数据论文

# 面向技术创新系统的专利数据集构建: 跨尺度结构对比与演化机制研究*

黄羿炜[1)] 徐舒琪[2)†] 吕琳媛[3)‡]

1) (电子科技大学基础与前沿研究院, 成都 611731)

2) (合肥综合性国家科学中心数据空间研究院, 合肥 230088)

3) (中国科学技术大学网络空间安全学院, 合肥 230026)

(2025 年 12 月 31 日收到; 2026 年 2 月 4 日收到修改稿)

社会经济复杂系统中, 网络拓扑结构往往呈现显著的尺度依赖性: 同一类经济活动在不同的时空尺度下可能呈现不同的组织形态, 而这种跨尺度结构差异的形成机制仍缺乏系统性解释. 本文以技术创新系统为研究对象, 构建并公开一套美国专利数据集, 覆盖 2000—2020 年申请的 1225373 件授权发明专利, 整合分类代码、受让主体与文本信息, 并依据引用信息对专利涉及的多项技术进行加权分配, 以量化技术活动强度. 基于该数据集, 本文在州、县、城市与企业四个尺度分别构建“创新主体-技术”二分网络及其对应的技术空间网络, 并通过计算模块度与嵌套性指标, 揭示网络结构的尺度依赖规律. 结果表明, 宏观尺度网络呈现出典型的全局嵌套结构, 而微观尺度网络呈现出更强的模块化与块内嵌套特征. 进一步地, 本文提出一个包含技术相关密度偏好与技术中心性偏好的演化模型, 用以描述主体选择技术的偏好机制. 模拟实验证明, 通过调节两类偏好参数, 该模型能够复现从全局嵌套到块内嵌套的结构转变. 本文公开的数据集 (https://doi.org/10.57760/sciencedb.j00213.00265) 与机制模型为理解跨尺度网络结构的差异提供了实证基础与可检验的解释框架.

**关键词:** 专利数据, 技术创新系统, 复杂网络, 嵌套性, 模块化

**DOI:** 10.7498/aps.75.20251802 **CSTR:** 32037.14.aps.75.20251802

## 1 引 言

复杂系统广泛存在于自然、社会与经济等多个领域, 其宏观功能往往由大量微观单元的相互作用涌现而来. 网络科学为刻画这类系统提供了统一框架: 将系统抽象为由节点与连边构成的网络, 进而通过分析网络的结构特征以揭示系统的组织方式. 大量实证研究表明, 复杂网络的拓扑结构呈现显著的尺度依赖性 (scale dependence): 当观测与建模的尺度发生改变时 (例如表征粒度由个体聚合为群体[1] 或者空间尺度由单点扩展到区域[2,3]), 同一系统往往呈现不同的连接模式与统计特征, 甚至出现系统性的结构形态转变[1–4]. 由此引出一个关键科学问题: 跨尺度的结构差异如何产生, 是否存在一类通用的演化规则, 能够在统一的框架下解释从微观到宏观的结构演化? 这一问题已成为复杂系统与复杂网络研究的重要议题之一.

要对尺度依赖的结构差异进行定量刻画, 首先需要选取能够表征网络组织方式的结构指标. 在

* 国家自然科学基金重大项目 (批准号: T2293771) 和国家自然科学基金青年科学基金 (批准号: 62306191) 资助的课题.

† 通信作者. E-mail: xushuqi@idata.ah.cn

‡ 通信作者. E-mail: linyuan.lv@ustc.edu.cn

© 2026 中国物理学会 Chinese Physical Society 

多类真实网络中, 模块化 (modularity) 与嵌套性 (nestedness) 是两类具有代表性的结构特征. 模块化刻画网络的分区组织, 表现为网络可被划分为若干内部连接密集而模块间连接相对稀疏的子模块[5,6], 反映系统的功能分工与组织边界. 嵌套性刻画网络的层级组织, 表现为低度节点的邻居集合倾向于被高度节点的邻居集合所包含, 从而形成由“核心”向“边缘”逐级收缩的拓扑结构. 嵌套性最初在生态系统研究中被提出, 随后在生态网络[7–9]、社交网络[10,11]与经济网络[12–14]等多类系统中被广泛识别与研究, 并被证实与系统鲁棒性与稳定性密切相关[15–17]. 总体而言, 模块化与嵌套性分别从“分区”与“层级”两个互补维度刻画网络组织方式, 共同构成理解复杂系统结构与演化规律的基础指标. 在跨尺度研究中, 将两者纳入同一指标体系, 可为不同尺度网络的结构对照提供统一量化框架, 并为揭示系统结构演化机制提供关键切入点.

早期研究常将嵌套性与模块化视为相互排斥的结构模式, 并分别将其归因于合作和竞争主导的动力学过程[18]. 然而, 越来越多的研究表明, 许多真实网络并非呈现单一结构, 而是表现为一种复合形态: 网络整体由多个模块构成, 同时在各模块内部存在清晰的嵌套层级组织. 该复合结构通常被称为块内嵌套结构 (in-block nestedness), 已在生态、社会与经济系统中得到验证[4,11,19], 并被认为是模块化分区与嵌套层级在同一系统中共同作用的结果[20]. 因此, 在开展跨尺度结构比较时, 有必要将模块化、嵌套性及其复合形态纳入同一指标体系, 以形成覆盖单一结构与复合结构的统一量化框架.

在社会经济系统中, 跨尺度结构差异同样尤为显著. Laudati 等[4]对国家-产品与企业-产品二分网络的对比研究发现, 即便源于相同的产品出口行为, 不同尺度主体形成的网络结构仍可能显著不同, 宏观尺度的国家-产品网络往往表现为明显的全局嵌套结构, 而微观尺度的企业-产品网络则呈现显著的块内嵌套性. 该研究将这种差异理解为多样化过程受主体规模与能力限制的结果, 从而提示全局嵌套与块内嵌套结构可能是同一演化规则在不同尺度下的差异化表现. 由此引出一个更具一般性的问题: 在同一社会经济系统中, 当主体尺度从宏观转向微观时, 网络结构为何会在全局嵌套与块内嵌套等形态之间发生转变, 这种转变能否由统一机制解释?

围绕块内嵌套结构的形成机制, 现有理论研究主要建立在生态系统的演化假设之上. Leung 和 Weitz[21]受病毒-宿主相互作用启发, 提出了基于冲突连接的二分网络增长模型; Pinheiro 等[22]从消费者-资源关系出发, 构建了包含专一化与泛化权衡的进化模型. 上述工作均能解释生态网络中的块内嵌套结构, 然而这些模型的核心机制通常依赖突变、遗传漂变与自然选择等长期被动演化过程[23], 因而难以直接迁移到社会经济系统. 相比之下, 社会经济系统中的主体 (如企业、城市) 会在信息、能力与制度等多重约束下进行有目的的决策与互动, 其连接结构的形成更多由选择过程与策略调整塑造[24,25]. 因此, 要解释社会经济系统中的跨尺度结构差异, 需要提出能够显式刻画主体约束与选择行为的机制模型, 以便在不同尺度上对机制解释进行一致检验.

除结构指标之外, 开展跨尺度结构的定量比较还需要高质量的实证数据支撑. 已有研究多依赖国际贸易数据[4,24], 其局限主要体现在两方面: 首先, 出口数据更多呈现经济产出的多样化结果, 而支撑其背后的核心能力难以被直接观测和量化, 通常只能通过产出结构间接推断[26,27]; 其次, 在数据粒度上, 微观层面的出口记录在覆盖范围和时间连续性上通常弱于宏观统计, 这限制了从宏观到微观的系统性对照研究. 相比之下, 以专利为代表的技术创新数据提供了更理想的数据基础: 一方面, 针对能力难以直接测度的问题, 专利的分类代码与文本内容能够刻画创新主体所涉及的知识构成与技术组合, 从而为其技术能力结构的衡量刻画提供可检验的分析基础. 另一方面, 针对跨尺度数据不一致的问题, 专利受让人的地理与企业标识使技术活动能够在州、县、城市与企业尺度上被一致映射, 为开展稳健的跨尺度比较奠定了基础. 然而, 尽管专利数据具有上述潜力, 但面向技术创新系统的跨尺度研究仍面临一个关键的数据瓶颈: 支撑跨尺度比较需要经过专利数据收集、专利数据清洗、主体地理定位、企业归属匹配、技术定义以及多分类专利的归属计量等环节, 而目前尚缺乏公开、严格处理且可重复使用的基准数据集与标准化流程.

基于此, 本文构建并发布一套面向技术创新系统的多维专利数据集, 整合专利分类信息、受让主体信息与文本信息, 并在统一框架下将受让人映射至州、县、城市与企业四个层级. 针对涉及多项技

术的专利, 本文进一步基于引用信息对技术活动进行加权计量, 以获得更贴近实际的技术活动强度量化, 从而为跨尺度构建主体-技术二分网络与技术空间网络提供一致的数据基础. 在该数据集之上, 本文从结构分析和机制模拟两个层面展开分析: 其一, 使用模块化与嵌套性结构指标系统揭示网络结构的尺度依赖规律; 其二, 提取主体选择技术发展的一般规律, 并据此构建了一个包含技术相关密度偏好与中心性偏好的网络生成模型, 通过调节两类偏好参数来检验模型是否能够在统一框架下复现不同尺度下从全局嵌套到块内嵌套的结构转变.

总体而言, 本文通过构建可复用的多尺度专利数据集与标准化处理流程, 并提供跨尺度结构比较与机制检验的分析框架, 为理解技术创新系统中不同尺度主体的结构差异及其形成机制提供数据与方法支撑, 并为后续关于技术系统结构演化与创新政策含义的研究奠定基础.

## 2 数据与方法

### 2.1 专利数据集介绍

专利作为技术创新活动的正式产出, 系统记录了技术产生的时间、归属主体及其技术内容, 是刻画技术系统演化过程的核心数据来源. 专利数据不仅包含专利申请的时序信息, 还通过专利分类代码与文本描述刻画了专利所涉及的技术, 从而为后续构建技术空间及其结构分析提供了基础.

本文围绕技术创新系统的结构与演化特征, 基于美国专利数据构建了一个多尺度专利数据集. 本文使用的数据来源于 PatentsView 数据库[①], 该数据库整合并整理了美国专利商标局 (USPTO) 授权专利的公开数据. 为保证数据的完整性与可重复性, 本文设置如下筛选条件: 1) 专利类型限定为美国专利商标局授权的发明专利 (Utility patent); 2) 专利申请年份在 2000 年至 2020 年之间; 3) 专利至少存在一个受让人[②]的地理位置位于美国 50 个州之一. 在上述条件下, 最终获得有效专利共 1225373 件.

图 1(a) 给出了按申请年份统计的授权专利数量分布. 从整体上看, 自 2009 年起年度授权专利数量呈现出持续增长趋势. 需要指出的是, 2018—2020 年专利数量的回落主要源于专利申请与授权之间普遍存在的时间滞后, 并不意味着技术创新活动的发生频率出现实质性下降[③].

专利的分类代码用于标识其所涉及的技术, 是连接专利与技术系统的核心要素. 本文采用联合专利分类体系 (Cooperative Patent Classification, CPC) 对技术进行界定. CPC 体系采用多层级结构, 从技术部 (section) 逐级细化至小组 (subgroup), 其层次结构示例如表 1 所列. 一般而言, 分类层级越细, 技术定义越精确, 但同时也会导致技术数量增加并降低不同技术之间的可比性. 为兼顾技术定义

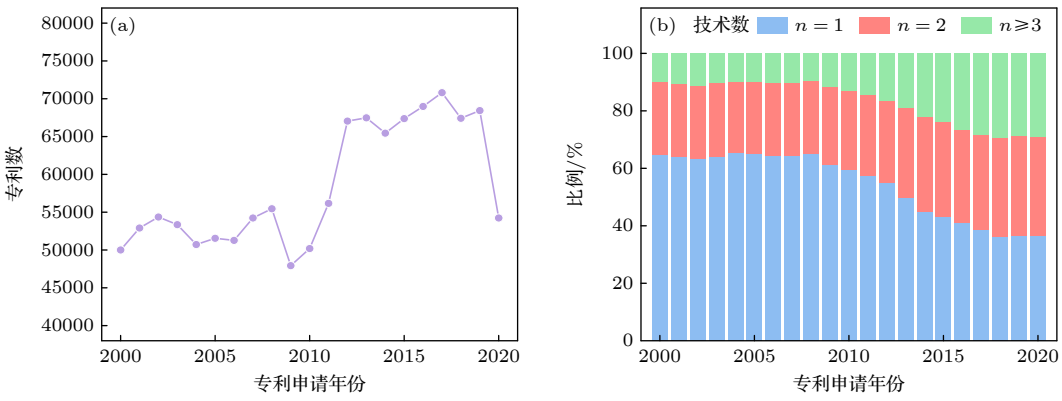



图 1 专利数据统计特征 (a) 按申请年份统计的授权发明专利数量; (b) 不同申请年份中涉及不同数量技术的专利占比

Fig. 1. Descriptive statistics of the patent dataset: (a) Number of granted USPTO utility patents by filing year; (b) proportion of patents involving different numbers of technologies by filing year.

① PatentsView 数据库链接: https://patentsview.org/download/data-download-tables.

② “受让人”(assignee) 指依法享有专利所有权的个人或机构, 其所在地可作为界定专利地理归属的关键依据.

③ 截至 2024 年 7 月, 美国专利商标局从专利申请提交之日起到发出第一次审查意见通知 (first office action) 的平均时间超过 19.7 个月, 详见报告: https://www.uspto.gov/sites/default/files/documents/ppac-2024-annual-report.pdf.

的准确性与区分度, 本文选取 CPC 分类代码的前四位 (即小类层级) 来表示一项技术. 需要指出的是, CPC 分类代码可进一步区分为发明信息 (invention information) 与额外信息 (additional information) 两类①. 其中, 发明信息直接反映专利相对于现有技术的实质性创新内容, 而额外信息主要用于补充检索与分类. 为确保技术能够真实刻画创新活动的核心内容, 本文仅保留代表发明信息的分类代码. 经处理, 最终数据集中共涉及 624 项不同的技术.

表 1 CPC 专利分类体系的层次结构示例

Table 1. An illustrative hierarchical structure of the CPC classification system.

| 分类代码 | 级别 | 含义 |
|---|---|---|
| G | 部 | 物理 |
| G06 | 大类 | 计算; 计数 |
| **G06F** | **小类** | **数据处理** |
| G06F18/ | 大组 | 模式识别 |
| G06F18/10 | 小组 | 预处理; 数据清洗 |

在完成专利筛选与技术定义后, 本文进一步整合了专利的地理归属信息与企业匹配信息, 构建用于跨尺度分析的标准化数据表. 专利与受让人企业之间的匹配关系来源于 Stoffman 等 [28] 的研究②, 该匹配数据将专利受让人与 CRSP (Center for Research in Security Prices) 数据库中的企业唯一识别号 (PERMCO) 对应, 为企业尺度的技术分析提供了可靠基础.

表 2 汇总了本文构建的数据集中各字段的含义及其示例. 其中, 专利在不同技术上的份额 (share) 的计算方法将在 2.2 节详细说明.

## 2.2 专利技术份额计算

技术创新活动日益呈现出多技术交叉融合的特征. 基于本文所构建的专利数据集统计发现, 自 2013 年起, 涉及两项及以上技术的专利在年度授权专利中的占比已超过 50%, 如图 1(b) 所示. 在此背景下, 需要为涉及多项技术的专利分配其在各技术上的权重, 从而将原始专利数据转化为可直接用于技术层面分析的结构化数据.

为此, 本文采用 Shen 和 Barabási[29] 提出的贡献度分配方法. 该方法最初用于衡量论文中不同作者的相对贡献, 其核心思想是利用共同被引信息刻画论文之间的相似性, 并依据相似论文中作者的贡献结构, 对目标论文中的作者贡献进行加权分配. 将该思想引入专利数据分析中, 可以借助专利在引用层面的关联结构, 对涉及多项技术的专利在不同技术上的归属份额进行量化.

具体而言, 专利 $i$ 在技术 $x$ 上的归属份额计算如下:

$$s_{i,x} = \sum\nolimits_{j \in U_i} n_{i,j}\, r_{j,x}, \tag{1}$$

其中 $U_i$ 表示与专利 $i$ 共同被引用的专利集合 (包含 $i$ 本身), $n_{i,j}$ 表示同时引用 $i$ 和 $j$ 的专利数量. $r_{j,x}$ 表示技术 $x$ 在专利 $j$ 中的占比, 例如, 若专利 $j$ 涉及技术组合 $\{x, x, y\}$ ③, 则 $r_{j,x} = 2/3$, $r_{j,y} = 1/3$. 当专利 $i$ 不存在与其共同被引用的其他专利时, 其

表 2 本文构建的专利数据集字段说明

Table 2. Schema and field definitions of the constructed patent dataset.

| 字段名 | 含义 | 示例(多个值由“;”分隔) |
|---|---|---|
| fyear | 专利申请年份 | 2000 |
| patnum | 专利在PatentsView数据库中的唯一识别号 | 6102874 |
| code | 专利涉及的技术 | A61B; G16H |
| share | 专利在对应技术上的归一化份额 | 0.8334; 0.1666 (与code字段顺序一致) |
| state | 专利受让人所在美国州的FIPS代码 | 27 |
| county | 专利受让人所在美国县的FIPS代码 | 27053(前两位表示州, 后三位表示县) |
| city | 专利受让人所在美国城市名称 | 27_Minneapolis(前两位为州代码, 用于区分同名城市) |
| firm | 专利受让人企业的唯一识别号PERMCO | 2850 |
| title | 专利标题 | Implantable medical device for tracking patient functional status |
| abstract | 专利摘要文本 | An implantable medical device...(示例为摘要片段, 完整摘要文本见数据集) |

① 关于两类信息的详细说明参见《Guide to the CPC》: https://www.cooperativepatentclassification.org/sites/default/files/attachments/212f75e9-e9d4-4446-ad7f-b8e943588d1b/Guide+to+the+CPC.pdf.

② 匹配数据下载链接: https://github.com/mwoeppel/patent-crsp-permco-match.

③ 同一专利可能被赋予多个分类代码; 当将分类代码按前四位映射为技术时, 不同代码可能映射到同一技术, 从而在技术层面出现重复.

技术份额退化为该专利自身的技术占比, 即 $s_{i,x}=r_{i,x}$. 该方法通过共同被引频率对技术占比进行加权, 隐含假设为: 与目标专利 $i$ 共同被引次数越高的专利, 在技术构成上与其越相近, 因此其技术分布对量化专利 $i$ 的技术份额具有更大的参考价值. 为保证技术份额在专利层面上的可比性, 最终对 $s_{i,x}$ 进行归一化处理, 使得每项专利在所有技术上的份额之和为 1, 即 $\sum_x s_{i,x}=1$.

## 2.3 网络构建

### 2.3.1 主体-技术二分网络

根据前文描述的专利数据集, 本文构建了覆盖州、县、城市与企业四种尺度的主体-技术二分网络. 在原始数据中, 专利作为中介, 分别与主体和技术建立关联关系: 一方面, 专利与技术之间的连边由专利的 CPC 分类代码确定, 表示该专利涉及相应的技术; 另一方面, 专利与主体之间的连边表示该主体为专利的受让人, 或专利受让人所在的地理位置与该主体相对应. 由此形成如图 2(a) 所示的"主体-专利-技术"三层关联结构.

在此基础上, 本文以专利为媒介, 将主体与技术直接关联, 从而构建主体-技术二分网络, 如图 2(b) 所示. 该网络由两类节点组成, 分别表示主体与技术; 主体与技术之间的连边表示该主体在对应技术上申请专利, 而主体之间及技术之间不直接相连.

在数学形式上, 主体-技术二分网络表示为一个含权二分图:

$$G_{\text{Weighted}}=(A,B,E,W),$$

其中, $A=\{a_1,a_2,\cdots,a_n\}$ 表示主体节点集合, $B=\{b_1,b_2,\ldots,b_m\}$ 表示技术节点集合, $E=\{(a,b)\mid a\in A,\ b\in B\}$ 为连边集合, $W$ 为连边权重集合.

每条连边的权重刻画主体在对应技术上的申请专利数量, 对于主体 $a$ 与技术 $b$ 之间的连边 $(a,b)$, 其权重定义为

$$w_{a,b}=\sum\nolimits_{i\in P_a}\frac{s_{i,b}}{NA_i}, \tag{2}$$

其中, $P_a$ 表示以主体 $a$ 作为受让人的专利集合, $s_{i,b}$ 表示专利 $i$ 在技术 $b$ 上的技术份额 (由 (1) 式给出), $NA_i$ 表示专利 $i$ 关联的主体数量. 该定义通过对多主体专利进行均分处理, 避免专利对多个主体的技术活动强度产生重复计量.

由于含权二分网络中的连边权重容易受到主体规模与技术普遍性差异的影响, 直接使用 $w_{a,b}$ 可能导致对主体技术结构的刻画产生系统性偏差. 为进一步识别主体在特定技术上的相对优势, 本文引入显性比较优势指数 (revealed comparative advantage, RCA)[30], 其定义为

$$\text{RCA}_{a,b}=\frac{w_{a,b}}{\dfrac{\sum_{a'}w_{a',b}}{\sum_{b'}w_{a,b'}}{\sum_{a',b'}w_{a',b'}}}, \tag{3}$$

其中, 分子刻画主体 $a$ 在技术 $b$ 上所占的技术活动比重, 反映其在该技术上的参与强度; 分母刻画主体 $a$ 在所有技术上的总体技术活动比重, 反映其平均技术参与水平. 由此, $\text{RCA}_{a,b}$ 衡量的是主体 $a$ 在技术 $b$ 上的相对专业化程度, 即主体在该技术上的活动强度是否高于其平均水平.

本文依据 $\text{RCA}_{a,b}$ 对主体-技术连边进行过滤[4]: 当 $\text{RCA}_{a,b}\geqslant 1$ 时, 认为主体 $a$ 在技术 $b$ 上具有显性比较优势, 保留对应连边; 当 $\text{RCA}_{a,b}<1$ 时, 则删去该连边. 由此得到过滤后的不含权主体-技术二分网络 $G_{\text{Binary}}$, 其邻接矩阵 $\boldsymbol{M}$ 定义为

$$\boldsymbol{M}_{a,b}=\begin{cases}1, & \text{if RCA}_{a,b}\geqslant 1,\\ 0, & \text{if RCA}_{a,b}<1.\end{cases} \tag{4}$$

后续所有分析均基于过滤后的不含权主体-技术二分网络 $G_{\text{Binary}}$ 展开.

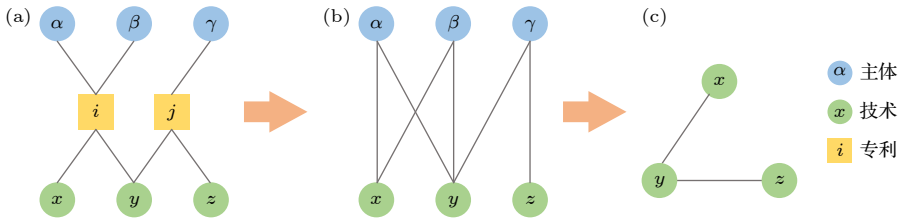



图 2 网络构建过程示意图 (a) "主体-专利-技术"三层关联结构; (b) 主体-技术二分网络; (c) 技术空间网络

Fig. 2. Illustration of network construction: (a) Entity-patent-technology relationship; (b) entity-technology bipartite network; (c) technology space network.

#### 2.3.2 技术空间网络

如图 2(c) 所示, 技术空间网络用于刻画不同技术之间的相似性结构, 其中节点表示技术, 节点之间的连边表示技术之间的相关关系, 连边权重反映其相似程度. 基于 2.3.1 节构建的主体-技术二分网络 $G_{\text{Binary}}$, 本文首先构建基于共现关系的技术空间网络. 技术 $b$ 与 $b'$ 之间的相似性定义为

$$\phi_{b,b'} = \frac{\sum_a \boldsymbol{M}_{a,b}\boldsymbol{M}_{a,b'}}{\max(k_b, k_{b'})}, \tag{5}$$

其中, $k_b = \sum_a \boldsymbol{M}_{a,b}$ 表示在技术 $b$ 上具有显性比较优势的主体数量, 即技术 $b$ 的流行度. 该指标通过刻画技术在主体层面的共现概率来度量其相似性: 若在主体集合中, 一个主体在技术 $b$ 上具有显性比较优势的条件下, 其同时在技术 $b'$ 上具有比较优势的概率较高, 则表明这两项技术在主体技术组合中具有更为紧密的关联[31]. 本文在州、县、城市与企业四个尺度上分别基于对应的 $G_{\text{Binary}}$ 构建技术空间网络, 反映了不同主体尺度视角下的技术关联结构, 用于后续跨尺度结构对比分析.

除基于共现关系的技术空间外, 本文进一步利用专利文本信息, 从技术内容相似性的角度构建另一类技术空间网络. 具体而言, 将每篇专利的标题与摘要拼接形成专利文本, 随后采用 Sentence-BERT 模型[32] 对文本进行向量化表示, 将每篇专利映射至一个 128 维向量空间, 记为 $\boldsymbol{v}_i = (v_1, v_2, \ldots, v_{128})$. 在此基础上, 技术 $b$ 的向量表示定义为其关联专利向量的加权平均:

$$\boldsymbol{v}_b = \frac{\sum_i s_{i,b}\,\boldsymbol{v}_i}{\sum_i s_{i,b}}, \tag{6}$$

其中, 权重 $s_{i,b}$ 即为 (1) 式计算得到的关联专利在该技术上的份额. 进一步通过计算技术向量之间的余弦相似度, 即可得到基于专利文本数据的技术空间网络. 与基于共现关系构建的技术空间相比, 基于专利文本数据得到的技术相似性不依赖于主体的技术发展模式, 而更侧重于技术内容本身在语义层面的相似性. 两类技术空间从不同视角刻画技术系统的结构特征, 为后续分析提供了互补的技术相似性度量.

由于完整的技术空间网络展示了所有技术节点对之间的相似性, 网络极为稠密, 不利于结构特征的识别. 为提取并研究技术空间的主干结构, 本文对技术空间网络进行过滤处理, 具体步骤如下[31]: 1) 基于连边权重提取网络的最大生成树; 2) 选取连边权重最高的前 20% 的连边. 通过叠加上述两部分连边, 得到过滤后的技术空间网络. 在后续结构分析中, 仅保留连边的拓扑关系而不考虑权重, 以突出原始技术空间中最重要的技术关联. 需要说明的是, 该过滤网络仅用于技术空间的结构特征分析, 其余分析环节仍采用完整的技术空间网络.

### 2.4 网络结构指标

#### 2.4.1 模块度

模块度是刻画网络模块化结构程度的常用指标, 用于衡量网络是否可以被划分为若干内部连接相对密集、模块之间连接相对稀疏的子结构. 本文采用 Barber 提出的二分网络模块度指标[33], 其形式为

$$Q = \frac{1}{|E|}\sum_{i=1}^{N_{\text{R}}}\sum_{\alpha=1}^{N_{\text{C}}}(\boldsymbol{M}_{i,\alpha} - P_{i,\alpha})\,\delta(c_i, c_\alpha), \tag{7}$$

式中, $|E|$ 表示网络中的连边总数; $\boldsymbol{M}_{i,\alpha}$ 为二分网络邻接矩阵中的元素; $N_{\text{R}}$ 与 $N_{\text{C}}$ 分别表示两类节点的数量; $P_{i,\alpha} = k_i k_\alpha/|E|$ 表示在保持节点度序列不变的随机二分网络中连边 $(i,\alpha)$ 出现的近似概率, 其中 $k_i$ 与 $k_\alpha$ 分别表示节点 $i$ 与节点 $\alpha$ 的度; $c_i$ 与 $c_\alpha$ 分别表示节点 $i$ 与节点 $\alpha$ 所属的模块; $\delta(c_i, c_\alpha)$ 为 Kronecker delta 函数, 仅当两个节点被划分到同一模块时取值为 1. 该指标通过累加同一模块内部实际连边数与随机期望连边数之间的差异, 对网络的模块化程度进行量化. $Q$ 值越大, 表明网络的模块结构相对于随机情形越为显著.

在一般 (单模) 网络中, 模块度定义为[33]

$$Q = \frac{1}{2|E|}\sum_{i,j}\left(\boldsymbol{M}_{i,j} - \frac{k_i k_j}{2|E|}\right)\delta(c_i, c_j), \tag{8}$$

其中, $\boldsymbol{M}_{i,j}$ 表示节点 $i$ 与 $j$ 之间的邻接关系, 其余符号含义与上述定义一致.

在实际分析中, 为获得网络模块化结构最为显著的节点划分结果, 本文采用 Louvain 算法[34] 对网络进行社区划分. 该算法通过迭代优化模块度函数, 能够在保证计算效率的同时识别出较高质量的模块结构, 因而被广泛应用于大规模网络的模块化分析.

#### 2.4.2 嵌套性

嵌套性是复杂网络中一种典型的有序结构特

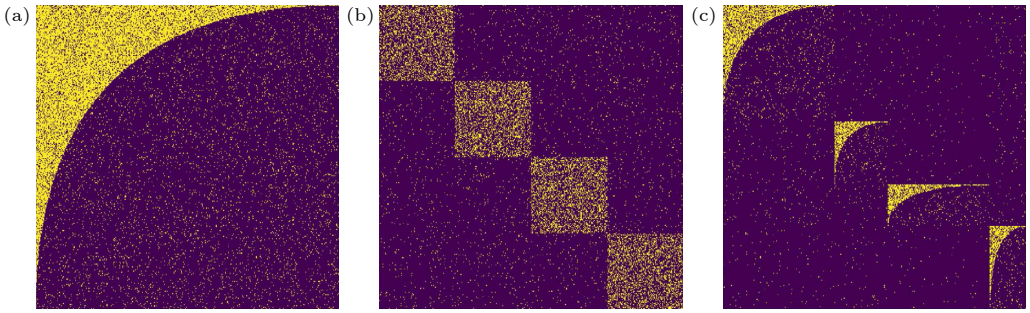



图 3 网络结构示意图 (a) 嵌套网络; (b) 模块化网络; (c) 块内嵌套网络

Fig. 3. Illustration of different network structures: (a) Nested network; (b) modular network; (c) in-block nested network.

征. 理想情况下, 完美嵌套网络满足如下条件: 对于网络中的任意两个节点 $i$ 和 $j$, 若节点 $i$ 的度小于节点 $j$ 的度, 则节点 $i$ 的邻居集合应完全包含于节点 $j$ 的邻居集合之中. 该定义既适用于一般网络, 也适用于二分网络, 但在二分网络情形下, 仅允许在同一节点类型内 (即行节点或列节点) 进行比较[15].

然而, 完美嵌套结构对网络连边模式的约束极为严格, 在真实系统中几乎不可能被严格满足. 因此, 现有研究通常关注网络结构在多大程度上接近嵌套状态, 即对嵌套性的"强弱"进行量化. 进一步的研究表明, 在同时具有模块化特征的网络中, 嵌套性往往并非全局属性, 而是作为一种局部结构特征主要出现在模块内部, 从而形成所谓的块内嵌套 (in-block nestedness) 结构. 如图 3 所示, 嵌套网络、模块化网络与块内嵌套网络在结构形态上存在显著差异.

本文采用 Solé-Ribalta 等[35] 提出的块内嵌套性量化指标. 该指标在模块划分给定的条件下, 通过比较同一模块内部节点对的实际邻居重叠程度与随机期望值, 对块内嵌套性进行量化. 网络的块内嵌套性指标定义为

$$\mathcal{I}=\frac{2}{N_{\mathrm{R}}+N_{\mathrm{C}}}\left\{\sum_{i,j}\frac{O_{i,j}-\langle O_{i,j}\rangle}{k_j\left(N_i-1\right)}\Theta(k_i-k_j)\delta(c_i,c_j)+\sum_{\alpha,\beta}\frac{O_{\alpha,\beta}-\langle O_{\alpha,\beta}\rangle}{k_\beta\left(N_\beta-1\right)}\Theta(k_\alpha-k_\beta)\delta(c_\alpha,c_\beta)\right\},\tag{9}$$

其中, $i,j$ 表示行节点; $\alpha,\beta$ 表示列节点; $O_{i,j}$ 表示节点对 $i,j$ 在同一模块内的共同邻居数量; $\langle O_{i,j}\rangle=k_ik_j/N_{\mathrm{C}}$ 表示在保持节点度的随机网络中节点对 $i,j$ 的期望共同邻居数量; $N_i$ 表示节点 $i$ 所在模块中与 $i$ 同类型的行节点数量; $\Theta(k_i-k_j)$ 为阶跃函数, 当 $k_i>k_j$ , 函数值为 1, 否则函数值为 0.

当将整个网络视为单一模块时, (9) 式退化为网络的全局嵌套性指标, 其表达式为

$$\mathcal{N}=\frac{2}{N_{\mathrm{R}}+N_{\mathrm{C}}}\left\{\sum_{i,j}\frac{O_{ij}-\langle O_{i,j}\rangle}{k_j(N_{\mathrm{R}}-1)}\Theta(k_i-k_j)+\sum_{\alpha,\beta}\frac{O_{\alpha\beta}-\langle O_{\alpha,\beta}\rangle}{k_\beta(N_{\mathrm{C}}-1)}\Theta(k_\alpha-k_\beta)\right\}.\tag{10}$$

通过比较网络的块内嵌套性指标 $\mathcal{I}$ 与全局嵌套性指标 $\mathcal{N}$, 可以区分嵌套性在网络中的表现形式: 当 $\mathcal{I}/\mathcal{N}\approx1$ 时, 嵌套性作为全局属性分布于整个网络, 对应于全局嵌套结构 (图 3(a)); 当 $\mathcal{I}/\mathcal{N}>1$ 时, 嵌套性主要作为局部属性出现在模块内部, 对应于块内嵌套结构 (图 3(c))[4,11,35].

在实际计算中, 为获得块内嵌套结构最为显著的节点划分, 本文采用极值优化算法[36], 通过最大化块内嵌套性指标 $\mathcal{I}$, 得到网络的最优模块划分结果.

## 3 结果与讨论

### 3.1 不同尺度下的网络结构差异

#### 3.1.1 主体-技术二分网络的差异

基于第 2.1 节介绍的 2000—2020 年的专利数据, 本文在州、县、城市与企业四个尺度上分别构建主体-技术二分网络, 所有网络均为经 RCA 指数阈值过滤后的不含权网络. 剔除孤立节点后, 网络的基本统计特征如表 3 所列. 随着主体尺度由宏观向微观细化, 主体节点数显著增加而网络密度持续下降, 表明在更细尺度下主体-技术连边相对更稀疏. 与此同时, 主体节点的平均度随尺度降低而减小, 而技术节点的平均度持续上升, 说明单个主体覆盖的技术范围在收缩, 但每项技术平均关联的主体数在增加.

表 3 主体-技术二分网络的统计数据

Table 3. Statistical properties of entity-technology bipartite networks.

| 网络 | 主体节点数 | 技术节点数 | 密度 | 主体平均度 | 技术平均度 |
|---|---|---|---|---|---|
| 州-技术 | 49 | 602 | 0.17 | 100.71 | 8.20 |
| 县-技术 | 764 | 602 | 0.03 | 17.73 | 22.50 |
| 城市-技术 | 1094 | 599 | 0.03 | 14.98 | 27.36 |
| 企业-技术 | 3866 | 601 | 0.01 | 7.08 | 45.51 |

为检验不同尺度下网络结构特征是否显著偏离随机基准, 本文采用二分网络配置模型 (bipartite configuration model, BiCM)[37,38] 作为零模型. 该模型基于最大熵原则构造随机二分网络的概率集合, 仅要求主体与技术节点的期望度序列与真实网络一致, 从而在零模型层面控制节点度异质性所带来的规模效应. 在该约束下, 每一条主体-技术潜在连边被建模为相互独立的伯努利随机变量, 其出现概率由两侧节点的度约束共同决定, 并由此得到连边概率矩阵作为随机基准①.

基于该连边概率矩阵, 本文进一步通过独立伯努利抽样生成 BiCM 零模型下的随机二分网络[38]: 对每一对主体-技术节点对, 按照其对应的连边概率进行一次随机试验 (即生成一个均匀分布于 $[0,1]$ 的随机数, 若不超过该概率, 则该节点对在随机网络中生成一条连边; 否则该节点对在该次采样中不连边), 遍历所有节点对即可得到一个随机二分网络. 重复上述过程可从零模型所定义的分布中获得随机二分网络集合. 需要指出的是, 由于采用期望度约束, 单个随机二分网络的度序列可能与真实度序列存在随机波动, 但在均值意义下与真实网络的度约束一致. 基于该随机二分网络集合, 本文计算模块度、全局嵌套度和块内嵌套度等结构指标在零模型下的均值与波动, 并据此开展显著性检验.

不同尺度主体-技术网络的结构特征如图 4 所示. 结果表明, 随着主体尺度降低, 二分网络的模块度升高, 网络的模块化结构不断增强. 在州与县尺度下, 嵌套度比值接近 1, 网络整体呈现出以全局嵌套为主的结构特征; 而在城市与企业尺度下, 嵌套度比值大于 1, 表明嵌套性不再以全局形式存在, 而主要表现在模块内部, 呈现出块内嵌套结构. 相比之下, 四个尺度下 BiCM 零模型生成的随机网络嵌套度比值均接近 1, 未呈现块内嵌套特征. 这表明, 存在超出度约束的额外组织机制导致了真实网络中的跨尺度结构差异.

为定量评估上述结构特征的显著性, 本文计算真实网络相对于 BiCM 随机网络的 $z$ 分数. 以模块度为例, 其定义为

$$Z_Q = \frac{Q - \mu_{\mathrm{BiCM}}}{\sigma_{\mathrm{BiCM}}}, \tag{11}$$

其中, $\mu_{\mathrm{BiCM}}$ 与 $\sigma_{\mathrm{BiCM}}$ 分别表示随机网络中对应指标的均值与标准差. 各结构指标对应的 $z$ 分数结果列于表 4.

表 4 主体-技术二分网络结构指标相对于 BiCM零模型的 $z$ 分数

Table 4. $Z$-scores of structural metrics of entity-technology bipartite networks relative to the BiCM null model.

| 网络 | 模块度 ($Z_Q$) | 全局嵌套度 ($Z_N$) | 块内嵌套度 ($Z_I$) | 嵌套度比值 ($Z_{I/N}$) |
|---|---|---|---|---|
| 州-技术 | 15.72 | –1.61 | 2.34 | 2.94 |
| 县-技术 | 53.42 | 4.20 | 4.00 | 3.25 |
| 城市-技术 | 87.17 | 7.04 | 6.53 | 4.90 |
| 企业-技术 | 196.29 | 10.75 | 46.11 | 38.69 |

结果表明, 不同尺度下主体-技术二分网络的多项结构指标均显著偏离 BiCM 零模型的随机预期: 州与县尺度网络已表现出显著的模块化, 但其嵌套性主要以全局形式存在, 块内嵌套特征相对有限; 而在城市与企业尺度, 网络模块度进一步增强, 同时块内嵌套度及嵌套度比值的 $z$ 分数显著增大, 表明嵌套结构主要集中于模块内部, 且强度远高于随机网络的预期水平.

总体而言, 主体尺度的变化伴随着主体-技术网络结构的系统性转变. 随着主体尺度的缩小, 网络在保持模块化的同时, 嵌套性逐渐由全局嵌套转变为块内嵌套. 这种随尺度变化出现的结构差异在不同统计指标下均保持稳定, 为后续分析尺度约束对技术发展过程的影响提供了可靠的经验依据. 此外, 补充材料 A (online) 给出了不同 RCA 筛选阈值下的对照结果, 表明上述跨尺度结构差异对阈值选择具有稳健性.

从统计物理视角看, 企业、城市、县与州四种主体尺度可理解为对同一技术创新过程施加不同粗粒化程度的观测. 粗粒化不仅改变节点含义与网

① 本文使用相关 Python 包求解 BiCM 模型参数并生成连边概率矩阵: https://github.com/mat701/BiCM.

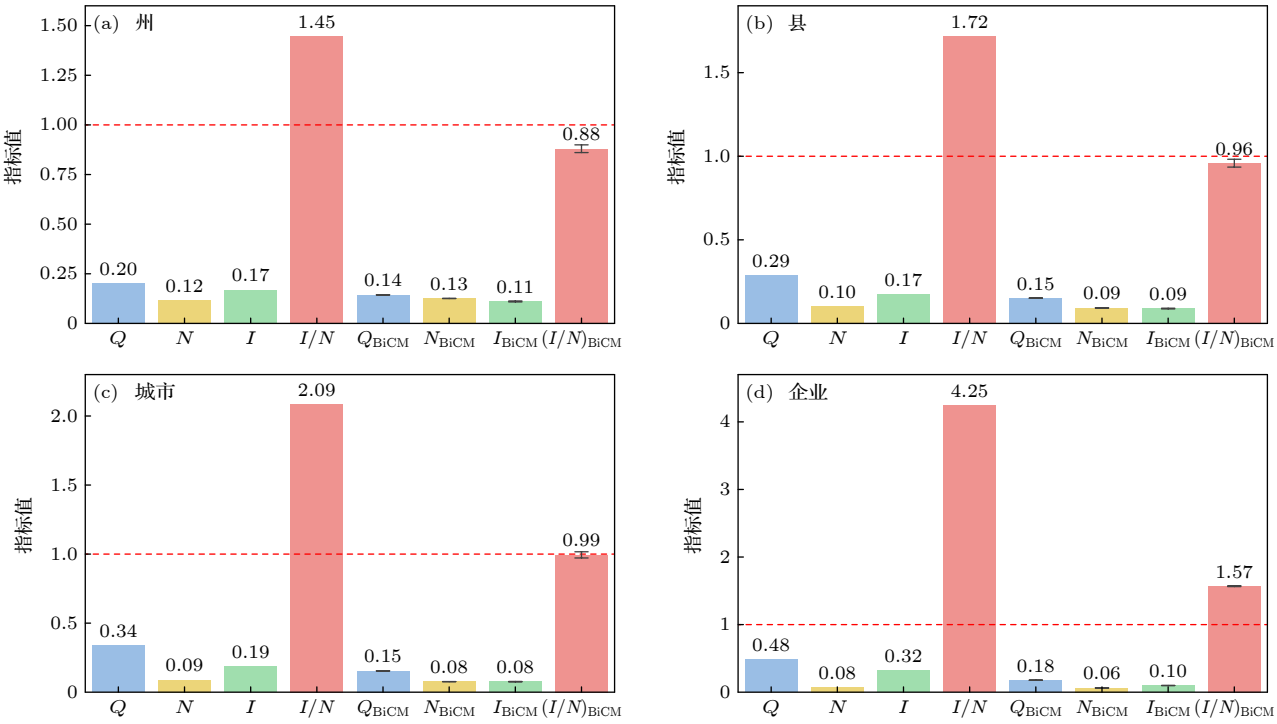



图 4 不同尺度主体-技术二分网络的结构特征, 包括模块度 ($Q$)、全局嵌套度 ($N$)、块内嵌套度 ($I$) 及嵌套度比值 ($I/N$). 其中, $Q$ 和 $I$ 取自 100 次优化过程中的最优值. 为检验结构特征的显著性, 对每个网络基于 BiCM 模型生成 100 个随机网络 (其结构特征以下标 BiCM 表示). 误差条表示随机网络结构指标的均值标准误

Fig. 4. Structural properties of entity-technology bipartite networks across different scales, including modularity ($Q$), global nestedness ($N$), in-block nestedness ($I$), and the nestedness ratio ($I/N$). The values of $Q$ and $I$ correspond to the optimal results obtained from 100 optimization runs. To assess the statistical significance of the structural properties, 100 randomized networks are generated for each empirical network based on the bipartite configuration model (BiCM), with the corresponding metrics denoted by the subscript BiCM. Error bars represent the standard error of the mean of structural metrics computed from the randomized networks.

络规模, 更会改变微观差异在聚合后的呈现方式, 从而影响网络结构特征的可见性与相对强度. 由于本文在四种尺度下采用统一的技术定义与建网流程, 跨尺度差异主要反映尺度变换所对应的统计模式变化, 而非由权重处理或筛选规则不一致造成的表面差异. 在宏观尺度上, 多个微观主体聚合为更高层级后, 宏观主体的技术组合呈现出更强的自平均效应: 微观主体在技术组合上的随机波动被平均化, 使得宏观主体对通用核心技术的覆盖更加稳定, 主体间在这些技术上的重叠程度普遍增强, 邻居集合更容易形成稳定的包含关系, 从而表现为典型的全局嵌套结构. 相反, 在微观尺度下, 主体受资源与能力边界约束更强, 可覆盖的技术更有限且更专一化, 网络更稀疏并呈现更强异质性. 主体间的技术重叠更可能局限在少数相近的技术簇内部, 而跨簇共享较弱, 因此更容易形成相对独立的子结构与清晰的分区边界, 表现为模块化增强. 与此同时, 每个分区内部仍可能存在由通用到专用的层级分化: 少数较通用技术被分区内更多主体共同占据, 而专用技术仅由少数主体掌握, 使嵌套性主要以模块内局部层级的形式保留并体现为块内嵌套. 需要强调的是, 在各尺度下采用零模型控制两侧节点的期望度序列后, 生成的随机网络并未呈现跨尺度结构差异. 这表明, 真实网络中的跨尺度结构差异不能简单归因于度序列变化, 而指向存在超出度约束的额外组织机制, 使网络在尺度变换下发生系统性的结构重组.

### 3.1.2 技术空间网络的差异

基于第 3.1.1 节构建的主体-技术二分网络, 结合 (5) 式, 本文进一步在不同主体尺度下构建了基于技术共现关系的技术空间网络, 用以刻画不同尺度主体视角下的技术相似性结构. 此外, 本文还构建了基于专利文本相似性的技术空间网络, 作为反映技术内容关联的对照.

对于过滤后的技术空间网络, 其结构指标汇总

于表 5. 在基于共现关系构建的技术空间中, 为检验模块化特征的显著性, 本文采用保持节点度序列不变的置换零模型作为对照. 通过反复随机选取两条连边并交换其端点, 在不引入自环和重边的前提下生成随机网络, 并对每个真实网络生成 100 个随机实例, 从而计算模块度的 $z$ 分数.

表 5 不同关系来源下技术空间网络的结构指标

Table 5. The structural metrics of technology space networks constructed from different relation sources.

| 关系来源 | 技术节点数 | 模块度 ($Q$) | 随机网络模块度 ($\bar{Q}_{\text{Swap}}$) | $Z_Q$ | AMI |
|---|---|---|---|---|---|
| 州-技术 | 602 | 0.26 | 0.06 | 125.84 | 0.06 |
| 县-技术 | 602 | 0.32 | 0.07 | 138.23 | 0.10 |
| 城市-技术 | 599 | 0.33 | 0.07 | 147.47 | 0.16 |
| 企业-技术 | 600 | 0.36 | 0.08 | 159.59 | 0.19 |
| 专利文本 | 624 | 0.31 | 0.05 | 204.18 | 0.35 |

为进一步比较不同技术空间中模块划分结果的差异, 本文以 CPC 分类体系的部层级划分作为基准, 将技术按照专利分类代码的首字母 (A—H) 划分为 8 个模块. 同时, 引入基于专利文本构建的技术空间作为内容相似性视角下的对照网络. 为了衡量不同技术空间的模块划分结果与 CPC 部层级划分之间的一致性, 采用调整互信息 (adjusted mutual information, AMI) 进行量化 [39]. AMI 的取值范围为 $[0,1]$, 较大的 AMI 值表示同一模块内的技术更可能来自相同的 CPC 部, 反之则说明模块划分与 CPC 分类之间的对应关系较弱.

剔除孤立节点后, 技术空间网络的结构指标如表 5 所列, 所有技术空间网络均表现出显著的模块化特征, 其模块度显著高于对应零模型的随机预期. 对于基于技术共现关系构建的技术空间, 随着主体尺度由州逐步细化至企业层面, 网络模块度呈现出小幅上升趋势, 表明技术空间的模块结构在微观尺度下更加清晰. 进一步与 CPC 部层级划分对比可以发现, 主体尺度越小, 技术空间模块划分与 CPC 分类之间的一致性越高, 说明在微观主体尺度下, 技术之间的关联主要由技术内容本身决定; 而在宏观主体尺度下, 技术空间中的模块结构可能同时受到技术内容以外因素的影响.

相比之下, 基于专利文本构建的技术空间与 CPC 部层级划分之间表现出更高的一致性, 其 AMI 值明显高于基于共现关系构建的技术空间网络. 这一结果表明, 专利文本所刻画的技术相似性能够较为直接地反映技术之间的内容关联, 为理解技术空间的模块结构提供了一个合理的对照基准.

### 3.2 基于实证数据的网络演化规律分析

为深入理解不同尺度下网络结构差异的形成机制, 本文基于实证数据刻画了主体在技术发展过程中的选择偏好. 具体而言, 按专利申请年份将数据集划分为四个连续且不重叠的五年时间窗口: 2000—2004 ($T_1$), 2005—2009 ($T_2$), 2010—2014 ($T_3$) 和 2015—2019 ($T_4$), 并在每个窗口内分别构建主体-技术二分网络与技术空间网络. 选择五年作为窗口长度, 是在统计稳定性与动态刻画能力之间的折中: 窗口过短容易受到单一年份专利产出与分类等短期波动的影响, 导致网络稀疏、噪声增大; 窗口过长则可能将较长时期内发生的多次技术发展事件聚合为一次观测, 从而削弱模型对技术发展动态过程的刻画能力. 对于前三个时间窗口 ($T_1$—$T_3$), 本文通过比较相邻时间窗口的主体-技术二分网络, 将在 $T_i$ 中尚未出现而在 $T_{i+1}$ 中出现的主体-技术连边定义为主体新发展的技术; 将 $T_i$ 和 $T_{i+1}$ 中均未出现的连边定义为主体未发展的技术. 为刻画主体发展新技术时的偏好, 本文进一步引入两个技术层面的指标——相关密度与接近中心性. 相关密度衡量潜在技术与主体已有技术组合之间的相关程度, 用以反映主体对技术相似性的依赖; 接近中心性刻画技术在整体技术空间中的位置, 用以反映主体对技术全局结构位置的偏好.

技术 $b$ 对于主体 $a$ 的相关密度定义为

$$\omega_{a,b}=\frac{\sum_{b'}\boldsymbol{M}_{a,b'}\phi_{b,b'}}{\sum_{b'}\phi_{b,b'}}, \tag{12}$$

其中相关密度较高的技术与主体已有技术组合具有更强的相似性.

技术 $b$ 在技术空间中的接近中心性计算为

$$CC_b=\frac{m-1}{\sum_{b'}d(b,b')}, \tag{13}$$

其中, $m$ 表示技术节点的总数, $d(b,b')$ 表示两技术节点之间的最短路径距离. 在完整的含权技术空间网络中, 若两技术节点之间的连边权重为正, 则其直接连边长度定义为权重的倒数, 即 $1/\phi_{b,b'}$; 若连边权重为零, 则认为两者之间不存在直接连边. 需要指出的是, 考虑到基于共现关系构建的技术空间

在不同主体尺度下可能受到行为模式差异的影响, 本文统一采用基于专利文本构建的技术空间来计算相关密度与接近中心性指标, 以保证不同尺度之间的可比性.

为检验相关密度、接近中心性与主体是否发展新技术之间的统计关系, 本文构建逻辑回归模型, 并以主体-技术对 $(a, b)$ 作为样本: 对所有在 $T_i$ 中尚未出现的主体-技术连边, 若其在 $T_{i+1}$ 中出现则记为 $y_{a,b}=1$, 否则记为 $y_{a,b}=0$, 将相关密度与接近中心性作为解释变量进行回归. 回归结果如表 6 所列. 表中报告的伪 $R^2$ (McFadden's pseudo $R^2$) 用于衡量模型相对于仅含截距项的零模型的拟合改进程度. 本文的估计结果显示, 该指标在各尺度与窗口设定下保持在 0.07—0.11 的量级, 表明模型相对于零模型存在稳定的拟合改进. 需要指出的是, 技术发展可能受到主体能力、制度环境等多种难以完全观测的因素共同影响, 本节回归分析的主要目的在于识别与检验技术发展偏好的关键驱动因素, 并为后续生成模型提供经验约束, 而非对单个技术发展事件进行精确预测; 因此, 相比伪 $R^2$ 的绝对大小, 本节更关注的是回归系数在不同尺度与时间窗口下是否呈现一致的符号方向, 并在统计意义上保持显著. 结果显示, 所有解释变量的回归系数均达到 0.1% 的显著性水平, 且在各主体尺度与时间窗口中, 相关密度与接近中心性均与主体是否发展该技术保持正向关联. 该结果从统计意义上支持这两类指标能够有效刻画主体在技术发展过程中所遵循的经验规律.

表 6 不同主体尺度与五年时间窗口下新技术发展概率的逻辑回归结果

Table 6. Logistic regression results for the probability of newly developed technologies across different entity scales and five-year time windows.

| 主体尺度 | 变量 | 2000—2004 | 2005—2009 | 2010—2014 |
|---|---|---|---|---|
| 州 | 相关密度 | 7.33 | 6.80 | 6.24 |
| | 接近中心性 | 6.71 | 6.64 | 7.83 |
| | 截距 | –7.04 | –6.79 | –7.42 |
| | 伪 $R^2$ | 0.09 | 0.08 | 0.07 |
| 县 | 相关密度 | 13.52 | 12.40 | 10.59 |
| | 接近中心性 | 7.23 | 7.12 | 6.86 |
| | 截距 | –8.25 | –8.04 | –7.79 |
| | 伪 $R^2$ | 0.11 | 0.10 | 0.08 |
| 城市 | 相关密度 | 16.37 | 14.95 | 12.92 |
| | 接近中心性 | 6.39 | 6.26 | 6.16 |
| | 截距 | –8.11 | –7.92 | –7.77 |
| | 伪 $R^2$ | 0.09 | 0.08 | 0.07 |
| 企业 | 相关密度 | 18.33 | 17.07 | 15.55 |
| | 接近中心性 | 6.12 | 5.40 | 5.21 |
| | 截距 | –8.35 | –7.94 | –7.90 |
| | 伪 $R^2$ | 0.08 | 0.08 | 0.07 |

由于逻辑回归的原始回归系数并不是对技术发展概率变化幅度的直接度量, 其数值大小除受解释变量本身作用影响外, 还会受到不同样本组中未观测因素离散程度的影响[40]. 因此, 即使模型形式完全一致, 不同主体尺度对应的样本组之间原始回归系数的大小也不宜直接用于比较解释变量作用的强弱. 为比较相关密度与接近中心性在不同尺度主体发展新技术过程中的相对重要性, 本文首先在全样本范围内对二者进行统一标准化处理, 使其单位变化在不同尺度下具有一致含义; 随后, 在各尺度样本组中分别计算两个解释变量的平均边际效应 (average marginal effect, AME). AME 刻画了解释变量增加一个单位时主体发展新技术概率的平均变化. 由于其定义在概率尺度上, 且基于统一标准化后的变量计算, 因此更适合用于不同尺度样本组之间的效应比较. 进一步地, 本文通过比较各尺度样本组中相关密度与接近中心性的 AME 比值, 衡量二者在主体发展新技术过程中的相对作用强度.

结果如表 7 所列, 随着主体尺度由宏观向微观细化, 相关密度与接近中心性的 AME 比值单调上

表 7 不同主体尺度与五年时间窗口下相关密度和接近中心性的平均边际效应 (AME) 及其比值

Table 7. Average marginal effects (AME) of relatedness density and closeness centrality, and their ratios, across different entity scales and five-year time windows.

| 主体尺度 | 指标 | 2000—2004 | 2005—2009 | 2010—2014 |
|---|---|---|---|---|
| 州 | 相关密度 | 0.0105 | 0.0105 | 0.0090 |
| | 接近中心性 | 0.0171 | 0.0171 | 0.0182 |
| | AME比值 | 0.6152 | 0.6150 | 0.4925 |
| 县 | 相关密度 | 0.0053 | 0.0051 | 0.0051 |
| | 接近中心性 | 0.0050 | 0.0049 | 0.0054 |
| | AME比值 | 1.0535 | 1.0456 | 0.9533 |
| 城市 | 相关密度 | 0.0045 | 0.0045 | 0.0044 |
| | 接近中心性 | 0.0031 | 0.0031 | 0.0034 |
| | AME比值 | 1.4422 | 1.4335 | 1.2960 |
| 企业 | 相关密度 | 0.0031 | 0.0031 | 0.0029 |
| | 接近中心性 | 0.0019 | 0.0017 | 0.0016 |
| | AME比值 | 1.6869 | 1.8969 | 1.8459 |

升, 表明相对于接近中心性, 相关密度的作用强度逐渐增强. 这一变化揭示了不同尺度的主体在技术发展模式上的不同: 宏观尺度的主体在技术发展过程中更倾向于选择位于技术空间核心位置的技术, 从而体现出对整体技术布局的关注; 而微观尺度的主体在发展新技术时, 则更依赖于潜在技术与其已有技术组合在技术内容上的相关性, 表现出更强的局部一致性取向[4,41]. 这一行为层面的尺度差异与第 3.1.2 节的结构发现相一致: 在城市与企业尺度下, 技术空间表现出更高的模块化水平, 且模块划分与 CPC 分类的对应关系更为接近. 为检验结论的稳健性, 在补充材料 B (online) 中补充了三年与七年窗口的回归结果, 结论保持一致.

### 3.3 基于实证规律的生成模型构建与验证

基于第 3.2 节的实证结果, 本文提出一种创新主体-技术二分网络的生成模型, 用以刻画主体在技术空间中逐步扩展技术组合的过程. 模型的核心思想是: 主体对潜在技术的选择同时受到两类因素的影响, 即 (i) 潜在技术与主体已有技术组合的相关性 (相关密度), 以及 (ii) 潜在技术在整体技术空间中的全局位置 (接近中心性). 据此, 主体发展技术的过程可表示为在技术空间网络上逐步“激活”技术节点的过程[42].

模型采用基于 2000—2020 年的专利文本构建的技术空间网络作为外部环境. 在该技术空间上, 技术节点的接近中心性 $CC_b$ 为固定值; 主体 $a$ 对候选技术 $b$ 的相关密度 $\omega_{a,b}$ 则随主体已激活技术集合的变化而动态更新.

针对每一种主体尺度 (州、县、城市、企业), 模型使用对应尺度下真实主体的数量以及主体节点度序列作为约束条件, 并执行如下增长过程:

初始化: 对每个主体 $a$, 在技术空间中随机选择一个技术节点作为起点并激活. 迭代增长: 依次遍历主体. 对主体 $a$ 而言, 在其尚未激活的技术集合中选择一项技术 $b$ 进行激活, 其选择概率定义为

$$P(a,b) \propto CC_b^{x} \cdot \omega_{a,b}^{y}, \tag{14}$$

其中, 参数 $x$ 刻画主体对技术全局位置 (接近中心性 $CC_b$) 的偏好强度, 参数 $y$ 刻画主体对局部相关性 (相关密度 $\omega_{a,b}$) 的偏好强度. 每次迭代仅激活一个新技术节点, 并据此更新主体的技术组合与相关密度 $\omega_{a,b}$. 停止条件: 当主体 $a$ 激活的技术节点数达到其在真实主体-技术二分网络中的节点度时, 停止对该主体的增长; 所有主体完成增长后, 依据各主体的激活集合生成最终的主体-技术二分网络.

为系统考察两类偏好机制对生成结构的影响, 本文在 $(x,y)$ 参数平面 ($x,y \in [0,10]$) 上进行扫描, 分别对四种主体尺度生成模拟网络. 图 5—图 7 分别展示了模拟网络的模块度 $Q$、全局嵌套度 $N$ 以及嵌套度比值 $I/N$ 随 $(x,y)$ 变化的相图结果.

首先, 图 5 模块度相图表明, $Q$ 在四种尺度下均由 $y$ 主导, 随 $y$ 增大而上升, 而随 $x$ 增大而下降. 当主体更依赖相关密度 (较大 $y$) 时, 技术发展更倾向于高相关度的局部邻域, 从而更容易形成清晰的模块结构; 相反, 当主体更强烈偏好技术空间的核心位置 (较大 $x$) 时, 不同主体更容易在少数核心技术上产生重叠, 跨模块的“桥接”连边增多, 从而削弱了模块边界并降低模块度.

其次, 图 6 全局嵌套度相图显示, $N$ 在四种尺度下均由 $x$ 主导, 并随 $x$ 增大而升高. 当选择概率更偏向技术空间中的高中心性节点时, 不同主体的技术组合更容易在少数核心技术上发生集中重合, 从而增强主体之间技术组合的包含关系, 增强全局层面的嵌套结构.

最后, 图 7 嵌套度比值 $I/N$ 的相图显示, 随着 $y$ 增大, $I/N$ 在四种尺度下整体上升, 说明增强对相关密度的依赖会显著强化块内嵌套特征. 这一点与 3.1.1 节中“微观尺度下网络呈现显著块内嵌套结构”的结构对比结果, 以及 3.2 节中基于平均边际效应比较得到的“微观尺度对相关密度更依赖”的结论相一致.

综上, 在统一技术空间与真实度序列约束下, 模型相图揭示了两类偏好参数与结构指标之间的清晰对应关系: 增强相关密度偏好 ($y$) 会提高模块度并总体提升块内嵌套程度; 增强中心性偏好 ($x$) 会显著提高全局嵌套度 $N$. 结合前文的结构对比与实证结果, 上述模拟结果在定性层面与实证发现保持一致, 表明相关密度偏好和中心性偏好两种偏好机制能够为不同尺度下主体-技术网络结构由全局嵌套向块内嵌套转变提供合理的解释支撑.

从统计物理视角看, 本节模拟结果的意义在于揭示了跨尺度网络结构的涌现机制: 模型并未对嵌套性或模块度等结构指标施加任何显式目标或优化约束, 而是在仅引入两类技术发展偏好 (技术中心性偏好与技术相关密度偏好) 并保持经验约束的

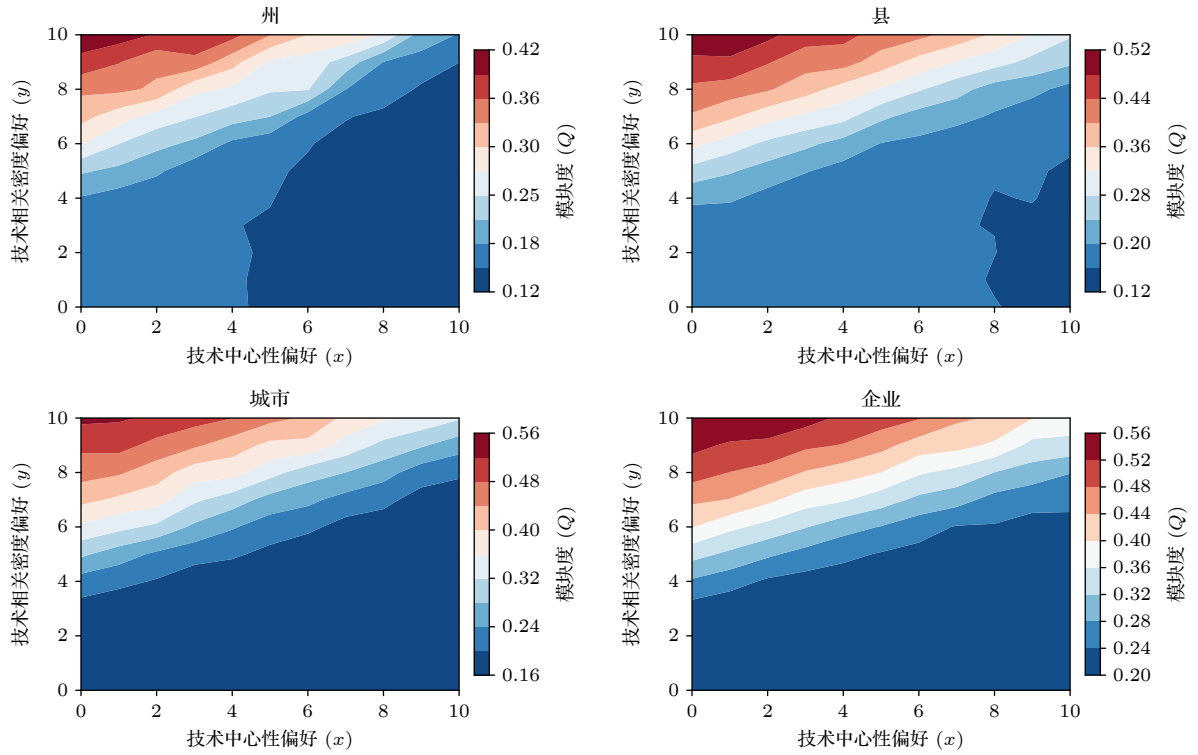



图 5 不同技术中心性偏好 $x$ 与相关密度偏好 $y$ 组合下主体-技术二分网络模块度 $Q$ 的模拟结果

Fig. 5. Simulation results of modularity $Q$ in entity-technology bipartite networks under different combinations of closeness centrality preference $x$ and relatedness-density preference $y$.

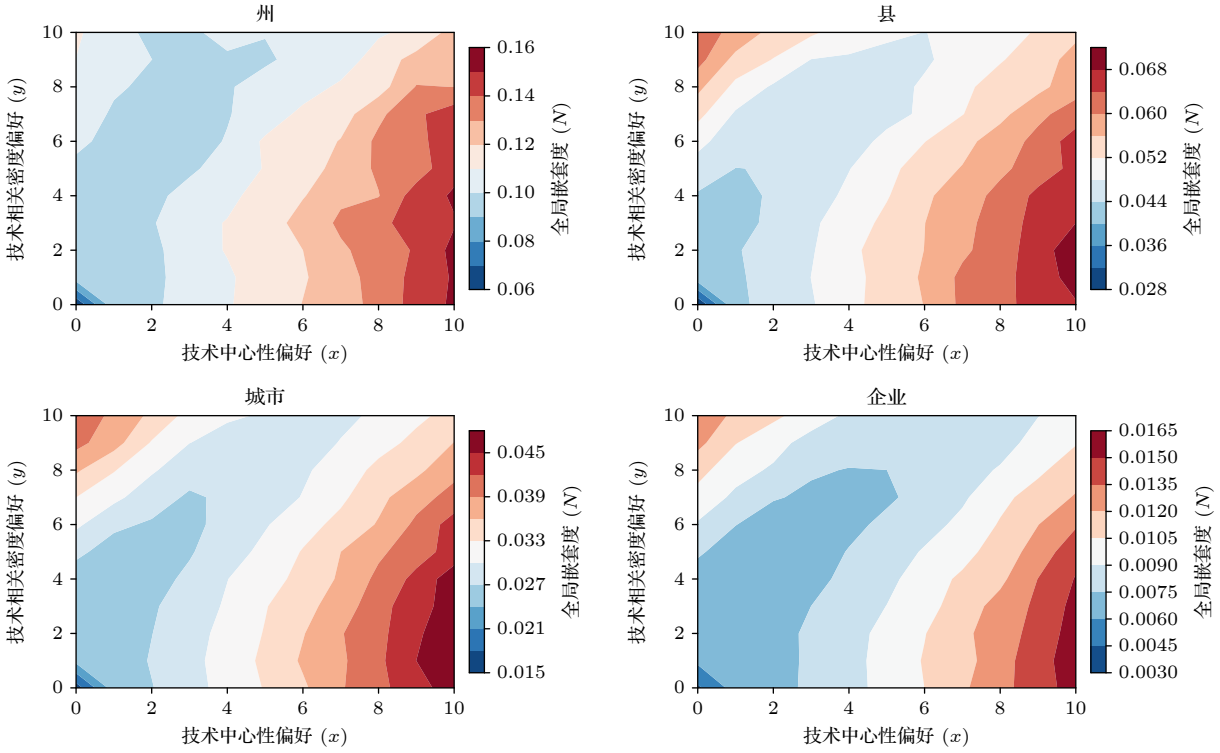



图 6 不同技术中心性偏好 $x$ 与相关密度偏好 $y$ 组合下主体-技术二分网络全局嵌套度 $N$ 的模拟结果

Fig. 6. Simulation results of global nestedness $N$ in entity-technology bipartite networks under different combinations of closeness centrality preference $x$ and relatedness-density preference $y$.

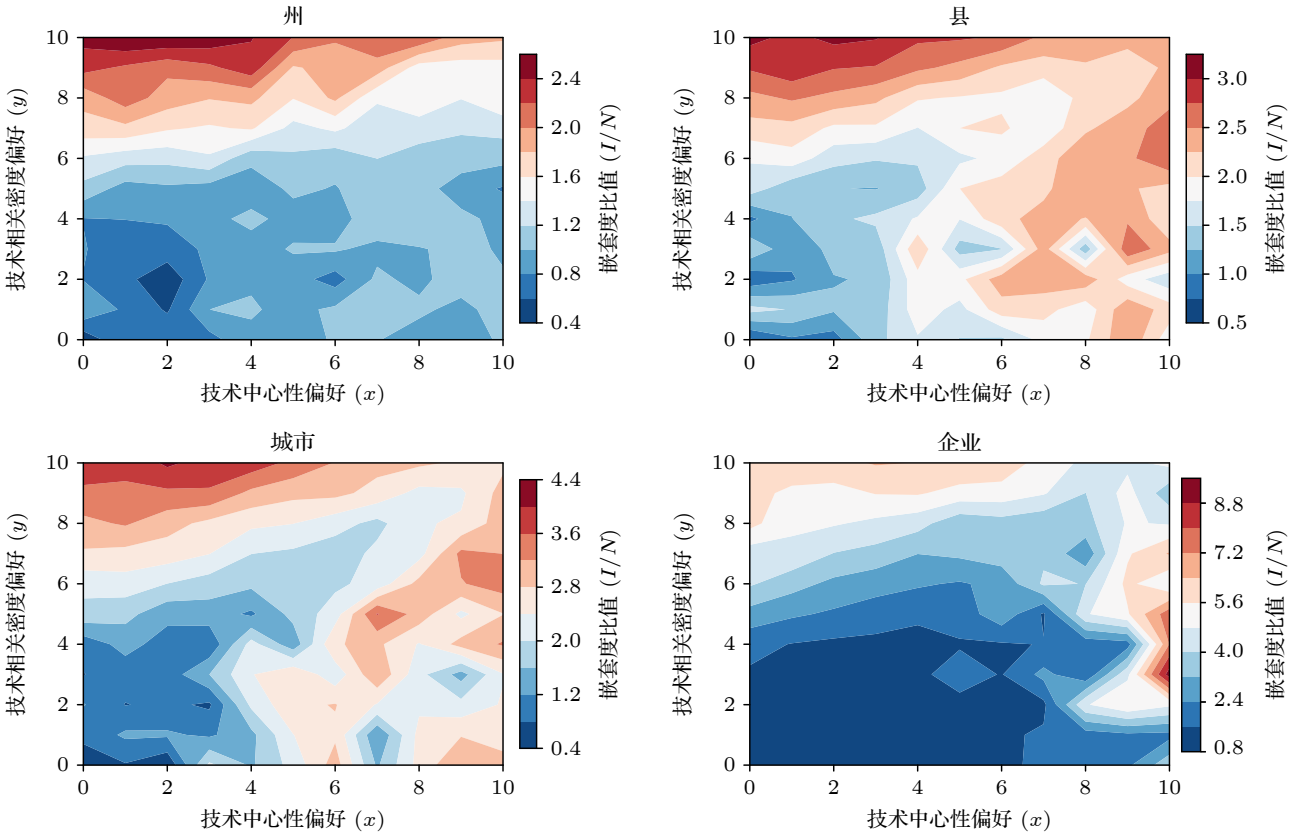



图 7 不同技术中心性偏好 $x$ 与相关密度偏好 $y$ 组合下主体-技术二分网络嵌套度比值 $I/N$ 的模拟结果

Fig. 7. Simulation results of the nestedness ratio $I/N$ in entity-technology bipartite networks under different combinations of closeness centrality preference $x$ and relatedness-density preference $y$.

前提下, 即可在参数平面上自发产生稳定的结构分区. 具体而言, 当技术中心性偏好占优时, 主体更倾向于发展通用核心技术, 网络呈现全局嵌套结构; 当技术相关密度偏好增强时, 主体的技术发展更受局部相关性约束, 跨域连接受到抑制, 模块边界随之强化, 嵌套性主要以模块内层级有序的形式保留 (即块内嵌套). 因此, 本节结果提供了一个利用微观规则调控宏观形态的映射: 少量具有明确行为含义的偏好参数即可刻画并复现不同尺度下的典型网络组织形态, 从而为实证发现的跨尺度结构转变提供了自组织的物理解释.

## 4 结 论

本文面向技术创新系统的跨尺度结构分析与机制检验, 构建并公开了一套可复用的美国专利数据集. 数据集在统一框架下整合了专利的分类信息、创新主体信息与文本信息, 并针对涉及多项技术的专利引入基于引用信息的技术份额分配方案, 以更合理地量化技术活动强度. 在保证数据完整性与可复用性的前提下, 该数据集为跨尺度刻画主体技术行为、比较网络结构差异及开展机制检验提供了稳健的数据支撑.

基于该数据集, 本文在州、县、城市与企业四个主体尺度上分别构建主体-技术二分网络及其对应的技术空间网络, 并在节点度约束的零模型基准下进行显著性对比. 分析结果显示: 各尺度主体-技术网络均显著偏离随机结构并呈现清晰的模块化组织, 同时模块化程度随主体尺度细化而系统性增强; 在嵌套结构方面, 州与县尺度网络更接近全局嵌套形态, 而城市与企业尺度网络则更突出块内嵌套特征. 与此一致, 技术空间网络在各尺度下均表现出显著的模块结构, 但在更微观的主体视角下, 模块内部技术在内容维度上的一致性更强, 表明尺度差异不仅改变主体技术组合的多样化水平, 也会影响技术关联的组织方式与模块边界的清晰度.

在结构比较的基础上, 本文进一步从动态视角刻画主体技术组合的演化过程. 通过时间窗口划分识别新技术发展行为, 并采用逻辑回归模型定量分

析新技术发展概率与技术相关密度及技术中心性之间的关系. 结果表明, 随着主体尺度由宏观向微观细化, 相关密度相对于接近中心性的平均边际效应 (AME) 比值逐渐上升, 说明相关密度相对于接近中心性的作用逐渐增强, 新技术发展更依赖于潜在技术与已有技术组合的局部相关性, 而对技术在整体技术空间中核心位置的依赖相对减弱. 这一尺度依赖的选择模式为模块化增强与块内嵌套凸显提供了行为层面的可检验证据, 并表明微观主体更倾向于沿局部技术邻域渐进扩展, 而宏观主体则更容易覆盖更广泛且包含核心技术的技术组合.

基于上述经验规律, 本文提出一个主体-技术网络的生成模型, 在经验度约束下引入“技术相关密度偏好”和“技术中心性偏好”两类选择机制, 用以描述主体在技术发展过程中的决策倾向. 模拟实验表明, 通过调节两类偏好参数, 模型能够在统一框架下复现从全局嵌套到块内嵌套的结构转变, 从而将跨尺度结构差异理解为统一演化规则在不同主体约束强度与选择偏好条件下的结果, 并为跨尺度技术系统的结构形成提供了可检验的机制解释.

总体而言, 本文公开的数据集与配套的构建及分析流程, 为从复杂系统视角研究技术创新提供了可复现、可比较的统一基准. 其贡献体现在两个层面: 在实证层面, 支持对不同尺度下的“主体-技术”网络进行一致性刻画与比较, 从而揭示网络结构随尺度演化的稳健规律; 在方法论层面, 流程化的数据清洗、整合与技术份额度量提升了研究的可重复性与可检验性, 为跨尺度对比研究与机制验证奠定了基础. 基于上述工作, 未来研究包括以下几个方面: 在理论建模上, 可在现有二分网络与技术空间网络的基础上, 引入超图、多层网络等更丰富的表示形式, 以刻画多技术协同、多主体互动等复杂关联; 在预测分析上, 可将网络结构特征与机器学习方法相结合, 开展技术发展路径预测与融合趋势识别, 推动研究从“结构描述”向“趋势预测”延伸; 在政策与风险评估上, 可基于多尺度网络开展机制模拟与冲击推演, 评估关键主体或关键技术节点受到扰动时对区域创新系统韧性、恢复力的影响, 为区域创新政策与风险治理提供量化依据. 综上, 本研究提供的数据资源与分析框架, 不仅有助于系统揭示技术创新复杂系统的跨尺度演化机制, 也为连接理论探索、趋势研判与政策设计提供了重要的数据基础与方法支撑.

## 数据可用性声明

支撑本研究成果的数据集可在科学数据银行 https://doi.org/10.57760/sciencedb.j00213.00265 访问获取.

## 参考文献


[1] Wang L L, Yang Y P, Duan Y W 2021 *Ecol. Evol.* **11** 17509
[2] Galiana N, Hawkins B A, Montoya J M 2019 *Ecography* **42** 1175
[3] Vizentin-Bugoni J, Tarwater C E, Foster J T, Drake D R, Gleditsch J M, Kelley J P, Sperry J H 2019 *Science* **364** 78
[4] Laudati D, Mariani M S, Pietronero L, Zaccaria A 2023 *J. Phys. Complex.* **4** 025011
[5] Newman M E 2006 *Proc. Natl. Acad. Sci. U.S.A.* **103** 8577
[6] Fortunato S, Hric D 2016 *Phys. Rep.* **659** 1
[7] Staniczenko P P, Kopp J C, Allesina S 2013 *Nat. Commun.* **4** 1391
[8] Mariani M S, Mazzilli D, Patelli A, Sels D, Morone F 2024 *Commun. Phys.* **7** 102
[9] Payrató-Borras C, Hernández L, Moreno Y 2019 *Phys. Rev. X* **9** 031024
[10] Borge-Holthoefer J, Baños R A, Gracia-Lázaro C, Moreno Y 2017 *Sci. Rep.* **7** 41673
[11] Palazzi M J, Solé-Ribalta A, Calleja-Solanas V, Meloni S, Plata C A, Suweis S, Borge-Holthoefer J 2021 *Nat. Commun.* **12** 1941
[12] Alves L G, Mangioni G, Cingolani I, Rodrigues F A, Panzarasa P, Moreno Y 2019 *Sci. Rep.* **9** 2866
[13] Ren Z M, Zeng A, Zhang Y C 2020 *Humanit. Soc. Sci. Commun.* **7** 156
[14] Bustos S, Gomez C, Hausmann R, Hidalgo C A 2012 *PLoS One* **7** e49393
[15] Burgos E, Ceva H, Perazzo R P, Devoto M, Medan D, Zimmermann M, Delbue A M 2007 *J. Theor. Biol.* **249** 307
[16] Rohr R P, Saavedra S, Bascompte J 2014 *Science* **345** 1253497
[17] Mariani M S, Ren Z M, Bascompte J, Tessone C J 2019 *Phys. Rep.* **813** 1
[18] Thébault E, Fontaine C 2010 *Science* **329** 853
[19] Mello M A, Felix G M, Pinheiro R B, Muylaert R L, Geiselman C, Santana S E, Tschapka M, Lotfi N, Rodrigues F A, Stevens R D 2019 *Nat. Ecol. Evol.* **3** 1525
[20] Palazzi M J, Borge-Holthoefer J, Tessone C J, Solé-Ribalta A 2019 *J. R. Soc. Interface* **16** 20190553
[21] Leung C Y, Weitz J S 2016 *Phys. Rev. E* **93** 032303
[22] Pinheiro R B, Felix G M, Dormann C F, Mello M A 2019 *Ecology* **100** e02796
[23] Yamamichi M, Ellner S P, Hairston Jr N G 2023 *Ecol. Lett.* **26** S16
[24] Hidalgo C A 2021 *Nat. Rev. Phys.* **3** 92
[25] Bardoscia M, Barucca P, Battiston S, Caccioli F, Cimini G, Garlaschelli D, Saracco F, Squartini T, Caldarelli G 2021 *Nat. Rev. Phys.* **3** 490
[26] Hidalgo C A, Hausmann R 2009 *Proc. Natl. Acad. Sci. U.S.A.* **106** 10570
[27] O' Clery N, Yıldırım M A, Hausmann R 2021 *Nat. Commun.* **12** 1479
[28] Stoffman N, Woeppel M, Yavuz M D 2022 *J. Account. Econ.* **74** 101492
[29] Shen H W, Barabási A L 2014 *Proc. Natl. Acad. Sci. U.S.A.*

**111** 12325
[30] Balassa B 1965 *Manch. Sch.* **33** 99
[31] Hidalgo C A, Klinger B, Barabási A L, Hausmann R 2007 *Science* **317** 482
[32] Reimers N, Gurevych I 2019 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* Hong Kong, China, November 3–7, 2019 p3982
[33] Barber M J 2007 *Phys. Rev. E* **76** 066102
[34] Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E 2008 *J. Stat. Mech.: Theory Exp.* **2008** P10008
[35] Solé-Ribalta A, Tessone C J, Mariani M S, Borge-Holthoefer J 2018 *Phys. Rev. E* **97** 062302
[36] Duch J, Arenas A 2005 *Phys. Rev. E* **72** 027104
[37] Saracco F, Di Clemente R, Gabrielli A, Squartini T 2015 *Sci. Rep.* **5** 10595
[38] Vallarano N, Bruno M, Marchese E, Trapani G, Saracco F, Cimini G, Zanon M, Squartini T 2021 *Sci. Rep.* **11** 15227
[39] Vinh N X, Epps J, Bailey J 2009 *Proceedings of the 26th Annual International Conference on Machine Learning* Montreal Quebec, Canada, June 14–18, 2009 p1073
[40] Mood C 2010 *Eur. Sociol. Rev.* **26** 67
[41] Pugliese E, Napolitano L, Zaccaria A, Pietronero L 2019 *PLoS One* **14** e0223403
[42] Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley H E, Havlin S 2019 *Nat. Commun.* **10** 3439

DATA PAPERS

# A multi-scale patent dataset for technological innovation systems: Cross-scale structural comparison and evolutionary mechanisms*

HUANG Yiwei [1)] XU Shuqi [2)†] LÜ Linyuan [3)‡]

1) (*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu* 611731, *China*)

2) (*Institute of Dataspace, Hefei Comprehensive National Science Center, Hefei* 230088, *China*)

3) (*School of Cyber Science and Technology, University of Science and Technology of China, Hefei* 230026, *China*)

( Received 31 December 2025; revised manuscript received 4 February 2026 )

**Abstract**

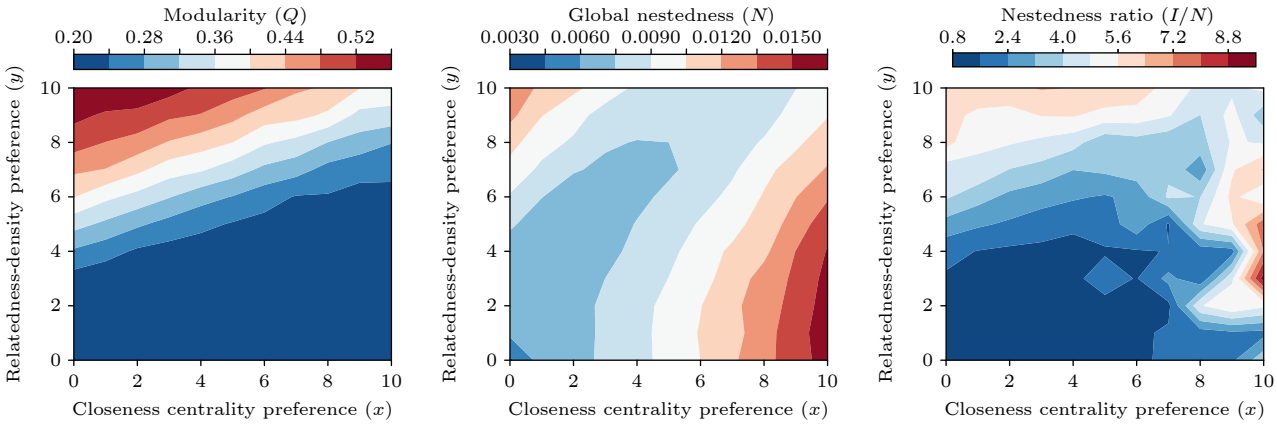
Scale dependence is a pervasive feature of socioeconomic networks: even when generated by the same class of economic activities, networks observed at different levels of entity aggregation can exhibit markedly different structural organizations, yet a systematic and testable explanation for such cross-scale divergence remains lacking. This paper presents a U.S. patent dataset together with a unified analytical framework that combines empirical network analysis with mechanism-based modeling to quantify and interpret structural differences across scales. The dataset contains 1,225,373 granted USPTO utility patents filed during 2000–2020 and integrates assignee geography (state/county/city), firm identifiers, CPC classification codes, and patent texts; technologies are defined at the 4-digit CPC level. To measure technology activity when patents involve multiple technologies, we use co-citation information to allocate each patent's technological shares across its associated CPC codes, thereby obtaining technology-share weights beyond naive equal counting. Using these weights, we construct entity-technology bipartite networks at four scales (state, county, city, and firm) and derive two technology space networks, one based on technology co-occurrence across entities and the other based on patent-text similarity. We characterize network structure using bipartite modularity ($Q$), global nestedness ($N$),

* Project supported by the Major Program of the National Natural Science Foundation of China (Grant No. T2293771) and the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62306191).

† Corresponding author. E-mail: xushuqi@idata.ah.cn

‡ Corresponding author. E-mail: linyuan.lv@ustc.edu.cn

and in-block nestedness ($I$), and evaluate statistical significance against degree-constrained null models based on the bipartite configuration model (BiCM). Empirically, modularity increases as the entity scale becomes finer; state- and county-level networks are closer to a globally nested organization, whereas city- and firm-level networks exhibit a pronounced shift toward in-block nestedness. Temporal analysis further shows that the formation of new entity-technology links reflects a scale-dependent balance between preference for globally central technologies and reliance on relatedness density to the existing technological portfolio, with smaller-scale entities exhibiting a stronger dependence on relatedness density. Finally, we propose an evolutionary model that incorporates both relatedness-density preference and technology-centrality preference under empirical degree constraints. Simulations demonstrate that tuning these two preferences reproduces the observed transition from global nestedness to in-block nestedness, providing a mechanism-based explanation for scale-dependent structural patterns in technological innovation networks. The dataset presented in this paper is openly available at https://doi.org/10.57760/sciencedb.j00213.00265.




**Keywords:** patent data, technological innovation systems, complex networks, nestedness, modularity

**DOI:** 10.7498/aps.75.20251802 **CSTR:** 32037.14.aps.75.20251802

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

# 面向技术创新系统的专利数据集构建：跨尺度结构对比与演化机制研究

黄羿炜　徐舒琪　吕琳媛

# A multi-scale patent dataset for technological innovation systems: Cross-scale structural comparison and evolutionary mechanisms

HUANG Yiwei　XU Shuqi　LÜ Linyuan

引用信息 Citation: Acta Physica Sinica, 75, 100003 (2026)　DOI: 10.7498/aps.75.20251802
CSTR: 32037.14.aps.75.20251802
在线阅读 View online: https://doi.org/10.7498/aps.75.20251802
当期内容 View table of contents: http://wulixb.iphy.ac.cn

**您可能感兴趣的其他文章**
**Articles you may be interested in**

基于复杂网络理论的供应链研究
Supply chain research based on complex network theory
物理学报. 2024, 73(19): 198901　https://doi.org/10.7498/aps.73.20240702

融合节点动态传播特征与局域结构的复杂网络传播关键节点识别
Identification of key spreaders in complex network by integrating dynamic characteristics and local structure of nodes
物理学报. 2025, 74(10): 108901　https://doi.org/10.7498/aps.74.20250179

基于信息熵与迭代因子的复杂网络节点重要性评价方法
Importance evaluation method of complex network nodes based on information entropy and iteration factor
物理学报. 2023, 72(4): 108901　https://doi.org/10.7498/aps.72.20221878

基于引力方法的复杂网络节点重要度评估方法
Node importance ranking method in complex network based on gravity method
物理学报. 2022, 71(17): 176401　https://doi.org/10.7498/aps.71.20220565

具有异质增益因子的超图上的演化公共品博弈
Evolutionary public goods games on hypergraphs with heterogeneous multiplication factors
物理学报. 2022, 71(11): 110201　https://doi.org/10.7498/aps.70.20212436

识别高阶网络传播中最有影响力的节点
Identifying influential nodes in spreading process in higher-order networks
物理学报. 2024, 73(4): 048901　https://doi.org/10.7498/aps.73.20231416