模的影响有限。在GR算法框架下，利用迭代寻优过程不断更新用户信誉和群组规模，直到信誉排序保持稳定。在两个数据集上测试算法，发现IGR算法没有明显的用户度偏好，不受用户活跃程度影响，稍微偏好追求热门产品的用户。IGR算法对用户信誉的整体排序更准确，能很好地检测恶意型和随机型作弊用户。特别地，当作弊用户比例较大时，IGR算法比GR算法在鲁棒性方面有显著的提高。

推荐系统为用户推荐产品，本质上也是社会经济系统的排序问题。在分析二部分网络结构的基础上，对用户可能购买的产品进行排序。如何实现准确的推荐排序，是信息过滤研究的重要问题。本章第4.3节提出了两种基于复杂网络结构特征的个性化推荐算法。首先，提出了一种新的网络节点相似性（CosRA）指标，保留了余弦相似性和资源分配指数的优点。基于此指标，提出了一种CosRA推荐算法。在四个真实评分数据集上测试算法，发现CosRA算法能很好地平衡推荐结果的准确性、多样性和新颖性，且CosRA算法在推荐效果上达到了最优。进一步，将用户信任关系引入到CosRA算法框架，提出了一种基于信任关系的CosRA+T推荐算法，利用标度参数调整受信任用户的资源总量。在两个数据集上测试算法，发现CosRA+T算法相比于CosRA算法提高了推荐准确性。另外，CosRA+T算法存在最优参数，说明过分依赖信任关系进行推荐将起负面作用。

第五章 宏观层面的社会经济结构建模研究

社会经济发展逐渐从简单统一向复杂多样过度，这个过程中社会经济水平不断提高。利用网络建模方法分析大规模真实数据，不仅能从结构角度感知经济态势和刻画发展过程中涌现的复杂性，还能利用网络结构特征预测经济发展趋势。本章从三个方面介绍宏观层面的社会经济结构建模研究。首先，介绍区域经济复杂性的刻画方法，基于二部分网络计算经济复杂性指标，分析其演化规律及其与社会经济指标的关联性。然后，介绍区域产业空间的建模和分析方法，基于人力和企业数据构建产业空间，分析其结构特征和演化规律。最后，介绍利用信息和人才流动推断区域经济水平的方法，基于关注关系和求职者简历数据分别构建信息和人才流动网络，分析两个网络的结构特征对区域经济水平的预测能力。

5.1 经济复杂性建模刻画与关联分析

设计有效的指标估计经济发展状态，对经济决策有非常重要的意义[305]。很多传统的宏观经济指标，如人均GDP、克强指数[102]、CPI和PPI[306]等，已经被广泛用于揭示经济发展状态[97]。然而，这些传统指标大多依赖经济普查和统计数据，整个过程耗费大量人力和物力，而且时间滞后很长。另外，这些指标仅能从单一维度估计经济发展所处的大致阶段，无法体现经济发展的复杂结构特征，也缺乏对未来发展的预测能力。例如，两个GDP总量相同的国家，可能在产业结构组成和进出口贸易类型上差别很大[25, 61]。产品和产业结构所展现的经济多样性和复杂性，恰恰是能用来刻画经济发展状态和未来发展潜力的重要特征。

最近的研究已经逐渐转变为依靠数据驱动的范式[9]，借助来自交叉学科领域的工具和分析方法[168]，定量刻画宏观经济结构和经济复杂性[68]。利用新方法分析大规模社会经济数据，提出新的社会经济指标，有希望更好地揭示社会经济发展状态[15]。例如，分析卫星图像数据来估计国家贫困程度[30]；分析手机通讯数据来预测社会经济水平[46]；分析在线搜索数据来预测金融市场中的交易行为[307]等。特别地，最近一些研究基于国际贸易网络数据，提出了非货币性指标来刻画经济复杂性和竞争力，能够量化国家未来的经济发展潜力[24, 308]。

本节使用中国企业注册信息数据研究区域经济复杂性，企业属于不同产业类型和位于不同地理位置。通过企业注册信息将产业类型与地理位置对应起来，可以构建“省份-产业”二部分网络，进而利用二部分网络的结构特征刻画区域经济复杂性。首先，介绍刻画经济复杂性的研究背景，简述两大类经济复杂性指标的

计算和分析框架。进一步，在分析企业注册信息数据的基础上，介绍刻画中国区域经济复杂性的建模和分析方法。最后，对比分析不同经济复杂性指标的特点，分析经济复杂性随时间的演化规律，及其对传统社会经济指标的预测能力。

5.1.1 研究背景与宏观经济复杂性

专业分工提高了经济生产效率，随着市场的扩大和经济水平的提高，经济多样性逐渐增加。例如，区域所发展的产业从单一技术向复杂技术转变，国际贸易从复杂性低的产品向复杂性高的产品转变[25]。经济发展中个体社会经济活动存在复杂的相互作用，其中涌现出来的复杂性与经济发展水平密切相关。在国际贸易中，经济活动有地域属性，不能简单的通过进出口来迁移。所以，国家的生产力能通过其所具有的非贸易“能力”（Capabilities）的多样性来体现[24]。国家之间的收入差异，也能通过经济复杂性的差异来解释。这里所说的复杂性，能通过国家所具有“能力”的多样性和他们之间的相互作用来刻画[68]。

Hidalgo和Hausmann[24]开创地开展了国家经济复杂性的刻画工作，将国际贸易抽象成“国家-产品”二部分网络，用一组线性迭代方程刻画二部分网络结构，提出了经济复杂性（ECI）指标。其中，国家出口产品的多样性与产品的普遍性相互耦合。国家经济复杂性由其出口产品的多样性和产品的普遍性共同决定：经济发展水平高的国家，出口的产品更多样，产品复杂程度也更高；复杂程度低的产品，能被很多国家出口；复杂程度高的产品，仅有少数国家能出口。受ECI指标启发，Tacchella等人[113]采用一组非线性迭代方程的不动点来刻画二部分网络结构，计算了国家竞争力（Fitness）指标和产品复杂性指标。如果产品被竞争力弱的国家出口，那么用国家的竞争力来限制产品的复杂性。在这两个工作的基础上，已经发展出了一系列刻画经济复杂性的方法[68, 309]。

首先，介绍经济复杂性ECI指标的计算方法[24]。利用比较优势（RCA）[310]刻画国家对产品的显著出口程度，即产品占国家出口量的比例相对于全球平均水平的情况。具体而言，国家c出口产品p的比较优势RCA_{cp}定义为

$$\mathrm{RCA}_{cp} = \frac{x_{cp}}{\sum_{p'} x_{cp'}} \Bigg/ \frac{\sum_{c'} x_{c'p}}{\sum_{p'}\sum_{c'} x_{c'p'}} \,. \tag{5–1}$$

其中，x_{cp}为国家c出口产品p的贸易总额。将国家贸易关系抽象为“国家-产品”二部分网络，以邻接矩阵M_{cp}表示。如果国家c显著地出口产品p，即$\mathrm{RCA}_{cp} \geq 1$，那么$M_{cp} = 1$；否则，$M_{cp} = 0$。进一步，基于M_{cp}矩阵定义国家c的经济复杂性：

$$\mathrm{ECI}_c = \frac{K_c - \langle \vec{K} \rangle}{std(\vec{K})} = \frac{N^2 K_c - N\sum_c K_c}{\sqrt{N\sum_c (NK_c - \sum_c K_c)^2}}. \tag{5–2}$$

其中，N为国家数量；$\langle\cdot\rangle$和$std(\cdot)$表示分别对向量$\vec{K}$求平均值和标准差。具体而言，$\vec{K}$为以下矩阵的第二大特征值所对应的特征向量[68]：

$$\tilde{M}_{cc'} = \sum_p \frac{M_{cp}M_{c'p}}{k_{c,0}k_{p,0}}. \tag{5–3}$$

事实上，矩阵$\tilde{M}_{cc'}$定义了一组线性迭代方程。出口相同产品的国家相互连接，以产品的普遍性$k_{p,0}=\sum_c M_{cp}$来加权，以国家的出口多样性$k_{c,0}=\sum_p M_{cp}$来归一化。这样一来，将产品平均普遍性方程$\mathrm{k}_{p,N}=\sum_c M_{cp}k_{c,N-1}/k_{p,0}$代入国家平均多样性方程$\mathrm{k}_{c,N}=\sum_p M_{cp}k_{p,N-1}/k_{c,0}$，得到耦合方程$k_{c,N}=\sum_{c'}k_{c',N-2}\sum_p(M_{cp}M_{c'p}/k_{c,0}k_{p,0})$。其中，$N\geq 2$ 为迭代次数。经济复杂性ECI指标通过$\mathrm{ECI}_c=\sum_{c'}\tilde{M}_{cc'}\mathrm{ECI}_{c'}$计算，其中$\mathrm{ECI}_{c'}$为前一步迭代得到的经济复杂性指标。

然后，介绍国家竞争力Fitness指标的计算方法[62]，即FCM (Fitness-Complexity Method) 分析框架。基于“国家-产品”二部分网络，使用一种偏置马尔可夫过程进行经济复杂性排序。提出的非线性迭代方法，采用两个参数的偏置来刻画二部分网络的结构特征，在经济复杂性概念上更自恰。类似的方法已经在搜索引擎[268]和在线信誉系统[51]中得到广泛应用。具体而言，马尔科夫过程定义如下

$$\begin{cases} w_c^{(N+1)}(\alpha,\beta)=\sum_p G_{cp}(\beta)w_p^{(N)}(\alpha,\beta) \\ w_p^{(N+1)}(\alpha,\beta)=\sum_c G_{pc}(\alpha)w_c^{(N)}(\alpha,\beta) \end{cases}. \tag{5–4}$$

其中，w_c为国家c的竞争力，w_p为产品p的复杂性，N为迭代步数，G为马尔科夫转移矩阵。具体而言，马尔科夫转移矩阵G由以下公式给出

$$\begin{cases} G_{cp}(\beta)=\frac{M_{cp}k_c^{-\beta}}{\sum_{c'}M_{c'p}k_{c'}^{-\beta}} \\ G_{pc}(\alpha)=\frac{M_{cp}k_p^{-\alpha}}{\sum_{p'}M_{cp'}k_{p'}^{-\alpha}} \end{cases}. \tag{5–5}$$

其中，α和β为两个参数。如果以向量形式表示FCM方法，那么国家c的竞争力为$\mathbf{w}_c^{(N+1)}(\alpha,\beta)=T(\alpha,\beta)\mathbf{w}_c^{(N)}(\alpha,\beta)$，其中遍历随机矩阵$T$中的元素可以定义为$T_{cc'}(\alpha,\beta)=\sum_p G_{cp}(\beta)G_{pc'}(\alpha)$；产品$p$的复杂性$\mathbf{w}_p^{(N+1)}(\alpha,\beta)=S(\alpha,\beta)\mathbf{w}_p^{(N)}(\alpha,\beta)$，其中遍历随机矩阵$S$中的元素定义为$S_{pp'}(\alpha,\beta)=\sum_c G_{pc}(\alpha)G_{cp'}(\beta)$。国家经济多样性和产品普遍性之间存在非线性耦合，共同决定国家竞争力和产品复杂性。

最后，介绍经济复杂性ECI指标和竞争力Fitness指标的变体方法。Tacchella等人[113]提出了一种FCM统计分析框架，采用非线性迭代定义国家竞争力和产品复杂性。具体而言，国家c的竞争力F_c和产品p的复杂性Q_p之间的迭代方程为：

$$\begin{cases} \tilde{F}_c^{(N)} = \sum_i M_{cp}Q_p^{(N-1)} \\ \tilde{Q}_p^{(N)} = \frac{1}{\sum_c M_{cp}\frac{1}{F_c^{(N-1)}}} \end{cases}. \tag{5–6}$$

其中，初始条件为$F_c^{(0)}=1$和$Q_p^{(0)}=1$，每一迭代步都对$\tilde{F}_c^{(N)}$和$\tilde{Q}_p^{(N)}$进行归一化，即$F_c^{(N)}=\tilde{F}_c^{(N)}/\langle\tilde{F}_c^{(N)}\rangle$和$Q_p^{(N)}=\tilde{Q}_p^{(N)}/\langle\tilde{Q}_p^{(N)}\rangle$。当迭代达到稳态时[311]，$F$为国家的竞争力指标，$Q$为产品的复杂性。从公式（5–5）看到，一方面，国家经济竞争力与其出口产品的数量和产品复杂性的乘积成正比。另一方面，产品复杂性与所出口国家竞争力的相反数的乘积成反比。FCM方法的核心思想是，出口产品单一的国家，更可能出口复杂性低的产品；经济多样的国家，在刻画产品复杂性上能提供的信息有限。所以，有必要对国家竞争力和产品复杂性进行非线性迭代。Cristelli等人[308]进一步分析了二部分网络的结构与非线性迭代依赖之间的对应关系，将FCM方法拓展到含权网络，发现FCM方法在概念上与经济学观点更一致。

最近一些研究分析了经济复杂性的数学框架[312]，并将其应用到预测未来经济发展上。例如，Mariani等人[309]比较了ECI方法和FCM方法在国家和产品排序上的效果。进一步，引入可调参数控制FCM方法中的非线性耦合，提出了MFCM方法，能在促进出口产品多样的国家和惩罚被大量国家出口的产品之间进行权衡。Wu等人[313]分析了“国家-产品”二部分网络的嵌套结构，提出了一种简化版本的MEM方法，产品的复杂性等于出口该产品的国家所具有的最小竞争力。在特定参数情况下，MEM方法能退化为MFCM方法。Cristelli等人[22]分析了经济复杂性的异质性动力学，发现国家竞争力处在较高水平时，才对国家收入水平有很好的预测能力。Tacchella等人[63]将经济发展表示为二维动力系统，提出的模型对未来五年GDP的预测能力比传统IMF方法在效果上提升25%。Mealy等人[314] 对经济复杂性ECI指标进行了细致解读，发现ECI方法相当于一种图的谱聚类算法。

5.1.2 企业注册数据刻画经济复杂性

尽管已有一些利用国际贸易数据刻画经济复杂性的工作[37, 314]，可观经济复杂性（OEC）网站也提供相关数据的可视化[315]，但是仍然缺乏利用企业数据对中国区域经济复杂性进行定量刻画的工作。一方面，以往的大部分工作仅关注国家层面的经济复杂性，对区域层面的经济复杂性研究不足[316]。另一方面，以往研究大多使用国际贸易数据，即产品进出口数据，没有考虑一些没有产品输出的产业，如服务业。事实上，产品和服务都对经济复杂性的刻画非常重要。特别地，服务行业的增长及其复杂性，能为经济增长提供额外的路径[116]。

国家经济增长的过程，同时关注经济总量的增加和经济结构的转变。中国经济实现飞速发展的事实仍然令人不解，例如很多研究关注中国是如何实现经济增长的[317]，以及中国区域内的经济发展历经了什么过程[318]。为理解国家经济增长的潜在动力，需要分析区域经济的组成结构和复杂性，这对数据粒度和质量提出

了更高需求。有幸的是，国家工商部门和证券交易机构，记录了涵盖所有行业类型的企业注册信息数据，有助于实现对中国区域经济复杂性的刻画，为分析经济复杂性对传统社会经济指标的预测能力提供了基础。

企业注册信息数据来自锐思数据库（http://www.resset.com），涵盖中国沪深A股上市公司的注册和财务信息，时间从1990年到2015年。字段包括：注册日期、注销日期、注册地址和产业分类。清洗之后的数据涵盖2690家企业，位于中国大陆31个省份，属于70个行业类别（按照证监会公布的2011年国民经济产业分类）。另外，从《国家统计年鉴》收集和估计了省份层面的宏观经济数据，包括：人均GDP、人口总数、城市化水平、教育水平、创新能力和贸易水平等。具体而言，人均GDP单位为元，刻画经济发展水平；人口总数为年末常住人口数；城市化水平通过城市面积所占比例估计；教育水平通过高等教育学生比例估计；创新能力通过授权专利数量估计；贸易水平通过进出口总额估计。另外，经过2010年的购买力（PPP）调整的城市和农村的相对收入分别为RICU和RICR[319]；通过RICU和RICR的比值计算相对收入差异（RICD），以此估计收入不平等性。

在计算区域经济复杂性ECI指标时[183]，首先构建“省份-产业”二部分网络。每个节点为省份，节点之间的连边权重为省份内的企业数量，图5–1(a)给出了二部分网络示意图。自然地，将二部分网络表示为邻接矩阵$M_{p,i}$。其中，如果省份p在产业i中有比较优势，即$\mathrm{RCA}_{p,i} \geq 1$，那么$M_{p,i}=1$；否则，$M_{p,i}=0$。具体而言，省份$p$在产业$i$中的比较优势定义为[310]

$$\mathrm{RCA}_{p,i} = \frac{x_{p,i}}{\sum_{i'} x_{p,i'}} \Bigg/ \frac{\sum_{p'} x_{p',i}}{\sum_{i'}\sum_{p'} x_{p',i'}} \,. \tag{5–7}$$

其中，$x_{p,i}$为省份p中在产业i中的公司数量。进一步，将省份内有比较优势的产业数量，定义为省份的产业多样性。对省份p，其多样性为

$$\mathrm{Diversity} = k_{p,0} = \sum_i M_{p,i}. \tag{5–8}$$

类似地，将有比较优势的省份数量定义为产业的普遍性。对于产业i，其普遍性为

$$\mathrm{Ubiquity} = k_{i,0} = \sum_p M_{p,i}. \tag{5–9}$$

最后，根据省份的产业多样性和产业的普遍性定义省份的经济复杂性。具体而言，将省份p的经济复杂性指标（ECI_p）定义为

$$\mathrm{ECI}_p = \frac{K_p - \langle \vec{K} \rangle}{std(\vec{K})}. \tag{5–10}$$

其中，m为省份的总数量；$\langle\cdot\rangle$ 和$std(\cdot)$表示对向量进行均值和标准差运算；向

量$\vec{K}$为矩阵$\tilde{M}_{p,p'}=\sum_i M_{p,i}M_{p',i}/k_{p,0}k_{i,0}$的第二大特征值所对应的特征向量[62]。实际上，矩阵$\tilde{M}_{p,p'}$连接了有相似产业的省份，通过产业普遍性（$k_{i,0}$）的倒数进行加权平均，以及通过省份的多样性（$k_{p,0}$）进行归一化计算[68]。

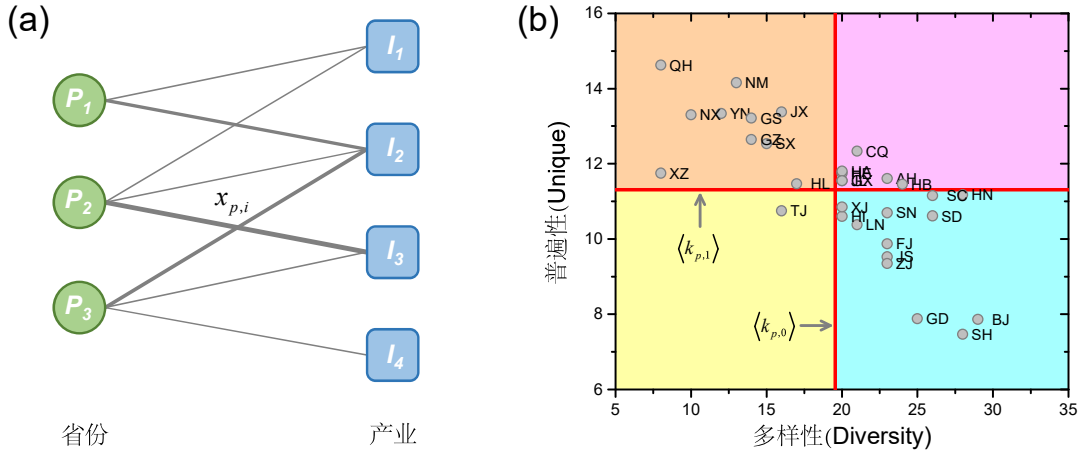

图 5–1 “省份-产业”二部分网络示意图和“多样性-普遍性”相图

经济复杂性ECI指标通过将省份多样性和产业的普遍性相结合，对区域经济结构进行刻画。为了验证ECI指标所依赖的基本假设，即复杂程度高的经济体更多样且有更低的普遍性，图5–1(b)给出了省份经济多样性$k_{p,0}=\sum_i M_{p,i}$和省份所拥有的、有比较优势的产业的平均普遍性$k_{p,1}=\sum_i(k_{i,0}M_{p,i})/\sum_i M_{p,i}$之间的关系。其中，省份字母编码与省份名称的对应关系由表5–1给出。可以看到，多样性$k_{p,0}$与普遍性$k_{p,1}$呈现非常显著的负相关，两者之间的皮尔森关联为$r=-0.777$，显著性水平p-value$=2.8\times10^{-7}$。换句话说，省份的经济多样性越高，那么省份内产业的平均普遍性越低。也就意味着，经济多样性高的省份，拥有的产业很独特，只有很少的省份能支撑这些产业；经济多样性低的省份，拥有的产业都很普遍，很多省份都能支撑这些产业。这些结果支撑了经济复杂性的基本假设，即多样性高的省份拥有普遍性低的产业。

表 5–1 中国省份名称与两位字母编码对应表

字母编码	省份名称	字母编码	省份名称	字母编码	省份名称	字母编码	省份名称
BJ	北京	SH	上海	HB	湖北	YN	云南
TJ	天津	JS	江苏	HN	湖南	XZ	西藏
HE	河北	ZJ	浙江	GD	广东	SN	陕西
SX	山西	AH	安徽	GX	广西	GS	甘肃
NM	内蒙古	FJ	福建	HI	海南	QH	青海
LN	辽宁	JX	江西	CQ	重庆	NX	宁夏
JL	吉林	SD	山东	SC	四川	XJ	新疆
HL	黑龙江	HA	河南	GZ	贵州		

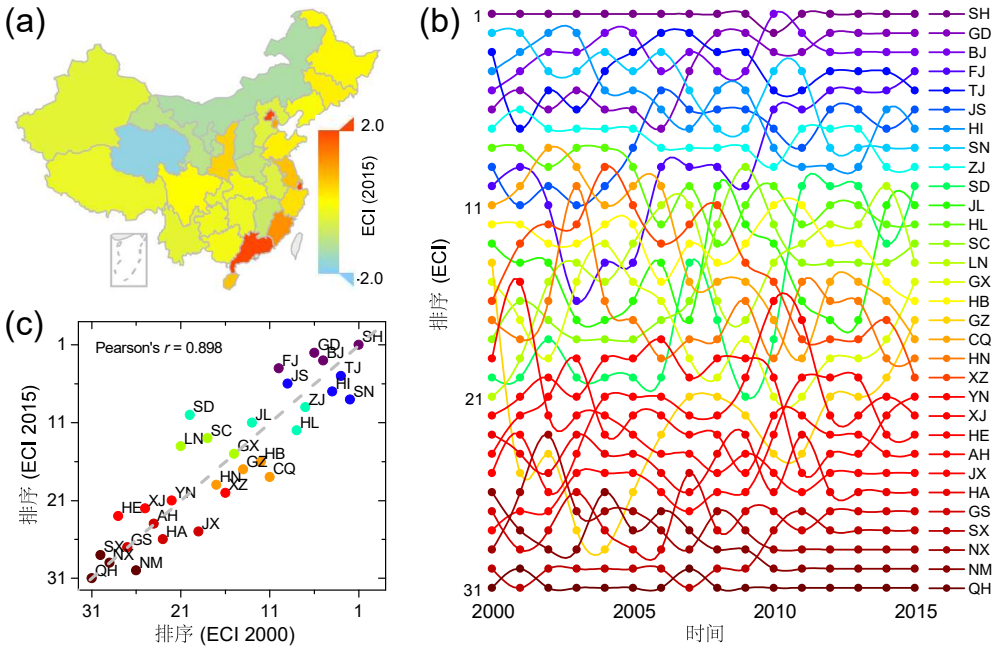

图 5–2 中国区域经济复杂性及省份排序的时间演化分析

进一步，基于企业注册信息数据计算各省经济复杂性指标。图5–2展示了区域经济复杂性及省份排序随时间的演化。其中，省份字母编码与省份名称的对应关系由表5–1给出。具体而言，图5–2(a)给出了2015年各省经济复杂性指标的数值。可以看到，位于沿海的省份有更高的经济复杂性，西南和东北地区的省份紧随其后。图5–2(b)给出了不同省份经济复杂性排序从2000年到2015年的变化。整体上看，经济复杂性很高和很低的省份，在这段时间的排序更稳定。例如，上海（SH）和北京（BJ）始终排在顶端，青海（QH）和宁夏（NX）始终排在底端。经济复杂性在中间的省份排序变化较大，一些省份的经济复杂性提高，如山东（SD）和福建（FJ）；一些省份的经济复杂性降低，如陕西（SN）和重庆（CQ）。图5–2(c)给出了省份2000年排序和2015年排序之间的对应关系。可以看到，省份在两个时间的经济复杂性排序显著正相关，皮尔森关联为$r = 0.898$。这说明，省份在经济复杂性排序上保持相对稳定和缓慢演化。

5.1.3 经济复杂性关联分析与比较

已有研究发现位于沿海的省份经济复杂性较高，进一步分析经济复杂性与经济发展之间的关系。图5–3展示了中国区域经济复杂性与经济发展水平的关联性以及演化关系。其中，省份字母编码与省份名称的对应关系由表5–1给出。具体而言，图5–3(a)和(b)分别给出了2015年和2000年省份经济复杂性与省份人均GDP之间的关系。可以看到，省份人均GDP与经济复杂性呈现显著的正关联。其中，2000年的皮尔森关联为$r = 0.554$，2015年的皮尔森关联为$r = 0.667$。图5–3(c)给出

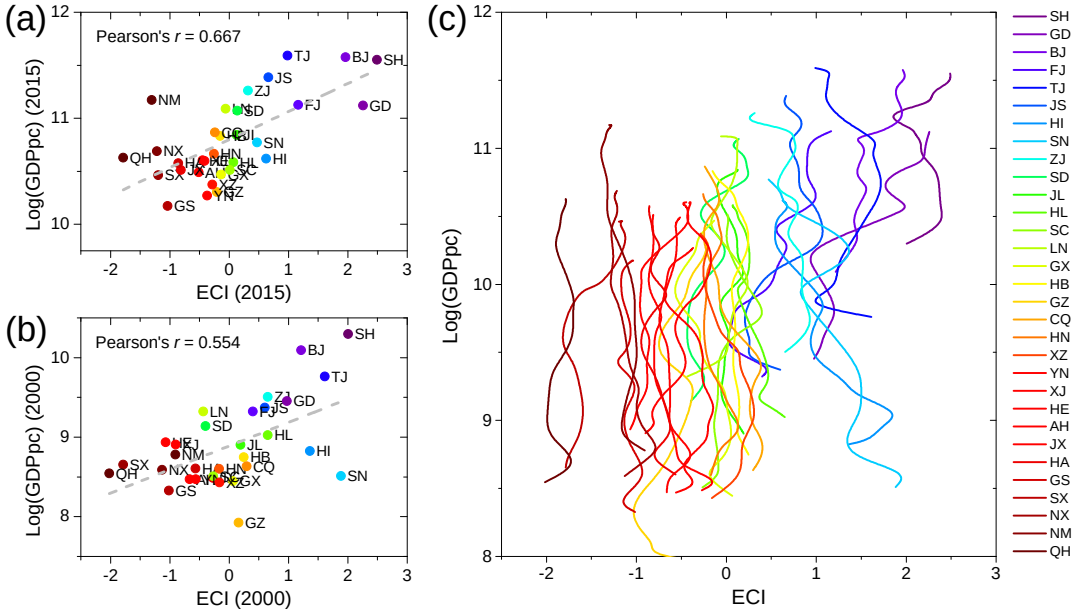

图 5–3 区域层面经济复杂性与经济发展水平之间的关系

了所有省份在“ECI-GDPpc”相图中的时间演化。总体来说，省份在相图中的演化有很强的异质性，以经济复杂性0.5左右为界限，相图大致分为两个区域。在相图的左侧是平流层区域，经济复杂性与人均GDP的关联性较强。在该区域，省份表现出缓慢和稳定的经济增长。由于经济发展受多方面复杂因素的影响，个别国家对应曲线的趋势不明显，这里主要关注整体发展趋势和规律。在相图的右侧是混沌层区域，省份的经济发展难以预测。在该区域，省份的经济发展速度很快，经济发展水平也高。例如，陕西省比其他相同经济水平的省份有更高的经济复杂性，在这15年中其经济总量增长了9.6倍，人均GDP排序从23位跃升到14位，而其他省份经济总量仅增长了7.3倍。这些结果说明，经济复杂性对中国发展省份经济走势的预测能力强，而对发达省份的预测效果不理想。

经济发展不平衡性一直以来都是经济学理论和政策研究所关注的热点问题。Hartmann等人[320]利用国际贸易数据计算全球经济复杂性，分析发现经济复杂性与收入不平等性显著负相关。类似地，图5–4展示了中国区域经济复杂性与收入不平等性之间的关系。其中，省份字母编码与省份名称的对应关系由表5–1给出。图5–4(a)和(b)分别给出了2010年省份经济复杂性与城市相对收入（RICU）和农村相对收入（RICR）之间的关系。可以看到，经济复杂性与相对收入都显著正相关，皮尔森关联分别为$r=0.531$和$r=0.589$。这说明，不论城市还是农村，经济发展水平都与经济复杂性有很强的关联性。图5–4(c)给出了收入不平等性（RICD）与经济复杂性之间的关系。可以看到，收入不平等性与经济复杂性存在显著的负相关，皮尔森关联为$r=-0.413$，与之前国际层面的分析结果一致[320]。尽管中国

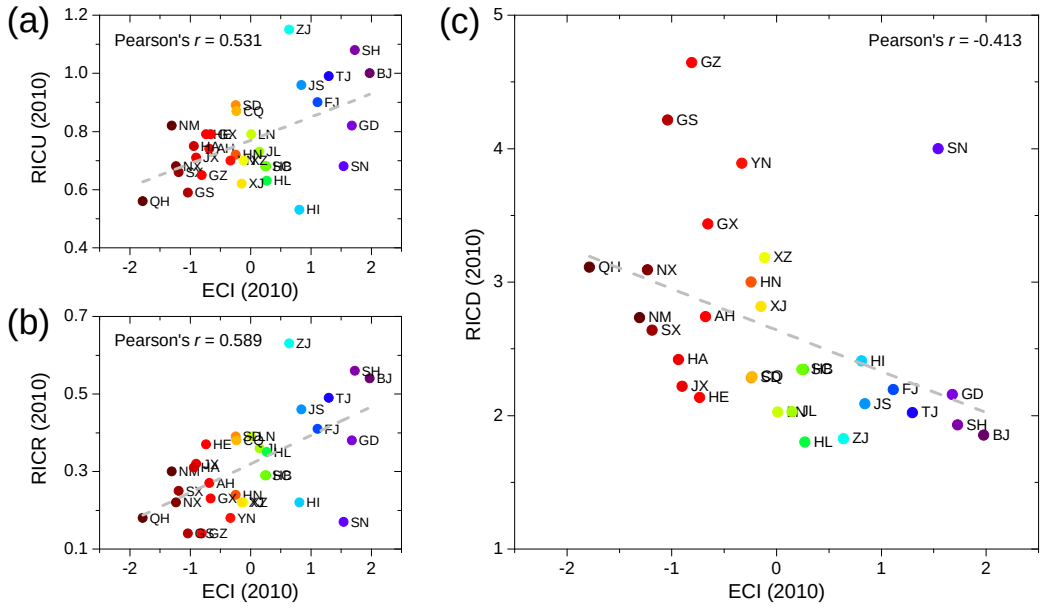

图 5–4 区域经济复杂性与收入不平等性之间的关系

经济飞速发展扩大了地区差距[321]，经济复杂性仍然对区域不平等性有解释能力。这些结果说明，中国区域发展不均衡，需要同时考虑不同层面的经济发展情况。

进一步，对比分析不同复杂性指标对区域经济复杂性的刻画能力，包括：经济复杂性ECI指标[24]、竞争力Fitness指标[113]、多样性Diversity指标[24]和信息熵Entropy指标[322]。首先，基于“省份-产业”二部分网络计算竞争力Fitness指标，核心思想是产业类型多样的省份对产业复杂性的贡献非常有限，产业不多样的省份倾向于拥有特定低复杂程度的产业[113]。所以，需要用非线性迭代过程，以低竞争力省份的竞争力指数来限制产业的复杂性[22, 308]。具体而言，省份p的竞争力指数F_p对于产业i的复杂性Q_i的非线性迭代过程为

$$\begin{cases} \tilde{F}_p^{(n)} = \sum_i M_{p,i} Q_i^{(n-1)} \\ \tilde{Q}_i^{(n)} = \dfrac{1}{\sum_p M_{p,i} \frac{1}{F_p^{(n-1)}}} \end{cases} . \tag{5–11}$$

其中，$F_p^{(n)} = \tilde{F}_p^{(n)} / \langle \tilde{F}_p^{(n)} \rangle$和$Q_i^{(n)} = \tilde{Q}_i^{(n)} / \langle \tilde{Q}_i^{(n)} \rangle$在每个迭代步之后都进行归一化，初始条件设置为$F_p^{(0)} = 1$和$Q_i^{(0)} = 1$。不断进行非线性迭代直到稳态，最终的省份竞争力指数F_p刻画其经济复杂性。其次，利用公式（5–8）计算多样性Diversity指标[183]，表示省份内有比较优势的产业数量。最后，利用香农熵公式计算信息熵Entropy指标[322]，刻画省份内有比较优势的产业的多样性。

在比较复杂性指标时，首先分析经济复杂性ECI指标和竞争力Fitness指标对省份的排序。图5–5展示了两类复杂性指标对省份排序的比较和两者的相关性随时间的演化关系。其中，图5–5(a)和(b)分别给出了ECI指标在2005年和2015年对

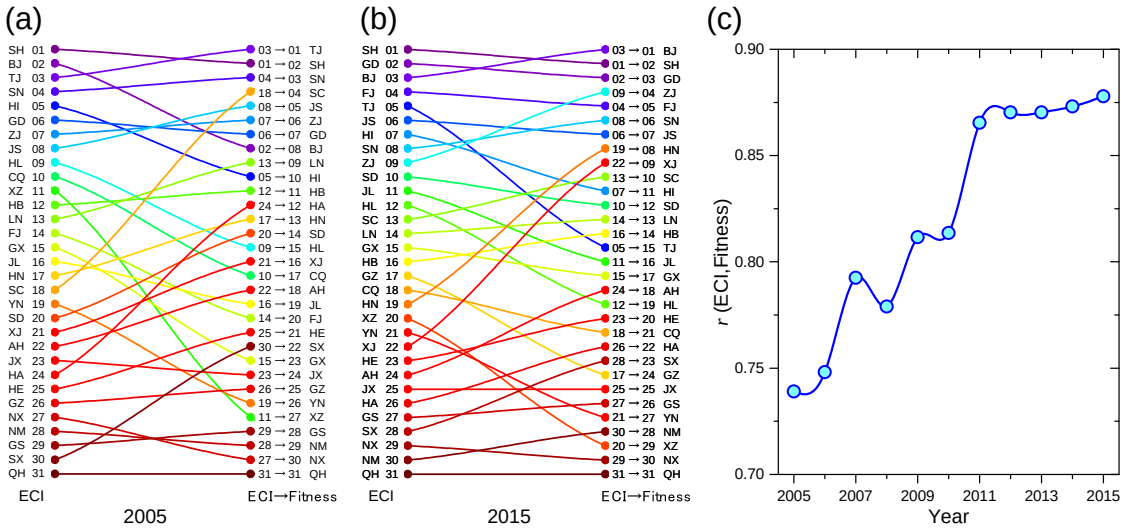

图 5–5 经济复杂性指标对省份的排序比较及两者的关联性分析

应于Fitness指标的情况。其中，省份字母编码与省份名称的对应关系由表5–1给出。总体而言，ECI指标与Fitness指标对于较高和较低排序的省份比较一致，两种指标对于居中省份的排序存在较大差异。例如，ECI指标在2015年将海南和新疆分别排在19位和22位，而Fitness指标将两者分别排在8位和9位，排序相差较大。以2015年为例，相比于Fitness指标，有一些省份的经济复杂性被ECI指标高估了，例如天津（TJ：5→15）、黑龙江（HL：12→19）和西藏（XZ：20→29）。图5–5(c)给出了省份根据ECI指标和Fitness指标排序得到的相关性随时间的变化。可以看到，两个指标对省份的排序显著正相关。从2011年以来，两者的皮尔森关联系数稳定在$r = 0.871$左右，说明两种指标所给出的省份排序基本保持一致。这有别于基于国际贸易数据的结果，即ECI指标和Fitness指标表现不一致[308, 309]。

图5–6展示了四种经济多样性指标（ECI、Fitness、Diversity和Entropy）与四种社会经济指标（GDPpc、RICU、RICR和RICD）之间的关联分析结果。从前四行看到，所有的四种经济多样性指标彼此之间都显著正相关。其中，ECI指标与Fitness指标强相关，Diversity指标与Entropy指标强相关。从第五列看到，与Diversity指标和Entropy指标相比，ECI指标和Fitness指标对人均GDP有更强的解释能力。其中，ECI指标的关联系数为$r = 0.665$，Fitness指标的关联系数为$r = 0.662$。从第六列和第七列看到，ECI指标和Fitness指标对城市和农村相对收入（RICU和RICR）的解释能力也更强。值得注意的是，经济多样性与RICR的关联性强于RICU，说明经济多样性更能解释农村相对收入的差异。第八列给出了收入不平等性RICD的关联性分析结果，发现所有的经济多样性指标都与RICD显著负关联。结果表明，经济越发达和越多样的省份，收入不平等性越小。总的来说，ECI指标和Fitness指标在与其他四种社会经济指标关联上的结果比较接近（第一行

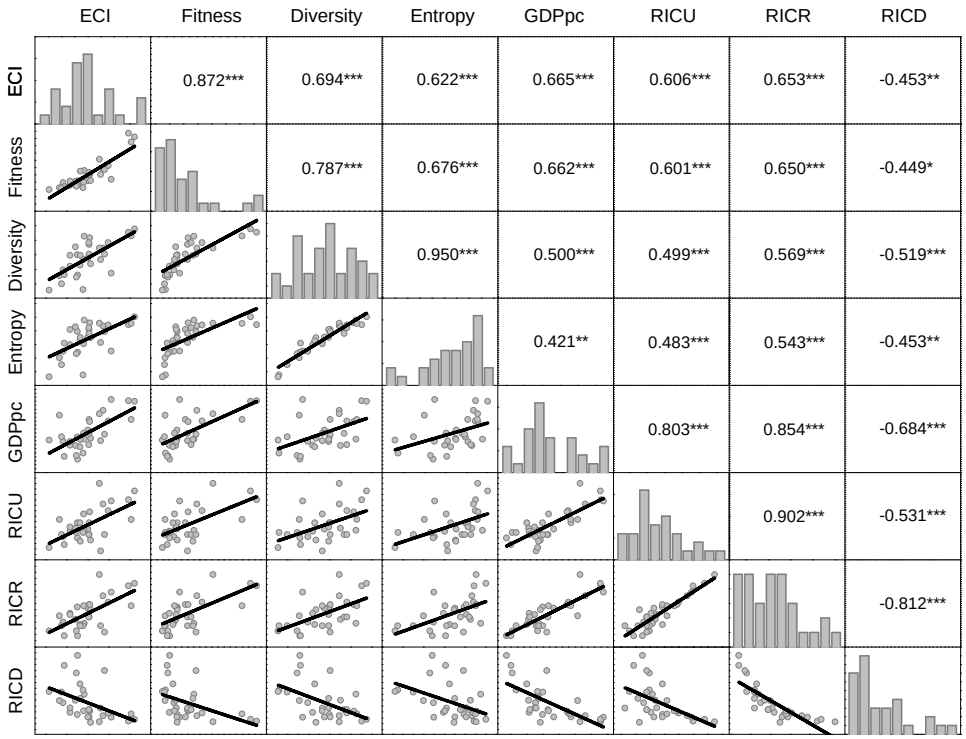

图 5–6 经济多样性指标对经济发展和收入不平等性的解释能力

和第二行），说明他们在解释区域经济发展水平和不平等性方面的能力相当。

进一步，使用最小二乘回归分析（OLS）模型，分析经济复杂性ECI指标和Fitness指标对区域经济发展水平$\log(GDPpc)$的预测能力，控制其他社会经济指标的影响，包括人口数量（Population）、城市化水平（Urbanization）、教育水平（Schooling）、创新能力（Innovation）和国际贸易水平（Trade）。如果经济复杂性对经济发展有很好的预测能力，那么经济复杂性指标（ECI和Fitness）应当与经济发展水平$\log(GDPpc)$显著正相关。表5–2展示了经济复杂性指标对$\log(GDPpc)$的OLS回归分析结果（完整的标准回归分析结果表在文献[183]中给出）。其中，第（1-3）列对应于复杂性ECI指标的回归结果，第（4-6）列对应于竞争力Fitness指标的回归结果。

从表格前三列看到，ECI指标始终与人均GDP显著正相关。其中，ECI指标与人口和城市化水平一起能解释55.8%的人均GDP变化，但人口的作用不显著；ECI指标与教育水平和创新能力能解释62.8%的人均GDP变化，教育水平的作用非常明显；贸易水平也能解释经济水平的变化，与ECI指标一起能解释57.9%的人均GDP变化，这些结果说明教育对经济发展有重要影响。事实上，教育能提高人们的知识、技能、生产力和创造力，这些要素对经济发展非常关键[323, 324]。表格后三列给出了竞争力Fitness指标对经济发展的解释能力。可以看到，Fitness指标与人均GDP之间始终保持显著正相关；城市化、教育和国际贸易也对区域经济发

表 5–2 经济复杂性指标对区域经济发展的多元线性回归分析结果

变量	(1)	(2)	(3)	(4)	(5)	(6)
ECI/Fitness	0.170^{***}	0.104^{***}	0.275^{***}	0.171^{***}	0.127^{***}	0.145^{***}
Log(Population)	0.015			0.001		
Urbanization	0.731^{***}			0.635^{***}		
Schooling		21.49^{***}			22.17^{***}	
Log(Innovation)		0.046^{*}			0.031	
Log(Trade)			0.111^{***}			0.110^{***}
Obs.	186	186	186	186	186	186
Adj. R^2	0.558	0.628	0.579	0.530	0.640	0.585

统计显著性水平：$^{*}p<0.1$；$^{**}p<0.05$；$^{***}p<0.01$

展有显著影响。另外，比较经济复杂性指标的回归系数，发现ECI指标和Fitness指标对人均GDP解释能力相当，都对经济发展有预测能力。

5.2 区域产业空间结构建模与特征分析

在感知经济发展态势方面，传统方法大多依赖单一经济指标，如最为常用的GDP。然而，宏观经济指标只能估计经济发展所处的大概阶段，无法体现经济发展的结构特征和变化情况[59]。GDP水平相同的两个省份，可能有完全不同的产业结构，在未来的经济发展潜力也不同[24, 25]。近年来，大规模高质量的社会经济数据的逐渐丰富，为刻画社会经济发展状态奠定了基础，例如国际贸易数据[25]、企业数据[199]、手机数据[325]和社会媒体数据[97]等。另一方面，借助交叉学科的新分析工具，如复杂网络[33]、机器学习[30]和统计力学[21]等，能更好地刻画和分析经济发展的结构信息。实际上，借助复杂网络分析方法，国际贸易数据和企业信息数据已经被用来刻画国家和区域的产业结构[131, 157]。

从网络结构的视角出发，人们对区域产业结构和经济发展之间的联系逐渐有了新认识[326]。特别地，技术之间的接近性对于区域产业的发展非常重要[154]。类比于国家产品空间[25]，构建区域产业空间，以产业之间的接近性为基础。例如，相比于制造厨具的产业，制造衬衫的产业与制造袜子的产业有更高的接近性，两者在产业空间中的位置也更接近。产业空间网络的结构特征体现区域的产业发展情况，及其在未来发展新产业的可能性[199]。从产业空间结构的角度理解经济发展，也是新结构经济学所重点关注的问题[59]。

本节研究中使用巴西劳动力市场数据和中国企业注册信息数据，介绍区域产业空间的建模和分析方法。首先，基于劳动力市场数据刻画巴西区域产业结构，分析产业空间的“核心-边缘”结构。然后，基于企业注册信息数据刻画中国区域

产业结构，分析产业空间的“哑铃型”结构。最后，针对不同的建模方法，分析产业空间结构的鲁棒性，探究区域之间产业的协同发展和竞争关系。

5.2.1 劳动市场数据刻画巴西产业结构

刻画区域经济结构的基本思想，是在度量产业之间接近性的基础上，构建区域产业空间网络。对于产品空间，如果两种产品经常共同出现在国家出口篮子里，那么该国生产这两种产品的能力相近，产品的基本的物质和人力基础类似，这两种产品的接近性应该很高[25]。类似地，产业之间的接近性，能通过产业在地域上的共同出现来度量。如果两种产业经常共同在不同地域出现，那么这两种产业之间的接近性应该很高。另一方面，考虑到产业对人才和技能的需求，如果两种产业所提供的就业岗位非常类似，那么这两种产业的接近性也应该很高[327]。

首先，介绍所使用的巴西劳动力市场数据的概况。数据来自巴西年度社会信息报告（RAIS），涵盖巴西全国97%的人力资源市场，由巴西人力资源部编纂[158, 327]，获取网址为http://www.rais.gov.br。数据中涵盖7662万员工，时间从2006年到2013年。其中，涉及到501个职业，对应于2002年巴西职业分类表。企业所处区域分为四个层级，涵盖558个Microregion，聚合到137个Mesoregion、27个State和5个Region。企业归属产业分为三个层级，涵盖669个Class，聚合到87个Division和21个Section。表5–3给出了巴西劳动力市场数据的基本统计信息。同时，图5–7(a)给出了数据中字段之间的关系。例如，企业属于某个区域；员工既属于某个企业，又在某个职业中工作；企业提供某些职业，属于某个产业。

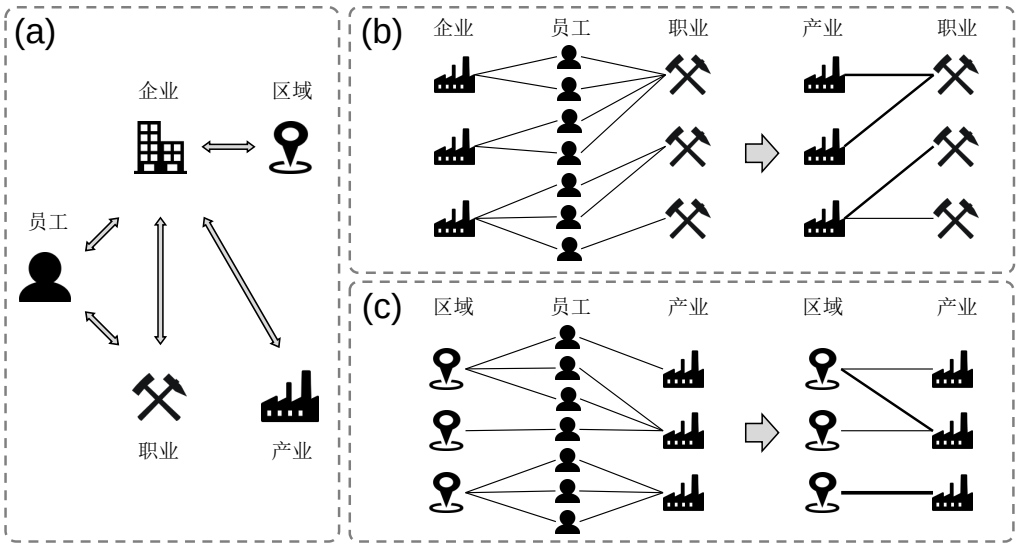

图 5–7 巴西人力资源数据字段之间的关系

基于巴西劳动力市场数据，根据产业所提供职业的相似性来计算产业之间的接近性。在计算接近性时，仅考虑产业中具有比较优势的职业[310]。如图5–7(b)所

表 5–3 巴西劳动力市场数据的基本统计信息

数据字段	数据数量	年份	员工数量	年份	员工数量
区域数量	558	2006	42.21M	2010	33.70M
产业小类	669	2007	44.34M	2011	36.00M
产业大类	21	2008	45.53M	2012	37.76M
职业数量	501	2009	47.00M	2013	39.47M

示，通过员工将职业和产业联系起来，构建“产业-职业”二部分网络。其中，连边权重为相应的员工数量。产业α中职业i在时间t的比较优势$\mathrm{RCA}_{i,\alpha,t}$，定义为实际员工数量与预期员工数量的比例：

$$\mathrm{RCA}_{i,\alpha,t} = \frac{x_{i,\alpha,t}}{\sum_{\alpha} x_{i,\alpha,t}} \Big/ \frac{\sum_{i} x_{i,\alpha,t}}{\sum_{\alpha}\sum_{i} x_{i,\alpha,t}} . \tag{5–12}$$

其中，$x_{i,\alpha,t}$为产业α中职业i的员工在时间t的人数。如果$\mathrm{RCA}_{i,\alpha,t} \geq 1$，那么职业$i$在产业$\alpha$中有比较优势。如图5–7(c)所示，构建“区域-产业”二部分网络，连边权重为相应的员工数量。类似地，定义区域中有比较优势的产业[199]。如果$\mathrm{RCA}_{i,\alpha,t} \geq 1$，那么区域$i$中的产业$\alpha$在时间$t$有比较优势。

根据“产业-职业”二部分网络，以$x_{i,\alpha}$表示产业α中职业i的员工数量，利用公式（5–12）计算职业在产业中的比较优势。然后，利用余弦相似性计算产业之间的接近性。具体而言，定义两个向量$x_{i,\alpha,t} = \ln(\mathrm{RCA}_{i,\alpha,t}+1)$和$x_{i,\beta,t} = \ln(\mathrm{RCA}_{i,\beta,t}+1)$，计算产业$\alpha$和产业$\beta$之间的接近性为

$$\phi_{\alpha,\beta,t} = \frac{\sum_{i} x_{i,\alpha,t} x_{i,\beta,t}}{\sqrt{\sum_{i}(x_{i,\alpha,t})^2}\sqrt{\sum_{i}(x_{i,\beta,t})^2}}. \tag{5–13}$$

根据产业接近性矩阵ϕ，分三步构建产业空间网络[25]。首先，构建最大生成网络。利用最大生成树算法，将所有产业以最少的连边数、最大的连边权重进行连接，得到一个连通的产业网络。然后，构建最大权重网络。以一定阈值保留产业网络中最大权重的连边，使用6倍网络节点数量确定连边数量。最后，构建叠加网络。将最大生成网络和最大权重网络叠加，得到产业空间网络。利用Gephi软件（http://gephi.github.io）中的ForceAtlas和Fruchterman-Reingold布局算法，对产业空间进行可视化，节点之间的距离与连边权重成反比，呈现产业空间的网络结构。

图5–8展示了2013年巴西区域产业空间网络可视化。产业空间中的节点表示产业，节点大小为归属该产业的员工总数。节点越大，表示在该产业中工作的员工数量越多。节点之间的连边表示产业之间的相似关系，连边权重为产业之间的接近性数值。连边权重越大、连边颜色越深，表示两个产业越相似。节点的不同颜色，表示产业所属的不同分类，总共涵盖产业分类的21个Sections，包括制造业（Processing Industries）、贸易（Trade）和信息通讯（Information and

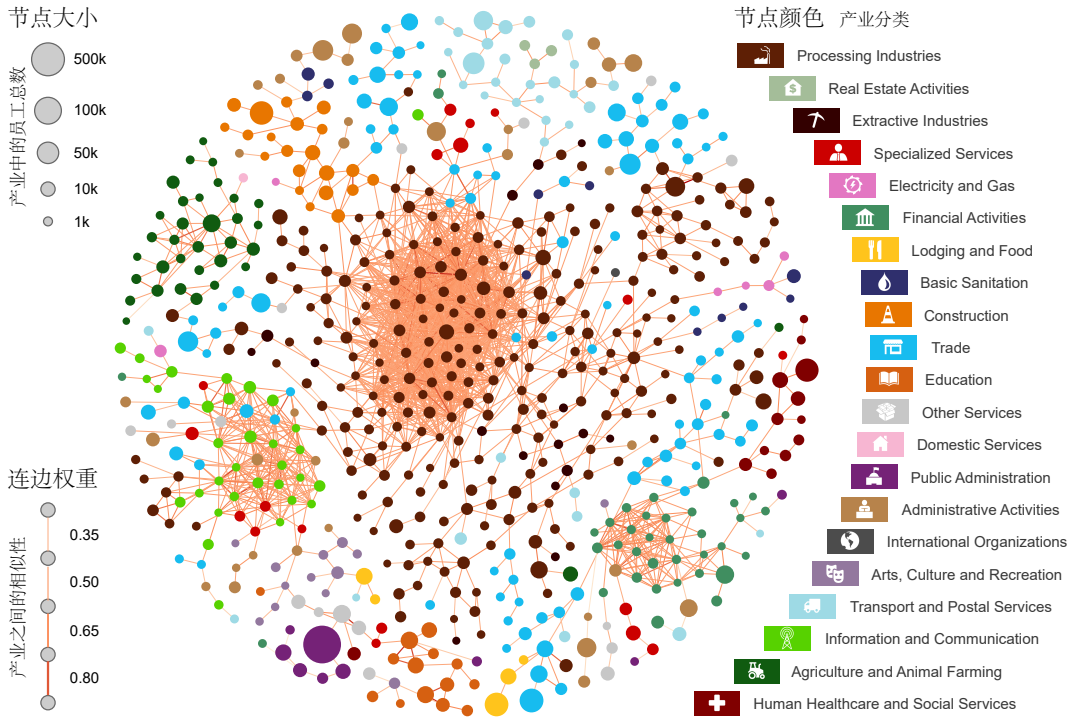

图 5–8 巴西区域产业空间网络可视化

Communication）等。注意到，相同颜色的节点在产业空间中更倾向于彼此连接，位置上也相互更接近，说明分类相同的产业之间有更大的接近性。

从可视化上看，产业空间有明显的“核心-边缘”结构：一些产业紧密的连接，集中在产业空间的核心，例如位于最核心区域的制造业（Processing Industries）；大部分产业连接相对松散，位于产业空间的边缘，例如公共管理（Public Administration）和教育（Education）；也有一些相对较小的、紧密连接的产业，位于产业空间的核心外侧，例如金融活动（Financial Activities）和信息与通讯（Information and Communication）。另外，类似于产品空间的特点[25]，经济复杂程度高的产业位于产业空间的核心，例如制造业；经济复杂程度低的产业位于产业空间的边缘[24, 183]，例如公共管理和教育。

进一步，定量分析产业空间的结构特征。图5–9(a)展示了产业之间接近性的概率分布。可以看到，所有产业接近性呈现类似于对数正态分布。很少产业之间有很高的接近性，大部分产业之间的接近性不高，大多集中在0.3左右。图5–9(b)展示了产业之间的接近性矩阵。其中，行列按照产业编号排列，编号相近的产业属于相同或相近产业大类。相图中颜色表示产业接近性，颜色越深表示越相似。可以看到，产业接近性矩阵存在明显的分块，产业编号相近的产业形成深颜色的分块，说明分类相同或相近的产业有更高的接近性。另外，产业编号在300附近的制造业呈现出最大的分块，说明制造业在产业空间中形成大的核心。

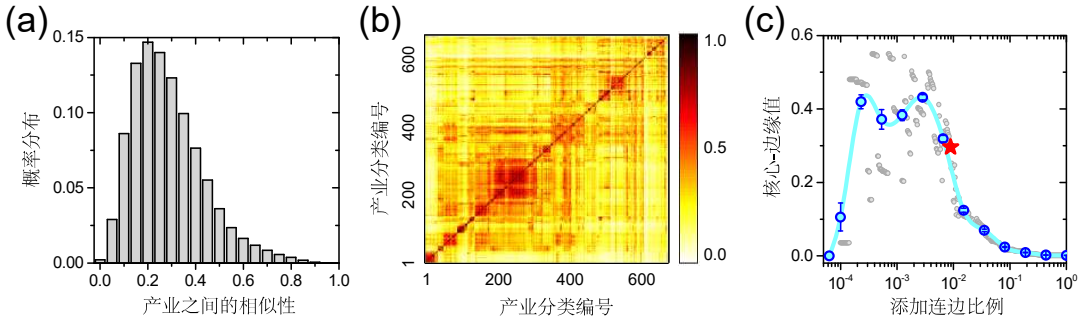

图 5–9 产业之间的接近性以及产业空间的“核心-边缘”结构量化分析

相比之下，其他产业大多位于产业空间的边缘，仅能形成规模小的团簇。

接下来，采用“核心-边缘”（core-periphery）度量指标的一种简单变体方法[328, 329]，定量刻画产业空间的“核心-边缘”结构[330, 331]。简单来说，该方法基于网络的k-核分解过程[104, 332]，将网络逐渐按层分解。如果节点被放入度数至少为k_s的层，那么该层节点的核数为k_s。对于一个有显著CP结构的网络，应当有很多节点有小的核数k_s值（边缘），有很少节点有大的核数k_s值（核心）。基于这些考虑，将“核心-边缘”结构的度量指标定义为

$$\lambda = (\tau_{max} - \tau_{min}) \frac{S_{\tau_{min}}}{S_{\tau_{max}}}. \tag{5–14}$$

其中，$S_{\tau_{min}}$为有最小核数值τ_{min}的节点的总数量，$S_{\tau_{max}}$为有最大核数值τ_{max}的节点的总数量。度量指标λ的数值越大，表示网络的“核心-边缘”结构越明显。

图5–9(c)展示了产业空间的CP值（λ）随着网络中添加连边比例的变化。可以看到，曲线呈现倒U型：当产业空间添加连边比例居中时，CP值达到最大。具体而言，当产业空间中添加极少最大权重连边时，CP值接近于0，因为此时的产业空间为最大生成网络；当添加产业之间的所有连边时，CP值也接近于0，因为此时的产业空间为完全图；当产业空间中添加适当比例的最大权重连边时，网络涌现出显著的CP结构，CP值达到最大。特别地，图5–8所展示的巴西区域产业空间，CP值在0.3附近（红星标记），有显著的“核心-边缘”结构。

5.2.2 企业注册数据刻画中国产业结构

在过去三十年，中国经济飞速发展。从1990年到2015年，中国总体GDP增长了30倍，人均GDP增长了近10倍。相比而言，世界GDP仅增长了3倍，人均GDP仅增长了2倍。在惊叹中国经济发展奇迹的同时[317]，人们开始关注如何解释中国经济发展。一种理论认为中国飞速经济增长依赖于出口特定产品，这些产品的经济复杂性能支撑更高的收入水平[333, 334]。中国出口大量的电子产品和其他高科技制造品，这些产品主要由人均收入水平远高于中国的国家生产[59]。通过不断出口复

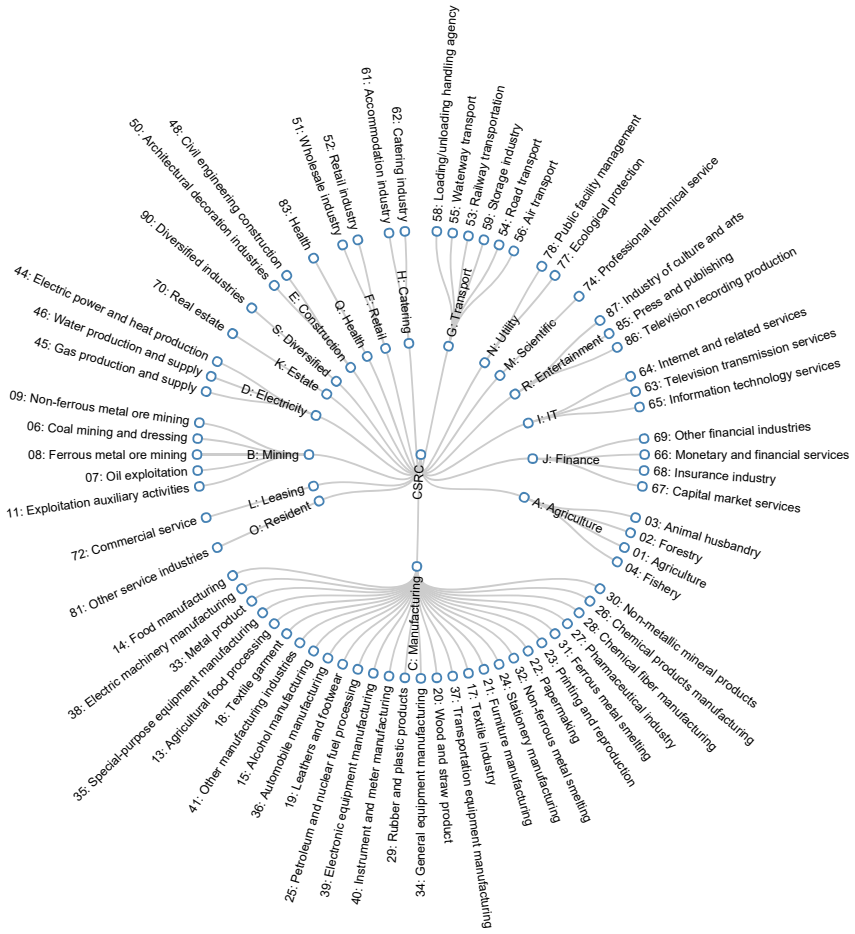

图 5–10 中国上市公司行业分类结构和产业分类代码

杂程度高的产品，中国得以参与全球市场，支撑高工资和高收入。这些解释中国经济增长的理论，依赖于对国家产品空间和经济结构的理解[25]，以及对区域经济结构和产业结构所涌现出复杂性的刻画和分析[24, 183]。

利用企业注册信息数据，对中国区域产业结构进行刻画，有两方面的优势。一方面，企业数据弥补了区域层面缺少国际贸易数据的不足。另一方面，企业数据涵盖所有产业类型，包括没有产品输出的服务行业。所使用的企业注册信息数据来自锐思经济和金融研究数据库（http://www.resset.com），涵盖1990年到2015年期间所有沪深A股上市公司的注册信息和财务信息，包括上市日期、退市日期、注册地址、产业分类、年度利润和员工数量等[199]。企业的注册地址，涵盖中国大陆31个省份。企业归属的产业类别，依据2011年银监会公布的《上市公司行业分类指南》划分为两个层级，涵盖18个大类和70个小类。图5–10给出了产业分类结构图，包括大类的字母代码和小类的阿拉伯数字代码。

首先，基于企业注册信息数据计算产业之间的接近性。类似于产品空间[25]，如果两个产业经常共同在不同地区出现，那么他们之间的接近程度很高。企业的注册地址归属一个省份，企业属于一个产业类别。所以，可以通过企业注册将省

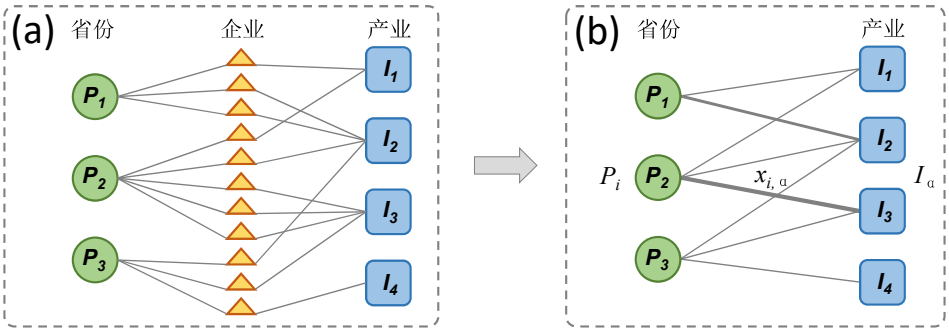

图 5–11 利用企业注册信息构建“省份-产业”二部分网络

份和产业联系起来（如图5–11所示），构建“省份-产业”二部分网络，连边权重表示省份内归属于产业的企业总数。用$x_{i,\alpha,t}$和$x_{i,\beta,t}$表示省份i中在t时间分别属于产业α和产业β的企业数量，利用余弦相似性计算产业α和产业β之间的接近性。具体而言，产业α和产业β之间在t时间的接近性$\phi_{\alpha,\beta,t}$定义为

$$\phi_{\alpha,\beta,t}=\frac{\sum_i x_{i,\alpha,t}x_{i,\beta,t}}{\sqrt{\sum_i (x_{i,\alpha,t})^2}\sqrt{\sum_i (x_{i,\beta,t})^2}}. \tag{5–15}$$

对于每个时间t，都能计算得到产业接近性矩阵ϕ_t。注意，构建“省份-产业”二部分网络时，仅考虑新增和存续的企业，剔除已经退市的企业。

得到产业接近性矩阵ϕ之后，利用生成产品空间的方法[25]，构造中国区域产业空间。简单来说，首先构造最大生成网络，然后构造最大权重网络，最后叠加两个网络，使用布局算法可视化得到产业空间网络。图5–12展示了2015年中国区域产业空间网络可视化。其中，节点表示产业，涵盖70个产业小类，产业小类编号（节点标签）与产业结构分类图5–10中的行业分类代码一致。节点大小，表示属于产业小类的企业总数。节点颜色，表示不同产业小类所归属的产业大类，涵盖18个产业大类，包括农业（Agriculture）、制造业（Manufacturing）和信息技术（IT）等。连边权重和颜色，表示产业之间接近性的大小。两个产业节点之间的连边权重越大、颜色越深，表示产业之间的接近性越大。

从图5–12中看到，中国区域产业空间有非常特殊的结构。一方面，产业空间有“核心-边缘”结构。制造业（Manufacturing）相关的产业（红色社团）组成紧密连接的核心团簇，其他产业大多散布于产业空间的边缘，例如农业（Agriculture）、采矿业（Mining）和娱乐（Entertainment）等。另一方面，中国区域产业空间还有“哑铃型”结构。左侧是以制造业为主的核心团簇，包括各类设备制造和重工业等。右侧是以信息和服务类为主的核心团簇，包括信息技术、电子和金融服务等。这两类核心产业，通过位于产业空间中间的一些产业连接，包括某些制造业（Manufacturing）、房地产（Estate）、零售（Retail）和餐饮（Catering）等。另外，类似于产品空间[25]和巴西区域产业空间[327]，经济复杂程度

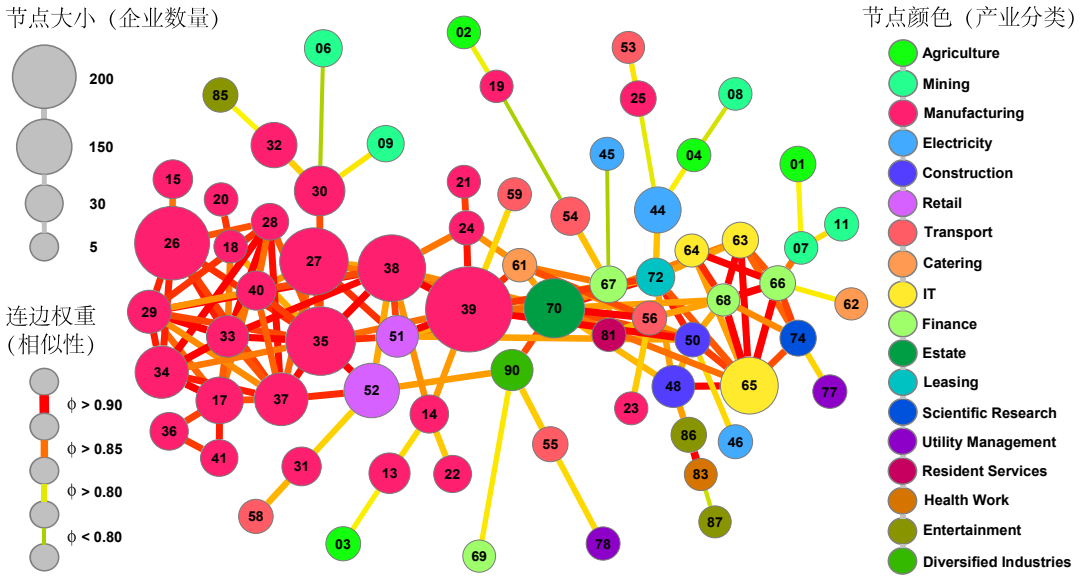

图 5–12 中国区域产业空间网络可视化

高的产业占据产业空间的核心位置，例如制造业和信息技术。经济复杂程度低的产业占据产业空间的边缘位置，例如农业和采矿业。

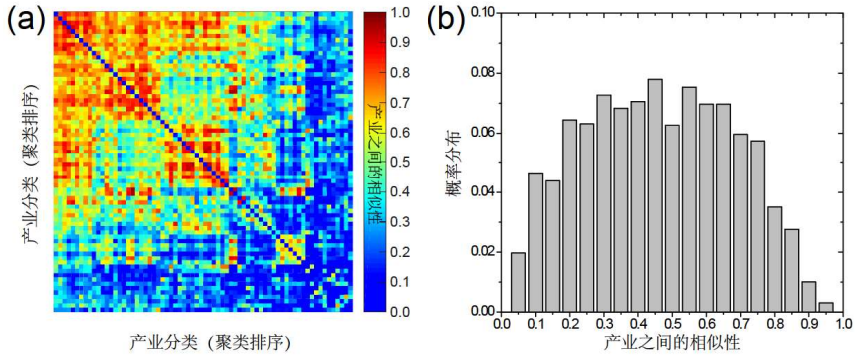

图 5–13 产业接近性矩阵层次聚类和产业接近性的概率分布

进一步，定量分析中国区域产业空间的结构特征。利用谱聚类方法[335]，对产业接近性矩阵ϕ聚类，按照分块密度排列结果。图5–13(a)展示了层次聚类后接近性矩阵的可视化结果图。其中，矩阵行和列对应于聚类后的小类产业，颜色表示产业之间的接近性数值。可以看到，经过聚类后的产业接近性矩阵呈现出两个清晰的对角分块：大分块是制造业为主的核心团簇，小分块是信息和服务类为主的核心团簇。这一结果，验证了产业空间有“核心-边缘”结构和“哑铃型”结构。图5–13(b)展示了产业接近性的概率分布。可以看到，产业接近性呈现出类似于正态分布，体现出余弦相似性对不同产业之间接近性的区分能力。

5.2.3 产业结构特征及区域间产业竞争

从结构的角度分析经济发展，产业空间以直观的方式展示了产业之间的接近

性，有助于理解经济结构与经济发展之间的关系。特别地，基于劳动力市场数据构建的巴西区域产业空间和基于企业注册信息数据构建的中国区域产业空间，都有显著的“核心-边缘”结构，对解释区域所面临的不同发展机会有重要意义。事实上，已有研究发现产品空间的结构限制了国家经济发展，跨越产品空间的发展非常困难[25]。经济发达的国家占据产品空间的核心区域，经济落后的国家占据产品空间的边缘区域。国家在发展经济时，更容易发展与已有产品接近性高的产品，国家逐渐从产品空间的边缘向核心拓展，不断提高经济发展水平。类似地，刻画和分析产业空间结构，对理解区域经济发展的规律有很大帮助。

产业之间的接近性是构建产业空间的基础，所以度量产业接近性的方法尤为重要。针对巴西劳动力市场数据，不但产业与职业之间存在对应关系，而且产业与区域之间也存在对应关系[327]。对于中国企业注册信息数据，仅存在产业与区域之间的对应关系[199]。为了更全面地分析产业空间的结构特征，验证产业接近性计算方法不显著地影响产业空间结构，以巴西劳动力市场数据为例，采用三种新方法构建产业空间。在此基础上，对比分析利用不同方法计算得到的产业接近性分布。进一步，给出经过层次聚类后的产业接近性矩阵的团簇结构，展现产业空间的“核心-边缘”结构，计算产业空间的“核心-边缘”值。

产业接近性的三种计算方法，包括共同招聘职业的接近性方法、比较优势的余弦相似性方法和员工数量的余弦相似性方法。首先，共同招聘职业的接近性方法（Co-hiring Proximity），基于产业中有比较优势的职业，通过公式（5–13）定义职业i在产业α中的比较优势$\mathrm{RCA}_{i,\alpha}$。如果$\mathrm{RCA}_{i,\alpha} \geq 1$，那么产业$\alpha$显著地招聘职业$i$。参考产品接近性的计算方法[25]，将产业$\alpha$和产业$\beta$之间的接近性定义为

$$\phi_{\alpha,\beta} = \min\left\{P(RCA_\alpha|RCA_\beta), P(RCA_\beta|RCA_\alpha)\right\}. \tag{5–16}$$

其中，$P(RCA_\alpha|RCA_\beta)$为职业被产业β显著招聘的情况下被产业α显著招聘的条件概率。其次，比较优势的余弦相似性方法（Co-location Cosine RCA）方法，仅考虑区域内有比较优势的产业，利用余弦相似性计算产业接近性。参考职业在产业中的比较优势的计算公式（5–13），定义产业α在区域i中的比较优势$\mathrm{RCA}_{i,\alpha}$[199]。利用余弦相似性计算公式，将产业α和产业β之间的接近性定义为

$$\phi_{\alpha,\beta} = \frac{\sum_i x_{i,\alpha}x_{i,\beta}}{\sqrt{\sum_i (x_{i,\alpha})^2}\sqrt{\sum_i (x_{i,\beta})^2}}. \tag{5–17}$$

其中，如果产业α在区域i中有比较优势，即$\mathrm{RCA}_{i,\alpha} \geq 1$，那么$x_{i,\alpha} = 1$；否则，$x_{i,\alpha} = 0$。最后，员工数量的余弦相似性方法（Co-location Cosine Labor），仅考虑产业中的员工总数，利用余弦相似性计算产业接近性。产业α和产业β之间的接近性$\phi_{\alpha,\beta}$通过公式（5–17）计算，其中$x_{i,\alpha}$表示区域i在产业α中工作的员工总数。

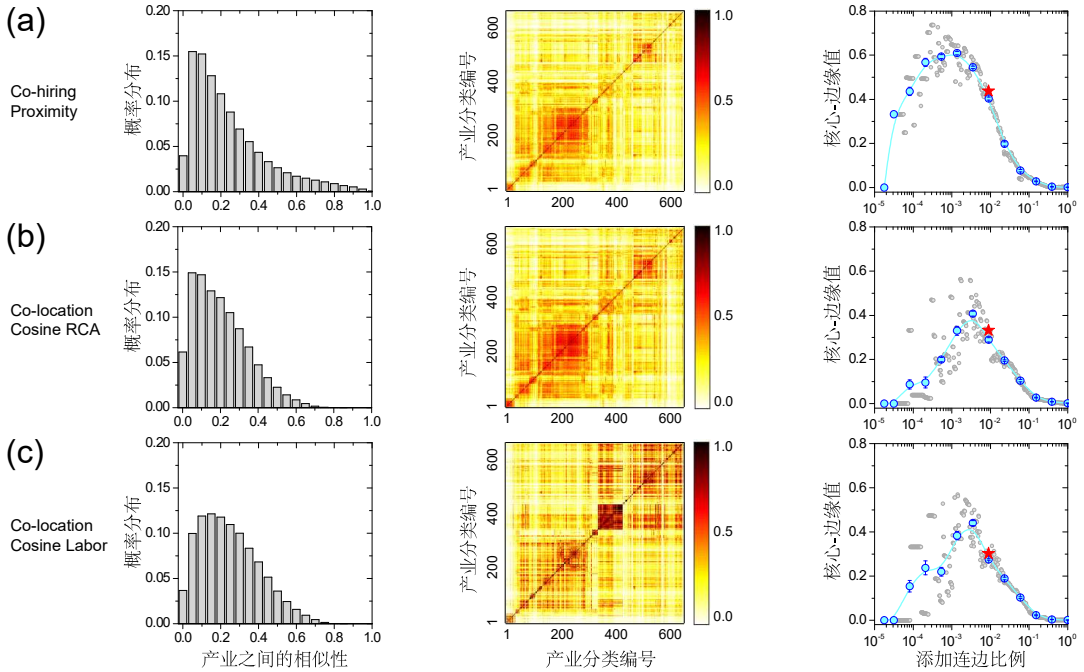

图 5–14 三种计算方法给出的产业接近性概率分布、接近性矩阵和“核心-边缘”值

图5–14（左列）展示了三种产业接近性计算方法得到的产业接近性的概率分布。可以看到，三种情况的分布都呈现出类似于对数-正态分布的形态，这与公式（5–13）得到的结果类似，体现出这三种相似性计算方法同样对产业接近性有很好的区分能力。图5–14（中列）展示了产业接近性矩阵。其中，行和列代表产业小类，颜色代表产业接近性的数值。可以看到，产业接近性矩阵都展现出对角分块结构，存在一个大分块和一些小分块，说明产业空间中存在一个核心的大团簇和很多边缘的小团簇。图5–14（右列）展示了根据公式（5–14）计算得到的产业空间的“核心-边缘”（CP）值随添加最大权重连边的变化。可以看到，曲线都为倒U型。基于三种产业接近性构建的产业空间，都有显著的“核心-边缘”结构，CP值都相对较大（红星标记），说明相似性计算方法不显著影响产业空间结构。

在构建巴西产业空间的基础上，分析产业结构随时间的演化，仅考虑区域内有比较优势的产业。如果$\mathrm{RCA}_{i,\alpha,t} \geq 1$，表示$t$时间产业$\alpha$在区域$i$中有比较优势。由此，对每个区域都计算出有比较优势的产业，将其展现在产业空间中。图5–15(a)以巴西的三个区域为例，展示了优势产业从2006年到2013年的变化情况。可以看到，经济发达的Sao Paulo地区有很多优势产业，位于产业空间的核心位置。经济相对不发达的Brasilia 地区仅有少数优势产业，位于产业空间的边缘位置。另外，区域所新发展的优势产业大多与已有优势产业在产业空间中彼此相邻，暗示产业接近性对发展新产业有重要作用。图5–15(b)展示了三个地区在2013年的产业组成情况。其中，颜色表示不同产业，面积表示员工数量的占比。可以看到，

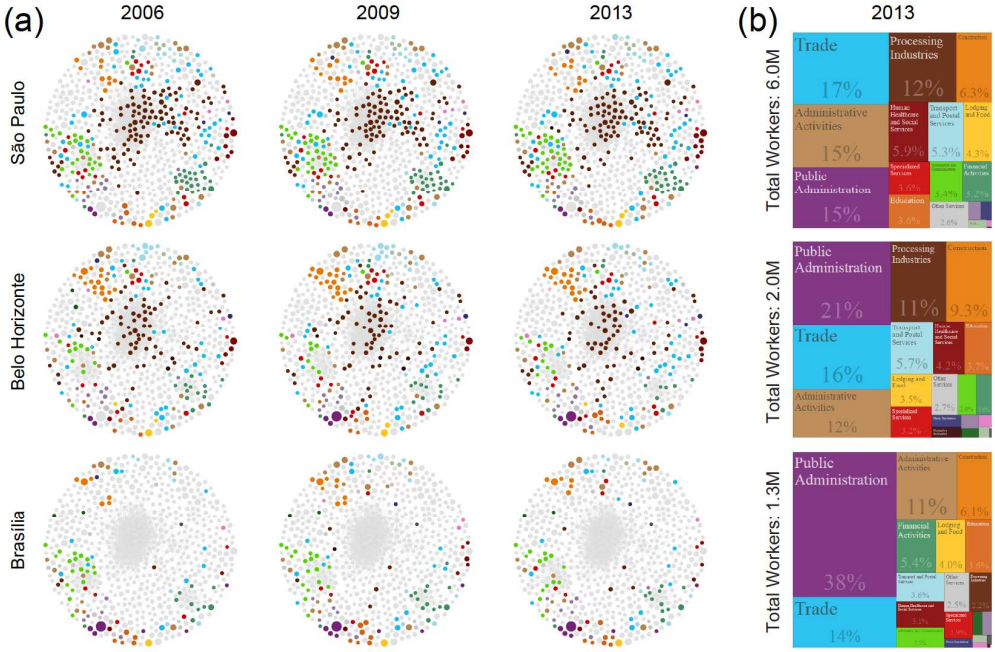

图 5–15 巴西区域产业空间的时间演化及具体的产业组成

Sao Paulo地区大多从事贸易和制造业，其他两个地区大多从事管理和行政活动。

类似地，分析中国区域产业结构的时间演化，重点关注省份产业结构的协同发展和竞争关系。图5–16分别以北方（北京、河北和天津）和南方（上海、江苏和浙江）的三个省份为例，展示了产业空间中的优势产业从1994年到2015年的变化。可以看到，在过去的二十年，北京逐渐占据产业空间中的大部分产业，逐渐从制造业（左侧核心）向信息服务业（右侧核心）转移，最终占据了大部分的信息服务等高收入的行业。相比而言，作为近邻的河北和天津，逐渐向制造业（左侧核心）发展，逐渐在制造业中取得产业优势。三个省份的地理近邻形成了产业

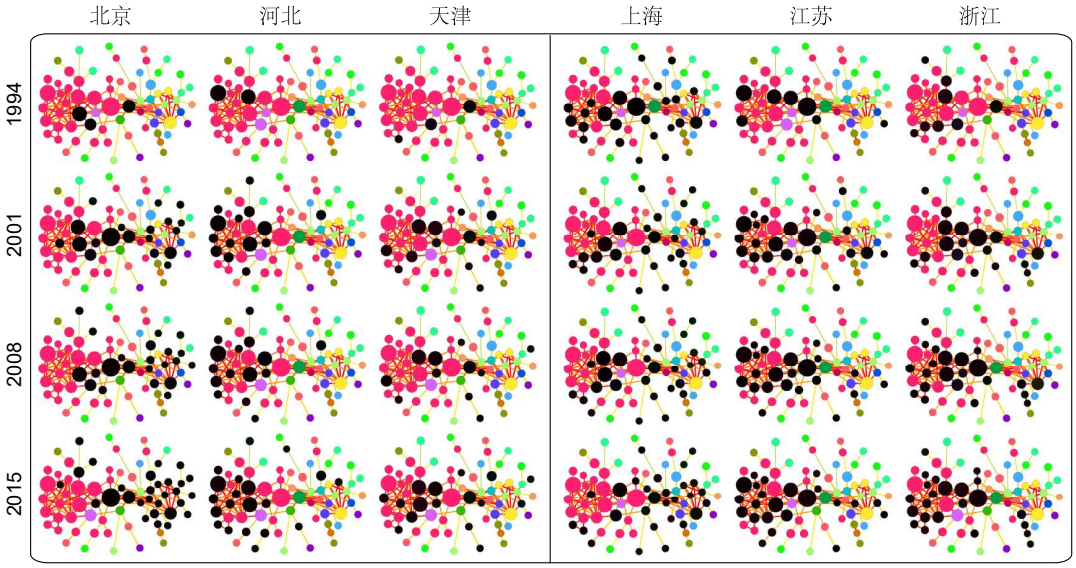

图 5–16 中国的产业空间随时间演化及产业竞争分析

竞争，北京经济发展好了，信息和人才有流动性，在信息服务产业加强了产业优势。河北和天津的优势产业，逐渐向制造业转移。类似地，上海逐渐在信息服务行业取得比较优势，作为其地理近邻的江苏和浙江，在制造业逐渐取得比较优势。总体而言，区域在产业发展上存在协同发展和竞争关系，近邻区域在产业结构上互补，例如北京发展信息服务业，河北发展制造业。

5.3 信息和人才流动推断区域经济状况

精准和及时地感知和推断区域的社会经济状态，对解决很多与经济发展有关的问题都有重要意义[9, 117]。例如，制定脱贫政策[336]，预测区域失业率[79, 176]，优化经济发展策略[12]。然而，传统方法大多依靠统计数据计算宏观统计指标，整个过程消耗大量人力和物力，而且时效性很差[15]。随着信息技术的进步，近年来很多新数据已经被用来推断经济发展状况[9, 30]。例如，Elvidge等人[101]利用夜间光亮数据估计经济状况，绘制高分辨率的世界贫困地图；Liu等人[97]利用微博用户注册数据刻画社交活跃度，推断城市的经济发展状况。

在感知社会经济态势上，一方面关注社会网络结构与经济状态的联系[337]。例如，Eagle等人[66]分析了英国手机通讯网络数据，发现社会网络多样性与社区经济发展指标之间强相关，通讯多样性越高的社区，社会经济水平越高。类似地，Mao等人[96]发现打入和打出电话的比例能预测地区的收入水平；Holzbauer等人[98]发现美国社会网络上跨洲的长程连边数量与GDP强相关。另一方面关注人类移动模式与社会经济状态的联系[338]。人才移动也对经济发展非常重要[339]，不同社会经济水平的个体有独特的移动模式[340, 341]。例如，Frias-Martinez等人[342]基于分析手机数据揭示了移动模式对个体经济水平的预测能力；Pappalardo等人[343]发现人类移动的多样性能预测社会经济指标。然而，这些推断社会经济状态的工作，仅关注社会网络结构或人类行为模式，缺乏对两者预测能力的比较研究。

近年来，大规模社会经济数据的可用性提高，数据的时空分辨率更好，方便直接比较信息流动和人类移动对经济状况的推断能力。本节研究中使用两个来自大规模在线社会平台的数据集，表5–4给出了数据的基本统计信息。一个是微博

表 5–4 信息流动网络和人才流动网络的基本统计信息

数据集	分辨率	区域数量	连边总数	网络密度	平均连边权重
信息流动	省份	31	961	1	1.277×10^7
	城市	336	112,896	1	1.087×10^5
人才流动	省份	31	818	0.8512	347.7
	城市	287	9,746	0.1183	29.18

平台的关注关系数据，用来估计在线的信息流动。另一个是招聘网站的匿名简历数据，用来估计离线的人才流动。基于这两个大规模真实数据集，首先构建信息流动网络和人才流动网络，然后计算和分析网络基本结构特征，最后利用网络结构特征预测区域的经济发展水平[170]。

5.3.1 社交媒体数据构建信息流动网络

在线社会媒体平台为人们的信息分享、资讯交换和日常沟通提供了极大的便利[344]。社交媒体所承载的在线社会网络，促进了信息的沟通交流，也为估计区域之间的信息流动提供了途径。本节研究中使用新浪微博数据，涵盖大约4.33亿注册用户，包括用户的基本信息和关注关系[97]。首先，根据基本信息确定用户所处的地理位置，涵盖336个地级市，归属31个省份。然后，利用关注关系构建区域之间的信息流动网络（OIF, Online Information Flow）[170]，记为G^I。其中，区域既可以是城市，又可以是省份，依据研究的尺度而定。

将信息流动网络G^I表示为含权邻接矩阵形式A^I，其中矩阵元素$a^I_{i,j}$为区域i流入区域j的信息总量。如果区域j关注区域i，那么区域j会看到区域i的推文和信息，实际上信息由区域i流入区域j。所以，根据位于区域j中用户关注位于区域i中用户的数量，估计从区域i流入区域j的信息总量。考虑到区域内的用户彼此之间也可能存在关注关系，信息相当于在区域内部流动。所以，信息流动网络G^I存在自环，即邻接矩阵A^I中$a^I_{i,i} \neq 0$。表5–4给出了信息流动网络的基本统计信息。

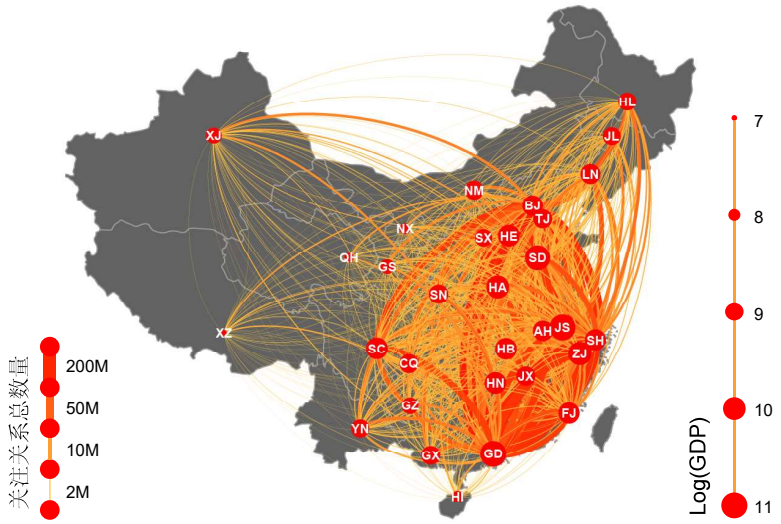

图 5–17 中国省份层面的在线信息流动网络

图5–17展示了中国省份层面的在线信息流动网络。其中，节点为省份，根据省会城市地理坐标布局。节点大小为自然对数下的省份GDP总量，以2016年国家统计局公布的数据为例。节点之间的连边为省份之间的信息流动关系，连边粗细

和颜色表示关注关系总量，连边越粗、颜色越深表示信息流动总量越大。可以看到，省份层面的信息流动网络非常稠密。沿海和经济发达省份的信息流动总量相对更大；西部地区省份的信息流动总量相对较小。另外，经济相对落后的省份，更倾向于关注经济发展好的省份。例如，黑龙江更多关注北京，四川更多关注广东，而反向关注数量相对较少。这些结果表明，社会网络的关注方向，即信息流动方向，一定程度上反映区域经济发展水平。

为了分析信息流动网络的结构特征与区域经济发展水平的关系，首先定义一些直接的网络结构指标，包括：S_{out}，S_{in}和S_{loop}。具体而言，给定一个区域i，$S_{out}(i)=\sum_j a_{i,j}$为出向连边权重的总和，即所有从区域$i$流出信息总量；$S_{in}(i)=\sum_j a_{j,i}$为入向连边权重的总和，即所有流入区域$i$的信息总量；$S_{loop}(i)=a_{i,i}$为自环的连边权重总和，即区域$i$内部流动的信息总量。然后，定义一些相对的网络结构指标，包括R_{io}，R_{lo}和R_{li}。具体而言，$R_{io}(i)=S_{in}(i)/S_{out}(i)$为流入和流出信息量的比值，体现保持自身信息的比率；$R_{lo}(i)=S_{loop}(i)/S_{out}(i)$为自环和流出信息的比值，体现信息流失的程度，$R_{lo}=0$和$R_{lo}=1$分别表示所有信息都流失和都保持；$R_{li}(i)=S_{loop}(i)/S_{in}(i)$为自环和流入信息的比值，体现获取信息的能力，$R_{li}=0$表示获得所有新信息，$R_{li}=1$表示保持所有自身的信息。

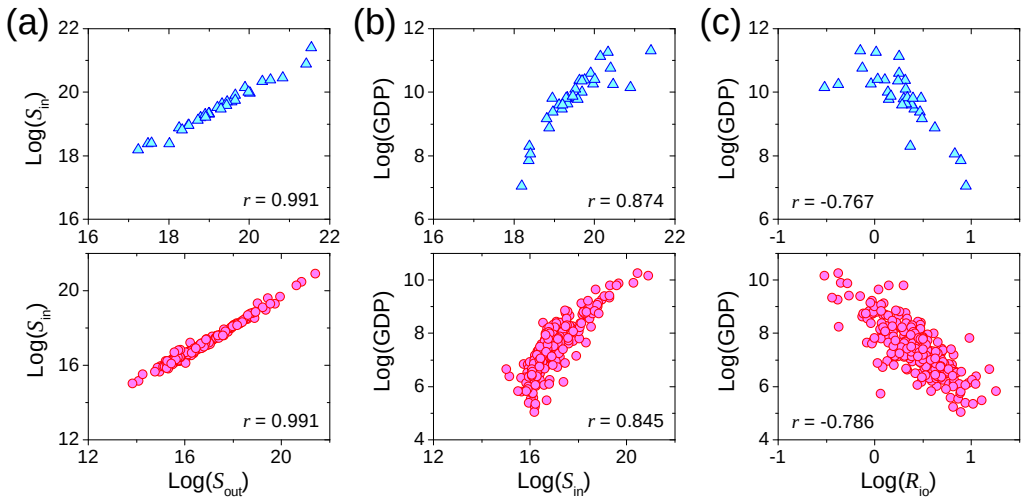

图 5–18 信息流动网络结构特征与区域经济发展指标的关联

图5–18展示了信息流动网络结构特征与经济发展指标之间的关系。从图5–18(a)看到，不论在省份层面（第一行），还是在城市层面（第二行），区域的信息流入总量（S_{in}）和信息流出总量（S_{out}）都非常相关，皮尔森关联高达$r=0.991$。这说明，有能力向外传播信息的省份，也有从外部吸引信息的能力，反之亦然。从图5–18(b)看到，信息流入总量（S_{in}）与GDP强相关，说明信息流入多的区域，经济发展水平越好。信息流入总量越大，经济发展越好。考虑

到信息流入和流出之间的强关联，吸引更多关注的区域，经济发展也越好。从图5–18(c)看到，自身信息保持比率（R_{io}）与GDP之间有很强的负相关，暗示信息保持能力越强的区域，经济发展水平越差。反之，信息流入更少、信息流出更多的区域，有更大的信息扩散能力，表现出更好的经济发展水平。

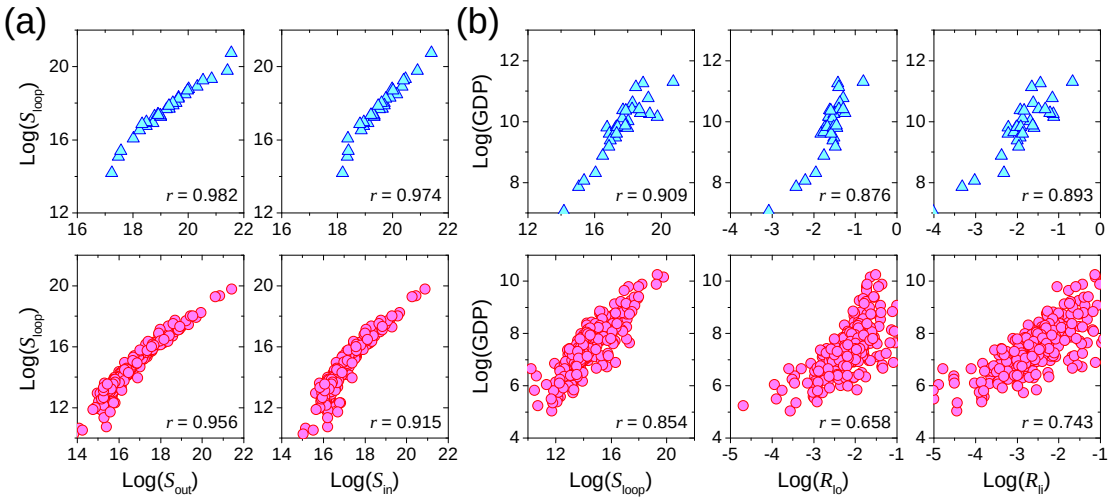

图 5–19 自环信息流动结构特征与区域经济发展指标的关联

图5–19展示了自环相关的信息流动网络结构特征与经济发展指标之间的关系。由图5–19(a)可知，信息自环流动总量与信息流入总量和流出总量都强相关。其中，省份层面的关联性强于城市层面的关联性。这说明，信息流动量大的区域，自环信息流动总量也大。由图5–19(b)左列可知，自环信息流动总量与GDP强相关，皮尔森关联在省份层面达到$r = 0.909$和在城市层面达到$r = 0.854$，说明自环信息量大的城市，其经济发展水平高。由图5–19(b)中列和右列可知，信息流失程度（R_{lo}）和信息获取能力（R_{li}）都与GDP很相关，体现出信息流动网络的两种相对结构特征对GDP的预测能力。此外，省份层面的关联性强于城市层面的关联性，说明相对结构特征对省份经济发展预测能力更好。

5.3.2 求职简历数据构建人才流动网络

人类移动和行为模式与社会经济状况之间存在紧密的联系[341, 345]。例如，Pappalardo等人[343]利用手机通讯数据刻画个体的移动行为模式，发现移动多样性与收入水平和初等教育比例显著相关。特别地，人才流动对经济发展至关重要，人才富集和人才流入量大的区域经济发展好[346, 347]。人才带来了经济发展机遇，经济快速发展也反过来吸引更多人才。传统方法不易获得人才流动数据，阻碍了从人才流动角度理解经济发展。近年来，新方法获取的大规模高质量社会经济数据，为估计区域间人才流动提供了便利。例如，State等人[348]通过分析LinkedIn上求职者的职业记录，发现亚洲技术移民比例在过去二十年逐渐增加。

本节研究中使用来自前程无忧和中华英才等在线招聘网站的求职者匿名简历数据，涵盖大约14.2万大专及以上教育水平的求职者[184]。简历数据中包含求职者的出生地、工作地和期望工作地等位置信息。综合考虑求职者从出生地到工作地的移动，以及从工作地到期望工作地的移动，粗略地估计区域之间人才的移动总量。地理位置涵盖287个地级市，归属31个省份。不考虑数据中没有涉及到的城市，以保证得到的人才流动网络是最大联通网络，不存在孤立的城市节点。

根据求职者的移动方向，构建区域之间的人才流动网络（OTM, Offline Talent Mobility）[170]，记为G^T。将人才流动网络G^T表示为含权邻接矩阵形式A^T，其中矩阵元素$a^T_{i,j}$为区域i流入区域j的人才总量。考虑到求职者可能出生、居住和工作在相同区域，人才相当于在区域内部流动。所以，人才流动网络G^T中存在自环，即邻接矩阵A^T中$a^T_{i,i} \neq 0$。表5–4给出了人才流动网络的基本统计信息。可以看到，人才流动网络节点数量仅为信息流动网络的大约三千分之一，其网络密度也相对较小。尤其在城市层面，人才流动网络相对比较稀疏，平均连边权重较小。

图 5–20 中国省份层面的离线人才流动网络

图5–20展示了中国省份层面的离线人才流动网络。其中，节点为省份，根据省会城市地理坐标布局。节点大小为自然对数下省份2016年的GDP总量。节点之间的连边表示人才流动关系，连边粗细和颜色表示人才流动总量。可以看到，人才流动活跃的区域大多集中在东南沿海省份，以及北京、四川和湖北等经济发展相对较好的省份；东北地区和中西部地区的省份，人才流动数量较少，并且存在很强的非对称性。例如，人才大部分由东北三省流入北京，而反向流动的数量较少；人才大部分由四川、湖北、湖南和广西流入广东，而广东省向外流出的数量微乎其微。人才在区域之间的非对称流动，体现出区域经济发展水平的差异。普遍而言，人才倾向于流入经济发展好的区域。

进一步，类似于信息流动网络分析方法，使用三种直接网络结构指标（S_{out}，S_{in}和S_{loop}）和三种相对网络结构指标（R_{io}，R_{lo}和R_{li}），量化分析人才流动网络的结构特征与区域经济发展水平之间的关系。具体而言，对于任意一个区域i，$S_{out}(i)=\sum_j a_{i,j}$表示所有从区域$i$流出的人才总量；$S_{in}(i)=\sum_j a_{j,i}$表示所有流入区域$i$的人才总量；$S_{loop}(i)=a_{i,i}$表示区域$i$内部流动的人才总量。$R_{io}(i)=S_{in}(i)/S_{out}(i)$体现区域保持自身人才的比率；$R_{lo}(i)=S_{loop}(i)/S_{out}(i)$体现区域人才流失的程度；$R_{li}(i)=S_{loop}(i)/S_{in}(i)$体现区域获得人才的能力。

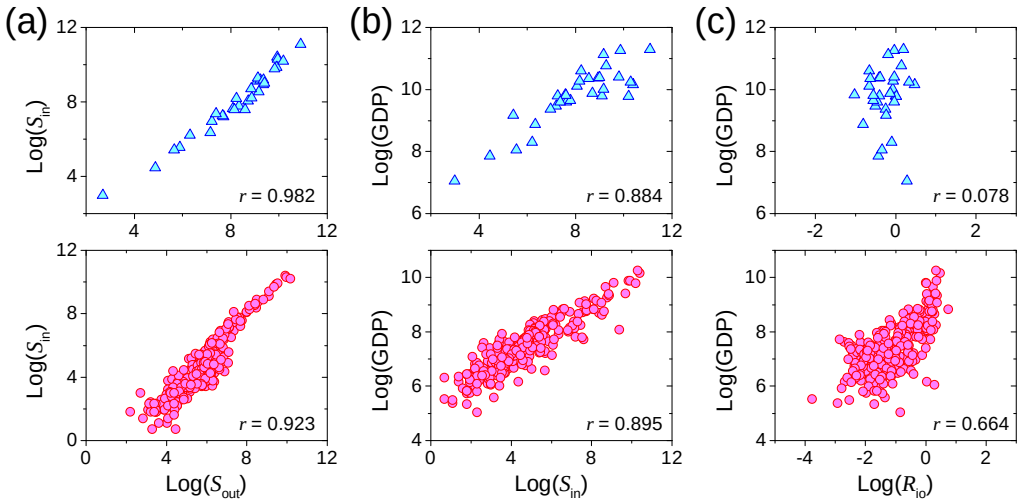

图 5–21 人才流动网络结构特征与区域经济发展指标的关联

图5–21展示了人才流动网络的结构特征与经济发展指标之间的关系。从图5–21(a)看到，区域的人才流入和流出总量之间强相关。其中，省份层面的关联性$r=0.982$明显强于城市层面的关联性$r=0.923$。从图5–21(b)左列看到，区域的人才流入总量与GDP之间强相关，省份和城市层面的关联程度相近，暗示人才流入量能很好地预测经济发展水平。考虑到人才流入和流出总量之间的强关联，区域的人才流出量也与经济发展水平很相关。从图5–21(b)右列看到，区域保持自身人才的比率仅在城市层面与GDP正相关，在省份层面的关联性接近于零，也就是说关联性消失了。可能因为省份中城市的经济发展水平不同，省内人才流动抵消了城市发展的不平衡，导致保持人才的比率与经济发展水平的关联性不强。

图5–22展示了自环人才流动相关的网络结构特征与经济发展指标之间的关系。其中，网络结构指标和GDP进行自然对数运算（本文中统一使用log符号表示）。由图5–22(a)可知，自环人才流动总量与区域的人才流出总量（左列）和人才流入总量（右列）都强相关，皮尔森相关系数接近1，说明区域内部的人才流动与总体的人才流动在趋势上保持一致。由图5–22(b)可知，不论在省份还是城市层面，区域的自环人才流动总量（S_{loop}）都与GDP强相关（左列），皮尔森关联达

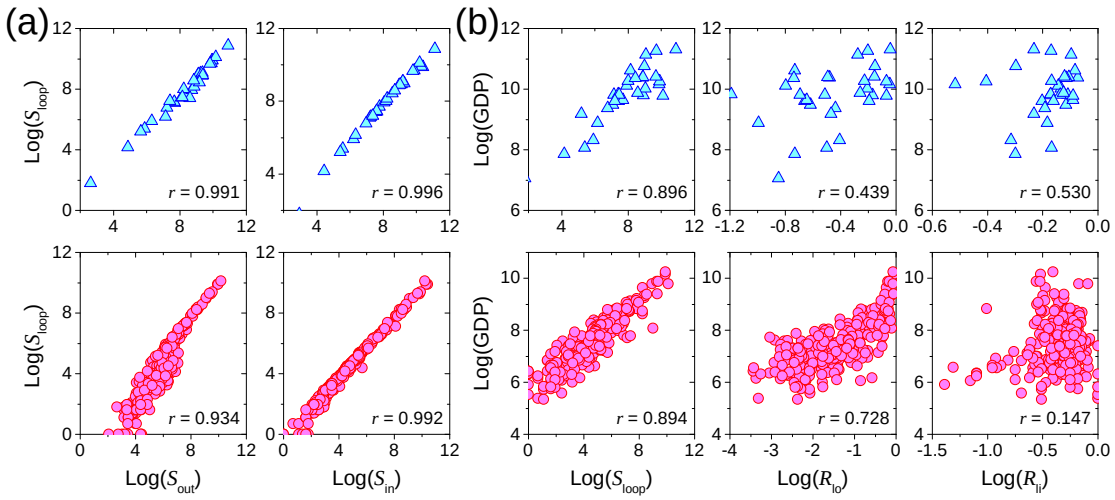

图 5–22 自环人才流动结构特征与区域经济发展指标的关联

到$r \approx 0.89$，体现自环人才流量对经济发展水平的预测能力。另外，区域获得人才的能力（R_{io}）在城市城市层面与GDP的关联性更强（中列），在省份层面的关联性较弱；区域人才流失的程度（R_{li}）仅在省份层面与GDP存在关联性（右列），在城市层面的关联性很弱。总体而言，与人才自环流动相关的网络结构指标与区域经济发展水平有关联，但研究尺度显著影响结果，在省份层面的关联性更强。

5.3.3 利用网络结构特征预测经济水平

利用社会网络的结构特征能分析区域社会经济状态，建立预测模型利用网络结构特征能预测区域经济发展水平。Eagle等人[66]开创性的研究了英国通讯网络结构特征与社区经济发展水平之间的关系。其中，通讯网络数据涵盖2005年英国超过90%的手机用户；社会经济发展指标采用2004年的复合剥夺指数。利用香农熵计算网络的两种结构多样性指标，关联分析社区的复合剥夺指数。发现复合剥夺指数的排序与通讯网络结构多样性指标之间强相关，皮尔森关联性高达0.73。这说明，利用社会网络的结构特征能一定程度上预测区域的经济发展水平。

类似地，根据信息流动网络和人才流动网络的结构特征，有希望准确地推断区域的经济发展水平。除了已有的三个直接的网络结构特征和三个相对的网络结构特征，参考Eagle等人[66]的工作，进一步定义网络结构的多样性指标。考虑到信息和人才流动的权重和方向，网络结构多样性指标包括两个拓扑多样性指标（H_{out}和H_{in}）和两个空间多样性指标（D_{out}和D_{in}）[170]。入向和出向的网络结构拓扑多样性（Topological Diversity）通过香农熵定义，分别对应于信息或人才流动入向和出向。具体而言，对于区域i，出向拓扑多样性$H_{out}(i)$定义为

$$H_{out}(i) = -\sum_{j \neq i} p_{i,j} \log(p_{i,j}). \tag{5–18}$$

其中，$p_{i,j}=a_{i,j}/\sum_j a_{i,j}$。区域$i$的出向空间多样性指标$D_{out}(i)$，通过将$H_{out}(i)$使用所涉及到的区域个数归一化来定义，即

$$D_{out}(i)=\frac{H_{out}(i)}{\log(k_{out}(i))}. \tag{5–19}$$

其中，$k_{out}(i)$为区域i的出度。类似地，将区域i的入向拓扑多样性$H_{in}(i)$定义为

$$H_{in}(i)=-\sum_{j\neq i}p_{j,i}\log(p_{j,i}). \tag{5–20}$$

其中，$p_{j,i}=a_{j,i}/\sum_j a_{j,i}$。区域$i$的入向空间多样性指标$D_{in}(i)$，通过将$H_{in}(i)$使用所涉及到的区域个数归一化来定义，即

$$D_{in}(i)=\frac{H_{in}(i)}{\log(k_{in}(i))}. \tag{5–21}$$

其中，$k_{in}(i)$为区域i的入度。信息流动网络在省份和城市层面都为完全图（见表5–4），区域i涉及到所有其他区域。所以，信息流动网络的拓扑多样性和空间多样性相等。后续仅关注信息流动网络的空间多样性，拓扑多样性分析结果相同。

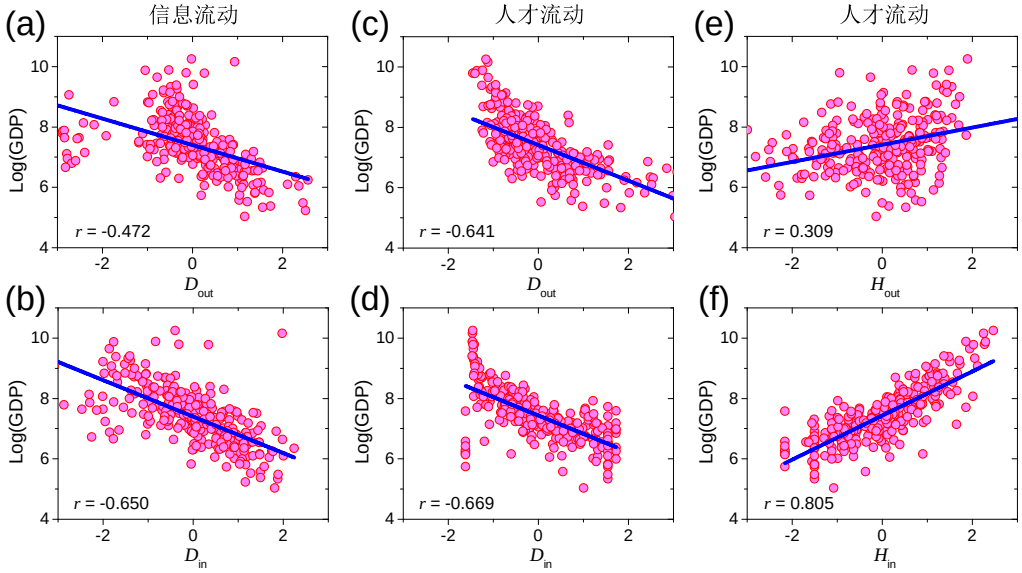

图 5–23 城市层面信息和人才流动网络结构多样性与区域经济发展指标的关联

图5–23展示了城市层面信息流动网络和人才流动网络的四种结构多样性指标与区域经济发展指标之间的关系。对于信息流动网络，图5–23(a)和图5–23(b)分别展示了出向（D_{out}）和入向（D_{in}）空间多样性与GDP之间的关系。可以看到，空间多样性都与经济发展水平之间有强负相关。其中，D_{out}的关联性为$r=-0.472$，D_{in}的关联性为$r=-0.650$。类似地，图5–23(d)和图5–23(d)展示了人才流动网络空间多样性与GDP之间的负相关，关联系数为$r\approx-0.66$。D_{in}在两个网络中都与GDP呈现更强的关联性。另外，空间多样性与区域的经济发展水平负相关，

这不同于Eagle等人[66]在英国社区层面发现的社会网络空间多样性与经济发展水平正相关。对于人才流动网络，图5–23(e)和图5–23(f)分别展示了出向（H_{out}）和入向（H_{in}）拓扑多样性与GDP之间的关系。可以看到，H_{in}与GDP的关联性（$r=0.805$）显著强于H_{out}与GDP的关联性（$r=0.309$），这说明入向拓扑多样性对经济发展水平的预测能力更强。

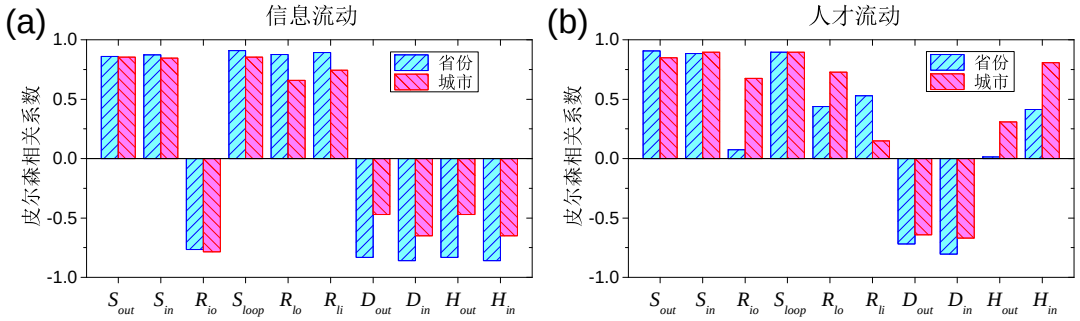

图 5–24 信息和人才流动网络的结构特征与经济发展指标的皮尔森关联系数

图5–24总结了信息和人才流动网络的所有结构特征指标与GDP之间的皮尔森关联系数。从图5–24(a)看到，对于信息流动网络，除R_{io}之外的简单网络结构指标与GDP正相关，网络结构多样性指标与GDP负相关。另外，省份层面的网络结构指标比城市层面的网络结构指标与GDP的关联性更强。特别地，关联性最强的网络结构特征，包括省份层面的信息自环特征和信息流动多样性指标，以及城市层面的信息自环特征和信息流动强度指标。从图5–24(b)看到，对于人才流动网络，在省份和城市层面与GDP关联性最强的网络结构指标比较一致，包括S_{out}、S_{in}和S_{loop}，体现出网络结构指标与经济发展水平之间的强关联。

进一步，验证网络结构特征对区域经济发展预测能力的鲁棒性，使用普通最小二乘法（OLS）模型，在城市层面利用网络结构特征对GDP进行回归分析。具体而言，所使用的OLS回归方程为

$$\begin{aligned}\log(GDP) = \beta_0 + \beta_1 S_{out} + \beta_2 S_{in} + \beta_3 R_{io} + \beta_4 S_{loop} + \beta_5 R_{lo} \\ + \beta_6 R_{li} + \beta_7 H_{out} + \beta_8 H_{in} + \beta_9 D_{out} + \beta_{10} D_{in} + \varepsilon.\end{aligned} \tag{5–22}$$

其中，$\{\beta_0, \beta_1, \ldots, \beta_{10}\}$为网络结构特征的回归系数，$\varepsilon$为误差项。除多样性结构特征之外，所有的网络结构特征参量都进行自然对数运算。判断共线性的变量时，计算容忍度（Tolerance）和方差膨胀因子（VIF, Variance Inflation Factor）。采用逐步回归方法筛选变量，解决变量多重共线性问题，保留重要网络结构特征。

表5–5总结了回归分析的结果，其中“−−”表示因共线性而忽略的变量。从第（1）列和第（2）中看到，包含信息流动和人才流动网络结构特征的模型能分别最多解释76.2%和80.2%的GDP变化。特别地，信息流动网络中的D_{out}和D_{in}分

表 5–5 网络结构特征对经济水平的预测能力

变量	信息流动	人才流动	两个网络	
	(1)	(2)	(3-1)	(3-2)
S_{out}	0.823^{***}	0.587^{***}	$--$	0.266^{***}
S_{in}	$--$	$--$	0.363^{***}	$--$
R_{io}	0.217^{**}	0.300^{***}	0.051	0.128^{**}
S_{loop}	$--$	$--$	$--$	$--$
R_{lo}	0.216^{***}	$--$	0.087^{*}	$--$
R_{li}	$--$	0.041	$--$	0.059^{**}
D_{out}	0.208^{***}	-0.010	0.203^{***}	-0.096
D_{in}	-0.291^{***}	0.013	-0.192^{***}	0.010
H_{out}	$--$	0.067^{**}	$--$	0.016
H_{in}	$--$	0.103	$--$	0.120^{*}
Obs.	290	280	280	
Adj. R^2	0.762	0.802	0.832	

统计显著性水平：$^{*}p<0.1$；$^{**}p<0.05$；$^{***}p<0.01$

别对GDP有显著地正向和负向的预测能力，人才流动网络中的H_{out}对GDP有正向预测能力。第（3）列同时包含了两个网络的结构特征，其中第（3-1）列对应于信息流动网络，第（3-2）列对应于人才流动网络。可以看到，两个网络的结构最多能解释83.2%的GDP变化，体现出最强的解释能力。另外，自环是能最好地解释GDP变化的网络结构特征。对于信息流动网络，解释能力为$R^2=74.3\%$；对于人才流动网络，解释能力为$R^2=79.8\%$。结果表明，人才流动网络的结构特征对区域经济发展水平的预测能力更强。

基于回归分析的结果，构建网络结构特征的复合指标，实现对区域经济发展水平的最佳预测。具体而言，通过将网络结构指标根据回归系数加权来计算复合指标。对于区域i，复合指标$CI(i)$定义为

$$CI(i)=\sum_{j=1}^{10}\beta_j^I M_{j,i}^I+\sum_{j=1}^{10}\beta_j^T M_{j,i}^T. \tag{5–23}$$

其中，$M=\{\vec{S}_{out},\vec{S}_{in},\vec{R}_{io},\vec{S}_{loop},\vec{R}_{lo},\vec{R}_{li},\vec{D}_{out},\vec{D}_{in},\vec{H}_{out},\vec{H}_{in}\}$为计算得到的10种网络结构特征指标，$\vec{\beta}=\{\beta_1,\ldots,\beta_{10}\}$为表5–5中给出的相应网络结构指标的回归系数。其中，M^I和β^I 对应于信息流动网络，M^T和β^T对应于人才流动网络。在构建复合指标之前，所有的网络结构特征指标都采用Z-score[221]方法进行标准化。

图5–25展示了城市层面复合网络结构指标与经济发展水平之间的关系。其中，图5–25(a)和图5–25(b)分别对应于信息流动网络和人才流动网络。可以看到，对于两个网络，复合网络指标都与GDP之间强相关。特别地，人才流动网络的

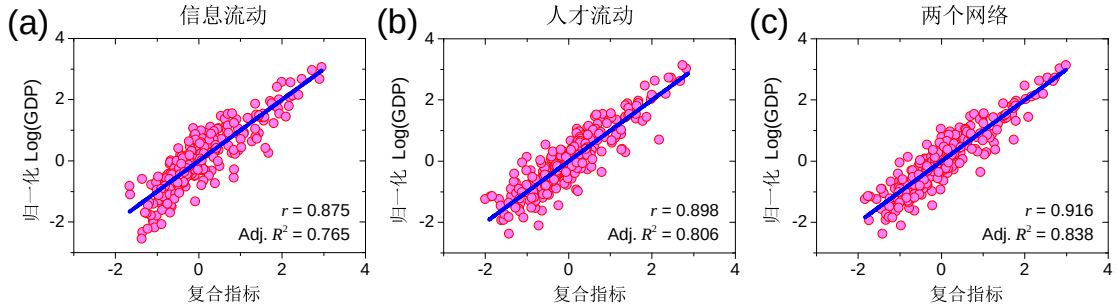

图 5–25 城市层面复合网络结构指标与经济发展水平指标之间的关联

复合指标与GDP之间的关联性（$r = 0.898$）稍微强于信息流动网络的复合指标与GDP之间的关联性（$r = 0.875$）。信息和人才流动网络的复合指标能分别最多解释76.5%和80.6%的GDP变化。这些结果表明，信息流动和人才流动都对区域经济发展水平有很强的解释和预测能力。进一步，图5–25(c)展示了复合网络结构指标与GDP之间的关系。可以看到，复合结构指标与GDP之间有最强的关联性（$r = 0.916$），能最多解释83.8%的GDP变化。这说明，综合信息流动和人才流动网络的结构特征，能提高对区域经济发展水平的推断能力。

5.4 本章小结

随着大规模社会经济数据可用性的提高，借助网络分析工具能对经济发展结构和多样性进行建模，从结构上理解区域经济发展潜力，甚至利用结构特征推断经济状态。本章从宏观层面介绍了社会经济系统的结构建模研究。基于区域内产业的组成情况，建模刻画区域的经济复杂性，分析复杂性指标对经济发展的预测能力。进一步，利用劳动力市场数据和企业注册信息数据构建产业空间，分析产业空间的结构特征以及区域的产业竞争。最后，基于在线平台数据构建信息和人才流动网络，建立预测模型利用网络结构特征推断区域经济发展水平。

经济发展伴随着多样性的提高和经济结构的转变，传统方法无法刻画经济发展的结构特征，也难以预测未来经济发展。本章第5.1节介绍了区域经济复杂性建模方法，比较了复杂性对传统指标的预测能力。首先，介绍了利用贸易数据计算复杂性ECI指标和竞争力Fitness指标。然后，基于企业注册信息数据构建“省份-产业”二部分网络，计算中国区域经济复杂性ECI指标和区域竞争力Fitness指标。结果发现，沿海省份的经济复杂性更高；省份在经济复杂性排序上保持相对稳定和缓慢的演化。最后，分析经济复杂性ECI指标与人均GDP的演化，发现省份的“复杂性ECI-人均GDP”相图分为两个区域，经济复杂性对发展省份的预测能力强。另外，关联分析和回归分析结果显示，收入不平等性和经济复杂性之间显著负相关，ECI指标和Fitness指标对GDP的解释能力相当。

从网络视角分析产业结构，能更直观地展现经济发展的结构特点，感知经济发展的态势。本章第5.2节介绍了区域产业空间建模和分析方法。首先，基于巴西劳动力市场数据构建“产业-职业”二部分网络，利用余弦相似性计算产业接近性，构建巴西区域产业空间。发现巴西产业空间有显著的“核心-边缘”结构，复杂程度高和低的产业分别占据核心和边缘。然后，基于中国企业注册信息数据构建“省份-产业”二部分网络，利用类似的方法构建中国区域产业空间。发现中国产业空间有“核心-边缘”结构和“哑铃型”结构，复杂程度高的制造业和信息服务业分别占据核心。进一步，利用三种方法计算产业接近性来构造产业空间，发现产业空间的“核心-边缘”结构有鲁棒性。分析中国产业空间演化，发现近邻区域存在产业竞争，北京和上海占据信息服务业，周围省份占据制造业。

在线社会网络是信息流动的载体，人才流动体现经济发展的趋势，信息和人才对区域经济发展至关重要。本章第5.3节介绍了信息流动和人才流动网络结构对区域经济发展水平的预测能力。首先，基于微博关注关系数据构建信息流动网络，发现经济相对落后的省份，更倾向于关注经济发展好的省份；信息保持能力越强的区域，经济发展水平越差。然后，基于匿名简历数据构建人才流动网络，发现人才倾向于流入经济发展水平好的区域；区域保持自身人才的比率，仅在城市层面与GDP呈现正相关；两个网络的空间多样性指标都与GDP呈现很强的负关联，但只有人才流动网络的入向拓扑多样性指标与GDP呈现很强的关联。结合两个网络的结构特征来构建复合指标，能最多解释大约83.8%的GDP变化，体现出信息流动和人才流动网络的结构特征对区域经济发展水平的推断能力。

第六章 经济结构演化路径与发展策略研究

理解经济结构演化和发展路径，揭示产业发展和升级的普遍规律，对制定区域产业策略至关重要。本章从三个方面介绍经济结构演化规律与发展策略研究。首先，介绍社会经济系统的空间网络模型，分析空间网络上的信息传播动力学过程，研究网络的空间结构对信息传播的影响。然后，介绍经济发展过程中的两种学习途径，即产业空间网络上的相似技术学习和地理空间网络上的近邻区域学习，分析两种学习途径的相互作用。最后，介绍基于空间网络的最优经济发展学习策略，分析高铁对近邻学习效果的促进作用，提出基于空间网络的两种最优产业发展策略，介绍国际双边贸易中的知识扩散和三种促进贸易的策略。

6.1 社会经济空间网络结构与传播动力学

现实中的很多网络都有空间嵌入信息，形成有特殊结构的空间网络（Spatial Network）[110]，例如一些社会网络和经济网络[117, 349]。已有研究分析在线社会网络[350]、邮件网络[351]和手机通讯网络[111]等，发现真实社会网络中普遍存在空间标度律（Spatial Scaling Law）。具体而言，长度为r的连边的概率密度函数服从标度$P(r) \sim r^{\alpha}$，幂指数α的数值接近于-1[112]。事实上，早在实证发现这一标度律之前，Kleinberg[109]就已经提出了一种空间网络模型，利用在二维方格网络上添加长程连边的方式构造空间网络。分析发现，当概率密度分布服从$P(r) \sim r^{-1}$时，空间网络的结构能实现最优导航。最近，Hu等人[112]提出了一种空间标度律的解释，认为这种特殊的空间网络结构与最优信息收集密切相关。

网络结构对网络上信息传播的影响，能解释很多社会经济现象，例如社会网络上的信息传播[352]、模型网络上的疾病传播[353, 354]、新产品推广和社会行为采纳[355, 356]、国家发展新产品和产业扩散[12, 357]等。在研究网络上的传播动力学上[354, 358]，已经提出了很多理论模型，例如线性阈值模型[70]、疾病传播SIR模型[359]、各类相关的变体模型[354]等。其中，靴襻渗流（Bootstrap Percolation）模型[202, 203]是最具代表性的传播动力学过程，有很大的社会经济意义。网络结构和传播模型都对信息扩散有很大影响，所以选择有代表性的空间网络和传播模型，分析网络结构对信息扩散、新产品采纳和新产业发展的影响。

本节针对一般的社会经济系统，构建有代表性的空间网络模型，分析网络结构对信息传播的影响。首先，介绍空间网络模型的构造方法和靴襻渗流模型，研究空间网络中长程连边的分布对信息传播范围和传播速度的影响。进一步，介绍

空间网络上信息传播的临界现象，给出相变类型和相变点的判断方法。最后，研究更一般的空间网络结构对信息传播的影响，分析长程连边的数量和分布对相变类型和相变点数值的影响。

6.1.1 社会经济系统中的空间网络模型

空间结构能改变网络的维度[360–362]，导致网络的很多物理性质发生改变，包括网络的脆弱性、抗毁性和鲁棒性等[363–365]。Moukarzel等人[366]研究了长程连边空间网络上的k-核渗流过程（k-Core Percolation）。其中，空间网络通过在二维方格网络中添加长程连边来构造，长度为r的长程连边的概率密度分布服从$P(r) \sim r^{\alpha}$。结果发现，当幂指数$\alpha > -1.75$时，3-核渗流为一级相变；否则，相变类型为二级相变[367]。实际上，k-核渗流过程与靴襻渗流过程在很多方面有紧密的联系[368]，两种传播在不同网络结构上又有一些区别[203, 369]。

首先，介绍一种典型的空间网络模型，称之为Kleinberg模型[109]，该模型已经被社会经济系统中的很多真实数据所验证[111, 351]。类似于Moukarzel等人[366]构建空间网络的方法，基于有周期边界的二维方格网络，构建包含$N = L \times L$个节点的无向Kleinberg空间网络，其中L为方格网络的边长。除了周围连接的四个近程节点之外，每个节点i还以概率$Q_i(r_{ij})$随机连接到一个远程节点：

$$Q_i(r_{ij}) \sim r_{ij}^{\alpha-1}. \tag{6–1}$$

其中，α 为可调参数，控制长程连边的长度；r_{ij}为曼哈顿距离，度量节点i和节点j之间的最短路径长度。在d-维方格网络中，到给定节点距离为r的节点数量正比于r^{d-1}，得到的概率函数$Q(r_{ij})$ 能转换为概率密度函数：

$$P(r) \sim r^{d-1} \cdot Q(r) = r^{d-1} \cdot r^{\alpha-1} = r^{\alpha+d-2}. \tag{6–2}$$

对于二维方格网络，$d = 2$，得到的概率密度函数能表示为$P(r) \sim r^{\alpha}$。

在构建空间网络时，任意两个节点之间都添加一条无向的长程连边。具体步骤如下[204]：第一步，在长度2和$L/2$之间以概率$P(r) \sim r^{\alpha}$随机生成一个长度r，以此保证空间标度律。第二步，将长度r随机拆分成两个整数长度Δx和Δy，需要满足$|\Delta x| + |\Delta y| = r$，以确定候选节点。如果节点$i$的坐标为$(x, y)$，那么候选节点的坐标为$(x + \Delta x, y + \Delta y)$，以此保证到目标节点$i$距离为$r$的所有候选节点是随机分布的。这样能从未配对的候选节点中随机选取节点，构建长度为r的长程连边。当α很小时，存在候选节点都已经被匹配的情况，需要使用距离粗粒化方法[370]，随机选取候选节点的近邻节点，直到能完成匹配。第三步，重复以上过程，直到所有节点都添加了一条长程连边，空间网络中每个节点的度都为5。

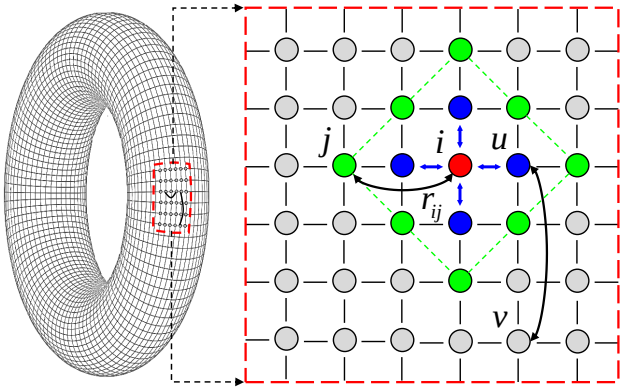

图 6–1 基于二维方格网络构建的空间网络模型

图6–1展示了一个基于二维方格网络构建且带有长程连边的空间网络模型示意图。左图为带有周期边界的二维方格网络，每个节点仅与周围的四个节点相连。右图为加入长程连边的空间网络局部示意图，除了周围四个节点之外，每个节点还有一条长程连边，将其与远距离的节点相连。图中位于中心的红色节点i为目标节点，远程蓝色节点到节点i的距离为$r=2$，远程绿色节点到节点i的距离为$r=3$。节点i与节点j之间添加的长程连边的长度为$r_{ij}=2$；节点u和节点v之间添加的长程连边的长度为$r_{uv}=3$。

然后，介绍靴襻渗流模型[202, 203]。谚语“三人成虎”体现出周围邻居的影响[355]，如果周围很多朋友向你传递一个信息，你更容易相信[371]，这是靴襻渗流的一种直观解释。靴襻渗流可以看做是网络上的节点激活过程：（i）节点处于活跃态或非活跃态；（ii）节点一旦被激活，就保持活跃态；（iii）初始时刻，每个节点以概率p处于活跃态；（iv）如果一个非活跃态节点有至少k个邻居处于活跃态，那么该节点被激活；（v）以迭代方式遵照步骤（iv）激活节点，直到没有新节点能被激活为止。通过这样的传播过程，信息从局部向全局扩散。也有一些广义靴襻渗流模型，例如阈值k由邻居比例替代的Watts阈值传播模型[70]。

利用靴襻渗流模型分析空间网络中长边分布对信息传播范围和时间的影响，重点关注三个指标：（1）终态下第一极大活跃连通子图的相对规模S_{gc}，即一个随机选取的节点最终属于极大活跃连通子图的概率；（2）达到终态所需的迭代步数NOI，能用来判定一级相变点的位置[372]；（3）终态下第二极大活跃连通子图的相对规模S_{gc2}，能用来判断二级相变点的位置[373]。图6–2展示了不同结构的空间网络上的靴襻渗流结果。其中，随着长边分布的幂指数α从-5增加到5，空间网络中长边的长度逐渐增加。如果下文中不单独指出，默认空间网络中每个节点仅添加一条长边，并且使用参数为$k=3$的靴襻渗流模型。

从图6–2(a)看到，当$\alpha \geq -1$时，第一极大活跃连通子图的相对规模$S_{gc}(p)$曲

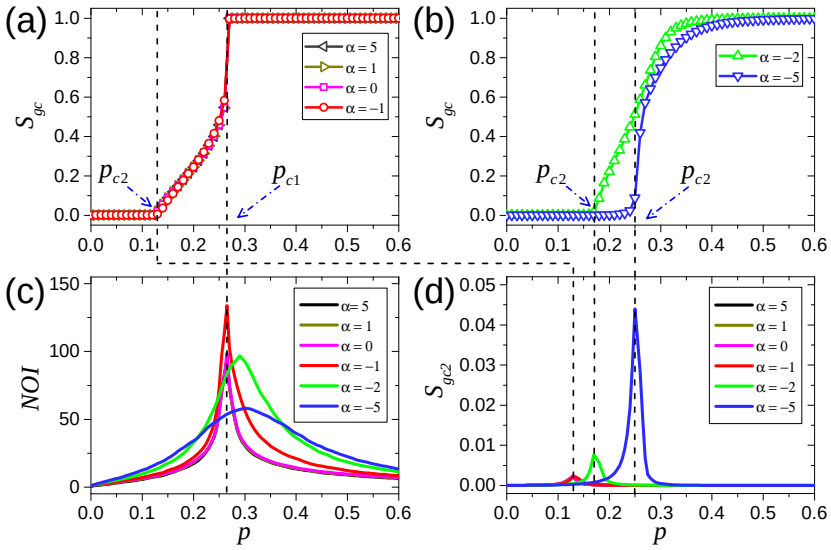

图 6–2 空间网络中长程连边分布对靴襻渗流传播范围和时间的影响

线几乎重叠在一起，呈现出包含二级相变（Second-Order）和混合（Hybrid）相变的双相变。具体而言，S_{gc}在到达$p_{c2} \approx 0.134$（二级相变点）时连续增加，相变类型为二级相变；S_{gc}在到达$p_{c1} \approx 0.263$（一级相变点）时发生从0.58到1的直接跳变，相变类型为混合相变。有趣的是，当$\alpha \geq -1$时，这两个相变点的数值似乎是保持不变的，因为观察到四条$S_{gc}(p)$曲线重叠在一起。从图6–2(b)看到，当$\alpha < -1$时，仅存在二级相变，二级相变点p_{c2}的数值随着α的减小而增大。具体而言，当$\alpha = -2$时，相变点$p_{c2} \approx 0.176$；当$\alpha = -5$，相变点增大到$p_{c2} \approx 0.256$。尽管当α减小时的S_{gc}在p超过p_{c2}后迅速上升，但数值模拟的结果显示$S_{gc}(p)$曲线仍是连续的，意味着$\alpha < -1$时的相变类型仍是二级相变。

通过模拟的方法确定相变类型往往很困难，需要巧妙的方法和很高的分辨率。当$\alpha \geq -1$时，双相变中的一部分是混合相变，能利用迭代次数NOI来判断一级相变点的位置[372]。对于一级相变而言，当p接近相变点p_{c1}时，NOI迅速增加到最大值。所以，通过绘制NOI随p的变化曲线，能确定p_{c1}的数值。从图6–2(c)看到，当$\alpha \geq -1$时，NOI在几乎相同的p处达到最大值，得到一级相变点数值为$p_{c1} \approx 0.263$。类似地，通过绘制第二极大活跃连通子图的相对规模S_{gc2}随p的变化曲线，能确定二级相变点p_{c2}的数值。对于二级相变，S_{gc2}在相变点p_{c2}处达到最大值。从图6–2(d)看到，二级相变点p_{c2}随α的减小而增大。当$\alpha \geq -1$时，$p_{c2} \approx 0.134$；当$\alpha = -2$时，$p_{c2} \approx 0.176$；当$\alpha = -5$时，$p_{c2} \approx 0.256$。

6.1.2 空间网络上信息传播的临界现象

空间网络的结构影响传播过程中序参量的相变类型。对于无向的Kleinberg空间网络，当$\alpha \geq -1$时，S_{gc}存在包含二级相变和混合相变的双相变；当$\alpha < -1$时，

S_{gc}仅存在二级相变。然而，对于有限系统，很难判断存在混合相变，也不容易利用模拟方法确定幂指数的临界值α_c。为了解决该问题，使用交叉验证方法同时确定一级相变点p_{c1}和幂指数的临界值α_c。具体而言，首先固定幂指数$\alpha=-1$来确定一级相变点p_{c1}。一方面，图6–3(a)展示了不同网络规模L下的S_{gc}随初始激活节点比例p的变化曲线。不同网络规模的$S_{gc}(p)$曲线相交在同一个点，交点的横坐标为$p'_{c1}\approx 0.2625$。根据有限尺度分析理论[374]，交点的横坐标p'_{c1}为一级相变点数值。另一方面，图6–3(b)展示了不同网络规模L下的迭代步数NOI随p的变化曲线。随网络规模的增大，$NOI-p$曲线的峰值变大。当$L=800$时，NOI最大值对应的$p''_{c1}\approx 0.2635$也作为一级相变点数值。结合这两个结果，将平均值$p_{c1}=(p'_{c1}+p''_{c1})/2\approx 0.263$作为一级相变点数值。

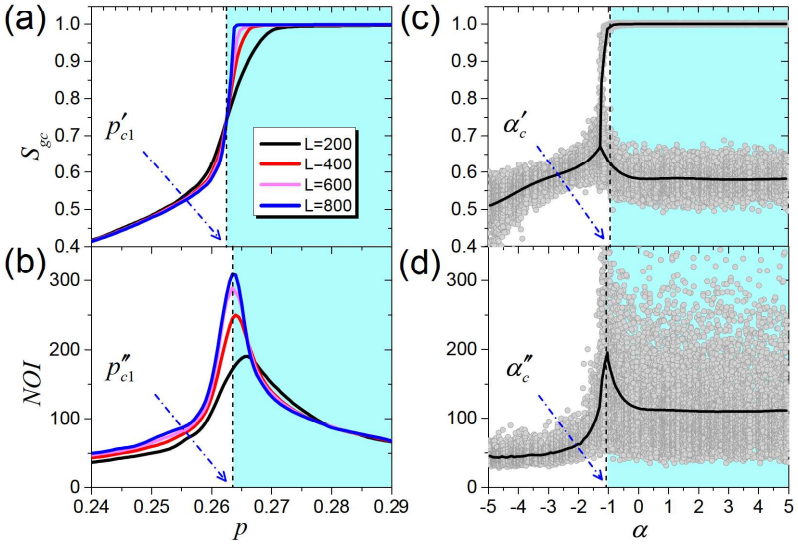

图 6–3 交叉验证一级相变点和改变相变类型的幂指数的临界值

反过来，将p固定为一级相变点的数值（即$p=0.263$）来确定幂指数的临界值α_c。一方面，图6–3(c)展示了S_{gc}随幂指数α的变化。从散点图中可以看到，当幂指数$\alpha\geq -0.95$时，S_{gc}分为两相：一个位于0.58附近，另一个接近达到1。这说明S_{gc}在此时发生了混合相变，出现了不连续的两种状态。如果S_{gc}是连续增加的，就观察不到两相之间的间隙。两相分离时对应的幂指数，作为临界值$\alpha'_c\approx -0.95$。另一方面，图6–3(d)展示了迭代步数NOI随p的变化。可以看到，NOI均值存在明显的峰值，该峰值所对应的横坐标为$\alpha=-1.05$。NOI均值达到峰值的横坐标，也作为幂指数的临界值$\alpha''_c\approx -1.05$。结合这两个结果，将平均值$\alpha_c=(\alpha'_c+\alpha''_c)/2\approx -1$作为幂指数的临界值。当$\alpha\geq\alpha_c$时，$S_{gc}$出现包含二级相变和混合相变的双相变；否则，$S_{gc}$仅出现二级相变。

值得注意的是，当$\alpha\geq\alpha_c$时，混合相变中的一级相变点数值p_{c1}应当是不变

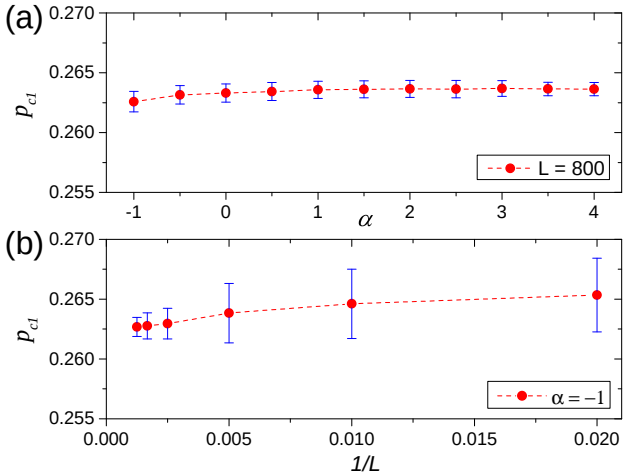

图 6–4 长程连边分布和网络规模对一级相变点数值的影响分析

的。否则，对于固定的$p=0.263$，就无法在图6–3(c)中观察到两个相的分离。为了验证这一猜想，改变长边分布的幂指数α，在$L=800$下估计一级相变点p_{c1}的误差。图6–4(a)展示了发生混合相变的情况下，一级相变点随幂指数α的变化。可以看到，当α逐渐接近临界值$\alpha_c\approx-1$时，一级相变点数值p_{c1}略微减小。尽管这样，当α在区间$[-1,4]$中变动时，p_{c1}的最大值和最小值之间仅相差0.0008，对于p的整个取值范围[0,1]而言非常小。这些结果表明，当$\alpha\geq-1$时，p_{c1}的数值对幂指数α不敏感。近似地认为p_{c1}是固定值，大小为0.2634，通过p_{c1}的均值计算。进一步，以$\alpha=-1$为例，考虑网络的有限尺度对结果的影响。从图6–4(b)看到，当网络规模L趋于无穷时，p_{c1}的平均值逐渐收敛到0.263附近的固定值，p_{c1}的标准差逐渐减小。用相同方法分析二级相变点，得到p_{c2}数值为0.134附近的固定值。

为了更系统地分析空间网络的结构对靴襻渗流相变类型的影响，在不同幂指数和初始激活比例的参数组合（α,p）下观察S_{gc}的变化。图6–5展示了靴襻渗

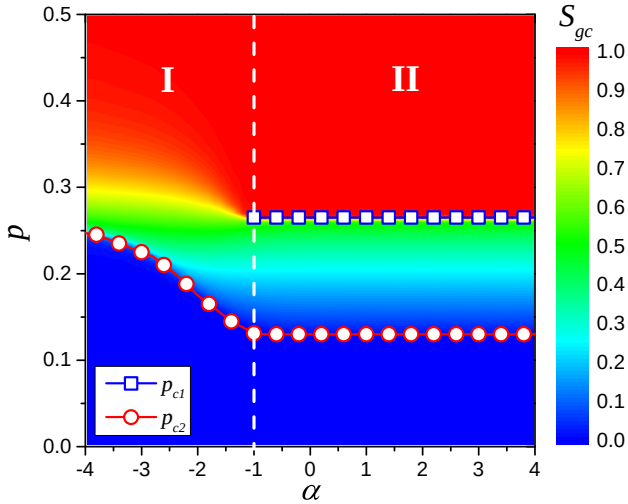

图 6–5 空间网络上靴襻渗流在$\alpha-p$平面上的相图

流在$\alpha-p$平面中的相图，其中颜色表示平衡状态下S_{gc}的数值。可以看到，当幂指数α变化时，S_{gc}的相变类型发生改变。总体而言，$\alpha_c \approx -1$为幂指数的临界值。当$\alpha \geq -1$时，S_{gc}出现一个双相变，如图中标记的II区域。特别地，$S_{gc}(p)$曲线是重叠的，说明这些空间网络有类似的靴襻渗流特征。当$\alpha < -1$时，S_{gc}仅出现二级相变，如图中标记的I区域。二级相变点数值随α的减小而增大，p_{c2}的最大值在0.259附近，在$\alpha \to -\infty$时估计得到，这时所有长边的长度都为2。

6.1.3 网络结构对信息传播的影响分析

现实中的空间网络比Kleinberg空间网络模型要复杂的多。为了进一步探究空间结构对靴襻渗流的影响，在更一般的空间网络上进行不同参数的靴襻渗流，关注长边分布的幂指数临界值$\alpha_c \approx -1$。具体而言，在参数空间（k, α, k_l）中进行数值模拟，确定相变点数值。其中，k为靴襻渗流的节点激活阈值：当有至少k个邻居处于活跃态时，节点被激活，默认$k=3$；α为控制长边分布的可调参数，长边的长度r随α的增大而增大；k_l为每个节点的无向长边的数量，默认$k_l=1$。

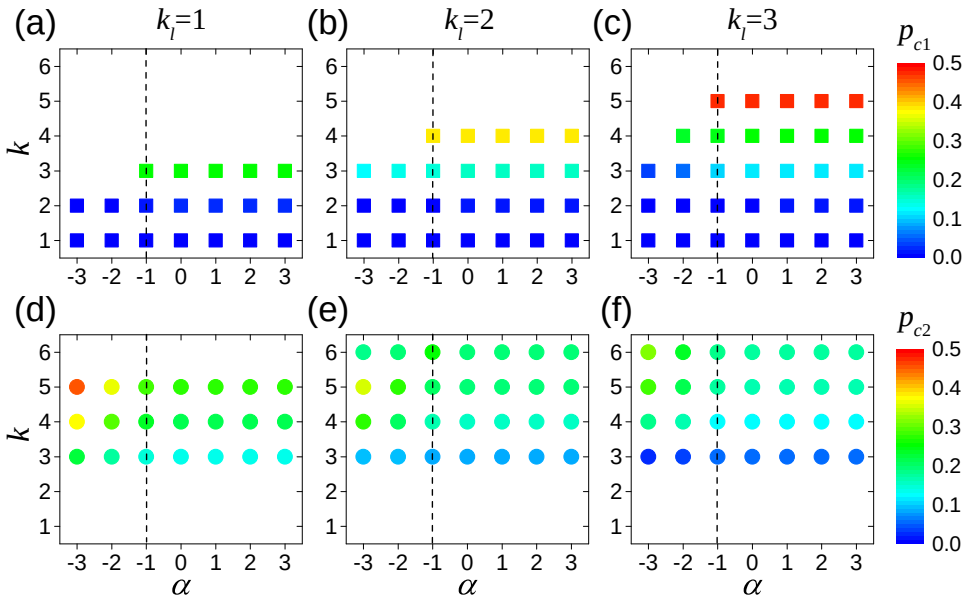

图 6–6 空间网络上靴襻渗流在参数空间（k, α, k_l）中的相图

图6–6展示了无向Kleinberg空间网络上靴襻渗流在参数空间（k, α, k_l）中的相图。其中，横坐标为长边分布的幂指数α；纵坐标为靴襻渗流的阈值参数k；每个节点的长边数量k_l逐渐增加（从左列到右列）；相图中的颜色表示一级相变点p_{c1}和二级相变点p_{c2}的数值。可以看到，相图被渗流阈值参数k和网络平均度的一半$\langle k_N\rangle/2$（其中$\langle k_N\rangle = k_l+4$）分为三个区域：

- 当k显著的小于$\langle k_N\rangle/2$时，例如$k=1$与$\langle k_N\rangle/2=2.5$相比，仅出现一级相变，一级相变点$p_{c1}\approx 0$，意味着极少初始激活节点就能激活整个网络。

- 当k在$\langle k_N\rangle/2$附近时，例如$k=3$与$\langle k_N\rangle/2=2.5$相比，存在一个幂指数的临界值α_c，当α超过α_c时出现双相变。临界值α_c的大小，受k和k_l共同影响。特别地，$\alpha_c^*\approx -1$是一个不依赖于参数的临界值，超过该临界值时，两个相变点的数值几乎保持不变。具体而言，当$\alpha\geq -1$时，对于相同的参数k而言，p_{c1}和p_{c2}的数值几乎是固定的。当$\alpha_c\leq\alpha<\alpha_c^*$时，随着$\alpha$的增加，$p_{c1}$减小，$p_{c2}$增大。值得注意的是，在一些参数组合下，$\alpha_c$等于$\alpha_c^*$。
- 当k显著的大于$\langle k_N\rangle/2$时，例如$k=5$与$\langle k_N\rangle/2=2.5$相比，双相变中的混合相变消失，S_{gc}仅出现二级相变。随着幂指数α的逐渐增大，二级相变点的数值p_{c2}逐渐减小。

进一步的实验结果表明，这些结果对于长程连边有向的Kleinberg空间网络也成立，$\alpha_c^*\approx -1$仍然是幂指数的临界值[204]。值得注意的是，当$\alpha\geq -1$时，S_{gc}仅出现一级相变，相变点p_{c1}几乎保持不变。这与长程连边无向的Kleinberg空间网络上的结果有所区别，不会观察到双相变中的二级相变过程。另外，使用基于无周期边界的方格网络构建的无向的Kleinberg空间网络，靴襻渗流的主要结果不会受到显著影响，说明方格网络的周期边界并不改变所观察到的相变类型和所得到的结论，体现出分析结果的鲁棒性和普适性。

为了理解这些现象背后的机制，在不同网络模型上进行数值模拟，对比分析相关的相变类型。使用的网络模型包括：简单二维方格网络（Lattice）、长程连边网络（LR）和无空间结构的网络（RR）[204]。其中，LR网络中每个节点都仅连接$k_l=5$条长边，没有初始连接的近程连边。图6–7展示了靴襻渗流在不同网络模型上的相变类型。可以看到，无向Kleinberg空间网络所对应的$S_{gc}(p)$曲线位于Lattice网络和RR网络之间。当$\alpha=-4$时，空间网络的$S_{gc}(p)$曲线与Lattice网络的类似，这时网络中长程连边很少，相变类型仍是二级相变。当$\alpha\geq -1$时，空

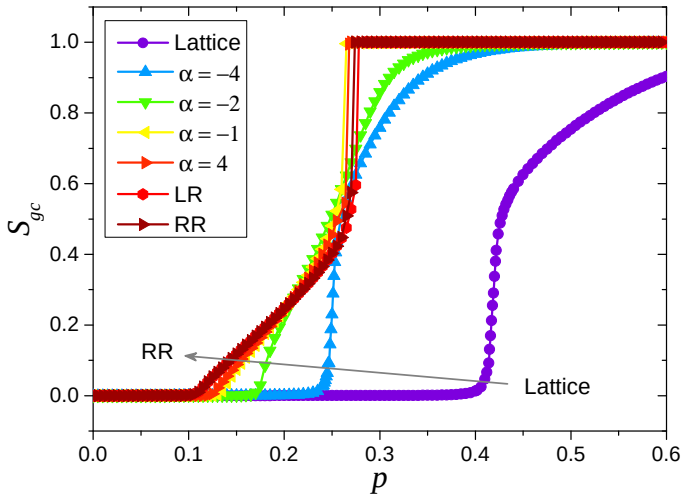

图 6–7 不同网络模型对$k=3$靴襻渗流相变类型的影响分析

间网络的$S_{gc}(p)$与RR网络的几乎重叠在一起，这时出现双相变。这些结果表明，通过调节幂指数α，能将空间网络的靴襻渗流特性从Lattice网络变为RR网络。具体而言，当$\alpha=-4$时，所有长边都被限制在局部，空间网络与Lattice网络类似；当$\alpha\geq-1$时，由于部分长边的存在，空间网络表现出RR网络的特性。

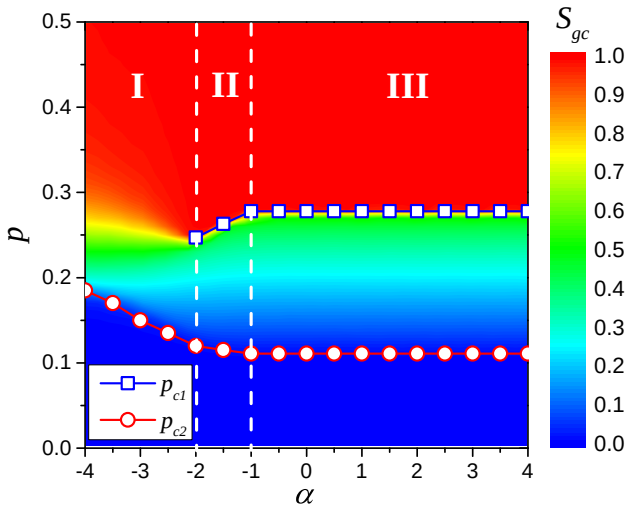

图 6–8 长程连边网络$k_l=5$对$k=3$靴襻渗流模型的相图

图6–8展示了$k_l=5$的LR网络上$k=3$靴襻渗流的相图，帮助更好地理解长边分布的幂指数α如何影响LR网络的相变类型。可以看到，相图被幂指数的临界值$\alpha_c\approx-2$和$\alpha_c^*\approx-1$分为三个部分。当$\alpha<\alpha_c\approx-2$时（区域I），S_{gc}仅出现二级相变，相变点p_{c2}数值随α的增大而减小。当$\alpha_c\approx-2\leq\alpha<\alpha_c^*\approx-1$时（区域II），$S_{gc}$出现双相变，随着$\alpha$的减小，一级相变点$p_{c1}$减小，二级相变点$p_{c2}$增大。当$\alpha\geq\alpha_c^*\approx-1$时（区域III），$S_{gc}$出现相变点数值几乎不变的双相变，一级相变点为$p_{c1}\approx0.278$，二级相变点为$p_{c2}\approx0.111$。

这些结果说明，空间网络结构的影响网络上信息传播，这对空间网络上信息的最优传播和控制有借鉴意义。例如，当$\alpha\geq\alpha_c$时，如果要让尽可能多的人获得信息，那么初始获得信息的人员比例应该为p_{c1}，因为更多的初始获得信息比例会花费更多成本，却不能带来更多收益。这些发现帮助更好地理解社会空间网络的自组织结构，有助于分析传播信息的最佳策略。另外，所使用的研究方法对于研究其他空间网络上的扩散问题有启发意义，例如利用靴襻渗流模型研究区域发展新产业受产业空间网络和地理近邻网络结构的影响。

6.2 经济发展过程的学习途径与路径依赖

经济发展是一个学习过程和路径依赖过程[375]，分析区域如何学习出口新产品和发展新产业是理解经济发展的关键[199]。已有一些研究发现，经济发展有协

同学习（Collective Learning）效应[376]，存在两条学习途径。一是相似技术学习（Inter-Industry Learning）途径，从同一区域的相似经济活动中学习[199]。例如，国家倾向于出口与已有出口产品相似的产品[25]；区域更容易发展与已有产业接近的新产业[69]。产业空间建模和网络结构分析，为研究相似技术学习提供了基础。二是近邻区域学习（Inter-Regional Learning）途径，从周围邻居区域的相同经济活动中学习[199]。例如，有很多邻居国家已经出口某个产品，那么这个国家更容易出口该产品[159]；周围邻居区域中有某个产业的区域，有更高的概率在未来发展或维持该产业[161]。空间网络结构和传播动力学分析，为研究近邻区域学习提供了基础。

协同学习理论已经被用来探究经济发展的微观机制，已有研究从国家到区域跨越不同尺度[25, 67]，涉及到北美、欧洲和亚洲等大陆[131, 156]，基于不同类型和体量的数据[161, 199]。然而，由于数据和方法的限制，经济发展途径仍然值得深入研究。首先，协同学习更可能源于经济的投入方面。相比体现经济产出的产品数据[24, 25]，劳动力市场数据能更好地体现经济发展的能力[377]。其次，协同学习是否在不同经济发展阶段的国家也适用，包括有不同经济和文化背景的中国和巴西等，这一点仍然缺乏基于大规模数据的分析和检验。

本节研究中使用巴西劳动力市场数据和中国企业注册信息数据，分析区域经济发展过程中的两条学习途径，即相似技术学习和近邻区域学习，以体现产业发展的路径依赖。首先，根据劳动力市场数据和企业注册信息数据，分别构建中国和巴西产业空间，分析已有相似的产业密度对区域发展新产业的影响。然后，基于产业空间计算不同区域的产业结构相似性，探究地理距离如何影响区域的产业相似性，分析已有产业的邻居区域密度对区域发展新产业的影响。最后，分析相似技术学习和近邻区域学习对区域发展新产业的共同影响，探究两条学习途径之间的相互作用，利用不同建模方法验证分析结果的鲁棒性。

6.2.1 区域经济发展的相似技术学习

相似技术学习（Inter-Industry Learning）在国家、区域和企业层面已有研究。在国家层面，国家在新产品上取得比较优势的概率依赖于国家已显著出口的相关产品数量[25, 68]。利用产品空间[25]计算与国家未出口产品相似的已出口产品密度，能预测国家在未来出口该产品的可能性。在区域层面，区域发展一个新产业的概率随区域内已有相关产业数量的增加而增大[67]。在企业层面，企业更容易拓展他们的产品组合，以包含相似产业所生产的产品[378]。这些结果表明，经济发展中的相似技术学习是一个路径依赖过程[375]，新经济活动的出现受局部相关经济活动密度的影响。下面，将分别介绍巴西和中国区域经济发展的相似技术学习。

巴西劳动力市场数据来自年度社会信息报告（RAIS）[158, 327]，涵盖来自501个职业的7662万员工，时间从2006年到2013年。涉及企业所处的区域涵盖558个Microregion，涉及产业涵盖669个Class。采用产业所提供职业的相似性计算产业之间的接近性，产业α与产业β在时间t的接近性为$\phi_{\alpha,\beta,t}$。进一步，利用产业接近性矩阵构建巴西区域产业空间，展示区域内有比较优势的产业（计算细节见第五章第5.2节）。简单而言，区域i中产业α在t时间的比较优势定义为

$$\mathrm{RCA}_{i,\alpha,t} = \frac{x_{i,\alpha,t}}{\sum_{\alpha} x_{i,\alpha,t}} \Big/ \frac{\sum_{i} x_{i,\alpha,t}}{\sum_{\alpha}\sum_{i} x_{i,\alpha,t}} . \tag{6–3}$$

其中，$x_{i,\alpha,t}$为区域i中产业α在t时间的员工数量。如果$\mathrm{RCA}_{i,\alpha,t} \geq 1$，那么产业$\alpha$在区域$i$中有比较优势，表示区域$i$中产业$\alpha$是活跃的。

在研究区域发展新产业的规律时，需要明确区域内出现新产业的定义。利用区域内在产业中工作的员工数量，限定$t+2$时间（相比于初始的t时间）区域内出现新产业需要符合两个条件：1）产业出现。区域在t时间没有员工在产业中工作，而在$t+2$时间至少有5位员工在产业中工作；2）产业持续。作为后向条件，区域需要在$t-1$时间没有员工在产业中工作。作为前向条件，区域需要在$t+3$时间保持至少有5位员工在产业中工作。采用这种方法定义新产业的出现，过滤区域中一些暂时存在的新产业，降低随机性对分析结果的影响。

为了度量区域内已经有多少相关产业处于活跃状态，对每个产业都计算活跃的相似产业密度（ω）[327, 379]。具体而言，通过$\mathrm{RCA} \geq 1$定义区域内的活跃产业。对于t时间的产业α和区域i而言，活跃的相似产业密度$\omega_{i,\alpha,t}$定义如下：

$$\omega_{i,\alpha,t} = \frac{\sum_{\beta} \phi_{\alpha,\beta,t} U_{i,\beta,t}}{\sum_{\beta} \phi_{\alpha,\beta,t}}. \tag{6–4}$$

其中，$\phi_{\alpha,\beta,t}$为产业α和产业β在t时间的接近性；如果$\mathrm{RCA}_{i,\beta,t} \geq 1$，那么$U_{i,\beta,t} = 1$；否则，$U_{i,\beta,t} = 0$。基于巴西劳动力市场数据计算每个产业的活跃的相似产业密度ω，并将ω与区域在未来两年内出现新产业的概率联系起来。

图6–9展示了巴西区域内活跃的相似产业密度ω与区域在未来两年内发展新产业的关系。其中，图6–9(a)将所有产业分为两类，即两年内出现新产业和两年内没有新产业，分别给出这两种情况对应的相似产业密度ω的概率分布。可以看到，两年内出现的新产业，所对应的ω均值大；两年内没有出现的产业，所对应的ω均值小。图6–9(b)给出了区域在两年内出现新产业的概率随该产业的活跃的相似产业密度的变化。可以看到，随着相似产业密度ω的增大，区域内新产业出现的概率增大。如果区域内已经存在很多相关的活跃产业，那么区域有更大的概率在未来发展这种产业，这说明了产业接近性对区域发展新产业的影响[67, 154]。

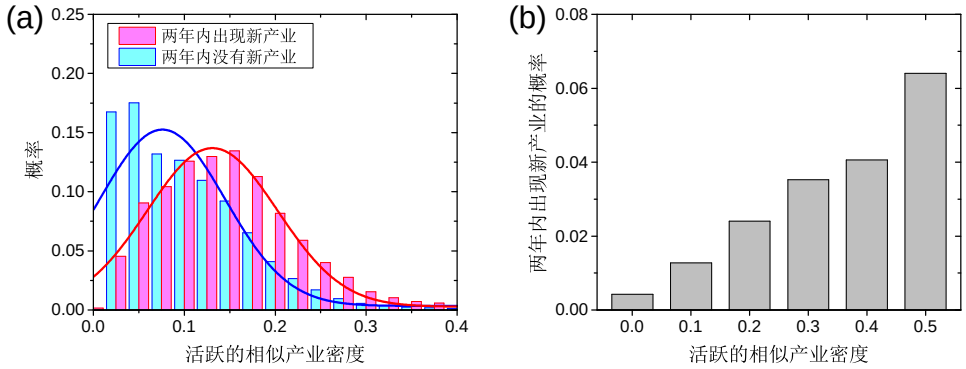

图 6–9 巴西区域内活跃的相似产业密度与未来两年出现新产业的关系

类似地，分析中国区域经济发展的相似技术学习[199]。中国企业注册信息数据涵盖沪深A股上市公司的注册信息和财务信息，时间从1990年到2015年。涉及企业所处的区域覆盖31个省份，涉及产业涵盖18个大类和70个小类。利用产业共同出现计算产业接近性，产业α与产业β在t时间的接近性为$\phi_{\alpha,\beta,t}$。根据产业接近性矩阵构建中国区域产业空间，展示区域内有比较优势的产业（计算细节见第五章第5.2节）。进一步，明确区域内出现新产业需要满足的两个条件：1）前项条件。开始时间t的前两年，产业在省份内的比较优势RCA小于1，即产业不活跃。2）后向条件。结束时间$t+5$的后两年，产业在省份内的RCA大于1，即产业保持活跃。采用该方法定义新产业出现，尽量降低随机性对分析结果的影响[159, 199]。

基于中国企业注册信息数据计算活跃的相似产业密度ω，将ω与区域在未来五年内出现新产业的概率联系起来。图6–10展示了省份内ω与省份在未来五年出现新产业的关系。其中，图6–10(a)给出了五年内出现和没有出现的产业对应的ω的概率分布。可以看到，五年内出现的新产业，所对应的ω的均值大；五年内没有出现的产业，所对应的ω的均值小；两个概率密度分布的均值存在显著差异，ANOVA检验的p-value为2.1×10^{-40}。图6–10(b)给出了省份在五年内出现新产业的概率随该产业的ω的变化。可以看到，五年内出现新产业的概率随着ω的增大而增大，说明省份内已有相似的活跃产业能提高省份发展新产业的可能性。

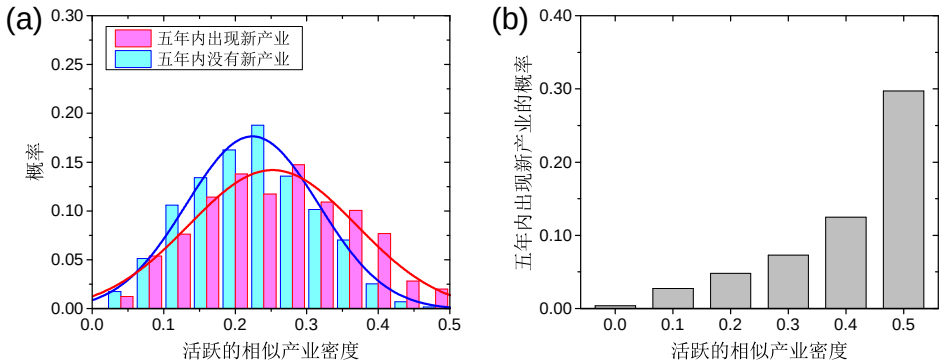

图 6–10 中国省份内活跃的相似产业密度与未来五年出现新产业的关系

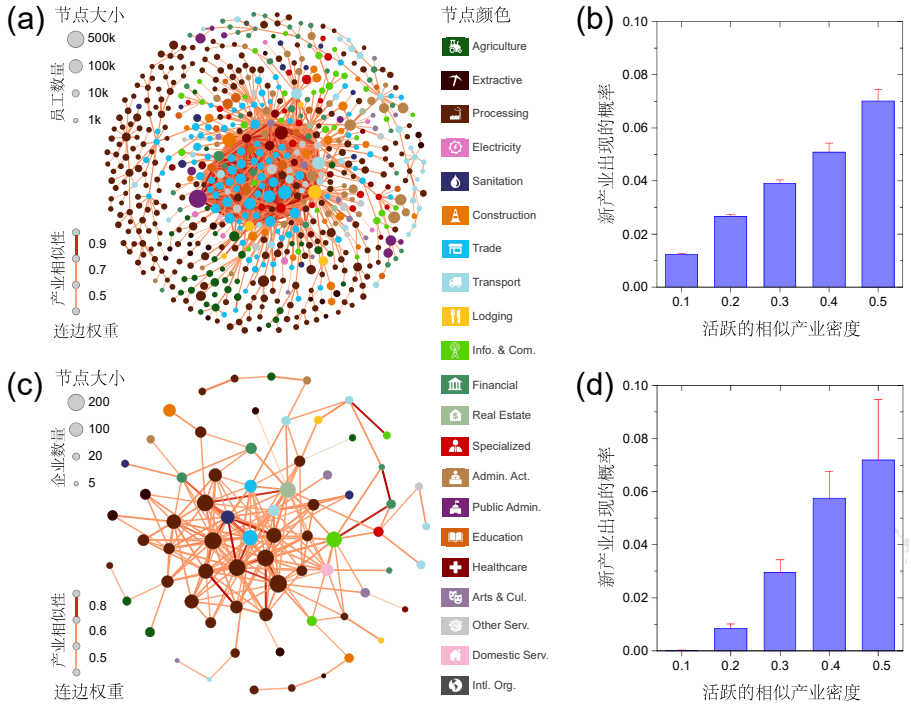

图 6–11 产业空间和区域经济发展的相似技术学习途径

为了更全面地阐述区域经济发展的相似技术学习途径，图6–11在统一分析框架下呈现产业空间和区域内活跃的相似产业密度ω对新产业出现概率的影响。构建产业空间时，产业接近性基于产业在区域内的共同出现，采用余弦相似性计算[199]。明确区域内新产业出现时，以产业在区域内的比较优势RCA为判断标准，采用前向和后向限定条件[199]。图6–11(a)和图6–11(c)分别展示了巴西和中国区域产业空间。其中，节点为产业，颜色为产业分类；节点大小为员工数量（巴西）或企业数量（中国）；连边粗细和颜色为产业接近性的大小。图6–11(b)和图6–11(d)分别展示了新产业出现概率随ω的变化。可以看到，巴西和中国在相似技术学习途径上遵循相同规律，新产业出现的概率随ω的增大而增大。

6.2.2 区域经济发展的近邻区域学习

近邻区域学习（Inter-Regional Learning）研究经济发展过程中如何从周围邻居区域中学习。在国家层面，如果一个国家的邻居国家已经成功出口一种产品，那么这个国家在未来出口这种产品的概率显著增加[159]，在控制产品接近性的情况下也是如此。在区域层面，区域更有可能发展已经在周围区域中出现的产业[161]。在企业层面，新沃尔玛商店的位置倾向于接近已经有很多沃尔玛商店的区域[160]。这些结果表明，经济发展的近邻区域学习也是一个路径依赖过程[375]，周围邻居区域的经济活动密度影响区域内新经济活动的出现。下面，将分别介绍巴西和中国区域经济发展的近邻区域学习。

首先，分析地理上邻近的区域是否有相似的产业结构。对于任意两个区域，采用余弦相似性度量他们产业结构之间的相似性。具体而言，利用$y_{i,\alpha,t}=\ln(\mathrm{RCA}_{i,\alpha,t}+1)$和$y_{j,\alpha,t}=\ln(\mathrm{RCA}_{j,\alpha,t}+1)$表示$t$时间产业$\alpha$在区域$i$和区域$j$中的比较优势程度。其中，$\mathrm{RCA}_{i,\alpha,t}$为产业$\alpha$在区域$i$中的比较优势，由公式（6–3）给出。将$t$时间区域$i$和区域$j$之间的产业相似性$\varphi_{i,j,t}$定义为

$$\varphi_{i,j,t}=\frac{\sum_{\alpha}y_{i,\alpha,t}y_{j,\alpha,t}}{\sqrt{\sum_{\alpha}(y_{i,\alpha,t})^2}\sqrt{\sum_{\alpha}(y_{j,\alpha,t})^2}}. \tag{6–5}$$

基于巴西劳动力市场数据，首先计算产业在区域中的比较优势RCA，然后根据公式（6–5）计算区域之间的产业相似性，最后把区域之间的产业相似性和地理距离联系起来，分析地理距离对区域产业结构之间相似性的影响。

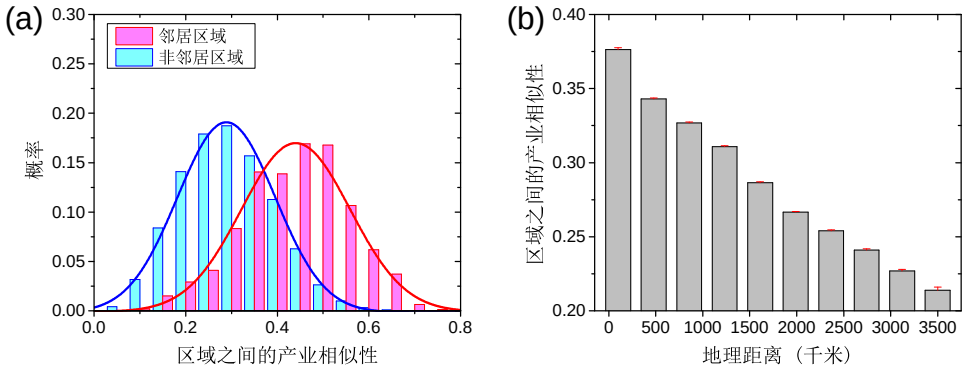

图 6–12 区域近邻关系和地理距离对产业相似性的影响分析

图6–12展示了区域近邻关系和地理距离对产业相似性的影响。其中，图6–12(a)中将区域分为邻居区域和非邻居区域，分别展示区域产业相似性的概率分布。邻居区域指区域在地理位置上相邻，非邻居区域指区域不共享边界。可以看到，邻居区域之间的产业接近性有很大的均值，非邻居区域之间的产业接近性有显著小的均值。这说明，邻居区域之间的产业结构更相似。图6–12(b)给出了区域产业相似性与地理距离的关系。可以看到，区域产业结构相似性与地理距离负相关，说明知识扩散随地理距离的增加而强烈衰减[159]。

然后，分析经济发展的近邻区域学习。已经发现，近邻区域更容易存在相似的产业结构，这意味着区域发展新产业会受到周围邻居区域的影响。为了度量产业已经在周围多少区域中处于活跃状态，对每个区域计算一个活跃的邻居区域密度（Ω）[161, 199]。同样地，通过$\mathrm{RCA}\geq 1$定义区域内的活跃产业。对于t时间的产业α和区域i，将活跃的邻居区域密度$\Omega_{i,\alpha,t}$定义为

$$\Omega_{i,\alpha,t}=\sum_{j}\frac{U_{j,\alpha,t}}{D_{i,j}}\Bigg/\sum_{j}\frac{1}{D_{i,j}}. \tag{6–6}$$

其中，$D_{i,j}$为区域i和区域j之间的地理距离。如果$\mathrm{RCA}_{i,\beta,t}\geq 1$，那么$U_{i,\beta,t}=1$；否

则，$U_{i,\beta,t}=0$。基于巴西劳动力市场数据，计算每个区域的活跃的邻居区域密度Ω，将Ω与区域在未来两年内出现新产业联系起来。

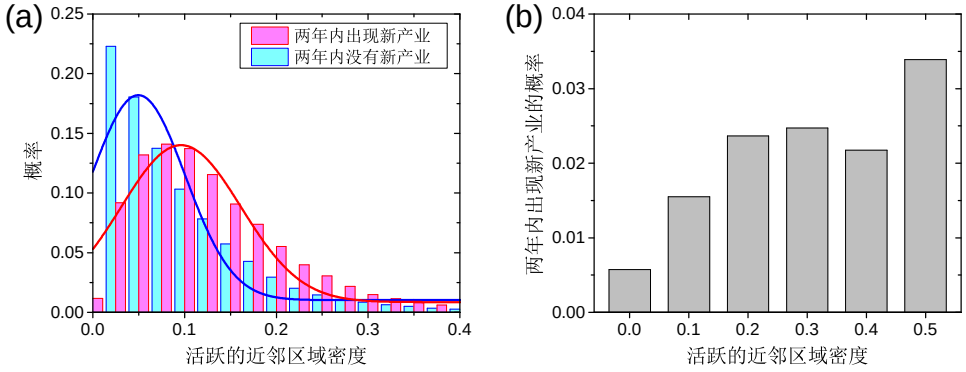

图 6–13 巴西区域周围活跃的邻居区域密度与未来两年内出现新产业的关系

图6–13展示了巴西区域周围活跃的邻居区域密度Ω与区域在未来两年内出现新产业的关系。其中，图6–13(a)给出了两年内出现的新产业和两年内没有出现的产业对应的Ω的概率分布。可以看到，两年内出现的新产业对应的Ω的均值比两年内没有出现的产业对应的Ω的均值大，说明将要出现的新产业在周围邻居区域已经有很大的产业密度。图6–13(b)给出了区域在两年内出现新产业的概率随该区域Ω的变化。可以看到，新产业出现的概率随Ω的增大而增大。如果区域周围已经有很多区域发展了一种产业，那么该区域有更大的概率在未来发展这种产业，这说明了周围邻居区域对发展新产业的影响[159, 161]。

类似地，基于中国企业注册信息数据计算省份周围活跃的邻居区域密度Ω，分析Ω与新产业发展之间的关系。图6–14展示了中国省份的Ω与省份在未来五年内出现新产业的关系。其中，图6–14(a)给出了五年内出现和没有出现的产业所对应的Ω的概率分布。可以看到，五年内出现的新产业所对应的Ω的均值比五年内没有出现的产业所对应的Ω的均值大，说明五年内出现的新产业已经在足够多的周围省份内活跃。图6–14(b)给出了省份在五年内出现新产业的概率随该省份的Ω的变

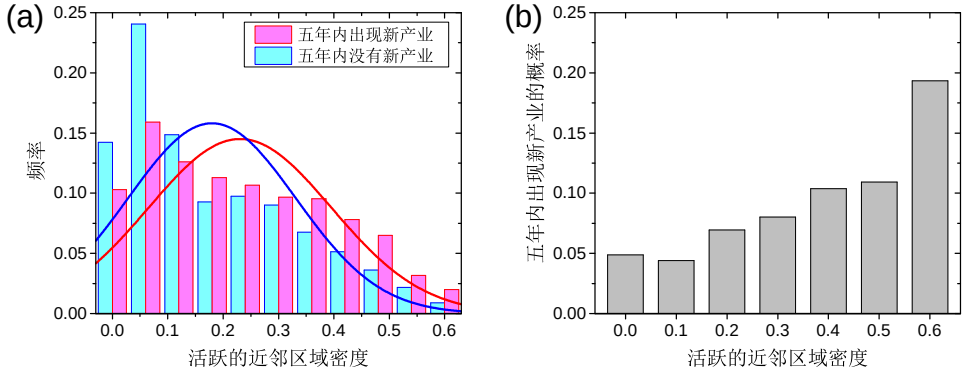

图 6–14 中国省份周围活跃的邻居区域密度与未来五年内出现新产业的关系

化。可以看到，五年内出现新产业的概率随着Ω的增大而增大，说明有足够多活跃的邻居区域能增大新产业发展的可能性。

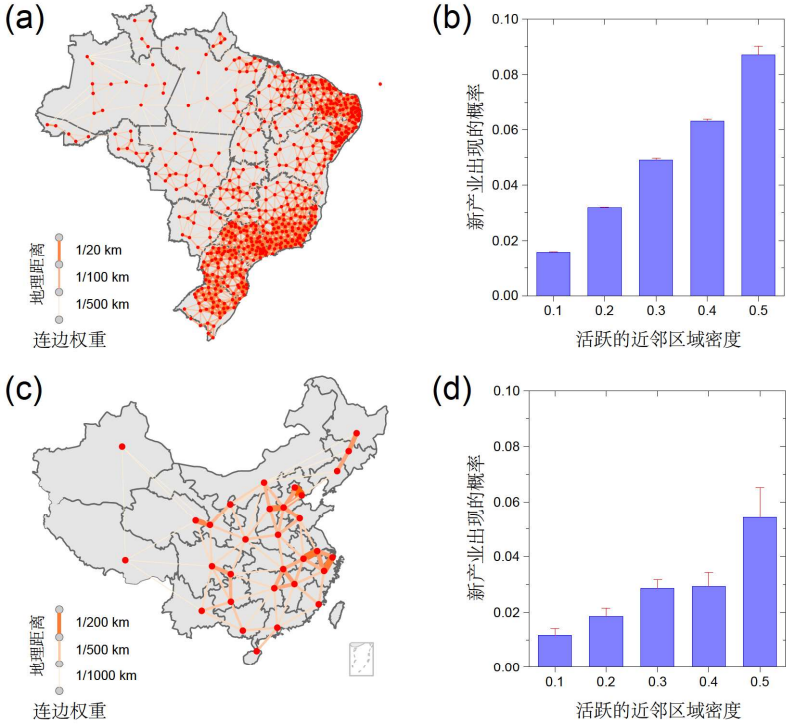

图 6–15 地理近邻网络和区域经济发展的近邻区域学习途径

为了更全面地阐述经济发展的近邻区域学习途径，图6–15在统一分析框架下呈现地理近邻网络和活跃的邻居区域密度Ω对新产业出现的影响。构建地理近邻网络时，使用地理距离的倒数作为连边权重。距离越近的两个省份，互相影响的程度越大[199]。明确区域内新产业出现时，以产业在区域内的比较优势RCA为判断标准，采用前向和后向限定条件[199]。图6–15(a)和图6–15(c)分别展示了巴西和中国地理近邻网络。其中，节点为区域（巴西）或省份（中国），对应中心城市地理坐标；连边粗细和颜色为区域之间地理接近性的大小。图6–15(b)和图6–15(d)分别展示了新产业出现概率随Ω的变化。可以看到，巴西和中国在近邻区域学习途径上遵循同样规律，新产业出现的概率随Ω的增大而增大。

6.2.3 两条学习途径的相互作用分析

分析巴西劳动力市场数据和中国企业注册信息数据，发现区域经济发展存在协同学习（Collective Learning）的两条途径，即相似技术学习和近邻区域学习[199, 327]。进一步，分析这两条学习途径的相互作用，在区域经济发展中是起联合作用，还是起替代作用。在分析时，首先利用图形统计方法给出直观的结果展示，然后利用多变量回归验证分析结果的鲁棒性。

利用公式（6–3）计算产业在区域中的比较优势RCA，确定区域中有比较优势的活跃产业。对于巴西劳动力市场数据，产业激活要求两年内区域中员工数量从0到不少于5，开始前一年员工数量为0，结束后一年员工数量不少于5（计算细节见第6.2.1节）。对于中国企业注册信息数据，产业激活要求五年内省份中产业的比较优势由RCA <0变为RCA ≥ 0，开始前两年保持RCA <0，结束后两年保持RCA ≥ 0（计算细节见第6.2.2节）。在此基础上，利用公式（6–4）计算区域内活跃的相似产业密度ω，利用公式（6–6）计算区域周围活跃的近邻区域密度Ω。进一步，同时考虑ω和Ω，计算新产业在区域内出现的联合概率。

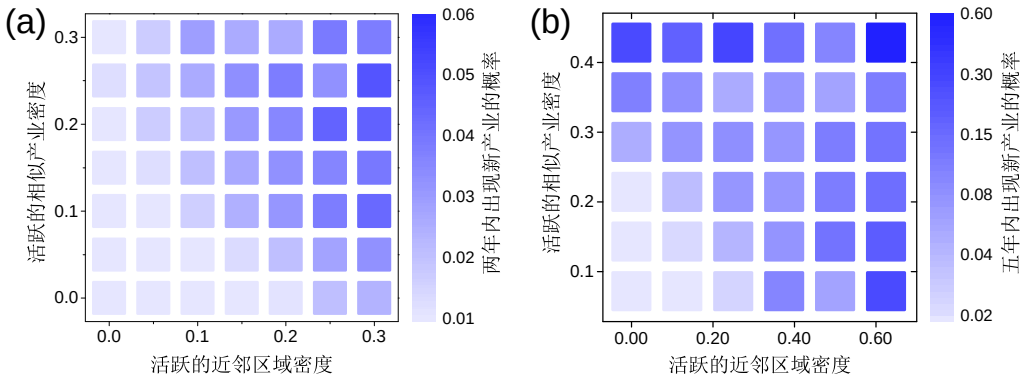

图 6–16 新产业出现相对于相似产业密度和近邻区域密度的联合概率

图6–16展示了区域出现新产业随ω和Ω变化的联合概率。其中，图6–16(a)为巴西区域在两年内出现新产业，图6–16(b)为中国区域在五年内出现新产业。可以看到，不论巴西还是中国，随着ω和Ω的共同增大，新产业出现的联合概率增大。特别地，巴西区域在两年内出现新产业的最大联合概率为0.06，中国区域在五年内出现新产业的联合概率达到0.60，这暗示中国区域经济发展中催生新产业的能力更强。另外，从图6–16(b)中看到，只要ω和Ω中有一个足够大，新产业出现的概率就已经很大，这暗示两种学习途径可能存在替代作用。

为了验证分析结果的鲁棒性，使用不同产业激活判别方法和活跃的相似产业（近邻区域）密度计算方法。对于巴西劳动力市场数据，使用产业在区域内的比较优势RCA定义新产业的出现。产业激活要求区域中产业的比较优势在两年内由RCA <0变为RCA ≥ 0，开始前两年保持RCA <0，结束后两年保持RCA ≥ 0（计算细节见第6.2.2节）。对于中国企业注册信息数据，使用区域内活跃的相似产业比例替代ω（基于2015年中国产业空间），使用区域周围活跃的近邻区域比例替代Ω。在新定义基础上，重新计算巴西和中国区域内新产业出现的概率。

图6–17展示了巴西区域经济发展中协同学习途径的鲁棒性分析结果。尽管使用RCA定义新产业的出现，两条学习途径的结果仍然保持。具体而言，从图6–17(a)看到，新产业在两年内出现的概率随ω和Ω增大而增大，两者共同起作

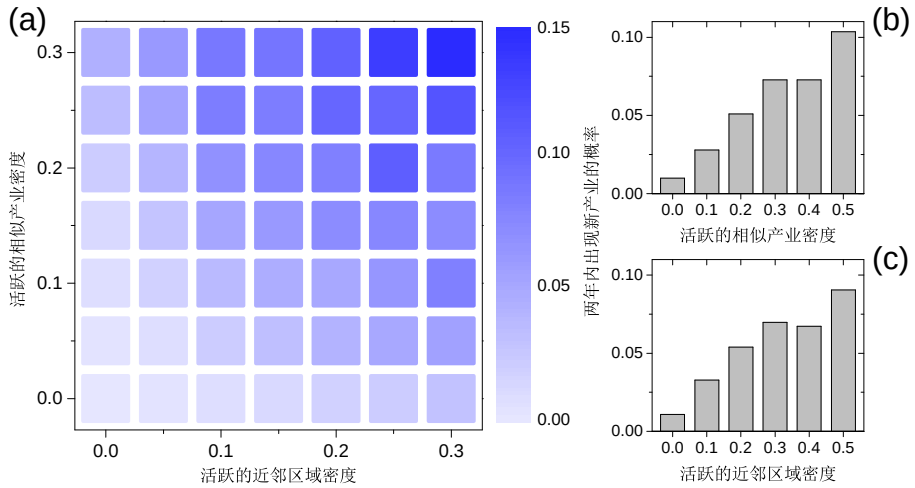

图 6–17 巴西区域经济发展中协同学习作用的鲁棒性分析结果

用。从图6–17(b)看到，新产业在两年内出现的概率随ω增大而增大，说明了巴西区域经济发展中相似产业学习的作用。从图6–17(c)看到，新产业在两年内出现的概率随Ω增大而增大，说明了巴西区域经济发展中近邻区域学习的作用。

图6–18展示了中国区域经济发展中协同学习途径的鲁棒性分析结果。尽管使用产业空间（地理近邻）网络中活跃产业（区域）的比例来替换相应的密度，两条学习途径的结果仍然保持。具体而言，从图6–18(a)看到，新产业在五年内出现的概率随活跃的相似产业（近邻区域）比例的增大而增大，两者共同起作用。从图6–18(b)看到，新产业在五年内出现的概率随区域内活跃的相似产业比例的增大而增大，说明了中国省份经济发展中相似产业学习的作用。从图6–17(c)看到，新产业在五年内出现的概率随区域周围活跃的近邻区域比例的增大而增大，说明了中国省份经济发展中近邻区域学习的作用。

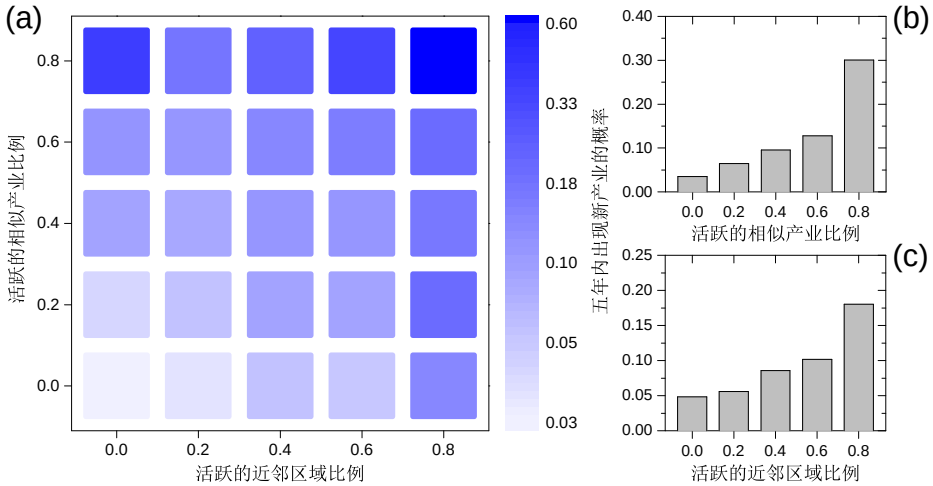

图 6–18 中国区域经济发展中协同学习作用的鲁棒性分析结果

进一步，使用多变量回归分析验证分析结果的鲁棒性。利用统一分析框架

处理巴西劳动力市场数据和中国企业注册信息数据：使用产业在区域内共同出现的方法计算产业接近性，即公式（5–15）；使用比较优势的方法判断区域内产业的活跃情况，即公式（6–3）；使用比较优势RCA ≥ 1的方法定义区域内新产业的出现，使用前向和后向限定（计算细节见第6.2.2节）。在使用统一规则的基础上，计算区域内活跃的相似产业密度ω和区域周围活跃的近邻区域密度Ω。然后，利用Probit模型在考虑ω和Ω的情况下对依赖变量$U_{i,\alpha,t+\Delta t}$（区域i中新产业α在时间Δt后出现）进行回归分析。具体而言，使用的Probit模型为

$$U_{i,\alpha,t+\Delta t} = \beta_0 + \beta_1 \omega_{i,\alpha,t} + \beta_2 \Omega_{i,\alpha,t} + \beta_3 \omega_{i,\alpha,t} \Omega_{i,\alpha,t} + \mu_t + \varepsilon_{i,\alpha,t}. \tag{6–7}$$

其中，$\omega_{i,\alpha,t}\Omega_{i,\alpha,t}$为两个密度的交叉项，$\mu_t$为时间固定效应，$\varepsilon_{i,\alpha,t}$为误差项。特别注意，$\Delta t$为发展的时间差，对于巴西$\Delta t = 2$，对于中国$\Delta t = 5$。

表6–1给出了区域内发展新产业受ω和Ω影响的回归分析结果。可以看到，不论巴西还是中国，回归分析结果都支持了区域经济发展的相似技术学习和近邻区域学习途径。具体而言，从第（1）列看到，ω对新产业的出现有显著正面的作用，这一结果支持了第6.2.1节的分析。中国的回归系数大于巴西的回归系数，暗

表 6–1 相似技术学习和近邻区域学习对新产业出现的相互作用效果分析

变量	Probit模型：区域发展新产业			
	(1)	(2)	(3)	(4)
巴西				
ω	0.240^{***}		0.193^{***}	0.234^{***}
	(0.002)		(0.002)	(0.003)
Ω		0.255^{***}	0.225^{***}	0.261^{***}
		(0.002)	(0.002)	(0.002)
$\omega\Omega$				-0.071^{***}
				(0.002)
Observations	1,301,335	1,301,335	1,301,335	1,301,335
Pseudo R^2	0.040	0.062	0.086	0.092
中国				
ω	0.453^{***}		0.457^{***}	0.503^{***}
	(0.022)		(0.023)	(0.023)
Ω		0.186^{***}	0.172^{***}	0.281^{***}
		(0.019)	(0.020)	(0.024)
$\omega\Omega$				-0.150^{***}
				(0.020)
Observations	19,835	19,835	19,835	19,835
Pseudo R^2	0.119	0.046	0.137	0.147

统计显著性水平：$^{*}p < 0.1$；$^{**}p < 0.05$；$^{***}p < 0.01$

示ω对中国发展新产业有更大的影响。从第（2）列看到，Ω对新产业的出现也有显著正面的作用，这一结果支持了6.2.2节的分析。巴西的回归系数大于中国的回归系数，暗示Ω对巴西发展新产业有更大影响。

进一步，分析ω和Ω的共同作用。从第（3）列看到，当两条学习途径共同存在时，都对发展新产业有显著正面的作用。对比两者作用时，巴西的Ω有更大的作用，中国的ω有更大的作用。第（4）列中增加了两个密度的交叉项$\omega\Omega$，用来分析他们之间的相互作用。结果发现，不论中国还是巴西，交叉项$\omega\Omega$的回归系数都为负，说明两个密度出现了收益递减。这说明，两条学习途径在区域经济发展中存在相互替代的作用，一条足够活跃的学习途径将削弱另一条学习途径的作用，同时拥有两条学习途径不能显著地提高区域发展新产业的能力。

6.3 基于空间网络的最优经济发展学习策略

分析经济结构和揭示经济发展路径，其中一个核心应用是指导经济发展策略的制定。一方面，经济发展是一个路径依赖过程[69, 161, 375]，区域当前的经济和产业结构影响未来的发展潜力[25, 67]。所以，在制定经济发展策略时，应当充分考虑区域当前所具有的产业结构，如产业空间中的活跃产业类型、数量和所处位置等，以帮助判断当前状态下区域应当优先发展的产业[12, 25]。另一方面，空间网络的结构影响信息传播，存在临界的长边分布幂指数使信息传播达到最优[204]。所以，利用长边调整地理空间网络的结构，如开通新航线，有希望促进区域发展新产业。

基于大规模社会经济数据，结合空间网络信息传播和计量经济分析方法，研究最优的区域经济学习和发展策略。首先，基于巴西劳动力市场数据，分析地理距离对近邻区域学习效果的影响。基于中国企业注册信息数据，以高铁引入作为工具变量，实证分析高铁对近邻区域学习的促进作用。然后，基于巴西劳动力市场数据，构建产业空间网络和地理邻近空间网络。利用靴襻渗流模型，分析两种基于空间网络的最优产业发展策略。最后，基于国际双边贸易数据，从协同学习的角度理解国际贸易中的知识扩散，分析三种学习策略对国际贸易的影响。

6.3.1 实证分析高铁对于近邻学习的影响

近邻区域学习的分析结果表明，区域发展新产业的概率随区域周围近邻区域密度的增大而增大，体现了周围相似技术和知识对新产业发展的作用[199]。另一方面，区域产业结构相似性随区域地理距离的增大而减小，因为知识传播随地理距离的增大而迅速衰减[159]。基于这两点可以推断，随着区域尺度的增加，地理距离变大，局部知识传播衰减，近邻区域学习效果会下降。类似地，地理尺度的增

加，会削弱产业相对聚集程度，降低根据地理接近性计算的产业相似性，相似技术学习效果也会下降。为了验证区域尺度对协同学习效果的影响，基于巴西劳动力市场数据，分析不同尺度的区域发展新产业的可能性。

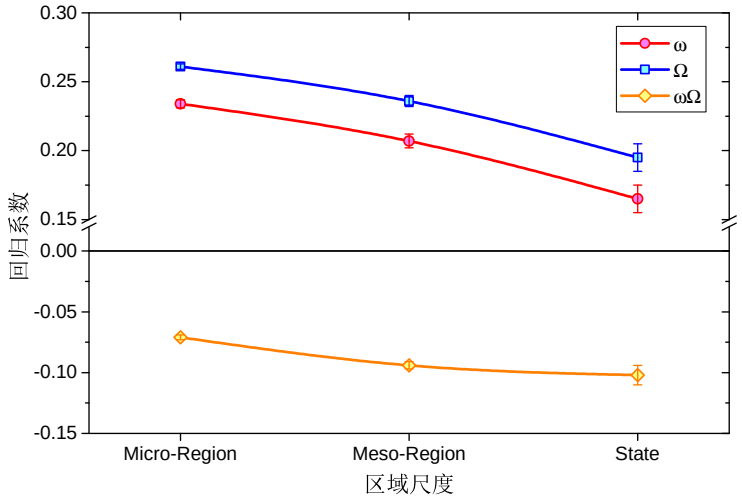

图 6–19 巴西区域经济发展中协同学习效果受区域尺度的影响

图6–19展示了巴西区域经济发展中协同学习效果受区域尺度的影响。其中，横坐标为研究尺度，从小到大分别为Micro-Region、Meso-Region和State；纵坐标为变量的回归系数，包括区域内活跃的相似产业密度ω、区域周围活跃的近邻区域密度Ω和两个密度的交叉项$\omega\Omega$。多变量回归分析使用Probit模型，由公式（6–7）给出。从图6–19看到，随着区域尺度的逐渐增大，两个密度ω和Ω的回归系数逐渐减小；两个密度的交叉项$\omega\Omega$保持为负，绝对值逐渐增大。这些结果说明，协同学习效果随着研究尺度的增大而减弱，两条学习途径的替代作用变强。反过来，随着研究尺度的减小，地理距离变近，两条学习途径的效果都会增强。

高速铁路缩短了区域之间的通勤时间，相当于减少了区域之间的地理距离。中国目前很多城市都被高铁连接[380]，使区域之间人流和物流变得更便捷。高铁能促进个体之间的面对面交流，可能提高区域之间的相互学习[381, 382]。新交通工具的引入（或者通勤成本的降低）作为工具变量，能分析通勤成本对社会互动的影响。例如，Catalini等人[383]发现机票价格的降低能促进航线连接城市中高校学者之间的合作。类似地，将高铁在中国省份之间的引入作为工具变量，能分析高铁开通对近邻区域学习的促进作用。对于协同学习而言，高铁引入是一个合适的工具变量，因为它仅影响近邻区域学习，而不直接影响相似技术学习。

分析高铁引入对近邻区域学习的影响，主要关注两个指标。一是区域之间的产业结构相似性，通过公式（6–5）计算，区域i和区域j在t时间的产业结构相似性为$\varphi_{i,j,t}$。二是区域共有产业的平均生产率，区域i和区域j在t时间的共有产业的

平均生产率为$\bar{p}_{i,j,t}$。生产率通过产业总产值与产业内员工总数的比值来计算。在此基础上，利用双重差分DID模型分析高铁引入对产业结构相似性和共有产业的平均生产率的影响[382]。以产业结构相似性为例，使用的DID回归分析模型为

$$\varphi_{i,j,t} = \beta_0 + \beta_1(Treat_{i,j} * After_t) + \beta_2 Treat_{i,j} + \beta_3 After_t + \mathbf{AX}' + \varepsilon_{i,j}. \qquad (6\text{–}8)$$

其中，$\varphi_{i,j,t}$为区域i和区域j在t时间的产业结构相似性，$\varepsilon_{i,j}$为误差项。$Treat_{i,j} * After_t$为双重差分项，其中哑变量$Treat_{i,j}$体现区域i和区域j之间是否有高铁。$After_t$体现t时间在高铁引入前还是后。向量$\mathbf{X}$为引力模型的控制变量，如省份在人口、人均GDP、城市化和国际贸易等方面的差异[199]。将公式（6–8）中的$\varphi_{i,j,t}$替换为$\bar{p}_{i,j,t}$，类似地分析高铁引入对区域共有产业的平均生产率的影响。

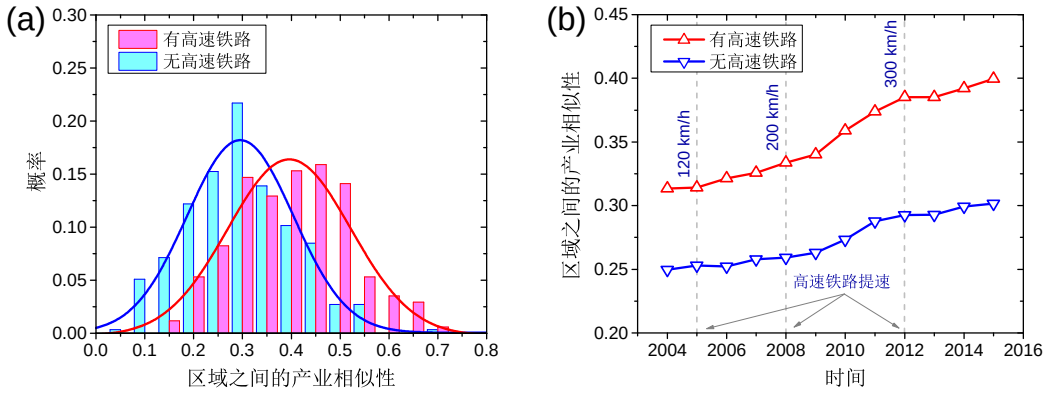

图 6–20 区域之间产业相似性的变化和高铁引入的影响分析

图6–20展示了区域之间产业相似性的变化和高铁引入的影响分析。其中，图6–20(a)中将区域分为两类，一类是2015年时区域之间有高速铁路，另一类是2015年时区域之间无高速铁路，分别给出了区域产业相似性的概率分布。可以看到，2015年高铁连接区域的产业相似性的平均值很大，2015年无高铁连接区域的产业相似性的平均值很小，这说明高铁连接区域有更大的产业相似性。图6–20(b)给出了区域产业相似性的平均值，按照有无高速铁路来划分区域。可以看到，随着高速铁路的三次提速，区域产业相似性相应提高。尤其对于有高速铁路的区域，其产业相似性受高铁提速的影响更大。为了定量地说明高铁引入对产业相似性和共有产业的生产率的影响，进行相应的双重差分DID归回分析。

图6–21展示了高铁引入影响产业相似性和生产率的DID回归分析结果，没有控制其他变量。其中，图6–21(a)给出了高铁引入前后依赖变量（产业相似性φ）随控制变量（高铁引入$Treat$）的变化。可以看到，在高铁引入前（1997-2005），回归系数没有显著趋势；在高铁引入后（2005-2015），回归系数显著增加。这说明，高铁引入增大了区域之间的产业相似性，但没有预先趋势（pre-trend），符合DID回归分析的使用条件。图6–21(b)给出了高铁引入影响产业相似性的DID回

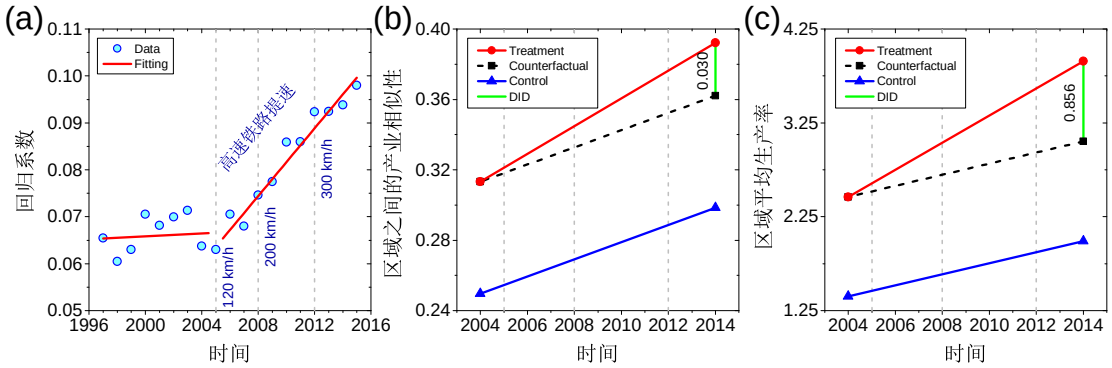

图 6–21 高铁引入影响产业相似性和生产率的双重差分DID回归分析结果

归分析结果。对照组（红色实线）与控制组的期望趋势（黑色虚线）的双重差分DID为0.029，说明高铁连接区域的产业结构变的更相似。图6–21(c)给出了高铁引入影响共有产业的平均生产率的DID回归分析结果。对照组与控制组的期望趋势的DID为0.856，说明高铁引入增大了连接区域共有产业的平均生产率。

表6–2给出了在控制变量情况下DID回归分析的结果。其中，控制变量包括：区域之间的人口差异（ΔPopulation）、人均GDP差异（ΔGDPpc (log)）、城市化水

表 6–2 基于双重差分模型回归分析高铁引入对产业相似性和生产率的影响

变量	简单最小二乘回归模型			
	产业相似性		产业生产率	
	(1)	(2)	(3)	(4)
HSR Entry	0.030**	0.027*	0.856***	0.838***
	(1.983)	(0.015)	(0.144)	(0.132)
Treatment	0.064***	0.053***	1.060***	0.808***
	(5.952)	(0.011)	(0.077)	(0.072)
After Entry	0.049***	0.042***	0.592***	0.527***
	(5.376)	(0.009)	(0.066)	(0.061)
Δ Population (log)		−0.024***		−0.396***
		(0.006)		(0.041)
Δ GDPpc (log)		−0.045***		−0.314***
		(0.013)		(0.106)
Δ Urbanization		0.048***		1.317***
		(0.018)		(0.188)
Δ Trade (log)		0.003		0.065**
		(0.003)		(0.026)
Observations	930	930	930	930
Adj-R^2	0.160	0.184	0.453	0.540
RMSE	0.111	0.109	0.957	0.878

统计显著性水平：$^{*}p<0.1$；$^{**}p<0.05$；$^{***}p<0.01$

平差异（ΔUrbanization）和国际贸易差异（ΔTrade）。其中，第（1）列和第（2）列展示了高铁引入对区域产业相似性的影响。可以看到，控制变量使回归系数由0.030减小到0.027，但高铁引入对区域产业相似性的影响仍然显著。另外，人口差异和人均GDP差异对区域产业相似性起负面作用，城市化水平差异对区域产业相似性起正面作用。第（3）列和第（4）列展示了高铁引入对区域共有产业的平均生产率的影响。可以看到，在控制变量的情况下，高铁引入仍然显著地提高区域共有产业的平均生产率。这些结果说明，高铁引入能促进近邻区域学习效果，不仅提高区域产业相似性，还提高共有产业的平均生产率。

6.3.2 基于空间网络的最优产业发展策略

利用交叉学科工具分析大规模社会经济数据，为理解经济发展规律和发展路径提供了便利[24, 183]。分析巴西劳动力市场数据和中国企业注册信息数据，发现经济发展存在两条学习途径[199, 327]。其中，相似技术学习是从区域内接近性高的产业中学习发展新产业，近邻区域学习是从周围区域的相同产业中学习发展新产业。由于产业接近性和区域地理距离的不同，经济发展的路径依赖使区域面临不同的发展机会。在考虑区域当前情况下，制定最优的产业发展策略非常关键，有助于最大化利用两条学习路径的效果，实现最快的经济发展速度。

研究产业的最优发展策略时，分别考虑相似技术学习和近邻区域学习途径，利用靴襻渗流模型探究实现最快经济发展的策略。在相似技术学习方面，针对产品空间的研究发现，贫穷国家占据产品空间边缘的简单产品，不容易发展位于产品空间核心的复杂产品，跨越产品空间发展不相关的产业很困难[25]。这些结果暗示，在区域经济发展的初期阶段，应当考虑产业空间的“核心-边缘”结构（如图6–22(a)所示巴西区域产业空间），兼顾发展位于产业空间核心和边缘的产业。在近邻区域学习方面，针对空间网络的研究发现，长边分布影响信息传播的范围和效率，长边分布幂指数存在临界值实现最优传播[204]。这些结果暗示，在

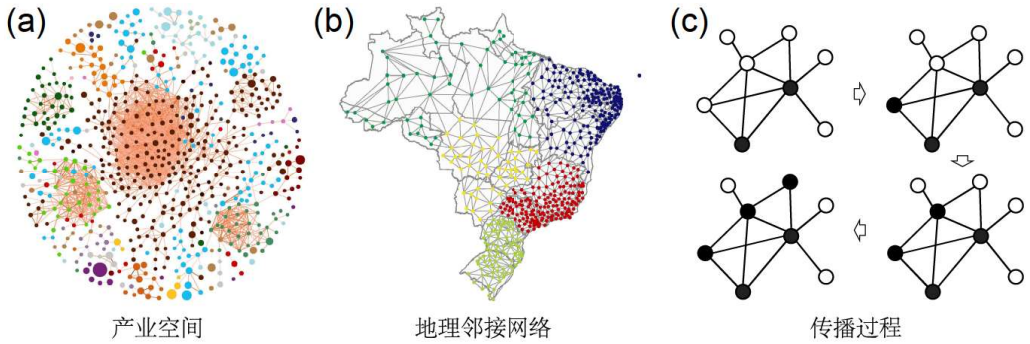

图 6–22 巴西区域产业空间网络、地理近邻网络和产业传播模型示意图

构建区域之间的长边时（如开通航线和修建高铁），应当考虑地理近邻网络（如图6–22(b)所示巴西区域近邻网络）的长边分布，兼顾连接近程和远程区域。

利用与靴襻渗流[204]类似的阈值传播模型，模拟产业在产业空间网络和地理近邻空间网络上的激活。网络中节点处于活跃态或非活跃态，节点一旦由非活跃态变为活跃态，将一直保持活跃态。产业传播过程如下：（i）初始时，有p比例的节点被激活；（ii）如果周围邻居中有超过一半的活跃节点，那么非活跃节点被激活；（iii）按照步骤（ii）中的规则逐渐激活节点，直到终态为止。产业传播过程中，关注两个指标：一是终态时活跃态节点的相对比例（S_a），体现激活产业的范围；二是达到终态所需的迭代步数（NOI），体现激活产业的速度。图6–22(c)所示的传播过程，初始时有2个节点被激活，即$p=2/8=1/4$；达到终态时，激活比例为$S_a=5/8$，迭代步数为$NOI=3$。

首先，针对相似技术学习途径，在产业空间网络上探索产业的最佳发展策略[384]。图6–23(a)展示了巴西产业空间网络（构建方法见第5.2.1节），包含669个Class层面的产业分类，网络平均度为6.5。产业空间存在“核心-边缘”结构，处在不同位置的产业面临不同的发展机会：位于核心的产业，经济复杂程度高，在产业空间网络中有更多的邻居节点，不容易被激活；位于边缘的产业，经济复杂程度低，在产业空间网络中有更好的邻居节点，更容易被激活。所以，可能存在最优的策略来选择初始激活节点，以最大化利用相似技术学习途径。

在模拟产业发展过程时，初始时刻激活产业比例为p，使用平衡指数q调节随机选择的激活节点位于产业空间核心或边缘的倾向，即策略选择。当$q=-1$时，

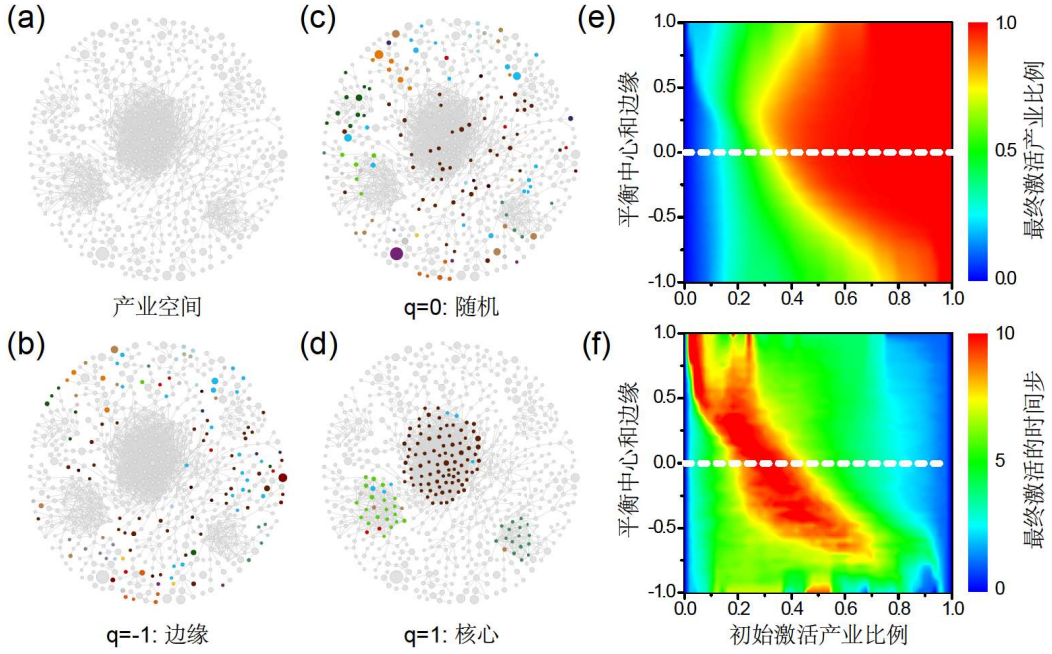

图 6–23 基于产业空间网络的相似技术学习最优发展策略分析

激活节点都位于产业空间的边缘（如图6–23(b)所示）；当$q=0$时，激活节点在产业空间中的位置随机（如图6–23(c)所示）；当$q=1$时，激活节点都位于产业空间的核心（如图6–23(d)所示）。通过改变平衡指数q，分析初始激活产业的策略对产业激活范围和时间的影响。

图6–23(e)和图6–23(f)分别展示了产业空间网络上初始激活产业比例p和平衡指数q对最终激活产业比例S_a和最终激活时间步NOI的影响。可以看到，S_a的数值随p增大而增大。当$p<0.3$或$p>0.8$时，q的取值对S_a的影响不大，说明不同策略的效果相当。当p的取值位于中间范围时，不同策略选择有不同影响。具体而言，当$0.3<p<0.5$时，初始激活核心产业（$q=1$）或边缘产业（$q=-1$）没有竞争力，因为仅有部分产业能被最终激活；初始随机激活产业的策略（$q=0$）表现最优，最终激活所有产业，且用时最少。当$0.5<p<0.8$时，初始激活边缘产业的策略（$q=-1$）表现最差，因为仅能最终激活大约一半的产业；初始激活核心产业的策略最优，在用时最少的情况下激活所有产业[384]。

然后，针对近邻区域学习途径，在地理近邻空间网络上探索产业的最佳发展策略[384]。以巴西区域之间的地理近邻关系为基础，构建巴西区域的地理近邻网络（详细计算方法见第6.2.2节）。如图6–24(a) 所示，地理近邻网络中的节点为558个Microregion层面的区域，网络的平均度大约为6。区域之间的邻近距离d，定义为两个区域相互到达所需跨越的最少区域个数。对于两个相邻区域，$d=1$。在远距离区域之间构建长边，能改变地理近邻空间网络的结构，为区域产业发展带来不同机遇，但连接近程和远程区域的成本不同。例如，修建高铁连接近程区

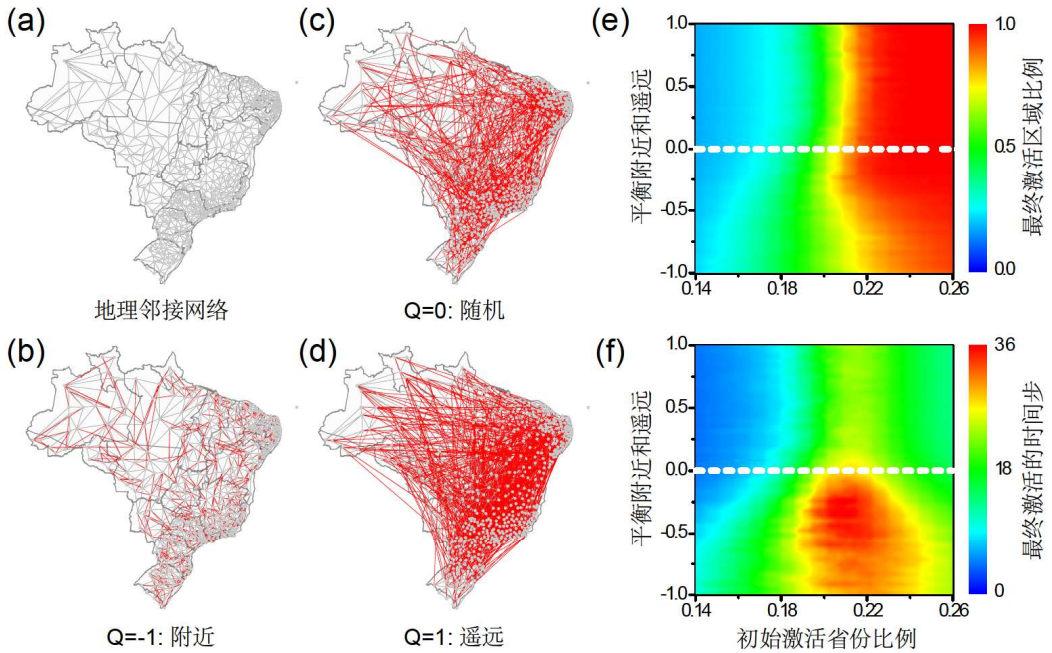

图 6–24 基于地理近邻空间网络的相邻区域学习最优发展策略分析

域，花费相对较少；开通新航线连接远程区域，花费相对较多。所以，可能存在最优的策略来选择长边的分布，以最大化利用近邻区域学习途径。

在模拟产业发展过程时，初始时刻激活产业比例为p，以平衡指数Q调节地理近邻空间网络中的长边分布，即倾向于连接远程还是近程区域的策略选择。为每个区域都添加一条无向长边，长度r范围从2到$D/2$，其中D为所有区域之间最大的邻近距离。长边长度r服从概率分布函数$P(r) \sim r^{5Q}$，其中衰减指数$5Q$用来逼近边界条件，这时长边达到最短或最长。具体而言，当平衡指数$Q = -1$时，仅连接近程区域（如图6–24(b)所示）；当平衡指数$Q = 0$时，随机连接近程和远程区域（如图6–24(c)所示）；当平衡指数$Q = 1$时，仅连接远程区域（如图6–24(d)所示）。通过改变平衡指数Q，分析连接区域的策略对产业激活范围和时间的影响。

图6–24(e)和图6–24(f)分别展示了地理近邻空间网络上初始激活产业比例p和平衡指数Q对最终激活产业比例S_a和最终激活时间步NOI的影响。可以看到，S_a的数值随p的增大而增大。当$p < 0.18$或$p > 0.24$时，Q的取值对S_a的影响不大，说明不同策略的效果相当。当p取值位于中间范围时，不同策略选择有不同的影响。具体而言，当$0.18 < p < 0.21$时，所有策略都仅能激活部分产业，这时随机连接区域的策略（$Q = 1$）和偏好连接远程区域的策略（$Q > 0$）更有效率，因为他们用时更少。当$0.21 < p < 0.24$时，偏好连接近程区域的策略（Q）效果最差；随机连接区域的策略相对最优，因为该策略（如开通一些长程航线和修建一些短程高铁）比偏好连接远程区域的策略（如全部开通航线）节省修建和运营成本[384]。

6.3.3 国际贸易中的知识扩散与发展策略

在国际贸易中，国家需要学习如何生产和出口产品。在进入一个新的出口目的地时，国家也要克服重要的信息摩擦[334]。例如，双边贸易的总量随着国际边界的出现而减少[385]，随着移民流动、相同语言和社会网络交流而增加[386, 387]，语言、社会网络和机构对差异化产品的影响更大[386]。这说明，影响知识扩散的因素在复杂产品贸易上扮演更重要的角色。此外，知识传播限制了国家进入新出口市场的能力，因为国家更容易出口与当前出口产品接近性高的产品[25, 379]，或者邻居区域中已经出口的产品[159]。总体而言，信息摩擦和知识传播决定国家所具有的知识、所能生产的产品和能够进行贸易的伙伴[388]。

本节研究中使用国际双边贸易数据，通过构建扩展的贸易引力模型（Gravity Model），分析国际贸易中的知识扩散和促进双边贸易的最佳发展策略。贸易数据来自麻省理工学院（MIT）的经济复杂性观察站（OEC）[315]，涵盖2000年到2015年的双边贸易总额。宏观经济数据来自世界银行的全球发展指数（WDI）；

产品复杂性数据来自OEC；地理和文化距离数据来自GEODIST，包括是否共有语言、城市之间的物理距离、是否共有边界和是否共有殖民历史；语言接近性数据来自全球语言网络[389]。基于国际双边贸易数据，构建三个相关性指标[388]。

第一个是产品相关性（Product Relatedness），刻画国家出口的产品与其他已经出口的产品之间的相似性。如图6–25所示，韩国已经出口产品1和产品2（衬衫和裤子）到智利，这将影响韩国未来向智利出口产品3（大衣）的贸易总额。产品相关性体现国家所拥有的制造产品所需的知识和能力[25]。以x_{opd}表示来源地国家o向目的地国家d出口产品p的贸易总额，将产品相关性定义为

$$\omega_{opd} = \sum_{p'} \frac{\phi_{pp'}}{\phi_p} \cdot \frac{x_{op'd}}{x_{od}}. \tag{6–9}$$

其中，$x_{od} = \sum_p x_{opd}$为来源地国家o和目的地国家d之间的贸易总额；$\phi_p = \sum_{p'} \phi_{pp'}$为产品之间的接近性，根据产品被共同出口的可能性估计[25]。

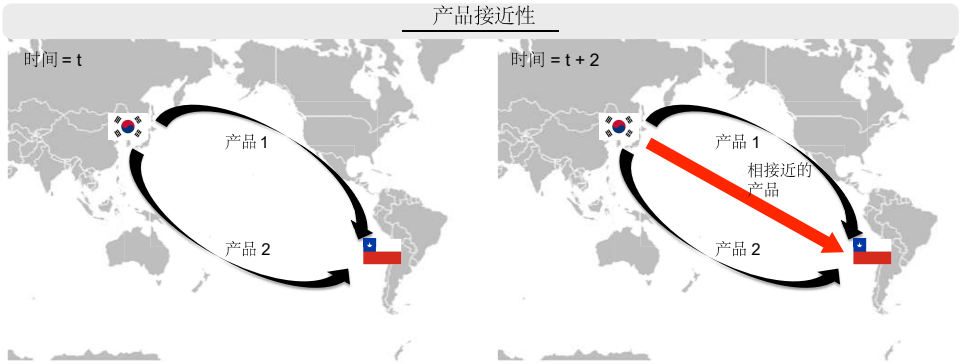

图 6–25 国际贸易中产品相关性的示意图

第二个是进口相关性（Importer Relatedness），刻画国家的邻居国家中从相同地区进口相同产品的可能性。如图6–26(a)所示，秘鲁和阿根廷已经从韩国进口产品1（衬衫），这将影响未来智利从韩国进口产品1（衬衫）的贸易总额，因为智利是秘鲁和阿根廷的地理近邻区域。进口相关性体现进口产品的知识在邻居进口国家之间的流动，或者出口产品到目的地周围国家的知识在出口国内的流动。类似于Chaney等人[390]的工作，将进口相关性定义为

$$\Omega_{opd}^{(d)} = \sum_{d'} \frac{1/D_{dd'}}{1/D_d} \cdot \frac{x_{opd'}}{x_{op}}. \tag{6–10}$$

其中，$x_{op} = \sum_d x_{opd}$为来源地国家o出口产品p的贸易总额；$1/D_d = \sum_{d'} 1/D_{dd'}$，其中$D_{dd'}$为目的地国家$d$与其邻居国家$d'$之间的地理距离。

第三个是出口相关性（Exporter Relatedness），刻画国家的邻居国家中出口相同产品到相同地区的可能性。如图6–26(b)所示，智利从中国和日本进口产品1（衬衫），这将影响未来智利从韩国进口产品1（衬衫）的贸易总额，因为韩国是

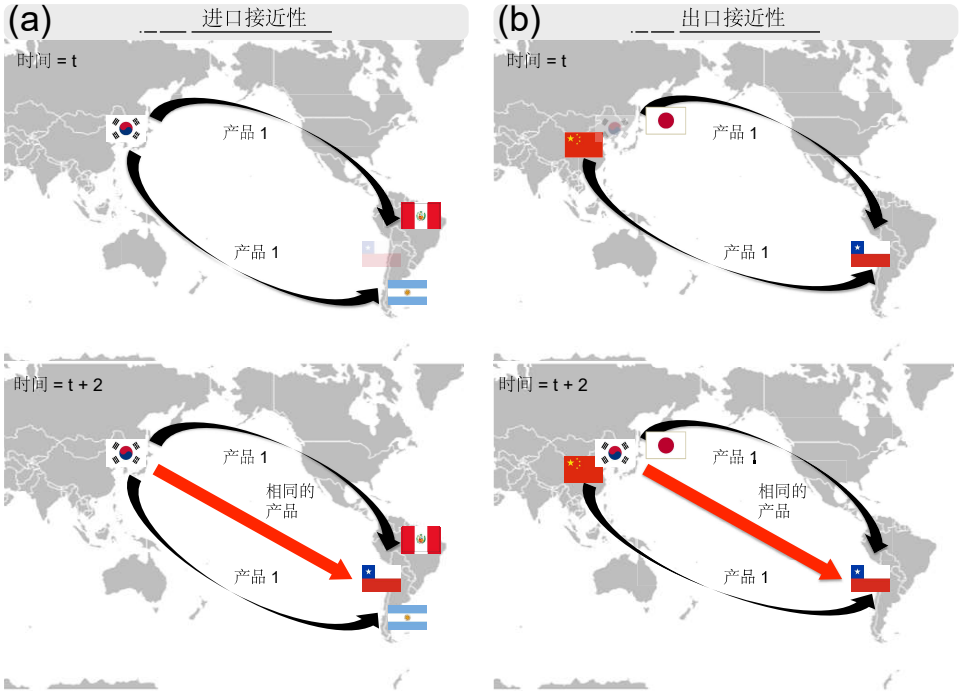

图 6–26 国际贸易中进口相关性和出口相关性的示意图

中国和日本的地理邻近区域。出口相关性体现出口产品的知识在邻居出口国之间的流动，或者从进口国周围国家进口的知识在进口国内的流动。类似于Bahar等人[159]等人的工作，将进口相关性定义为

$$\Omega_{opd}^{(o)} = \sum_{o'} \frac{1/D_{oo'}}{1/D_o} \cdot \frac{x_{o'pd}}{x_{pd}}. \tag{6–11}$$

其中，$x_{pd} = \sum_o x_{opd}$为出口到目的地国家d的产品p贸易总额；$1/D_o = \sum_{o'} 1/D_{oo'}$，其中$D_{oo'}$为来源地国家$o$和其邻居国家$o'$之间的地理距离。

基于国际双边贸易数据，结合文化和地理等因素，构建扩展的贸易引力模型，分析三个相关性指标对国家在未来的双边贸易总额的影响。所使用的引力模型为

$$\begin{aligned} x_{opd}^{t+2} = {} & \beta_0 + \beta_1 \omega_{opd}^t + \beta_2 \Omega_{opd}^{(d)t} + \beta_3 \Omega_{opd}^{(o)t} + \beta_4 x_{opd}^t + \beta_5 x_{op}^t \\ & + \beta_6 x_{pd}^t + \beta_7 D_{od} + \beta_8 gdp_o^t + \beta_9 gdp_d^t + \beta_{10} Population_o^t \\ & + \beta_{11} Population_d^t + \beta_{12} Border_{od} + \beta_{13} Colony_{od} \\ & + \beta_{14} Language_{od} + \beta_{15} Lang.Proximity_{od} + \varepsilon_{opd}^t \end{aligned}. \tag{6–12}$$

其中，依赖变量x_{opd}^{t+2}为两年后从来源地国家o到目的地国家d在产品p上的贸易总额。关心的变量为产品相关性ω_{opd}^t、进口相关性$\Omega_{opd}^{(d)t}$和出口相关性$\Omega_{opd}^{(o)t}$的回归系数。为了避免金融危机影响分析结果，将贸易数据按照时间分成三部分，分别进行回归分析。具体而言，金融危机前期，从2000年到2006年；金融危机期间，从2007年到2012年；金融危机恢复期，从2012年到2015年。

表 6–3 金融危机前、金融危机期间和金融危机恢复期国际双边贸易总量回归分析

变量	简单最小二乘回归分析模型		
	(1)	(2)	(3)
	2000-2006	2007-2012	2012-2015
产品相关性（ω_{opd}^{t}）	0.235***	0.215***	0.176***
	(0.000)	(0.000)	(0.000)
进口相关性（$\Omega_{opd}^{(d)}$）	0.169***	0.147***	0.155***
	(0.000)	(0.007)	(0.000)
出口相关性（$\Omega_{opd}^{(o)}$）	0.090***	0.105***	0.101***
	(0.000)	(0.000)	(0.000)
Constant	9.665***	9.844***	9.789***
	(0.000)	(0.000)	(0.000)
Observations	10,911,584	7,591,488	5,332,255
Adjusted R^2	0.512	0.540	0.579
Within R^2	0.358	0.380	0.421

统计显著性水平：$^{*}p < 0.1$；$^{**}p < 0.05$；$^{***}p < 0.01$

表6–3展示了最小二乘回归分析的结果，仅给出三个相关性指标的回归系数和标准误，完整的回归分析结果表在文献[388]中给出。考虑到误差之间可能存在组内的强关联性，使用三支聚类方法（Three-Way Clustering）[391]对产品、来源地国家和目的地国家的误差进行聚类。从表格中看到，三个相关性指标都与未来两年的双边贸易总额显著正相关。这说明，出口更多相关产品的国家、出口相同产品到目的地国家周围的国家、邻居国家已经出口相同产品的国家，这三种国家更倾向于提高出口贸易总额。分析其他控制变量的回归系数，发现共有语言和殖民历史也影响双边贸易，对技术复杂产品的影响更大[388]。这些分析结果，说明了国际双边贸易中知识扩散的作用，提供了发展国际贸易的三种策略，即优先出口与已有出口产品相似的产品到目的地、优先出口产品到已有目的地的邻居区域、优先出口附近区域已经出口的产品到目的地。

6.4 本章小结

以定量分析揭示经济发展规律，有助于制定科学的经济政策、选择合适的发展产业和实现最快的经济增长。本章研究了社会经济结构演化规律和最优产业发展策略。首先，介绍了社会经济空间网络的结构，分析了信息在空间网络上扩散的临界现象，探究了空间网络结构对信息扩散的影响。然后，介绍了区域经济发展中的协同学习途径，即相似技术学习途径和近邻区域学习途径，分析了两者的相互作用。最后，分析了高铁引入对近邻区域学习的促进作用，介绍了针对协同

学习途径的两种最优发展策略，提出了国际贸易的知识扩散和三种发展策略。

很多社会经济网络都有空间嵌入结构，理解网络的空间结构如何影响信息传播，探究促进信息传播的网络结构特征，对提出最佳经济发展策略有指导意义。本章第6.1节从理论的角度出发，分析了社会经济空间网络的结构及其对信息传播的影响。首先，介绍了无向Kleinberg空间网络模型和靴襻渗流模型。然后，研究了空间网络上的靴襻渗流，发现信息传播存在的临界现象。当长边分布的幂指数大于等于-1时，终态时活跃态节点比例出现双相变，一级和二级相变点不变；当幂指数小于-1时，仅出现二级相变，相变点随着幂指数的减小而增大。最后，分析了一般空间网络结构对信息传播的影响，发现长边分布能改变网络的渗流特征，幂指数-1为相变点不变的临界值。分析结果对信息的最优传播和控制有借鉴意义，例如通过调节空间网络的结构实现最快速度的信息传播。

经济发展是路径依赖的学习过程，当前的产业结构影响未来的经济发展潜力。分析经济发展中存在的学习途径，对理解经济发展和结构演化有很大帮助。本章第6.2节从实证的角度出发，分析了经济发展过程中的协同学习途径，包括相似技术学习和近邻区域学习。首先，基于产业空间网络分析了相似技术学习途径，区域发展新产业的概率随区域内活跃的相似产业密度的增大而增大。然后，计算区域之间的产业结构相似性，发现产业相似性随地理距离的增大而减小；基于地理近邻网络分析了近邻区域学习途径，发现区域发展新产业的概率随区域周围活跃的邻居区域密度的增大而增大。最后，分析了两条学习途径的相互作用，发现足够活跃的一条途径会抵消另一条途径的学习效果。分析结果揭示了区域经济发展的协同学习效应，帮助更好地理解经济发展的学习过程。

刻画经济结构和理解经济发展规律，有助于制定最优经济发展策略。考虑到经济发展的路径依赖，应当优先发展与当前产业相似的新产业，逐步发展复杂程度高的产业。本章第6.3节从模型结合实证的角度出发，研究了基于空间网络的最优经济发展策略。首先，分析了区域尺度对协同学习效果的影响，发现学习效果随地理距离的增大而减弱；分析了高铁引入对近邻区域学习的影响，发现高铁引入显著地提高连接区域的产业相似性和共有产业的平均生产率。然后，利用传播模型分析了产业空间网络和地理邻近空间网络上的产业激活，发现随机激活产业的策略和随机连接区域的策略，能快速地激活所有产业。最后，分析了知识扩散对国际贸易的影响，发现产品相关性、进口相关性和出口相关性能提高国际贸易总额，基于此提出了国际贸易的三种发展策略。

第七章 总结与展望

7.1 全文总结

新数据和新方法的应用催生了计算社会经济学这一新兴的交叉学科研究分支，旨在精准感知社会经济发展态势，理解社会经济运行规律。利用复杂网络刻画社会经济系统中的相互作用，研究社会经济的空间结构与动力学，为理解很多复杂社会经济现象提供更深刻洞见。一方面，网络的空间结构体现社会经济发展过程中涌现的复杂性，利用网络结构特征能推断社会经济发展态势。另一方面，利用网络的结构演化和网络上的传播动力学模型，能分析经济发展路径和学习过程，以理论结合实证探究最佳经济发展策略。本文在计算社会经济学研究框架下，分别从微观、中观和宏观层面研究了社会经济系统的空间结构，利用空间网络和传播模型研究了经济结构演化和发展策略。本文各章的内容总结如下：

第一章首先介绍了本文研究的背景与意义，其次介绍了国内外的研究现状和本文的主要创新点，最后介绍了本文的研究内容与具体章节安排。第二章介绍了计算社会经济学的基础知识。首先，简介了计算社会经济学的研究内容，包括对社会经济状态的感知和对社会经济发展规律的理解。然后，简介了计算社会经济学所使用的大规模社会经济数据，包括政府统计数据、社会媒体数据、非干预行为数据和其他类型数据。最后，简介了计算社会经济学常用的分析方法，包括回归分析、复杂网络分析和统计机器学习。大规模社会经济数据和交叉学科分析方法，是本文研究社会经济系统的空间结构与动力学的基础。

第三章研究了微观层面的社会经济预测性管理。在第3.1节中，利用校园卡记录的非干预行为数据定量刻画了个体行为规律性，分析了行为规律性对学生成绩的预测能力。基于寝室洗澡和食堂吃饭刷卡记录，采用时间序列真实熵量化个体行为规律程度，首次提出了谨严性指标。关联分析发现谨严性指标与学生成绩显著相关，谨严性高的学生成绩好。使用排序学习算法预测学生成绩，发现谨严性特征的引入能显著地提高预测准确性。研究结果对学生的个性化教育和预测性管理有借鉴意义。在第3.2节中，利用企业社会化平台记录的非干预行为数据构建互动网络和社会网络，研究了网络结构特征对员工职业发展的预测能力。发现互动网络有更高的连边互惠性，社会网络的中心性指标与绩效关联性强。处在两个网络中心位置的员工更容易升职和不容易离职，互动网络对升职和离职的预测能力强。研究结果有助于人力资源管理逐步转变为依靠分析非干预数据的预测性管理。在第3.3节中，利用大规模在线平台数据，以量化分析揭示社会经济现象。分

析企业社会化平台数据，发现了团队规模在8人以下时的沟通强度大和平均绩效高；分析手机通讯数据，在中国社会文化背景下验证了社交圈规模在150人左右；分析匿名求职者简历数据，揭示了职场中身高和性别等方面的不平等。研究结果为使用量化分析手段解决社会经济问题提供了新思路。

第四章研究了中观层面的社会经济系统排序问题。在第4.1节中，针对在线评分系统的信誉评价问题，提出了基于群组聚类的在线信誉排序GR算法。不依赖产品有唯一质量分数的传统假设，GR算法根据评分对用户聚类，利用所属群组规模计算用户信誉，稳定地属于大组的用户信誉水平高。在真实数据集上的实验结果表明，GR算法在评价用户信誉上比传统算法有更高的准确性和更强的鲁棒性，并且有算法复杂度低和不依赖传统假设等优点。在第4.2节中，利用迭代寻优求解过程改进GR算法，提出了基于迭代过程的群组聚类用户信誉排序IGR算法。拓展了群组聚类的思想，IGR算法中用户评分形成的群组规模由群组内的用户数量和用户信誉水平共同决定。实验结果表明，IGR算法没有明显的用户度偏好，对用户信誉的排序更准确，能有效检测恶意型和随机型作弊评分用户。尤其当应对大规模作弊评分用户时，IGR算法比GR算法在鲁棒性方面有显著提高。在第4.3节中，针对二部分网络刻画的推荐系统，研究了如何利用网络结构提高推荐效果。提出了一种新的网络节点相似性CosRA指标，基于此提出的CosRA推荐算法平衡了推荐结果的准确性、多样性和新颖性。进一步，将用户之间的信任关系引入CosRA算法框架，提出的CosRA+T算法提高了推荐结果的准确性，算法中存在最优的标度参数，暗示推荐时过度依赖用户信任关系将不利于提升推荐效果。

第五章研究了宏观层面的社会经济结构建模与分析。在第5.1节中，利用中国企业注册信息数据建模刻画区域经济复杂性，分析了经济复杂性对社会经济指标的预测能力。利用迭代方程刻画“省份-产业”二部分网络结构，计算区域经济复杂性ECI指标和竞争力Fitness指标。发现省份在“ECI指标-人均GDP”相图中分为两个区域，ECI指标对发展中区域的经济水平预测能力强；ECI指标和Fitness指标对中国区域经济发展的预测能力相当。在第5.2节中，利用巴西劳动力市场数据和中国企业注册信息数据建模刻画产业空间网络，分析了产业空间的结构特征和演化规律。基于巴西“产业-职业”和中国“省份-产业”二部分网络，利用余弦相似性计算产业接近性，分别构建巴西和中国区域产业空间。发现产业空间有“核心-边缘”结构，复杂程度高和低的产业分别占据产业空间的核心和边缘位置。中国区域产业空间还有“哑铃型”结构，其时间演化存在地区竞争。在第5.3节中，利用微博数据和匿名简历数据分别构建信息流动和人才流动网络，利用网络的结构特征推断区域的经济发展水平。发现信息和人才分别倾向于流出和流入经济发

展好的区域；两个网络的空间结构多样性都与GDP负相关，但仅人才流动网络的入向拓扑多样性与GDP强相关；人才流动对城市GDP的预测准确性更高，结合两个网络结构特征的复合指标能最多解释大约84%的GDP变化。

第六章研究了经济结构演化规律和产业发展策略。在第6.1节中，利用空间网络模型和传播动力学过程，从理论上研究了网络的空间结构对信息传播的影响。发现无向Kleinberg空间网络的长边分布能改变靴襻渗流的相变类型：当长边分布的幂指数大于等于-1时，终态时的活跃态节点比例出现双相变，一级和二级相变点保持不变；当幂指数小于-1时，仅出现二级相变，相变点随幂指数的减小而增大；长边分布能改变更一般的空间网络的渗流特性，幂指数-1仍为临界值。在第6.2节中，从实证上研究了经济发展的两条学习途径。基于产业空间网络分析相似技术学习途径，发现发展新产业的概率随区域内活跃的相似产业密度的增大而增大；基于地理近邻网络分析近邻区域学习途径，发现发展新产业的概率随区域周围活跃的邻居区域密度的增大而增大；分析两条学习途径的相互作用，发现两者存在收益递减。在第6.3节中，从理论结合实证上研究了经济发展过程中的最优学习策略。发现协同学习效果随距离的增大而降低；引入高铁能显著地提高近邻区域学习效果，高铁连接区域的产业相似性和平均生产率更高。利用靴襻渗流模型分析空间网络上的产业激活，发现两条学习途径都存在最优产业发展策略，能最快的激活所有产业；分析知识扩散对国际双边贸易的影响，提出了促进国际贸易的三种策略，即提高产品相关性、进口相关性和出口相关性。

本文对社会经济系统的空间结构与动力学的研究结果，有一定的理论价值和现实指导意义。借助现代手段收集大规模非干预行为数据，以量化方式分析个体行为规律和理解社会经济现象，有助于逐步实现微观层面的预测性管理。利用复杂网络刻画社会经济系统中的相互作用，提出排序算法利用网络结构和群体行为特征推断系统的整体状态，更好地揭示社会经济系统的结构与状态之间的联系。利用网络建模方法分析社会经济数据，从结构角度刻画经济发展所涌现的复杂性，更好地掌握宏观经济结构和预测经济发展。借助空间网络和传播动力学模型，以理论结合实证分析经济发展规律，帮助制定科学的最优产业发展策略。

7.2 研究展望

本文在计算社会经济学框架下研究了社会经济系统的空间结构与动力学，包括微观层面的社会经济预测性管理、中观层面的社会经济系统排序、宏观层面的社会经济结构建模、以及经济结构演化规律与发展策略。计算社会经济学是一个

充满活力的新兴交叉学科研究分支，在量化分析大规模真实数据、揭示社会经济发展规律时，面临很多新挑战和新问题，有希望在未来得到解决或部分解决。

融合不同来源的社会经济数据，提高数据的可获取性和代表性。计算社会经济学所依赖的大规模数据，包括手机通讯、遥感图像、街景图片、社交网络、文本内容等。一方面，由于商业利益和隐私约束，一些数据难以免费公开，如遥感和手机数据[177]。有必要明确能共享的数据范围，对隐私数据进行脱敏处理，保证数据的可获取性。另一方面，由于数据获取方式的限制，一些数据仅能反映特定区域的情况，受社会和文化背景的影响而缺乏代表性；一些数据获取方式存在采样偏差，导致数据不具代表性[392]。例如，贫困人群的手机普及率较低、可能不经常使用社交媒体，导致利用这些方式收集的数据无法覆盖这部分特别需要关注的人群[393]。通过融合不同来源的大规模社会经济数据，有希望获取个体层面的全方位数据，对推断个体社会经济状态和感知宏观经济发展态势有帮助。

结合数据驱动的新范式和传统随机对照试验，提高对社会经济问题的因果推断能力。计算社会经济学目前大多找到关联关系，缺乏对更本质的因果关系的挖掘。数据驱动的研究范式不是针对特定问题获取观测数据，导致天然地缺乏控制变量，不容易找到合适的工具变量解决内生性问题[394]。推断变量之间的因果关系，问题本身就存在很大的挑战，需要识别变量之间的统计相关性、识别潜在的因果方向、避免其他混淆因素的影响等。近年来发展了一些从真实数据中挖掘因果的方法[395]，例如，随机对照试验方法，利用互联网平台进行大规模实验[396]；准实验设计方法，从数据中寻找与随机对照实验一样满足因果推断条件的情况，如双胞胎比较研究；联合模型方法，通过自动估计联合概率分布从数据中推断因果关系，如虚拟事实模型。这些新方法为解决因果推断问题提供了新思路。

利用大规模数据检验传统社会经济理论，提高对现象的解释力和对政策的指导力。利用新数据能训练模型预测传统的社会经济指标，例如利用经济复杂性预测GDP[63]。这类工作虽然在精准和及时地感知社会经济态势上有重要意义，但新方法在效果上很难超过传统方法，也难以验证新维度上的结果。例如，ECI指标和Fitness指标在国家经济复杂性排名上有差异，对GDP的逼近无法作为新维度的评价标准[308]。另外，很多传统理论缺乏大规模数据、跨越社会文化背景、不同社会经济水平等普适场景的检验。实际上，不同学科有希望合作验证传统理论，甚至提出解释性更强的新理论。更重要的是，揭示的社会经济发展规律，应当指导相应的政策制定[12]，在实践中检验理论和分析结果，提高人们的生活水平。

计算社会经济学面临很多新机遇，是值得进一步关注的新研究分支。本文开展了对社会经济系统的空间结构与动力学的研究，但仅呈现了一些针对具体问题

的分析结果，在很多方面都值得深入研究。在微观层面，真实熵能反映个体行为的时间和序列特征，有希望设计一种综合不同行为数据的谨严性度量指标，更准确地刻画行为规律性。另外，采用深度学习算法从数据中直接抽取规律性特征，也有希望提高对成绩的预测准确性。在利用网络结构特征预测职业发展上，有希望扩展到更大规模的企业社会化平台，在跨越不同社会文化背景下验证分析结果的鲁棒性。在揭示社会经济现象上，有希望利用不同类型数据进行验证，将分析结果应用到管理实践，提高团队绩效水平和降低职场不平等性。在中观层面，基于群组聚类的信誉排序算法利用群组规模评价用户信誉，有必要设计在线评分实验对从众假设进行验证。虽然实验结果显示迭代信誉排序算法的收敛速度快，但仍然需要理论分析保证收敛性。此外，综合考虑节点相似性、用户信任关系和用户信誉水平，有希望设计一些效果更好的排序和推荐算法。

在宏观层面，使用千万量级的企业工商注册信息数据，有希望揭示经济复杂性的空间分布规律。借助统计力学和复杂网络等分析方法，有希望提出鲁棒性更好、解释力更强的新经济指标。在产业空间上，有待分析不同区域产业空间的共同特点，深入分析产业空间演化的竞合关系。另外，产业发展的尺度效应也是值得关注的问题，可以分析省、市、县的优势产业在类别和数量上的差异。在信息和人才上，值得研究人才政策和人才先行对经济发展的影响。在经济结构演化和发展策略方面，使用普适的传播模型在真实空间网络上研究产业激活过程，有希望发展出一套准确的理论方法。在协同学习效应上，当前仅得到中国和巴西的实证分析结果，值得研究对不同发展阶段经济体的泛化能力。经济结构与经济发展的因果关系，也是非常值得关注的研究问题。特别地，挖掘科技型产业结构和科技人才结构的共演化规律，有希望量化基础科学人才对产业升级的作用。

总之，计算社会经济学是一个新兴交叉学科研究分支，在数据和方法上都面临一些新挑战和新机遇。在融合不同来源的大规模数据、将新范式与传统方法相结合、使用新方法验证传统理论等很多方面，都还缺乏系统性的研究。对社会经济结构和动力学的研究还有待深入，利用研究结果指导政策制定也值得关注。在未来的研究中，有希望进一步提高对社会经济态势的感知和对社会经济运行规律的理解。长期而言，利用大规模真实数据和交叉学科分析工具，必将成为解决社会经济问题的主流方法论，也必将深刻地改变社会经济研究的图景。

致 谢

过完这个夏天，就在电子科技大学度过了整十一年。在即将走完博士五年的求学旅程之际，我特别感谢父母的养育之恩，他们带我来到这个世界上，给了我最朴实、最好的家庭教育，他们是我生活上的榜样。虽然小时候家里面生活有些艰难，但是父母仍然给我提供了读书的条件，让我有机会接受教育。感谢父母对我一直以来都选择读书的支持和理解，他们是我人生前进的最大动力。

在完成博士论文之际，由衷地感谢我的导师周涛教授。周老师治学严谨，生活朴素，不但是科研上值得尊敬的科学家，而且是在生活上值得学习的好榜样。周老师非常尊重学生，悉心指导和呵护学生。他既在把握重要科学问题上给了我极大启发，又在撰写和修改论文上给了我很多帮助。仍然记得周老师第一次给我修改论文，连标点符号、引文格式和句子结构都不放过，用不同颜色的笔细致地批注在打印文档上，令我非常惊讶和由衷地敬佩。在读博士的这五年，非常感谢周老师对我的指导和训练，不仅有研究上的严谨，还有生活上的态度。

科研上的学习和训练，离不开合作老师们的指导和帮助。在我博士学习的开始阶段，尚明生教授和蔡世民副教授在推荐和排序研究上给了我很多帮助，胡延庆副教授在渗流模型的研究上给了我很多指导，唐明教授在传播动力学研究方面给了我很多启发，荣智海教授在社会经济研究方面给了我很多建议，黄俊铭博士在因果推断方面给了我不少帮助。还特别感谢实验室的其他老师和同学们，包括邵俊明教授、张彦如研究员、郝东副教授、蒲剑苏副教授等老师，刘金虎、张千明、王伟、陈玲姣、赵志丹、王庆、高磊、崔鹏碧、朱郁筱、杨慧、舒盼盼、钟林峰、罗永劼、Ratha Pech等博士，以及樊超、王军、潘黎明、李艳丽、杨丹、刘权辉、陈小龙、曹奕、王心迪、李睿琪、岳仲涛、许雄锐、黄盼华、毛雅俊、赵倩、杨枭、刘亦丁、戴海星、苗丽莉等同学，谢谢他们给我带来的知识和快乐。

感谢麻省理工学院媒体实验室联合培养期间的合作导师César A. Hidalgo教授，他是一位热情洒脱、乐于助人的国际顶尖科学家，在科研上给予了我非常大的帮助，开阔了我的学术和生活视野，也启发了我对科学问题的思考。还非常感谢吕琳媛教授推荐我到波士顿大学物理系访问，感谢H. Eugene Stanley教授给我提供了学习和交流机会。特别感谢两次出国留学期间帮助我的伙伴们，包括Bogang Jun、Flávio L. Pinheiro、Cristian Candia、Kevin Hu、Takahito Ito、Aamena Alshamsi、Sanjay Guruprasad、Almaha Almalki、Christian Jara、Pablo

Astudillo、Mary Kaltenberg、Diana Orghian、Pachá、张静娴、董晓文、陈晓姣、段了了、冷妍、乔杨等。特别谢谢黄俊铭博士的帮助，让我度过难忘的留学时光。

感谢博士期间的三次实习经历，让我有机会将理论与实践相结合。特别感谢周俊临副教授和张琳艳师姐，在数联寻英实习期间对我的帮助和鼓励；感谢李倩师姐，在数联铭品实习期间对我的照顾和帮助；感谢赵明潇和曹宝林，在成都新经济发展研究院实习期间对我的关照。

特别感谢三个姐姐在生活上对我的照顾和帮助，大姐总能保障我的生活和学习条件，二姐总是关心我的学业情况，三姐总是关心我的生活状况，也感谢姐夫们也给我提供的很多帮助。特别感谢姐姐和姐夫们对父母的照顾，让我没有后顾之忧，得以安心地读书，否则我很难顺利完成学业。还要特别感谢生活上的很多好朋友，没有你们的陪伴，我的生活会缺少很多色彩，特别谢谢李美妍、叶谱英、廖宇星、蔡威威、李凡、张洪钒、韩志罡、吴斌杰、秦博、皮阳等朋友们。

在即将完成学业之际，感谢国家留学基金委对我出国交流的资助，感谢国家助学贷款保障了我的生活和学业，感谢研究生国家奖学金和学业奖学金对我的支持，感谢唐立新教育基金会和企业奖学金对我的资助，感谢学校提供了优越的学习条件，感谢实验室提供科研资助和设备保障。

最后，特别感谢评阅专家抽出宝贵的时间审阅我的论文，提出了特别好的修改意见和建议，对提高论文质量帮助很大。谢谢!

参考文献

[1] E. Ostrom. A general framework for analyzing sustainability of social-ecological systems[J]. Science, 2009, 325(5939):419–422

[2] J. H. Miller, S. E. Page. Complex adaptive systems: An introduction to computational models of social life[M]. Princeton, NJ, USA: Princeton University Press, 2009

[3] M. Weber. The theory of social and economic organization[M]. New York, NY, USA: Simon and Schuster, 2009

[4] 王飞跃. 人工社会、计算实验、平行系统–关于复杂社会经济系统计算研究的讨论[J]. 复杂系统与复杂性科学, 2004, 1(4):25–35

[5] Y.-Y. Liu, J.-J. Slotine, A.-L. Barabási. Controllability of complex networks[J]. Nature, 2011, 473(7346):167–173

[6] G. Weisbuch. Complex systems dynamics[M]. Boca Raton, FL, USA: CRC Press, 2018

[7] C. S. Holling. Understanding the complexity of economic, ecological, and social systems[J]. Ecosystems, 2001, 4(5):390–405

[8] R. W. Proctor, T. Van Zandt. Human factors in simple and complex systems[M]. Boca Raton, FL, USA: CRC Press, 2018

[9] L. Einav, J. Levin. Economics in the age of big data[J]. Science, 2014, 346(6210):1243089

[10] J. A. Kahl, J. A. Davis. A comparison of indexes of socio-economic status[J]. American Sociological Review, 1955, 20(3):317–325

[11] P. England. Households, employment, and gender: A social, economic, and demographic view[M]. London, UK: Routledge, 2017

[12] A. Alshamsi, F. L. Pinheiro, C. A. Hidalgo. Optimal diversification strategies in the networks of related products and of related research areas[J]. Nature Communications, 2018, 9:1328

[13] G. S. Becker, K. M. Murphy. Social economics: Market behavior in a social environment[M]. Cambridge, MA, USA: Harvard University Press, 2009

[14] J. Costa-i-Font, M. Macis. Social economics: Current and emerging avenues[M]. Cambridge, MA, USA: MIT Press, 2017

[15] 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(4):625–633

[16] R. J. Fisher. Social desirability bias and the validity of indirect questioning[J]. Journal of Consumer Research, 1993, 20(2):303–315

[17] P. S. Park, J. E. Blumenstock, M. W. Macy. The strength of long-range ties in population-scale social networks[J]. Science, 2018, 362(6421):1410–1413

[18] J. M. Wooldridge. Introductory econometrics: A modern approach[M]. Mason, OH, USA: Nelson Education, Ltd., 2015, 238–244

[19] M. O. Jackson, B. W. Rogers, Y. Zenou. The economic consequences of social-network structure[J]. Journal of Economic Literature, 2017, 55(1):49–95

[20] 张翼成, 吕琳媛, 周涛. 重塑：信息经济的结构[M]. 成都: 四川人民出版社, 2018

[21] M. Schulz. Statistical physics and economics: Concepts, tools, and applications[M]. New York, NY, USA: Springer, 2003

[22] M. Cristelli, A. Tacchella, L. Pietronero. The heterogeneous dynamics of economic complexity[J]. PLoS ONE, 2015, 10(2):e0117174

[23] E. Diener, E. Suh. Measuring quality of life: Economic, social, and subjective indicators[J]. Social Indicators Research, 1997, 40(1–2):189–216

[24] C. A. Hidalgo, R. Hausmann. The building blocks of economic complexity[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(26):10570–10575

[25] C. A. Hidalgo, B. Klinger, A.-L. Barabási, et al. The product space conditions the development of nations[J]. Science, 2007, 317(5837):482–487

[26] V. Mayer-Schönberger, K. Cukier. Big data: A revolution that will transform how we live, work, and think[M]. Boston, MA, USA: Houghton Mifflin Harcourt, 2013

[27] L. Einav, J. Levin. The data revolution and economic analysis[J]. Innovation Policy and the Economy, 2014, 14:1–24

[28] E. Jahani, P. Sundsøy, J. Bjelland, et al. Improving official statistics in emerging markets using machine learning and mobile phone data[J]. EPJ Data Science, 2017, 6:3

[29] Y. Kryvasheyeu, H. Chen, N. Obradovich, et al. Rapid assessment of disaster damage using social media activity[J]. Science Advances, 2016, 2(3):e1500779

[30] N. Jean, M. Burke, M. Xie, et al. Combining satellite imagery and machine learning to predict poverty[J]. Science, 2016, 353(6301):790–794

[31] M. Ettredge, J. Gerdes, G. Karuga. Using web-based search data to predict macroeconomic statistics[J]. Communications of the ACM, 2005, 48(11):87–92

[32] Y. Lecun, Y. Bengio, G. Hinton. Deep learning[J]. Nature, 2015, 521(7553):436–444

[33] A.-L. Barabási. Network science[M]. New York, NY, USA: Cambridge University Press, 2016

[34] A. De Martino, M. Marsili. Statistical mechanics of socio-economic systems with heterogeneous agents[J]. Journal of Physics A: Mathematical and General, 2006, 39(43):R465–R540

[35] J. Gao, Y.-C. Zhang, T. Zhou. Computational socioeconomics[J]. Physics Reports, 2019

[36] S. L. Shaw, M. H. Tsou, X. Ye. Editorial: Human dynamics in the mobile and big data era[J]. International Journal of Geographical Information Science, 2016, 30(9):1687–1693

[37] J. Felipe, U. Kumar, A. Abdon, et al. Product complexity and economic development[J]. Structural Change and Economic Dynamics, 2012, 23(1):36–68

[38] A. Nassif, C. Feijó, E. Araújo. Structural change and economic development: Is Brazil catching up or falling behind?[J]. Cambridge Journal of Economics, 2014, 39(5):1307–1332

[39] V. Frias-Martinez, J. Virseda-Jerez, E. Frias-Martinez. On the relation between socio-economic status and physical mobility[J]. Information Technology for Development, 2012, 18(2):91–106

[40] A. Dobra, N. E. Williams, N. Eagle. Spatiotemporal detection of unusual human population behavior using mobile phone data[J]. PLoS ONE, 2015, 10(3):e0120449

[41] N. Chung, S. J. Lee, H. Han. Understanding communication types on travel information sharing in social media: A transactive memory systems perspective[J]. Telematics and Informatics, 2015, 32(4):564–575

[42] Y. Cao, J. Gao, D. Lian, et al. Orderliness predicts academic performance: Behavioural analysis on campus lifestyle[J]. Journal of The Royal Society Interface, 2018, 15(146):20180210

[43] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039):207–211

[44] C. Song, Z. Qu, N. Blumm, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968):1018–1021

[45] X. Lu, L. Bengtsson, P. Holme. Predictability of population displacement after the 2010 Haiti earthquake[J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(29):11576–11581

[46] J. Blumenstock, G. Cadamuro, R. On. Predicting poverty and wealth from mobile phone metadata[J]. Science, 2015, 350(6264):1073–1076

[47] V. Kassarnig, E. Mones, A. Bjerre-Nielsen, et al. Academic performance and behavioral patterns[J]. EPJ Data Science, 2018, 7:10

[48] E. Bokányi, Z. Lábszki, G. Vattay. Prediction of employment and unemployment rates from Twitter daily rhythms in the US[J]. EPJ Data Science, 2017, 6:14

[49] A. G. Ingham, G. Levinger, J. Graves, et al. The Ringelmann effect: Studies of group size and group performance[J]. Journal of Experimental Social Psychology, 1974, 10(4):371–384

[50] K. Sohn. The height premium in Indonesia[J]. Economics and Human Biology, 2015, 16:1–15

[51] Y.-B. Zhou, T. Lei, T. Zhou. A robust ranking algorithm to spamming[J]. EPL (Europhysics Letters), 2011, 94(4):48002

[52] H. Liao, A. Zeng, R. Xiao, et al. Ranking reputation and quality in online rating systems[J]. PLoS ONE, 2014, 9(5):e97146

[53] M.-S. Shang, L. Lü, Y.-C. Zhang, et al. Empirical analysis of web-based user-object bipartite networks[J]. EPL (Europhysics Letters), 2010, 90(4):48006

[54] X.-L. Liu, Q. Guo, L. Hou, et al. Ranking online quality and reputation via the user activity[J]. Physica A: Statistical Mechanics and its Applications, 2015, 436:629–636

[55] L. Lü, M. Medo, C. H. Yeung, et al. Recommender systems[J]. Physics Reports, 2012, 519(1):1–49

[56] H. Liao, M. S. Mariani, M. Medo, et al. Ranking in evolving complex networks[J]. Physics Reports, 2017, 689:1–54

[57] X. Wang, Y. Liu, G. Zhang, et al. Diffusion-based recommendation with trust relations on tripartite graphs[J]. Journal of Statistical Mechanics: Theory and Experiment, 2017, 2017(8):083405

[58] 林毅夫. 新结构经济学[J]. 经济学(季刊), 2010, 1(1):1–32

[59] J. Y. Lin. New structural economics: A framework for rethinking development[J]. World Bank Research Observer, 2011, 26(2):193–221

[60] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012

[61] C. A. Hidalgo, R. Hausmann. A network view of economic development[J]. Developing Alternatives, 2008, 12(1):5–10

[62] G. Caldarelli, M. Cristelli, A. Gabrielli, et al. A network analysis of countries' export flows: Firm grounds for the building blocks of the economy[J]. PLoS ONE, 2012, 7(10):e47278

[63] A. Tacchella, D. Mazzilli, L. Pietronero. A dynamical systems approach to gross domestic product forecasting[J]. Nature Physics, 2018, 14(8):861–865

[64] C. Smith-Clarke, A. Mashhadi, L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks[C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 2014, 511–520

[65] 任晓龙, 朱燕燕, 王思云,等. 在线社交网络结构与区域经济关联性研究[J]. 电子科技大学学报, 2015, 44(5):643–651

[66] N. Eagle, M. Macy, R. Claxton. Network diversity and economic development[J]. Science, 2010, 328(5981):1029–1031

[67] F. Neffke, M. Henning, R. Boschma. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions[J]. Economic Geography, 2011, 87(3):237–265

[68] R. Hausmann, C. A. Hidalgo, S. Bustos, et al. The atlas of economic complexity: Mapping paths to prosperity[M]. Cambridge, MA, USA: MIT Press, 2014

[69] R. Martin, P. Sunley. Path dependence and regional economic evolution[J]. Journal of Economic Geography, 2006, 6(4):395–437

[70] D. J. Watts. A simple model of global cascades on random networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(9):5766–5771

[71] V. Soto, V. Frias-Martinez, J. Virseda, et al. Prediction of socioeconomic levels using cell phone records[C]. Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization, Girona, Spain, 2011, 377–388

[72] T. Gutierrez, G. Krings, V. D. Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets[J]. arXiv:1309.4496, 2013

[73] J. E. Blumenstock. Calling for better measurement: Estimating an individual's wealth and well-being from mobile phone transaction records[C]. Proceedings of the 20th ACM SIGKDD Workshop on Data Science for Social Good, New York, NY, USA, 2014, 1–6

[74] L. Lotero, R. G. Hurtado, L. M. Floría. Rich do not rise early: Spatio-temporal patterns in the mobility networks of different socio-economic classes[J]. Royal Society Open Science, 2016, 3(10):150654

[75] X. Yang, A. Belyi, I. Bojic, et al. Human mobility and socioeconomic status: Analysis of Singapore and Boston[J]. Computers, Environment and Urban Systems, 2018, 72:51–67

[76] M. Fixman, A. Berenstein, J. Brea, et al. A Bayesian approach to income inference in a communication network[C]. Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, San Francisco, CA, USA, 2016, 579–582

[77] S. Luo, F. Morone, C. Sarraute, et al. Inferring personal economic status from social network location[J]. Nature Communications, 2017, 8:15227

[78] E. Jahani, G. Saint-Jacques, P. Sundsøy, et al. Differential network effects on economic outcomes: A structural perspective[C]. Proceedings of the 9th International Conference on Social Informatics, Oxford, UK, 2017, 41–50

[79] A. Llorente, M. Garcia-Herranz, M. Cebrian, et al. Social media fingerprints of unemployment[J]. PLoS ONE, 2015, 10(5):e0128692

[80] J. L. Toole, Y.-R. Lin, E. Muehlegger, et al. Tracking employment shocks using mobile phone data[J]. Journal of the Royal Society Interface, 2015, 12(107):20150185

[81] P. Sundsøy, J. Bjelland, B.-A. Reme, et al. Towards real-time prediction of unemployment and profession[C]. Proceedings of the 9th International Conference on Social Informatics, Oxford, UK, 2017, 14–23

[82] A. Almaatouq, F. Prieto-Castrillo, A. S. Pentland. Mobile communication signatures of unemployment[C]. Proceedings of the 8th International Conference on Social Informatics, Bellevue, WA, USA, 2016, 407–418

[83] T. H. Feeley, J. Hwang, G. A. Barnett. Predicting employee turnover from friendship networks[J]. Journal of Applied Communication Research, 2008, 36(1):56–73

[84] N. Askitas, K. F. Zimmermann. Google econometrics and unemployment forecasting[J]. Applied Economics Quarterly, 2009, 55(2):107–120

[85] F. D'Amuri, J. Marcucci. The predictive power of Google searches in forecasting US unemployment[J]. International Journal of Forecasting, 2017, 33(4):801–816

[86] D. O. Olguín, B. N. Waber, T. Kim, et al. Sensible organizations: Technology and methodology for automatically measuring organizational behavior[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(1):43–55

[87] J.-i. Watanabe, M. Fujita, K. Yano, et al. Resting time activeness determines team performance in call centers[C]. Proceedings of the 4th International Conference on Social Informatics, Lausanne, Switzerland, 2012, 26–31

[88] P. Laureti, L. Moret, Y.-C. Zhang, et al. Information filtering via iterative refinement[J]. EPL (Europhysics Letters), 2006, 75(6):1006–1012

[89] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach[J]. PLoS ONE, 2013, 8(9):e73791

[90] S. C. Guntuku, W. Lin, J. Carpenter, et al. Studying personality through the content of posted and liked images on Twitter[C]. Proceedings of the 2017 ACM on Web Science Conference, Troy, NY, USA, 2017, 223–227

[91] C. Segalin, F. Celli, L. Polonio, et al. What your Facebook profile picture reveals about your personality[C]. Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, CA, USA, 2017, 460–468

[92] M. E. Larsen, T. W. Boonstra, P. J. Batterham, et al. We Feel: Mapping emotion on Twitter[J]. IEEE Journal of Biomedical and Health Informatics, 2015, 19(4):1246–1252

[93] S. M. Mohammad, F. Bravomarquez. Emotion intensities in tweets[C]. Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, Vancouver, BC, Canada, 2017, 65–77

[94] A. G. Reece, C. M. Danforth. Instagram photos reveal predictive markers of depression[J]. EPJ Data Science, 2017, 6:15

[95] H. Sueki. The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young Internet users in Japan[J]. Journal of Affective Disorders, 2015, 170:155–160

[96] H. Mao, X. Shuai, Y.-Y. Ahn, et al. Quantifying socio-economic indicators in developing countries from mobile phone communication data: Applications to Côte d'Ivoire[J]. EPJ Data Science, 2015, 4(1):15

[97] J.-H. Liu, J. Wang, J. Shao, et al. Online social activity reflects economic status[J]. Physica A: Statistical Mechanics and its Applications, 2016, 457:581–589

[98] B. O. Holzbauer, B. K. Szymanski, T. Nguyen, et al. Social ties as predictors of economic development[C]. Proceedings of the 12th International Conference and School on Network Science, Wroclaw, Poland, 2016, 178–185

[99] P. Salesses, K. Schechtner, C. A. Hidalgo. The collaborative image of the city: Mapping the inequality of urban perception[J]. PLoS ONE, 2013, 8(7):e68400

[100] N. Naik, J. Philipoom, R. Raskar, et al. Streetscore–Predicting the perceived safety of one million streetscapes[C]. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 2014, 793–799

[101] C. D. Elvidge, P. C. Sutton, T. Ghosh, et al. A global poverty map derived from satellite data[J]. Computers and Geosciences, 2009, 35(8):1652–1660

[102] X. Song, Z. He. The Keqiang index: A new benchmark for China's development[J]. Social Indicators Research, 2015, 123(3):661–676

[103] M. E. J. Newman. Assortative mixing in networks[J]. Physical Review Letters, 2002, 89(20):208701

[104] M. Kitsak, L. K. Gallos, S. Havlin, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11):888–893

[105] M. Girvan, M. E. J. Newman. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12):7821–7826

[106] A. Clauset, M. E. J. Newman, C. Moore. Finding community structure in very large networks[J]. Physical Review E, 2004, 70(6):066111

[107] R. I. M. Dunbar. The social brain hypothesis[J]. Evolutionary Anthropology, 1998, 6(5):178–190

[108] W.-X. Zhou, D. Sornette, R. A. Hill, et al. Discrete hierarchical organization of social group sizes[J]. Proceedings of the Royal Society of London B: Biological Sciences, 2005, 272(1561):439–444

[109] J. M. Kleinberg. Navigation in a small world[J]. Nature, 2000, 406(6798):845

[110] M. Barthélemy. Spatial networks[J]. Physics Reports, 2011, 499:1–101

[111] R. Lambiotte, V. D. Blondel, C. de Kerchove, et al. Geographical dispersal of mobile communication networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(21):5317–5325

[112] Y. Hu, Y. Wang, D. Li, et al. Possible origin of efficient navigation in small worlds[J]. Physical Review Letters, 2011, 106(10):108701

[113] A. Tacchella, M. Cristelli, G. Caldarelli, et al. A new metrics for countries' fitness and products' complexity[J]. Scientific Reports, 2012, 2:723

[114] R. Hausmann, C. A. Hidalgo. The network structure of economic output[J]. Journal of Economic Growth, 2011, 16(4):309–342

[115] S. Bustos, C. Gomez, R. Hausmann, et al. The dynamics of nestedness predicts the evolution of industrial ecosystems[J]. PLoS ONE, 2012, 7(11):e49393

[116] V. Stojkoski, Z. Utkovski, L. Kocarev. The impact of services on economic complexity: Service sophistication as route for economic growth[J]. PLoS ONE, 2016, 11(8):e0161633

[117] F. Schweitzer, G. Fagiolo, D. Sornette, et al. Economic networks: The new challenges[J]. Science, 2009, 325(5939):422–425

[118] N. Arinaminpathy, S. Kapadia, R. M. May. Size and complexity in model financial systems[J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(45):18338–18343

[119] A. G. Haldane, R. M. May. Systemic risk in banking ecosystems[J]. Nature, 2011, 469(7330):351–355

[120] Z. Liu, C. He, J. Wu. General spatiotemporal patterns of urbanization: An examination of 16 world cities[J]. Sustainability, 2016, 8(1):41

[121] G. Chi, J.-C. Thill, D. Tong, et al. Uncovering regional characteristics from mobile phone data: A network science approach[J]. Papers in Regional Science, 2016, 95(3):613–631

[122] N. J. Yuan, Y. Zheng, X. Xie, et al. Discovering urban functional zones using latent activity trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3):712–725

[123] V. Frias-Martinez, E. Frias-Martinez. Spectral clustering for sensing urban land use using Twitter activity[J]. Engineering Applications of Artificial Intelligence, 2014, 35(10):237–245

[124] Y. Zhi, H. Li, D. Wang, et al. Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data[J]. Geo-spatial Information Science, 2016, 19(2):94–105

[125] T. Shelton, A. Poorthuis, M. Zook. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information[J]. Landscape and Urban Planning, 2015, 142:198–211

[126] N. M. Yip, R. Forrest, X. Shi. Exploring segregation and mobilities: Application of an activity tracking app on mobile phone[J]. Cities, 2016, 59:156–163

[127] Z. Yang, D. Lian, N. J. Yuan, et al. Indigenization of urban mobility[J]. Physica A: Statistical Mechanics and its Applications, 2017, 469:232–243

[128] J. Hu, Q.-M. Zhang, T. Zhou. Segregation in religion networks[J]. EPJ Data Science, 2019, 8:6

[129] 贺灿飞, 董瑶, 周沂. 中国对外贸易产品空间路径演化[J]. 地理学报, 2016, 71(6):970–983

[130] 伍业君, 张其仔, 徐娟. 产品空间与比较优势演化述评[J]. 经济评论, 2012, 4:145–152

[131] Q. Guo, C. He. Production space and regional industrial evolution in China[J]. GeoJournal, 2017, 82(2):379–396

[132] M. R. Guevara, D. Hartmann, M. Aristarán, et al. The research space: Using career paths to predict the evolution of the research output of individuals, institutions, and nations[J]. Scientometrics, 2016, 109(3):1695–1709

[133] D. Acemoglu, U. Akcigit, W. R. Kerr. Innovation network[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(41):11483–11488

[134] A. Alabdulkareem, M. Frank, L. Sun, et al. Unpacking the polarization of workplace skills[J]. Science Advances, 2018, 4(7):eaao6030

[135] 周涛, 韩筱璞, 闫小勇,等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4):481–540

[136] 樊超, 郭进利, 韩筱璞,等. 人类行为动力学研究综述[J]. 复杂系统与复杂性科学, 2011, 8(2):1–17

[137] T. Zhou, H. A. T. Kiet, B. J. Kim, et al. Role of activity in human dynamics[J]. EPL (Europhysics Letters), 2008, 82(2):28002

[138] Z. Yang, Z.-K. Zhang, T. Zhou. Anchoring bias in online voting[J]. EPL (Europhysics Letters), 2013, 100(6):68002

[139] M. C. González, C. A. Hidalgo, A.-L. Barabási. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196):779–782

[140] X.-Y. Yan, X.-P. Han, B.-H. Wang, et al. Diversity of individual mobility patterns and emergence of aggregated scaling laws[J]. Scientific Reports, 2013, 3:2678

[141] 吕欣. 大数据技术在应急救援领域的应用及展望[J]. 中国计算机学会通讯, 2018, 14(9):56–62

[142] 尤伟杰, 高见, 周涛. 探索运营商数据在精准扶贫和应急救灾中的应用[J]. 电子科技大学学报(社科版), 2018, 20(6):83–88

[143] L. Gao, C. Song, Z. Gao, et al. Quantifying information flow during emergencies[J]. Scientific Reports, 2014, 4(2):3997

[144] T. Sakaki, M. Okazaki, Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors[C]. Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 2010, 851–860

[145] L. M. Bettencourt, J. Lobo, D. Helbing, et al. Growth, innovation, scaling, and the pace of life in cities[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(17):7301–7306

[146] M. B. Remi Louf. How congestion shapes cities: From mobility patterns to scaling[J]. Scientific Reports, 2014, 4:5561

[147] L. G. A. Alves, H. V. Ribeiro, R. S. Mendes. Scaling laws in the dynamics of crime growth rate[J]. Physica A: Statistical Mechanics and its Applications, 2013, 392(11):2672–2679

[148] W. Pan, G. Ghoshal, C. Krumme, et al. Urban characteristics attributable to density-driven tie formation[J]. Nature Communications, 2013, 4:1961

[149] T. Louail, M. Lenormand, O. G. C. Ros, et al. From mobile phone data to the spatial structure of cities[J]. Scientific Reports, 2014, 4:5276

[150] L. M. Bettencourt. The origins of scaling in cities[J]. Science, 2013, 340(6139):1438–1441

[151] R. Li, L. Dong, J. Zhang, et al. Simple spatial scaling rules behind complex cities[J]. Nature Communications, 2017, 8:1841

[152] J. Um, S.-W. Son, S.-I. Lee, et al. Scaling laws between population and facility densities[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(34):14236–14240

[153] C. A. Hidalgo, E. E. Castañer. The amenity space and the evolution of neighborhoods[J]. arXiv:1509.02868, 2015

[154] C. A. Hidalgo, P.-A. Balland, R. Boschma, et al. The principle of relatedness[C]. Proceedings of the Ninth International Conference on Complex Systems, Cambridge, MA, USA, 2018, 451–457

[155] 吴宗柠, 樊瑛. 复杂网络视角下国际贸易研究综述[J]. 电子科技大学学报, 2018, 47(3):469–480

[156] R. Boschma, S. Iammarino. Related variety, trade linkages, and regional growth in Italy[J]. Economic Geography, 2009, 85(3):289–311

[157] C. He, Y. Yan, D. Rigby. Regional industrial evolution in China[J]. Papers in Regional Science, 2018, 97(2):173–198

[158] C. Jara-Figueroa, B. Jun, E. L. Glaeser, et al. The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms[J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(50):12646–12653

[159] D. Bahar, R. Hausmann, C. A. Hidalgo. Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?[J]. Journal of International Economics, 2014, 92(1):111–123

[160] T. J. Holmes. The diffusion of Wal-Mart and economies of density[J]. Econometrica, 2011, 79(1):253–302

[161] R. Boschma, V. Martín, A. Minondo. Neighbour regions as the source of new industries[J]. Papers in Regional Science, 2017, 96(2):227–245

[162] I. Hong, M. R. Frank, I. Rahwan, et al. A common trajectory recapitulated by urban economies[J]. arXiv:1810.08330, 2018

[163] 林毅夫. 新结构经济学: 反思经济发展与政策的理论框架[M]. 北京: 北京大学出版社, 2014

[164] F. L. Pinheiro, A. Alshamsi, D. Hartmann, et al. Shooting high or low: Do countries benefit from entering unrelated activities?[J]. arXiv:1801.05352, 2018

[165] S. Zhu, C. He, Y. Zhou. How to jump further and catch up? Path-breaking in an uneven industry space[J]. Journal of Economic Geography, 2017, 17(3):521–545

[166] 金璐璐, 贺灿飞, 周沂,等. 中国区域产业结构演化的路径突破[J]. 地理科学进展, 2017, 36(8):974–985

[167] Y.-X. Zhu, J. Huang, Z.-K. Zhang, et al. Geography and similarity of regional cuisines in China[J]. PLoS ONE, 2013, 8(11):e79161

[168] D. S. Hamermesh. Six decades of top economics publishing: Who and how?[J]. Journal of Economic Literature, 2013, 51(1):162–172

[169] J. Spencer. 60+ Social Networking Sites You Need to Know About[EB/OL]. `https://makeawebsitehub.com/social-media-sites`, December 23, 2018

[170] J. Wang, J. Gao, J.-H. Liu, et al. Regional economic status inference from information flow and talent mobility[J]. EPL (Europhysics Letters), 2019, 125(6):68002

[171] R. Dong, L. Li, Q. Zhang, et al. Information diffusion on social media during natural disasters[J]. IEEE Transactions on Computational Social Systems, 2018, 5(1):265–276

[172] K. Leetaru, S. Wang, G. Cao, et al. Mapping the global Twitter heartbeat: The geography of Twitter[J]. First Monday, 2013, 18(5–6):1–33

[173] E.-K. Kim, J. H. Seok, J. S. Oh, et al. Use of hangeul Twitter to track and predict human influenza infection[J]. PLoS ONE, 2013, 8(7):e69305

[174] F. Toriumi, T. Sakaki, K. Shinoda, et al. Information sharing on Twitter during the 2011 catastrophic earthquake[C]. Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 2013, 1025–1028

[175] 高见, 张琳艳, 张千明,等. 大数据人力资源: 基于雇员网络的绩效分析与升离职预测[M]. 北京: 科学出版社, 2014, 38–56

[176] J. Yuan, Q.-M. Zhang, J. Gao, et al. Promotion and resignation in employee networks[J]. Physica A: Statistical Mechanics and its Applications, 2016, 444:442–447

[177] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, et al. D4D-Senegal: The second mobile phone data for development challenge[J]. arXiv:1407.4885, 2014

[178] Q. Wang, J. Gao, T. Zhou, et al. Critical size of ego communication networks[J]. EPL (Europhysics Letters), 2016, 114(5):58004

[179] X. Dong, E. Jahani, A. Morales-Guzman, et al. Purchase patterns, socioeconomic status, and political inclination[C]. Proceedings of the 2nd Annual International Conference on Computational Social Science, Evanston, IL, USA, 2016, 1–5

[180] B. Hashemian, E. Massaro, I. Bojic, et al. Socioeconomic characterization of regions through the lens of individual financial transactions[J]. PLoS ONE, 2017, 12(11):e0187031

[181] S. Aral, C. Nicolaides. Exercise contagion in a global social network[J]. Nature Communications, 2017, 8:14753

[182] 周涛, 李艳丽, 李倩,等. 利用网络数据预测企业失信行为[J]. 大数据, 2018, 4(5):2018049

[183] J. Gao, T. Zhou. Quantifying China's regional economic complexity[J]. Physica A: Statistical Mechanics and its Applications, 2018, 492:1591–1603

[184] X. Yang, J. Gao, J.-H. Liu, et al. Height conditions salary expectations: Evidence from large-scale data in China[J]. Physica A: Statistical Mechanics and its Applications, 2018, 501:86–97

[185] H. Choi, H. Varian. Predicting the present with Google Trends[J]. Economic Record, 2012, 88(s1):2–9

[186] T. Gebru, J. Krause, Y. Wang, et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(50):13108–13113

[187] N. Naik, S. D. Kominers, R. Raskar, et al. Computer vision uncovers predictors of physical urban change[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(29):7571–7576

[188] M. Haklay, P. Weber. OpenStreetMap: User-generated street maps[J]. IEEE Pervasive Computing, 2008, 7(4):12–18

[189] S. G. Donald, K. Lang. Inference with difference-in-differences and other panel data[J]. Review of Economics and Statistics, 2007, 89(2):221–233

[190] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13):1175–1197

[191] L. Lü, D.-B. Chen, X.-L. Ren, et al. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650:1–63

[192] L. C. Freeman. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1):35–41

[193] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1–7):107–117

[194] L. Lü, Y.-C. Zhang, C. H. Yeung, et al. Leaders in social networks, the Delicious case[J]. PLoS ONE, 2011, 6(6):e21202

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

[196] P. Erdös, A. Rényi. On random graphs I[J]. Publicationes Mathematicae Debrecen, 1959, 6:290–297

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

[198] T. Zhou, J. Ren, M. Medo, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4):046115

[199] J. Gao, B. Jun, A. S. Pentland, et al. Collective learning in China's regional economic development[J]. arXiv:1703.01369, 2017

[200] K. Gueorgi, D. J. Watts. Empirical analysis of an evolving social network[J]. Science, 2006, 311(5757):88–90

[201] D.-D. Zhao, A. Zeng, M.-S. Shang, et al. Long-term effects of recommendation on the evolution of online systems[J]. Chinese Physics Letters, 2013, 30(11):118901

[202] P. L. Leath, G. R. Reich. Bootstrap percolation on a Bethe lattice[J]. Journal of Physics C: Solid State Physics, 2001, 12(1):L31

[203] G. J. Baxter, S. N. Dorogovtsev, A. V. Goltsev, et al. Bootstrap percolation on complex networks[J]. Physical Review E, 2010, 82(1):011103

[204] J. Gao, T. Zhou, Y. Hu. Bootstrap percolation on spatial networks[J]. Scientific Reports, 2015, 5:14662

[205] C. M. Bishop. Pattern recognition and machine learning[M]. Berlin, Heidelberg: Springer, 2006

[206] D. G. Kleinbaum, M. Klein. Logistic regression: A self-learning text[M]. New York, NY, USA: Springer, 2010, 5–6

[207] D. R. Cox. The regression analysis of binary sequences[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1958, 20(2):215–232

[208] A. J. Smola, B. Schölkopf. A tutorial on support vector regression[J]. Statistics and Computing, 2004, 14(3):199–222

[209] T.-Y. Liu. Learning to rank for information retrieval[J]. Foundations and Trends in Information Retrieval, 2009, 3(3):225–331

[210] C. Burges, T. Shaked, E. Renshaw, et al. Learning to rank using gradient descent[C]. Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005, 89–96

[211] J. E. Blumenstock. Fighting poverty with data[J]. Science, 2016, 353(6301):753–754

[212] C. Baumann, H. Krskova. School discipline, school uniforms and academic performance[J]. International Journal of Educational Management, 2016, 30(6):1003–1029

[213] Y. Kim, J. Y. Park, S. B. Kim, et al. The effects of Internet addiction on the lifestyle and dietary behavior of Korean adolescents[J]. Nutrition Research and Practice, 2010, 4(1):51–57

[214] R. Wang, G. Harari, P. Hao, et al. SmartGPA: How smartphones can assess and predict academic performance of college students[C]. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 2015, 295–306

[215] M. Kosinski, D. Stillwell, T. Graepel. Private traits and attributes are predictable from digital records of human behavior[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(15):5802–2805

[216] C. Montag, E. Duke, A. Markowetz. Toward Psychoinformatics: Computer science meets psychology[J]. Computational and Mathematical Methods in Medicine, 2016, 2016:2983685

[217] S. Ghosh, S. K. Ghosh. Exploring the association between mobility behaviours and academic performances of students: A context-aware traj-graph (CTG) analysis[J]. Progress in Artificial Intelligence, 2018, 7(4):307–326

[218] Y. A. de Montjoye, L. Radaelli, V. K. Singh. Unique in the shopping mall: On the reidentifiability of credit card metadata[J]. Science, 2015, 347(6221):536–539

[219] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, et al. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text[J]. IEEE Transactions on Information Theory, 1998, 44(3):1319–1327

[220] P. Xu, L. Yin, Z. Yue, et al. On predictability of time series[J]. Physica A: Statistical Mechanics and its Applications, 2019, 523:345–351

[221] E. Kreyszig. Advanced engineering mathematics[M]. New York, NY, USA: John Wiley & Sons, 1979, 880

[222] C. Spearman. The proof and measurement of association between two things[J]. American Journal of Psychology, 1904, 15(1):72–101

[223] A. Vedel. The Big Five and tertiary academic performance: A systematic review and meta-analysis[J]. Personality and Individual Differences, 2014, 71:66–76

[224] J. A. Hanley, B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1):29–36

[225] F. D. Peter. The practice of management[M]. New York, NY, USA: Harper & Brothers, 1954

[226] A. Bryant. Google's quest to build a better boss[N]. The New York Times, March 12, 2011

[227] A. S. Pentland. The new science of building great teams[J]. Harvard Business Review, 2012, 90(4):60–69

[228] 张琳艳, 高见, 洪翔,等. 大数据导航人力资源管理[J]. 大数据, 2015, 1(1):2015012

[229] H.-B. Hu, X.-F. Wang. Unified index to quantifying heterogeneity of complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(14):3769–3780

[230] V. D. Blondel, J. L. Guillaume, R. Lambiotte, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10):P10008

[231] 尚可可, 许小可. 基于置乱算法的复杂网络零模型构造及其应用[J]. 电子科技大学学报, 2014, 43(1):7–20

[232] D. Garlaschelli, M. I. Loffredo. Patterns of link reciprocity in directed networks[J]. Physical Review Letters, 2004, 93(26):268701

[233] M. K. Ahuja, D. F. Galletta, K. M. Carley. Individual centrality and performance in virtual R&D groups: An empirical study[J]. Management Science, 2003, 49(1):21–38

[234] R. T. Sparrowe, R. C. Liden, S. J. Wayne. Social networks and the performance of individuals and groups[J]. Academy of Management Journal, 2001, 44(2):316–325

[235] M. C. Sturman, L. Shao, J. H. Katz. The effect of culture on the curvilinear relationship between performance and turnover[J]. Journal of Applied Psychology, 2012, 97(1):46–62

[236] T. H. Feeley, G. A. Barnett. Predicting employee turnover from communication networks[J]. Human Communication Research, 1997, 23(3):370–387

[237] K. W. Mossholder, R. P. Settoon, S. C. Henagan. A relational perspective on turnover: Examining structural, attitudinal, and behavioral predictors[J]. Academy of Management Journal, 2005, 48(4):607–618

[238] R. S. Burt. Structural holes: The social structure of competition[M]. Cambridge, MA, USA: Harvard University Press, 2009

[239] J. L. Herlocker, J. A. Konstan, L. G. Terveen, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1):5–53

[240] D. M. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation[J]. Journal of Machine Learning Technologies, 2011, 2(1):37–63

[241] A. Mao, W. Mason, S. Suri, et al. An experimental study of team size and performance on a complex task[J]. PLoS ONE, 2016, 11(4):e0153048

[242] G. Palla, A.-L. Barabási, T. Vicsek. Quantifying social group evolution[J]. Nature, 2007, 446(7136):664–667

[243] B. F. Jones, S. Wuchty, B. Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science[J]. Science, 2008, 322(5905):1259–1262

[244] R. I. M. Dunbar. Neocortex size as a constraint on group size in primates[J]. Journal of Human Evolution, 1992, 22(6):469–493

[245] R. A. Hill, R. I. M. Dunbar. Social network size in humans[J]. Human Nature, 2003, 14(1):53–72

[246] J. McAuley, J. Leskovec. Learning to discover social circles in ego networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 2012, 539–547

[247] B. Gonçalves, N. Perra, A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's number[J]. PLoS ONE, 2011, 6(8):e22656

[248] J. Zhao, J. Wu, G. Liu, et al. Being rational or aggressive? A revisit to Dunbar's number in online social networks[J]. Neurocomputing, 2014, 142:343–353

[249] Q. Guo, F. Shao, Z.-L. Hu, et al. Statistical properties of the personal social network in the Facebook[J]. EPL (Europhysics Letters), 2013, 104(2):28004

[250] J. Saramäki, E. A. Leicht, E. López, et al. Persistence of social signatures in human communication[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(3):942–947

[251] G. Heineck. Too tall to be smart? The relationship between height and cognitive abilities[J]. Economics Letters, 2009, 105(1):78–80

[252] T. A. Judge, D. M. Cable. The effect of physical height on workplace success and income: Preliminary test of a theoretical model[J]. Journal of Applied Psychology, 2004, 89(3):428–441

[253] A. Deaton, R. Arora. Life at the top: The benefits of height[J]. Economics and Human Biology, 2009, 7(2):133–136

[254] T. H. Kim, E. Han. Height premium for job performance[J]. Economics and Human Biology, 2017, 26:13–20

[255] I. B. Rosenberg. Height discrimination in employment[J]. Utah Law Review, 2009, 3:907–953

[256] J. Agerström. Why does height matter in hiring?[J]. Journal of Behavioral and Experimental Economics, 2014, 52:35–38

[257] 王军, 高见, 杨枭,等. 在线数据揭示预期薪金的影响因素[J]. 电子科技大学学报, 2019, 48(2):307–314

[258] Q.-M. Zhang, A. Zeng, M.-S. Shang. Extracting the information backbone in online system[J]. PLoS ONE, 2013, 8(5):e62624

[259] F. Ricci, L. Rokach, B. Shapira. Recommender systems handbook[M]. Boston, MA, USA: Springer, 2015

[260] N. Jindal, L. Bing. Review spam detection[C]. Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 2007, 1189–1190

[261] P. Massa, P. Avesani. Trust-aware recommender systems[C]. Proceedings of the 2007 ACM Conference on Recommender Systems, Minneapolis, MN, USA, 2007, 17–24

[262] K. Fujimura, T. Nishihara. Reputation rating system based on past behavior of evaluators[C]. Proceedings of the 4th ACM Conference on Electronic Commerce, New York, NY, USA, 2003, 246–247

[263] Y. Tian, J. Zhu. Learning from crowds in the presence of schools of thought[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012, 226–234

[264] L. Muchnik, S. Aral, S. J. Taylor. Social influence bias: A randomized experiment[J]. Science, 2013, 341(6146):647–651

[265] J. Gao, Y.-W. Dong, M.-S. Shang, et al. Group-based ranking method for online rating systems with spamming attacks[J]. EPL (Europhysics Letters), 2015, 110(2):28003

[266] B. S. Everitt. The Cambridge dictionary of statistics[M]. Cambridge, UK: Cambridge University Press, 1998

[267] J. H. Wilkinson. Rounding errors in algebraic processes[M]. Chelmsford, MA, USA: Courier Corporation, 1994

[268] J. M. Kleinberg. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5):604–632

[269] Q. Ou, Y.-D. Jin, T. Zhou, et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks[J]. Physical Review E, 2007, 75(2):021102

[270] T. Deguchi, K. Takahashi, H. Takayasu, et al. Hubs and authorities in the world trade network using a weighted HITS algorithm[J]. PLoS ONE, 2014, 9(7):e100338

[271] T. Zhou, L. Lü, Y.-C. Zhang. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4):623–630

[272] T. Zhou, Z. Kuscsik, J.-G. Liu, et al. Solving the apparent diversity-accuracy dilemma of recommender systems[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(10):4511–4515

[273] J. Gao, T. Zhou. Evaluating user reputation in online rating systems via an iterative group-based ranking method[J]. Physica A: Statistical Mechanics and its Applications, 2017, 473:546–560

[274] E. H. Simpson. Measurement of diversity[J]. Nature, 1949, 163:688

[275] J. Bobadilla, F. Ortega, A. Hernando, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46:109–132

[276] G. Linden, B. Smith, J. York. Amazon.com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003(1):76–80

[277] D. Billsus, C. A. Brunk, C. Evans, et al. Adaptive interfaces for ubiquitous web access[J]. Communications of the ACM, 2002, 45(5):34–38

[278] J.-H. Liu, T. Zhou, Z.-K. Zhang, et al. Promoting cold-start items in recommender systems[J]. PLoS ONE, 2014, 9(12):e113457

[279] L. Lü, T. Zhou. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6):1150–1170

[280] L. Lü, L. Pan, T. Zhou, et al. Toward link predictability of complex networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(8):2325–2330

[281] B. Sarwar, G. Karypis, J. Konstan, et al. Item-based collaborative filtering recommendation algorithms[C]. Proceedings of the 10th International Conference on World Wide Web, Hong Kong, 2001, 285–295

[282] D. Goldberg, D. Nichols, B. M. Oki, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12):61–70

[283] F. Yu, A. Zeng, S. Gillard, et al. Network-based recommendation algorithms: A review[J]. Physica A: Statistical Mechanics and its Applications, 2016, 452:192–208

[284] Y.-C. Zhang, M. Blattner, Y.-K. Yu. Heat conduction process on community networks as a recommendation model[J]. Physical Review Letters, 2007, 99(15):154301

[285] Y.-C. Zhang, M. Medo, J. Ren, et al. Recommendation model based on opinion diffusion[J]. EPL (Europhysics Letters), 2007, 80(6):417–429

[286] T. Zhou, L.-L. Jiang, R.-Q. Su, et al. Effect of initial configuration on network-based recommendation[J]. EPL (Europhysics Letters), 2008, 81(5):58004

[287] C.-X. Jia, R.-R. Liu, D. Sun, et al. A new weighting method in network-based recommendation[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(23):5887–5891

[288] J.-G. Liu, Q. Guo, Y.-C. Zhang. Information filtering via weighted heat conduction algorithm[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(12):2414–2420

[289] Q. Guo, W.-J. Song, J.-G. Liu. Ultra-accurate collaborative information filtering via directed user similarity[J]. EPL (Europhysics Letters), 2014, 107(1):18001

[290] K. Choi, Y. Suh. A new similarity function for selecting neighbors for each target item in collaborative filtering[J]. Knowledge-Based Systems, 2013, 37(1):146–153

[291] 陈玲姣. 基于社交网络个体行为特征的信息推荐算法研究[D]. 成都: 电子科技大学, 2018, 30–41

[292] E. A. Leicht, P. Holme, M. E. J. Newman. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2):026120

[293] L.-J. Chen, Z.-K. Zhang, J.-H. Liu, et al. A vertex similarity index for better personalized recommendation[J]. Physica A: Statistical Mechanics and its Applications, 2017, 466:607–615

[294] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2):163–175

[295] T.-Y. Liu, T. Qin, J. Xu, et al. LETOR: Benchmark dataset for research on learning to rank for information retrieval[C]. Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval, Amsterdam, Netherlands, 2007, 3–10

[296] T. Zhou, R.-Q. Su, R.-R. Liu, et al. Accurate and diverse recommendations via eliminating redundant correlations[J]. New Journal of Physics, 2009, 11(12):123008

[297] F. E. Walter, S. Battiston, F. Schweitzer. A model of a trust-based recommendation system on a social network[J]. Autonomous Agents and Multi-Agent Systems, 2008, 16(1):57–74

[298] M. Karsai, G. Iñiguez, R. Kikas, et al. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading[J]. Scientific Reports, 2016, 6:27178

[299] Z.-K. Zhang, C. Liu, X.-X. Zhan, et al. Dynamics of information diffusion and its applications on complex networks[J]. Physics Reports, 2016, 651:1–34

[300] J. O'Donovan, B. Smyth. Trust in recommender systems[C]. Proceedings of the 10th International Conference on Intelligent User Interfaces, San Diego, CA, USA, 2005, 167–174

[301] M. Jamali, M. Ester. Trustwalker: A random walk model for combining trust-based and item-based recommendation[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009, 397–406

[302] H. Ma, I. King, M. R. Lyu. Learning to recommend with explicit and implicit social relations[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):29

[303] X. Shen, H. Long, C. Ma. Incorporating trust relationships in collaborative filtering recommender system[C]. Proceedings of the 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Takamatsu, Japan, 2015, 1–8

[304] L.-J. Chen, J. Gao. A trust-based recommendation method using network diffusion processes[J]. Physica A: Statistical Mechanics and its Applications, 2018, 506:679–691

[305] R. E. Lucas Jr. On the mechanics of economic development[J]. Journal of Monetary Economics, 1988, 22(1):3–42

[306] S. Peltzman. Prices rise faster than they fall[J]. Journal of Political Economy, 2000, 108(3):466–502

[307] T. Preis, H. S. Moat, H. E. Stanley. Quantifying trading behavior in financial markets using Google Trends[J]. Scientific Reports, 2013, 3:1684

[308] M. Cristelli, A. Gabrielli, A. Tacchella, et al. Measuring the intangibles: A metrics for the economic complexity of countries and products[J]. PLoS ONE, 2013, 8(8):e70726

[309] M. S. Mariani, A. Vidmer, M. Medo, et al. Measuring economic complexity of countries and products: Which metric to use?[J]. European Physical Journal B, 2015, 88(11):293

[310] B. Balassa. Trade liberalisation and "revealed" comparative advantage[J]. Manchester School, 1965, 33(2):99–123

[311] E. Pugliese, A. Zaccaria, L. Pietronero. On the convergence of the Fitness-Complexity algorithm[J]. European Physical Journal Special Topics, 2016, 225(10):1893–1911

[312] G. Morrison, S. V. Buldyrev, M. Imbruno, et al. On economic complexity and the fitness of nations[J]. Scientific Reports, 2017, 7:15332

[313] R.-J. Wu, G.-Y. Shi, Y.-C. Zhang, et al. The mathematics of non-linear metrics for nested networks[J]. Physica A: Statistical Mechanics and its Applications, 2016, 460:254–269

[314] P. Mealy, J. D. Farmer, A. Teytelboym. Interpreting economic complexity[J]. Science Advances, 2019, 5(1):eaau1705

[315] A. J. G. Simoes, C. A. Hidalgo. The economic complexity observatory: An analytical tool for understanding the dynamics of economic development[C]. Proceedings of the 17th AAAI Conference on Scalable Integration of Analytics and Visualization, Menlo Park, CA, USA, 2011, 39–42

[316] A. Zaccaria, M. Cristelli, R. Kupers, et al. A case study for a new metrics for economic complexity: The Netherlands[J]. Journal of Economic Interaction and Coordination, 2016, 11(1):151–169

[317] Z. Song, K. Storesletten, F. Zilibotti. Growing like China[J]. American Economic Review, 2011, 101(1):196–233

[318] D. S. G. Goodman. China's regional development[M]. London, UK: Routledge, 2013

[319] Y. Xie, X. Zhou. Income inequality in today's China[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(19):6928–6933

[320] D. Hartmann, M. R. Guevara, C. Jara-Figueroa, et al. Linking economic complexity, institutions and income inequality[J]. World Development, 2017, 93:75–93

[321] G. Wan, M. Lu, Z. Chen. Globalization and regional income inequality: Empirical evidence from within China[J]. Review of Income and Wealth, 2007, 53(1):35–59

[322] C. E. Shannon. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(3):379–423

[323] G. Cainelli, R. Evangelista, M. Savona. Innovation and economic performance in services: A firm-level analysis[J]. Cambridge Journal of Economics, 2006, 30(3):435–458

[324] D. H. Autor. Skills, education, and the rise of earnings inequality among the "other 99 percent"[J]. Science, 2014, 344(6186):843–851

[325] J. E. Steele, P. R. Sundsøy, C. Pezzulo, et al. Mapping poverty using mobile phone and satellite data[J]. Journal of the Royal Society Interface, 2017, 14(127):20160690

[326] 刘金虎. 社会化网络的结构、演化和应用研究[D]. 成都: 电子科技大学, 2016, 96–105

[327] J. Gao, B. Jun, T. Zhou, et al. Revealing and maximizing the collective learning effects in Brazilian industrial diversification[EB/OL]. `http://gaocn.net/pdf/Brazil_Collective_Learning.pdf`, March 15, 2019

[328] T. Verma, N. A. M. Araújo, H. J. Herrmann. Revealing the structure of the world airline network[J]. Scientific Reports, 2014, 4:5638

[329] T. Verma, F. Russmann, N. A. M. Araújo, et al. Emergence of core-peripheries in networks[J]. Nature Communications, 2016, 7:10441

[330] H. Petter. Core-periphery organization of complex networks[J]. Physical Review E, 2005, 72(2):046111

[331] M. P. Rombach, M. A. Porter, J. H. Fowler, et al. Core-periphery structure in networks[J]. SIAM Journal on Applied Mathematics, 2012, 74(1):167–190

[332] S. Carmi, S. Havlin, S. Kirkpatrick, et al. A model of Internet topology using k-shell decomposition[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(27):11150–11154

[333] D. Rodrik. What's so special about China's exports?[J]. China and World Economy, 2006, 14(5):1–19

[334] C. A. Hidalgo. Why information grows: The evolution of order, from atoms to economies[M]. New York, NY, USA: Basic Books, 2015

[335] U. V. Luxburg. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4):395–416

[336] N. Birdsall, J. L. Londoño. Asset inequality matters: An assessment of the World Bank's approach to poverty reduction[J]. American Economic Review, 1997, 87(2):32–37

[337] M. Granovetter. The impact of social structure on economic outcomes[J]. Journal of Economic Perspectives, 2005, 19(1):33–50

[338] G. Carra, I. Mulalic, M. Fosgerau, et al. Modelling the relation between income and commuting distance[J]. Journal of the Royal Society Interface, 2016, 13(119):20160306

[339] S. P. Kerr, W. Kerr, Ç. Özden, et al. Global talent flows[J]. Journal of Economic Perspectives, 2016, 30(4):83–106

[340] X.-Y. Yan, X.-P. Han, B.-H. Wang, et al. Diversity of individual mobility patterns and emergence of aggregated scaling laws[J]. Scientific Reports, 2012, 3:2678

[341] L. Lotero, A. Cardillo, R. Hurtado, et al. Interconnected networks[M]. Cham, Switzerland: Springer, 2016, 149–164

[342] V. Frias-Martinez, C. Soguero-Ruiz, E. Frias-Martinez, et al. Forecasting socioeconomic trends with cell phone records[C]. Proceedings of the 3rd ACM Symposium on Computing for Development, Bangalore, India, 2013, 15

[343] L. Pappalardo, D. Pedreschi, Z. Smoreda, et al. Using big data to study the link between human mobility and socio-economic development[C]. Proceedings of the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA, 2015, 871–878

[344] W. G. Mangold, D. J. Faulds. Social media: The new hybrid element of the promotion mix[J]. Business Horizons, 2009, 52(4):357–365

[345] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, et al. Human mobility in advanced and developing economies: A comparative analysis[C]. Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development, Stanford, CA, USA, 2010, 79–84

[346] J. Benhabib, M. M. Spiegel. The role of human capital in economic development evidence from aggregate cross-country data[J]. Journal of Monetary Economics, 1994, 34(2):143–173

[347] E. Pelinescu. The impact of human capital on economic growth[J]. Procedia Economics and Finance, 2015, 22:184–190

[348] B. State, M. Rodriguez, D. Helbing, et al. Migration of professionals to the U.S.[C]. Proceedings of the 6th International Conference on Social Informatics, Barcelona, Spain, 2014, 531–543

[349] A. L. J. Ter Wal, R. A. Boschma. Applying social network analysis in economic geography: Framing some key analytic issues[J]. Annals of Regional Science, 2009, 43(3):739–756

[350] D. Liben-Nowell, J. Novak, R. Kumar, et al. Geographic routing in social networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(33):11623–11628

[351] L. Adamic, E. Adar. How to search a social network[J]. Social Networks, 2005, 27(3):187–203

[352] A. Guille, H. Hacid, C. Favre, et al. Information diffusion in online social networks: A survey[J]. ACM SIGMOD Record, 2013, 42(2):17–28

[353] R. Pastor-Satorras, A. Vespignani. Epidemic spreading in scale-free networks[J]. Physical Review Letters, 2000, 86(14):3200

[354] W. Wang, M. Tang, H.-F. Zhang, et al. Epidemic spreading on complex networks with general degree and weight distributions[J]. Physical Review E, 2014, 90(4):042803

[355] C. Damon. The spread of behavior in an online social network experiment[J]. Science, 2010, 329(5996):1194–1197

[356] F. Zhou, J.-R. Jiao, B. Lei. A linear threshold-hurdle model for product adoption prediction incorporating social network effects[J]. Information Sciences, 2015, 307:95–109

[357] S. K. Majumdar, S. Venkataraman. Network effects and the adoption of new technology: Evidence from the US telecommunications industry[J]. Strategic Management Journal, 1998, 19(11):1045–1062

[358] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, et al. Epidemic processes in complex networks[J]. Reviews of Modern Physics, 2015, 87(3):925

[359] M. J. Keeling, P. Rohani. Modeling infectious diseases in humans and animals[M]. Princeton, NJ, USA: Princeton University Press, 2008

[360] K. Kosmidis, S. Havlin, A. Bunde. Structural properties of spatially embedded networks[J]. EPL (Europhysics Letters), 2008, 82(4):283–286

[361] D. Li, K. Kosmidis, A. Bunde, et al. Dimension of spatially embedded networks[J]. Nature Physics, 2011, 7(6):481–484

[362] T. Emmerich, A. Bunde, S. Havlin, et al. Complex networks embedded in space: Dimension and scaling relations between mass, topological distance, and Euclidean distance[J]. Physical Review E, 2013, 87(3):032802

[363] J. Gao, S. V. Buldyrev, S. Havlin, et al. Robustness of a network of networks[J]. Physical Review Letters, 2011, 107(19):195701

[364] L. M. Shekhtman, B. Yehiel, M. M. Danziger, et al. Robustness of a network formed of spatially embedded networks[J]. Physical Review E, 2014, 90(1):012809

[365] Y. Berezin, A. Bashan, M. M. Danziger, et al. Localized attacks on spatially embedded networks with dependencies[J]. Scientific Reports, 2015, 5:8934

[366] C. F. Moukarzel, T. Sokolowski. Long-range k-core percolation[J]. Journal of Physics: Conference Series, 2010, 246(1):012019

[367] Y. Hu, B. Ksherim, R. Cohen, et al. Percolation in interdependent and interconnected networks: Abrupt change from second- to first-order transitions[J]. Physical Review E, 2011, 84(6):066116

[368] A. V. Goltsev, S. N. Dorogovtsev, J. F. F. Mendes. k-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects[J]. Physical Review E, 2006, 73(5):056101

[369] G. J. Baxter, S. N. Dorogovtsev, A. V. Goltsev, et al. Heterogeneous k-core versus bootstrap percolation on complex networks[J]. Physical Review E, 2011, 83(1):051134

[370] A. Zeng, D. Zhou, Y. Hu, et al. Dynamics on spatial networks and the effect of distance coarse graining[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(21–22):3962–3969

[371] S. Munik, M. Cristopher. Message-passing approach for threshold models of behavior in networks[J]. Physical Review E, 2014, 89(2):022805

[372] P. Roni, S. V. Buldyrev, H. Shlomo. Critical effect of dependency groups on the function of networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(3):1007–1010

[373] D. Li, G. Li, K. Kosmidis, et al. Percolation of spatially constraint networks[J]. EPL (Europhysics Letters), 2011, 93(6):68004

[374] R.-R. Liu, W.-X. Wang, Y.-C. Lai, et al. Cascading dynamics on random networks: Crossover in phase transition[J]. Physical Review E, 2012, 85(2):026110

[375] W. B. Arthur. Increasing returns and path dependence in the economy[M]. Ann Arbor, MI, USA: University of Michigan Press, 1994

[376] C. Lawson, E. Lorenz. Collective learning, tacit knowledge and regional innovative capacity[J]. Regional Studies, 1999, 33(4):305–317

[377] R. R. Nelson, E. S. Phelps. Investment in humans, technological diffusion, and economic growth[J]. American Economic Review, 1966, 56(1/2):69–75

[378] F. Neffke, M. Henning. Skill relatedness and firm diversification[J]. Strategic Management Journal, 2013, 34(3):297–316

[379] R. Boschma, A. Minondo, M. Navarro. The emergence of new industries at the regional level in Spain: A proximity approach based on product relatedness[J]. Economic Geography, 2013, 89(1):29–51

[380] J. Jiao, J. Wang, F. Jin, et al. Impacts on accessibility of China's present and future HSR network[J]. Journal of Transport Geography, 2014, 40(40):123–132

[381] S. Zheng, M. E. Kahn. China's bullet trains facilitate market integration and mitigate the cost of megacity growth[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(14):E1248–E1253

[382] Y. Qin. 'No county left behind?' The distributional impact of high-speed rail upgrades in China[J]. Journal of Economic Geography, 2017, 17(3):489–520

[383] C. Catalini, C. Fons-Rosen, P. Gaulé. Did cheaper flights change the direction of science?[R]. Cambridge, MA, USA: MIT Sloan Research Paper No. 5172-16, 2016

[384] J. Gao. Maximizing the collective learning effects in regional economic development[C]. 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 2017, 337–341

[385] J. McCallum. National borders matter: Canada-US regional trade patterns[J]. American Economic Review, 1995, 85(3):615–623

[386] J. E. Rauch, V. Trindade. Ethnic Chinese networks in international trade[J]. Review of Economics and Statistics, 2002, 84(1):116–130

[387] P.-P. Combes, M. Lafourcade, T. Mayer. The trade-creating effects of business and social networks: Evidence from France[J]. Journal of International Economics, 2005, 66(1):1–29

[388] B. Jun, A. Alshamsi, J. Gao, et al. Relatedness, knowledge diffusion, and the evolution of bilateral trade[J]. arXiv:1709.05392, 2017

[389] S. Ronen, B. Gonçalves, K. Z. Hu, et al. Links that speak: The global language network and its association with global fame[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(52):E5616–E5622

[390] T. Chaney. The network structure of international trade[J]. American Economic Review, 2014, 104(11):3600–3634

[391] A. C. Cameron, J. B. Gelbach, D. L. Miller. Robust inference with multiway clustering[J]. Journal of Business and Economic Statistics, 2011, 29(2):238–249

[392] D. Lazer, R. Kennedy, G. King, et al. The parable of Google Flu: Traps in big data analysis[J]. Science, 2014, 343(6176):1203–1205

[393] A. Wesolowski, N. Eagle, A. M. Noor, et al. Heterogeneous mobile phone ownership and usage patterns in Kenya[J]. PLoS ONE, 2012, 7(4):e35319

[394] T. D. Cook, D. T. Campbell, W. Shadish. Experimental and quasi-experimental designs for generalized causal inference[M]. Boston, MA, USA: Houghton Mifflin, 2002

[395] E. A. Stuart. Matching methods for causal inference: A review and a look forward[J]. Statistical Science, 2010, 25(1):1–21

[396] R. M. Bond, C. J. Fariss, J. J. Jones, et al. A 61-million-person experiment in social influence and political mobilization[J]. Nature, 2012, 489(7415):295–298

攻读博士学位期间取得的成果

发表论文：

[1] **J. Gao**, Y.-C. Zhang, T. Zhou. Computational socioeconomics[J]. Physics Reports, 2019 （第一作者, SCI期刊, 影响因子: 20.099, 已录用）

[2] Y. Cao‡, **J. Gao**‡, D. Lian, Z. Rong, J. Shi, Q. Wang, Y. Wu‡, H. Yao‡, T. Zhou‡. Orderliness predicts academic performance: Behavioural analysis on campus lifestyle[J]. Journal of The Royal Society Interface, 2018, 15(146): 20180210 （共同第一作者, SCI期刊, 影响因子: 3.355, 检索号: GV6VT）

[3] **J. Gao**, T. Zhou. Quantifying China's regional economic complexity[J]. Physica A: Statistical Mechanics and its Applications, 2018, 492: 1591–1603 （第一作者, SCI期刊, 影响因子: 2.132, 检索号: FT9TS）

[4] **J. Gao**, T. Zhou. Evaluating user reputation in online rating systems via an iterative group-based ranking method[J]. Physica A: Statistical Mechanics and its Applications, 2017, 473: 546–560 （第一作者, SCI期刊, 影响因子: 2.132, 检索号: EK6UO）

[5] **J. Gao**, T. Zhou. Stamp out fake peer review[J]. Nature, 2017, 546(7656): 33–33 （第一作者, SCI期刊, 影响因子: 41.577, 检索号: EW3CR）

[6] **J. Gao**, T. Zhou, Y. Hu. Bootstrap percolation on spatial networks[J]. Scientific Reports, 2015, 5: 14662 （第一作者, SCI期刊, 影响因子: 5.578, 检索号: CS4ZR）

[7] **J. Gao**, Y.-W. Dong, M.-S. Shang, S.-M. Cai, T. Zhou. Group-based ranking method for online rating systems with spamming attacks[J]. EPL (Europhysics Letters), 2015, 110(2): 28003 （第一作者, SCI期刊, 影响因子: 1.963, 检索号: CK1RJ）

[8] **J. Gao**. Maximizing the collective learning effects in regional economic development[C]. 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 2017, 337–341 （第一作者, EI国际会议, 检索号: 20183105637128）

[9] 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(4): 625–633 （第一作者, EI期刊, 检索号: 20163202699775）

[10] **J. Gao**, B. Jun, A. S. Pentland, T. Zhou, C. A. Hidalgo. Collective learning in China's regional economic development[J]. arXiv:1703.01369 （第一作者, 拟投稿Nature Communications，影响因子: 12.353）

[11] **J. Gao**, T. Zhou. Vertex similarity in complex networks[J]. 2019（第一作者, 拟投稿Applied Physics Reviews, 影响因子: 13.667）

[12] **J. Gao**, B. Jun, T. Zhou, C. A. Hidalgo. Collective learning among industries and regions in Brazilian economic development[J]. 2019 （第一作者, 拟投稿EPJ Data Science, 影响因子: 2.982）

[13] J. Wang, **J. Gao**†, J.-H. Liu, D. Yang, T. Zhou. Regional economic status inference from information flow and talent mobility[J]. EPL (Europhysics Letters), 2019, 125(6): 68002 （通讯作者, SCI期刊, 影响因子: 1.834）

[14] L.-J. Chen, **J. Gao**†. A trust-based recommendation method using network diffusion processes[J]. Physica A: Statistical Mechanics and its Applications, 2018, 506: 679–691 （通讯作者, SCI期刊, 影响因子: 2.132, 检索号: GL3XI）

[15] X. Yang, **J. Gao**†, J.-H. Liu, T. Zhou. Height conditions salary expectations: Evidence from large-scale data in China[J]. Physica A: Statistical Mechanics and its Applications, 2018, 501: 86–97 （通讯作者, SCI期刊, 影响因子: 2.132, 检索号: GC8FC）

[16] L.-J. Chen, Z.-K. Zhang, J.-H. Liu, **J. Gao**†, T. Zhou. A vertex similarity index for better personalized recommendation[J]. Physica A: Statistical Mechanics and its Applications, 2017, 466: 607–615 （通讯作者, SCI期刊, 影响因子: 2.132, 检索号: ED0QU）

[17] 王军, **高见**†, 杨枭, 刘金虎, 周涛. 在线数据揭示预期薪金的影响因素[J], 电子科技大学学报, 2019, 48(2): 307–314 （通讯作者, EI期刊）

[18] Q. Wang, **J. Gao**, T. Zhou, Z. Hu, H. Tian. Critical size of ego communication networks[J]. EPL (Europhysics Letters), 2016, 114(5): 58004 （第二作者, SCI期刊, 影响因子: 1.957, 检索号: DQ9JE）

[19] 尤伟杰, **高见**, 周涛. 探索运营商数据在精准扶贫和应急救灾中的应用[J]. 电子科技大学学报社科版, 2018, 20(6): 83–88 （第二作者, 核心期刊）

[20] 张琳艳, **高见**, 洪翔, 周涛. 大数据导航人力资源管理[J]. 大数据, 2015, 1(1): 2015012 （第二作者, 核心期刊）

[21] J. Yuan, Q.-M. Zhang, **J. Gao**, L. Zhang, X.-S. Wan, X.-J. Yu, T. Zhou. Promotion and resignation in employee networks[J]. Physica A: Statistical Mechanics and its Applications, 2016, 444: 442–447 （第三作者, SCI期刊, 影响因子: 2.243, 检索号: CZ0HO）

[22] L. Pan, L. Gao, **J. Gao**. Link prediction in weighted networks via structural perturbations[C]. 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 2017, 5–8 （第三作者, EI国际会议, 检索号: 20183105637045）

[23] B. Jun, A. Alshamsi, **J. Gao**, C. A. Hidalgo. Relatedness, knowledge diffusion, and the evolution of bilateral trade[J]. arXiv:1709.05392（第三作者, Journal of Evolutionary Economics第二轮审稿, SSCI期刊, 影响因子: 1.095）

图书章节：

[1] **高见**, 张琳艳, 张千明, 周涛. 大数据人力资源: 基于雇员网络的绩效分析与升离职预测[M]//刘怡君. 社会物理学–社会治理. 北京: 科学出版社, 2014, 38-56 （第一作者, 中文图书章节）

[2] Y. Cao, **J. Gao**, T. Zhou. Orderliness of campus lifestyle predicts academic performance: A

case study in Chinese university[M]. H. Baumeister, C. Montag. Mobile Sensing and Psychoinformatics. Springer: Berlin, Germany, 2019 （第二作者, 英文图书章节）

科研项目：

[1] 时空大数据可视分析中信息混淆问题研究, 2019, 国家自然科学基金面上项目, 项目编号: 61872066, 项目角色: 参与.

[2] 科技人力资源与产业结构或经济结构的关系研究, 2019, 中国科协创新战略研究院科研项目, 项目编号: 2018ysxh1-4-5-5, 项目角色: 参与.

[3] 基于复杂网络分析的文化演化研究及应用, 2018, 国家自然科学基金青年项目, 项目编号: 61703074, 项目角色: 参与.

[4] 社交网络中信息主体的行为模式分析及应用研究, 2018, 国家自然科学基金面上项目, 项目编号: 61673086, 项目角色: 参与.

[5] 基于城市出行的人类时空耦合行为研究, 2017, 国家自然科学基金青年项目, 项目编号: 61603074, 项目角色: 参与.

[6] 大数据语境下员工绩效与行为倾向预测研究, 2016, 国家社会科学基金青年项目, 项目编号: 15CGL029, 项目角色: 参与.

[7] 动态演化在线系统中的信息推荐问题研究, 2015, 国家自然科学基金面上项目, 项目编号: 61370150, 项目角色: 参与.

[8] 社会网络及其上的传播动力学集成研究, 2014, 国家自然科学基金重大研究计划培育项目, 项目编号: 91324002, 项目角色: 参与.

获奖情况：

[1] 国家建设高水平大学公派研究生项目奖学金, 电子科技大学-麻省理工学院（MIT）联合培养博士, 国家留学基金委, 2016年8月-2017年9月.

[2] Best Poster Award, 国际网络科学会议（NetSciX 2018）, 2018年.

[3] 优秀张贴报告奖, 中国复杂性科学研究会（第二次学术会议）, 2014年.

[4] 四川省优秀毕业研究生, 四川省教育厅, 2019年.

[5] 博士研究生国家奖学金, 国家教育部, 2016年.

[6] 唐立新奖学金, 电子科技大学, 2013年-2019年.

[7] 研究生学业一等奖学金, 电子科技大学, 2016年.

[8] 中国电科三十二所奖学金, 电子科技大学, 2015年.

[9] 研究生学业三等奖学金, 电子科技大学, 2015年.

[10] 博士研究生新生奖学金, 电子科技大学, 2014年.

学术活动：

[1] 第十五届中国网络科学论坛, 报告题目: 计算社会经济学：利用信息和人才流动推断区

域经济状况, 辽宁大连, 2019年5月11日.

[2] AI & Society学术沙龙第十三期, 数据时代的社会经济发展, 报告题目: 计算社会经济学的内容与方法, 四川成都, 2018年11月22日.

[3] 全国大数据与社会计算学术会议（BDSC 2018）, 社会计算与数字经济专题, 报告题目: 计算社会经济学-数据驱动的社会经济研究, 河北石家庄, 2018年8月29日.

[4] 中国数据挖掘会议（CCDM 2018）, 社交网络分析与挖掘论坛, 报告题目: 计算社会经济学-大数据驱动的社会经济洞察, 山东济南, 2018年8月7日.

[5] 国际网络科学会议（NetSciX 2018）, 报告题目: Evaluating Online Reputation via Group-Based Ranking Methods, 浙江杭州, 2018 年1月7日.

[6] 国际网络科学会议（NetSciX 2018）, 报告题目: Collective Learning in Regional Economic Development, 浙江杭州, 2018年1月6日.

[7] 美国地理学会年度会议（AAG 2017）, 报告题目: Collective Learning in Regional Economic Development, 美国波士顿, 2017年4月5 日.

[8] 智利BIT, BOTS, BRAIN & BEHAVIOR（B4 2017）, 报告题目: Collective Learning in China's Regional Economic Development, 智利圣地亚哥, 2017 年1月31日.

[9] 美国MIT Macro Connections学术研讨会, 报告题目: Knowledge Diffusion: Who Can We Learn From?, 美国波士顿, 2016年9月19日.

[10] 波士顿大学统计物理与复杂性研讨会, 报告题目: Symbiosis and Monopoly of Domestic Industry Structure, 美国波士顿, 2016年2月25日.

[11] 西南科技大学复杂性科学专题讨论会, 报告题目: Local Industry Structure in China: Symbiosis and Monopoly, 四川绵阳, 2015年10月13日.

[12] 首届全国经济复杂性跨学科研究会学术年会, 报告题目: Linking Companies and Economic Complexity, 山东烟台, 2015年6月27日.

[13] 第十届全国复杂网络大会, 报告题目: Bootstrap Percolation on Spatial Networks, 湖南长沙, 2014年10月17 日.

[14] 复杂性科学研究会第二次学术会议, 报告题目: Percolation on Random Spatial Networks, 浙江温州, 2014年7 月14日.

审稿服务：

[1] Knowledge-Based Systems, Elsevier, 2017-2019.

[2] Computer Standards and Interfaces, Elsevier, 2017-2019.

[3] IEEE Access, IEEE, 2019.

[4] Journal of the Royal Society Interface, Royal Society, 2018.

[5] Physica A: Statistical Mechanics and its Applications, Elsevier, 2016-2018.

[6] International Journal of Modern Physics C, World Scientific Publishing, 2018.

[7] Physica Scripta, IOP Publishing, 2018.

[8] Physics Letters A, Elsevier, 2017.

[9] Journal of Computational Methods in Sciences and Engineering, IOS Press, 2017.

[10] IEEE Transactions on Knowledge and Data Engineering, IEEE Press, 2016.

[11] 电子科技大学学报, 电子科技大学, 2014-2017.