系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
41(10) (2021, 10), 1–13

基于引文大数据的高阶网络建模及信息增益比较研究*

李佳旭[1] 蔡梦思[1] 谭索怡[1] 贾 韬[2] 吕 欣[1]

(1. 国防科技大学系统工程学院, 长沙 410073; 2. 西南大学计算机与信息科学学院软件学院, 重庆 400715)

摘要 传统一阶网络模型难以捕捉节点间的间接依赖关系, 高阶网络建模方法能有效提高模型对现实系统的表征精度. 文章使用美国物理学会电子期刊数据库 116 年间的引文数据, 在一阶网络的基础上, 构建以文献引用关系为节点、以路径长度为 2 的引用关系为边的二阶引文网络, 并进一步提出基于引用多样性信息熵的高阶网络表示信息增益指标. 结果表明, 期刊 *Rev Mod Phys* 和 *Phys Rev Lett* 的二阶网络表示信息增益最大, 这两种期刊的被引率受前序期刊的间接影响较大, 且在一阶与二阶网络中相差最大达到 0.38, 说明应用二阶网络开展引文数据分析的重要性. 高阶网络表示信息增益有效量化了真实系统在低阶网络模型中的高阶信息损失, 揭示了一阶和二阶网络中直接和间接依赖关系的差异.

关键词 高阶网络, 二阶引文网络, 高阶网络表示信息增益, 高阶马尔可夫模型, 信息熵.

MR(2000) 主题分类号 05C50, 60J10

A Comparison Study of Higher-Order Network Modeling and Information Gain Based on Big Citation Data

LI Jiaxu[1] CAI Mengsi[1] TAN Suoyi[1] JIA Tao[2] LU Xin[1]

(1. *College of Systems Engineering, National University of Defense Technology, Changsha* 410073;
2. *College of Computer and Information Science, Southwest University, Chongqing* 400715)

Abstract While traditional first-order network model is limited in capturing the indirect dependence relationships among multiple nodes, higher-order network modeling shows good capacity of effectively improving the accuracy of the representation of real systems. In this study, we construct a second-order citation network with citation relationships as nodes and the length-two citation path as edges based on

* 国家杰出青年科学基金 (72025405), 国家自然科学基金重大项目 (71790615), 国家自然科学基金基础科学中心项目 (72088101), 国家自然科学基金青年项目 (72001211), 湖南省科技计划项目 (2019GK2131, 2020JJ5679) 资助课题.

收稿日期: 2021-04-15, 收到修改稿日期: 2021-05-29.

通信作者: 吕欣, Email: xin_lyu@sina.com.

编委: 房勇.

the first-order citation network, using 116 years of citation data from the American Society of Physics Full-text Electronic Journal Database. To quantify the difference between the first-order network and the higher-order network, we propose the *higher-order network representation information gain* index based on information entropy of citation diversity, which represents the information loss of higher-order dependency in the first-order network structure. Results show that the second-order network representation information gains of *Rev Mod Phys* and *Phys Rev Lett* are the largest, the probabilities of these two journals being cited by other journals are largely indirectly affected by its preceding journals. Differences in these probabilities between the first-order and second-order networks is up to 0.38, indicating the importance of applying the second-order network to analyze citation data. The proposed information gain index can effectively quantify the higher-order information loss of the real system in the low-order network model, thereby revealing differences in direct and indirect relationships in the first-order and second-order networks.

Keywords Higher-order network, second-order citation network, higher-order network representation information gain, higher-order Markov model, entropy.

1 引 言

自WS小世界网络 [1, 2]、BA无标度网络 [3, 4]、BBV加权网络 [5]、确定性层次网络 [6, 7] 等模型被提出以来, 复杂网络成为研究大规模真实系统的一项重要工具. 利用复杂网络理论与方法对海量数据进行建模与分析成为复杂系统研究的重要手段 [8–10]. 传统网络建模通常采用真实数据来构建以系统中个体为节点、个体间关联关系为边的一阶网络模型. 该方法基于一阶马尔可夫假设, 认为网络中节点间的关联关系只与最近邻的两个节点有关 [11, 12]. 然而, 随着计算机运算能力的飞速提升, 科学家们发现传统网络模型难以捕获节点间的间接依赖关系, 从而导致对真实系统的整体认知存在偏差. 例如, 在软件开发者社交网络中, 节点表示同一开源项目的开发人员, 边表示开发人员间包含通讯时间的互动关系, 若不考虑节点间的时序依赖信息仅构建一阶网络, 则会对成员在团队中的重要性产生错判 [12]. 也有研究表明, 在真实的船运系统中, 一艘货船的前序停靠港口会对其后续航线产生影响, 与不考虑该依赖信息的一阶船运网络相比, 在引入时序依赖信息的高阶网络上得到的社团结构更能反映港口间的运输强度 [11].

高阶网络建模方法是捕捉节点间间接依赖关系的主要手段, 能有效弥补传统网络模型对现实系统准确表征能力的不足. 目前, 高阶网络模型主要分为多层高阶模型 [13, 14]、基于模体的高阶网络模型 [15–17]、高阶网络马尔可夫模型 [11, 18, 19] 三种. 在多层高阶模型方面, Scholtes 等 [18] 将具有非马尔可夫性连边的时序网络转化为具有马尔可夫性的二阶时序网络模型, 并进一步定义了该网络的基于路径的中心性指标, 提出了基于高阶马尔可夫链的多层图模型. 在基于模体的高阶网络模型方面, Bensor和Yin 等 [15–17] 认为网络模体可以视为一种网络高阶模式, 提出了基于模体的高阶网络矩阵表示、基于高阶网络的谱聚类算法 [15]、基于高阶模体结构的相似个性化PageRank (MAPPR)算法 [16]、基于高阶结构的网络聚集系数 [17]. 其中, MAPPR 算法根据模体阻断率提高了聚簇质量, 利用局部邻居节点集降低了算法复杂度. 在高阶网络马尔可夫模型方面, Xu等在Scholtes等学者的基础上提出了HON+ 高

阶网络建模方法, 其核心思想是将不同阶数的依赖关系嵌入到同一网络模型中, 此时, 每个高阶节点的阶数不再固定, 同一网络中可能出现 1 至k 阶种类型的节点; 该方法进一步提高了高阶网络模型的可扩展性和网络分析结果的准确度. Saebi 等 [19] 提出了高阶网络异常检测框架, 把可扩展的HON+高阶网络模型应用到异常检测领域, 发现与一阶网络相比, 高阶网络模型显著提高了识别异常节点行为的准确率. Lambiotte 等 [12] 从建模方法、社区检测、节点中心性、动力学过程等方面, 总结归纳出高阶网络模型的三个性质, 即高精度、可扩展性和兼容性, 并强调了该模型在各个领域的研究价值.

已有研究表明, 目前传统复杂网络模型主要基于一阶马尔可夫假设, 难以捕捉到节点间的间接依赖关系, 降低了其对现实系统的表征精度, 在某些特定场景下, 甚至会产生错误的分析结果. 此外, 与严重增加的时间和空间计算复杂度相比, 高阶网络分析是否必要, 也缺乏科学的指导指标. 为解决上述问题, 本文以一类典型的存在高阶依赖关系的网络 - 文献引用网络为示例, 探索其二阶网络表示与传统一阶网络表示的差异, 并设计度量指标来刻画高阶网络表示对网络依赖关系表达的准确程度. 引文网络通过以文献为节点、文献间的引用和被引用关系为边, 被广泛应用于描述科学领域的发展以及学科间的关系 [20]. 随着人类知识总量的不断增长, 引文网络已发展为一个超大规模的复杂网络系统, 基于引文大数据探索人类知识流动、转换与演变过程吸引了越来越多的关注. 在传统的引文分析中, 学者们主要构建以文献 (或期刊) 为节点、文献 (或期刊) 引用关系为边的一阶引文网络, 并结合网络分析方法探究节点关联关系及演化过程, 以得到知识在引文网络中的流动情况 [20]. 然而, 文献间的引用关系具有先后时序性, 只能是后期发表的文献引用前期的文献, 传统的一阶引文网络基于一阶马尔可夫假设则难以捕捉到节点间的间接依赖关系, 对现实引文网络系统中知识流动的准确表征能力不足. 目前, 已有少量研究开始将文献引用视为一种高阶依赖关系, 并初步表明基于高阶网络模型对引文数据进行分析是切实可行的 [20, 21]. 为进一步深入挖掘高阶网络分析方法在文献大数据中的应用价值, 本文使用美国物理学会全文电子期刊数据库 116 年间的文献数据, 在一阶引文网络的基础上, 构建以文献引用关系为节点、以路径长度为 2 的引用关系为边的二阶引文网络, 并进一步提出基于引用多样性信息熵的高阶网络表示信息增益指标, 全面比较一阶网络表示与二阶网络表示在文献引用模式上的规律差异, 验证信息增益指标在量化高阶网络信息表达上的必要性和可行性.

2 数据与方法

2.1 数据来源及范围

本文使用的文献数据 [22] 来源于1893年至2009年 (116 年) 美国物理学会全文电子期刊数据库中的9种物理学领域专业期刊, 共包含468, 291 篇文献, 906, 398 条文献引用关系, 包括引用文献和被引用文献的 DOI 号等字段. 文献数据统计结果如表 1 所示.

2.2 引文网络马尔可夫模型

在引文网络中, 文献间的引用关系在时间上具有单向性和时序性, 只能是后期的文献引用前期的文献, 即知识从发表较早的文献流入发表较晚的文献 [21, 23]. 高阶马尔可夫网络表示的关键优势在于它捕获了时间相关路径的拓扑结构, 其构造算法可以推广至任意阶的高

表 1 数据来源

(Table 1 Data source)

期刊名称	期刊代码	缩写	领域	文献数
Physical Review Letters	*Phys Rev Lett*	PRL	物理学各方面的基础研究	149766
Physical Review A	*Phys Rev A*	PRA	原子、分子和光学物理学	53655
Physical Review B	*Phys Rev B*	PRB	凝聚态物质与材料物理	137999
Physical Review C	*Phys Rev C*	PRC	核物理	29935
Physical Review D	*Phys Rev D*	PRD	粒子、场、重力与宇宙学	56616
Physical Review E	*Phys Rev E*	PRE	统计、非线性物理	35944
Physical Review Accelerators and Beams	*Phys Rev Accel Beams*	PRAB	粒子加速器与粒子束	1257
Reviews of Modern Physics	*Rev Mod Phys*	RMP	物理学综述	2926
Physics	*Physics*	PHY	前沿物理学科普	193

阶模型 [11, 14, 19]. 为完整记录文献间的引用关系, 本文基于高阶马尔可夫网络表示, 分别构造一阶和二阶引文网络马尔可夫模型. 本文提出的网络模型基于以下假设.

假设 1 存在一个非空集合 $S=\{p_1,p_2,\cdots,p_N\}$, 集合 S 包含 N 个独立的观测序列 p_i.

假设 2 一阶引文网络使用图 $G^{(1)}=\left(V^{(1)},E^{(1)}\right)$ 表示, 其中 $V^{(1)}$ 是网络中节点的集合, $E^{(1)}$ 是网络中有向边的集合, $E^{(1)}\subseteq V^{(1)}\times V^{(1)}$.

假设 3 观测序列 p_i 是图 G 中的一条路径, 长度为 l_i. p_i 表示引文网络中的一条知识流动记录, 由 l_i+1 个节点组成, 记为 $p_i=(v_0\to v_1\to\cdots\to v_{l_i})$, 其中 $(v_i,v_{i+1})\in E^{(1)}$, $i\in[0,l_i-1]$. p_i 被看作网络中多个节点组成的一条依赖关系, 存储在集合 S 中. 每个马尔可夫模型可以被定义为一个条件概率模型 [24].

1) 一阶网络马尔可夫模型

一阶引文网络基于无记忆性的一阶马尔可夫假设, 即网络中节点间的关联关系只与最近邻的两个节点有关, 与其他节点无关 [24]. 一阶网络马尔可夫模型可以表示为

$$P\left(v_{i+1}|\,v_0\to v_1\to\cdots\to v_i\right)=P\left(v_{i+1}|\,v_i\right). \tag{2.1}$$

当对一阶引文网络中的知识流动记录 p_i 建模时, 期刊 i 被期刊 $i+1$ 引用的概率只与节点 v_i 和 v_{i+1} 有关, 与该记录中的其他节点无关. $P\left(v_{i+1}\,|v_i\right)$ 被称为一阶马尔可夫链的转移概率, 表示从节点 v_i 移动到节点 v_{i+1} 的概率, 并且存储在转移概率矩阵 $P^{(1)}$ 中,

$$P^{(1)}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1|V^{(1)}|} \\ p_{21} & p_{22} & \cdots & p_{2|V^{(1)}|} \\ \vdots & \vdots & \ddots & \vdots \\ p_{|V^{(1)}|1} & p_{|V^{(1)}|2} & \cdots & p_{|V^{(1)}||V^{(1)}|} \end{bmatrix}, \tag{2.2}$$

其中 p_{ij} 表示期刊 j 对期刊 i 的文献引用量占期刊 i 总被引用量的比例. 转移概率矩阵 $P^{(1)}$ 可以反映期刊之间的文献知识流动特征 [25–27].

2) 二阶网络马尔可夫模型

二阶引文网络基于二阶马尔可夫假设, 网络中节点间的关联关系不仅与最近邻的两个

节点有关, 而且受到前序节点 v_{i-1} 的影响 [11, 18]. 二阶网络马尔可夫模型可以表示为

$$P\left(v_{i+1}|v_0 \to v_1 \to \cdots \to v_i\right) = P\left(v_{i+1}|v_{i-1} \to v_i\right). \tag{2.3}$$

当对二阶引文网络中的知识流动记录 p_i 建模时, 期刊 i 被期刊 $i+1$ 引用的概率受节点集 $\{v_{i-1}, v_i, v_{i+1}\}$ 的影响. $P\left(v_{i+1}|v_{i-1} \to v_i\right)$ 存储在条件概率矩阵 $P^{(2)}$ 中, 表示存在知识从期刊 $i-1$ 流向期刊 i 的前提下, 期刊 i 被期刊 $i+1$ 引用的概率. 由于二阶引文网络利用条件概率矩阵表征真实数据的高阶依赖信息, 因此其与一阶引文网络相比, 更能精确反映期刊之间的知识流动特征 [27].

$$P^{(2)} = \begin{bmatrix} P\left(v_1|v_1 \to v_1\right) & P\left(v_2|v_1 \to v_1\right) & \cdots & P\left(v_{|V^{(1)}|}|v_1 \to v_1\right) \\ P\left(v_1|v_1 \to v_2\right) & P\left(v_2|v_1 \to v_2\right) & \cdots & P\left(v_{|V^{(1)}|}|v_1 \to v_2\right) \\ \vdots & \vdots & \ddots & \vdots \\ P\left(v_1|v_1 \to v_{|V^{(1)}|}\right) & P\left(v_2|v_1 \to v_{|V^{(1)}|}\right) & \cdots & P\left(v_{|V^{(1)}|}|v_1 \to v_{|V^{(1)}|}\right) \\ P\left(v_1|v_2 \to v_1\right) & P\left(v_2|v_2 \to v_1\right) & \cdots & P\left(v_{|V^{(1)}|}|v_2 \to v_1\right) \\ \vdots & \vdots & \ddots & \vdots \\ P\left(v_1|v_{|V^{(1)}|} \to v_{|V^{(1)}|}\right) & P\left(v_2|v_{|V^{(1)}|} \to v_{|V^{(1)}|}\right) & \cdots & P\left(v_{|V^{(1)}|}|v_{|V^{(1)}|} \to v_{|V^{(1)}|}\right) \end{bmatrix}. \tag{2.4}$$

3) k 阶网络马尔可夫模型

k 阶引文网络基于 k 阶马尔可夫假设, 网络中相邻节点间的关联关系受多个前序节点的影响 [28]. k 阶网络马尔可夫模型可以表示为

$$P^{(k)} := P\left(v_{i+1}|\, v_0 \to \cdots \to v_i\right) = P\left(v_{i+1}|\, v_{i-k-1} \to \cdots \to v_i\right). \tag{2.5}$$

当对 k 阶引文网络中的知识流动记录 p_i 建模时, 期刊 i 被期刊 $i+1$ 引用的概率受到其前序期刊集 $\{v_{i-1}, v_{i-2}, \cdots, v_{i-k-1}\}$ 的影响. k 阶马尔可夫链的转移概率 $P\left(v_{i+1}|\, v_{i-k-1} \to \cdots \to v_i\right)$ 存储在条件概率矩阵 $P^{(k)}$ 中, 表示在前序期刊集 $\{v_{i-1}, v_{i-2}, \cdots, v_{i-k-1}\}$ 的影响下, 期刊 i 被期刊 $i+1$ 引用的概率.

2.3 基于引用多样性信息熵的高阶网络信息增益

在给定约束条件的情况下, 信息增益可以表示一个系统不确定性减少的程度. 低阶网络表示方法难以捕捉到系统中节点间的高阶依赖关系, 增加了系统的不确定性和随机性. 因此, 为全方位、多角度地量化低阶网络表示方法的高阶信息损失量, 本文在条件信息熵的基础上 [29–32], 结合引文大数据的背景, 定义基于引用多样性信息熵的高阶网络信息增益指标, 简称高阶网络表示信息增益.

首先, 一阶引文网络 $G^{(1)} = \left(V^{(1)}, E^{(1)}\right)$ 的定义如下, 其中, 矩阵元素 g_{ij} 表示期刊 i 中的文献被期刊 j 引用的频次.

$$G^{(1)} = \begin{bmatrix} g_{11} & \cdots & g_{1|V^{(1)}|} \\ \vdots & \ddots & \vdots \\ g_{|V^{(1)}|1} & \cdots & g_{|V^{(1)}||V^{(1)}|} \end{bmatrix}. \tag{2.6}$$

因此, 一阶引文网络中期刊 i 的被引用率可以表示为

$$q_i = \frac{\sum_{j=0}^{\left|V^{(1)}\right|} g_{ij}}{\sum_{i=0}^{\left|V^{(1)}\right|}\sum_{j=0}^{\left|V^{(1)}\right|} g_{ij}}. \tag{2.7}$$

引用多样性信息熵可以衡量一个期刊或者一种引文网络的随机性和不确定性, 信息熵值越大, 说明研究对象包含的有效信息量越少. 结合条件信息熵的概念 [29, 33], 本文定义一阶引文网络的引用多样性信息熵公式为

$$H_{FirstOrder} = -\sum_{v_i \in V^{(1)}} q_i \log q_i. \tag{2.8}$$

二阶引文网络 $G^{(2)} = \left(V^{(2)}, E^{(2)}\right)$ 由二阶节点集 $V^{(2)} = \{e_{ij} = (v_i, v_j) \mid v_i, v_j \in V\}$ 和二阶边集 $E^{(2)} = \left\{(e_{ij}, e_{jk}) \middle| e_{ij}, e_{jk} \in V^{(2)}\right\}$ 组成, 其定义如下, 其中, 矩阵元素 $g\left(v_k|v_i \to v_j\right)$ 表示在期刊 i 对 j 有知识输入的前提下, 期刊 j 中的文献被期刊 k 引用的频次.

$$G^{(2)} = \begin{bmatrix} g\left(v_1|v_1 \to v_1\right) & g\left(v_2|v_1 \to v_1\right) & \cdots & g\left(v_{\left|V^{(1)}\right|}|v_1 \to v_1\right) \\ g\left(v_1|v_1 \to v_2\right) & g\left(v_2|v_1 \to v_2\right) & \cdots & g\left(v_{\left|V^{(1)}\right|}|v_1 \to v_2\right) \\ \vdots & \vdots & \cdots & \vdots \\ g\left(v_1|v_1 \to v_{\left|V^{(1)}\right|}\right) & g\left(v_2|v_1 \to v_{\left|V^{(1)}\right|}\right) & \cdots & g\left(v_{\left|V^{(1)}\right|}|v_1 \to v_{\left|V^{(1)}\right|}\right) \\ g\left(v_1|v_2 \to v_1\right) & g\left(v_2|v_2 \to v_1\right) & \cdots & g\left(v_{\left|V^{(1)}\right|}|v_2 \to v_1\right) \\ \vdots & \vdots & \ddots & \vdots \\ g\left(v_1|v_{\left|V^{(1)}\right|} \to v_{\left|V^{(1)}\right|}\right) & g\left(v_2|v_{\left|V^{(1)}\right|} \to v_{\left|V^{(1)}\right|}\right) & \cdots & g\left(v_{\left|V^{(1)}\right|}|v_{\left|V^{(1)}\right|} \to v_{\left|V^{(1)}\right|}\right) \end{bmatrix}. \tag{2.9}$$

因此, 二阶引文网络中期刊 i 被期刊 j 引用的比例为

$$p\left(v_i \to v_j\right) = \frac{\sum_{k=0}^{\left|V^{(1)}\right|} g\left(v_k|v_i \to v_j\right)}{\sum_{i=0}^{\left|V^{(1)}\right|}\sum_{j=0}^{\left|V^{(1)}\right|}\sum_{k=0}^{\left|V^{(1)}\right|} g\left(v_k|v_i \to v_j\right)}. \tag{2.10}$$

由此可得, 二阶引文网络的引用多样性信息熵公式为

$$H_{SecondOrder} = -\sum_{v_i \in V^{(1)}}\sum_{v_j \in V^{(1)}}\sum_{v_k \in V^{(1)}} P\left(v_k|v_i \to v_j\right) \log P\left(v_k|v_i \to v_j\right). \tag{2.11}$$

高阶网络表示信息增益的计算公式为

$$Gain = H_{FirstOrder} - H_{SecondOrder}. \tag{2.12}$$

同理, 每个期刊在一阶引文网络中的高阶信息损失量 (即二阶网络信息增益指标) 的计算公式为

$$Gain\left(v_i\right) = H_{FirstOrder}\left(v_i\right) - H_{SecondOrder}\left(v_i\right). \tag{2.13}$$

其中, $H_{FirstOrder}\left(v_i\right)$ 和 $H_{SecondOrder}\left(v_i\right)$ 分别表示期刊 i 在一阶引文网络和二阶引文网络中的引用多样性信息熵, 可以通过式 (2.8) 和 (2.11) 分别计算得到.

3 结果分析

基于 468, 291 篇来自 9 种物理学领域期刊文献间的 906, 398 条文献引用数据, 得到期刊间的文献引用情况统计结果如表 2 所示, 对角线元素代表期刊的自引量. 比如, PRE 行与 PRA 列相交的数字为 1165, 表示期刊 *Phys Rev A* 引用了1165 篇来自 *Phys Rev E* 中的文献; 反之, *Phys Rev E* 被 *Phys Rev A* 引用了1165 篇文献.

表 2 期刊间的引用情况
(Table 2 Citation between journals)

期刊名	PRA	PRL	PRB	RMP	PRC	PRE	PRD	PRAB	PHY	总和
PRA	60932	12446	5762	1366	433	9533	602	59	15	**91148**[2]
PRL	42897	79422	120373	5440	15533	28412	34425	325	115	**326942**
PRB	5237	35716	216813	3055	148	6026	423	12	37	**267467**
RMP	3614	4270	9535	917	1605	2281	2682	26	5	**24935**
PRC	438	3636	109	580	36170	117	1706	4	8	**42768**
PRE	1165	5176	2047	464	78	24779	110	213	6	**34038**
PRD	1303	9031	1037	1435	5141	356	100217	19	7	**118546**
PRAB	12	69	2	8	9	47	12	395	0	**554**
PHY	0	0	0	0	0	0	0	0	0	**0**
总和	1155981[1]	149766	355678	13265	59117	71551	140177	1053	193	906398

[1] 添加下划线的数字是对应的列期刊对 9 种期刊的总引用量.
[2] 加粗数字为对应的行期刊被 9 种期刊的总被引频次.

3.1 一阶引文网络结构分析

一阶引文网络可以抽象成由一阶节点集 $V^{(1)}$ 和一阶边集 $E^{(1)}$ 组成的图 $G^{(1)}=\left(V^{(1)},E^{(1)}\right)$. 其中, $V^{(1)}=\{v_i|\,i=1,2,\cdots,9\}$ 表示期刊节点集合, $E^{(1)}=\{e_{ij}=(v_i,v_j)\mid v_i,v_j\in V\}$ 表示有向边集合, 边 e_{ij} 的权重 g_{ij} 表示期刊 j 中的文献对期刊 i 的引用量 [34, 35].

$$G^{(1)}=\begin{bmatrix} g_{11}\cdots g_{19}\\ \vdots\ \ddots\ \vdots\\ g_{91}\cdots g_{99}\end{bmatrix}. \tag{3.1}$$

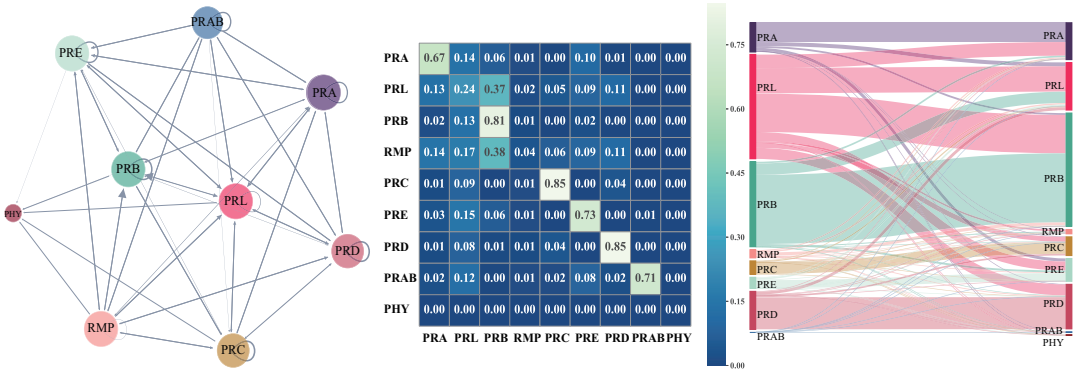
期刊之间基于文献引用关系实现不同学科之间的知识流动与交换, 知识的传播扩散促进了知识的利用和创新, 有利于学科的交叉、融合与发展 [25, 26, 36]. 本文采用网络度指标来量化期刊间的知识流动情况: 节点 v_i 的入度 k_i^{in}, 表示期刊 i 引用了 k_i^{in} 个期刊中的文献, 其吸收了 k_i^{in} 个期刊的知识; 节点 v_i 的出度 k_i^{out} 表示有 k_i^{out} 个期刊引用了期刊 i 收录的文献, 期刊 i 向 k_i^{out} 个期刊实现了知识输出. 同时, 定义期刊 i 被期刊 j 引用的概率 $CitedRate\,(i,j)$ (即期刊间的文献被引率) 为期刊 i 被期刊 j 引用的次数占期刊 i 被 9 种期刊引用的总次数的比例, 公式如下

$$CitedRate\,(i,j)=\frac{g_{ij}}{\sum_{j=1}^{9}g_{ij}}. \tag{3.2}$$

一阶引文网络拓扑结构如图 1(a) 所示, 节点大小和度呈正比, 节点越大, 说明该期刊的知识流动越频繁; 边的粗细与其权重呈正比, 边越粗, 说明期刊间的知识交流活动越活跃. 图

1(b) 展示了以期刊间的文献被引率为元素构建一阶引文网络的转移概率矩阵, 可见, 大部分期刊的自被引率 (主对角线元素) 较高, 知识的自我继承性较强, 倾向于自身知识内部利用, 相同研究方向的文献间有较强的知识交流, 如 *Phys Rev B*、*Phys Rev C*、*Phys Rev D* 的自被引率分别为 0.81、0.85、0.85. 值得注意的是, 综合性期刊 *Phys Rev Lett* 被 *Phys Rev B* 引用比较频繁, 高达 0.37, 超过了其自被引率 (0.24), 其向除自身以外的期刊转移和渗透了接近 80% 的知识. 同样, 分析发现 *Rev Mod Phys* 的自被引率仅为 0.04, 但其被其他期刊引用的概率高达 0.96, 这也是因为 *Rev Mod Phys* 主要发表前沿热点领域的权威综述性论文, 为各个研究方向提供指导, 其在物理学领域影响力广泛, 是各个前沿热点研究方向的风向标. 除期刊 *Phys Rev Lett*、*Rev Mod Phys*、*Physics* 之外, 其他 6 种期刊被 *Phys Rev Lett* 引用的概率均是除自被引率外最高的, 表明 *Phys Rev Lett* 在整个网络结构中相对处于知识存储地位. 而由于 *Physics* 主要收录在 *Physical Review* 系列期刊中文献的科普解读, 其被引率均为 0.

在图 1(a) 和图 1(b) 的基础上, 本文采用桑基图来直观地展示一阶网络中期刊间的知识流动规律和特征, 如图 1(c) 所示. 可以发现, 大部分期刊倾向于自身知识内部利用, 相同研究方向的期刊间知识交流活动频繁. 其中, 期刊 *Phys Rev Lett* 的知识吸收度和知识溢出度都比较大, 既吸收了来自 8 个期刊的知识, 同时向 9 个期刊实现了知识输出, 说明该期刊的知识吸收范围和溢出范围较为广泛, 其综合性和基础性较强.



(a) 一阶引文网络拓扑图 (b) 一阶引文网络转移概率矩阵 (c) 一阶引文网络桑基图

图 1 一阶引文网络示意图

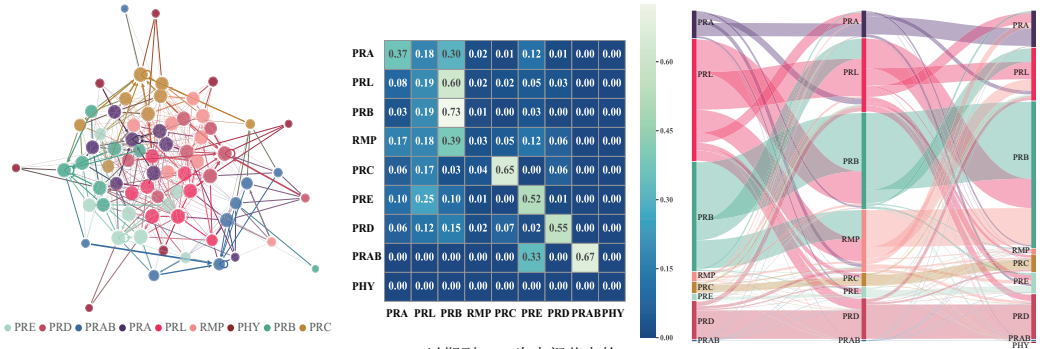
(Figure 1 First-order citation networks)

3.2 二阶引文网络结构分析

二阶引文网络可以表示为 $G^{(2)} = \left(V^{(2)}, E^{(2)}\right)$, 由二阶节点集 $V^{(2)} = \{e_{ij} = (v_i, v_j) \mid v_i, v_j \in V\}$ 和二阶边集 $E^{(2)} = \left\{(e_{ij}, e_{jk}) \,\middle|\, e_{ij}, e_{jk} \in V^{(2)}\right\}$ 组成. 每个二阶节点等价于一阶引文网络的一条边 e_{ij}, 表示期刊 j 对期刊 i 有文献引用, 即 i 对 j 有知识输出. 二阶边集是一阶引文网络中所有可能长度为 2 的路径组成的集合 [18]. 例如, 二阶边 $v_j \,|v_i \to v_k$ 表示在期刊 i 对 j 有知识输入的前提下, 部分存储在期刊 j 的知识流向期刊 k. 从二阶节点 $v_j|v_i$ 到节点 v_k 的转移概率表示为

$$P\left(v_k \,|(v_j|v_i)\right) = \frac{W\left(v_j|v_i \to v_k\right)}{\sum_{k=1}^{9} W\left(v_j|v_i \to v_k\right)}, \tag{3.3}$$

其中, $W(v_j|v_i \to v_k)$ 表示二阶边 $v_j|v_i \to v_k$ 的权重. 该转移概率等价于公式 (2.3) 的条件概率, 表示在期刊 i 对 j 有知识输入的前提下, 存储在期刊 j 的知识流向期刊 k 的比例. 二阶引文网络拓扑结构如图 2(a) 所示, 共包含二阶节点 81 个, 有向边 458 条. 图中二阶节点 $v_j|v_i$ 的颜色与期刊 j 相对应, 如果两个节点的颜色相同, 则表明两个节点的 v_j 表示的是同一种期刊. 图 2(b) 展示了以 $P(v_k|(RMP|v_i))$ 为元素的二阶引文网络的转移概率矩阵. 可见, 在期刊 *Phys Rev D* 对 *Rev Mod Phys* 有知识流输入的前提下, 从期刊 *Rev Mod Phys* 流向 *Phys Rev B* 的知识量占 *Rev Mod Phys* 总知识流出量的 15%. 仔细观察可以发现, 绝大部分主对角线元素是对应行元素的最大值, 说明期刊内部知识循环流动强劲, 倾向于知识内部利用与重构 [26, 37]. 例如, 受期刊 *Phys Rev C* 的影响, 期刊 *Rev Mod Phys* 被 *Phys Rev C* 引用的比例为 0.65, 超过了其自被引率 (0.04).



(a) 二阶引文网络拓扑图

(b) 以期刊RMP为中间节点的二阶引文网络转移概率矩阵

(c) 二阶引文网络桑基图

图 2 二阶引文网络示意图

(Figure 2 Second-order citation networks)

基于图 2(a) 和图 2(b), 二阶网络中期刊间的知识流动情况如图 2(c) 所示. 可以发现, 大部分期刊的研究聚焦于某一个或几个方向, 其内部存在较强的知识交流, 跨领域间的交流合作较少. 如聚焦于粒子、场、重力与宇宙学等物理学领域的 *Phys Rev D*, 在该期刊向自身渗透了 85% 的知识的前提下, 其中 87% 的知识又流回了 *Phys Rev D*, 即 $P(PRD|(PRD|PRD)) = 0.87$, 说明该期刊内部形成了非常密切的知识流动体系. 进一步观察可以发现, 图 2(c) 中存在大量 “上凸下凹” 的现象. 例如, 由 *Phys Rev B* 输入到 *Phys Rev Lett*, 并再次被 *Phys Rev B* 引用的路径展示了明显的 “上凸” 的现象. 在期刊 *Phys Rev B* 对 *Phys Rev Lett* 有知识流输入的前提下, 其中 70% 的知识扩散回 *Phys Rev B*, 即 $P(PRB|(PRL|PRB)) = 0.70$, 说明一篇刊登在 *Phys Rev Lett* 且参考文献主要来自凝聚态物质与材料物理研究领域的论文, 引用它的文献也将主要来自该领域中的其它论文. 这一发现较好地验证了文献 [38] 的结论, 即发表在 *Nature* 上的论文主要被其参考文献所属学科领域中的论文引用. 考虑文献来源都为 *Phys Rev B* 的条件下, 论文被其他期刊引用后再次被 *Phys Rev B* 引用的概率, 以 *Phys Rev B* 自身, 以及 *Rev Mod Phys* 和 *Phys Rev Lett* 为中间节点的文献在最终 *Phys Rev B* 总引用来源的比例分别为 0.53、0.25 和 0.20. 如果只使用一阶网络进行建模分析, *Phys Rev B* 对 *Rev Mod Phys* 和 *Phys Rev Lett* 的引用率分别为 0.03 和 0.34, 会得到与 *Rev Mod Phys* 相比, 来自 *Phys Rev Lett* 的知识流比重更大的错误结论, 这也体现了在文献数据挖掘过程中, 使用高阶网络建模与分析的必要性.

图 3 展示了以 $P(v_k|(RMP|v_i))$ 为元素的二阶网络的转移概率矩阵与一阶转移概率矩阵 RMP 行元素的差值. 可以发现, 超过 25% 的元素值大于 0.1 (或者小于 0.1), 两种不同阶数网络的知识流动规律和模式表现出较大差异. 如在期刊 *Phys Rev Lett* 对 *Rev Mod Phys* 有知识输入的前提下, 二阶网络中期刊 *Rev Mod Phys* 发表的文献被 *Phys Rev B* 引用的比例增加了 22%, 即 *Rev Mod Phys* 发表的文献有 60% 都会被 *Phys Rev B* 引用. 这一比例远高于仅分析从 *Rev Mod Phys* 到 *Phys Rev B* 的一阶引用关系 (38%), 表明发表在 *Phys Rev Lett* 的文献通过 *Rev Mod Phys* 引用后对 *Phys Rev B* 的知识转移与渗透能力较强. 相比而言, 当 *Rev Mod Phys* 受 *Phys Rev C*、*Phys Rev E*、*Phys Rev D*、*Phys Rev Accel Beams* 输入的影响时, *Rev Mod Phys* 被 *Phys Rev B* 引用的比例大幅下降, 最大值为 0.38. 然而, 这四个前序期刊在一阶引文网络中对应行的主对角元素是最大值, 说明 *Phys Rev C*, *Phys Rev E*, *Phys Rev D*, *Phys Rev Accel Beams* 分别在其内部知识交流密切, 而跨领域间的知识联系并不明显.

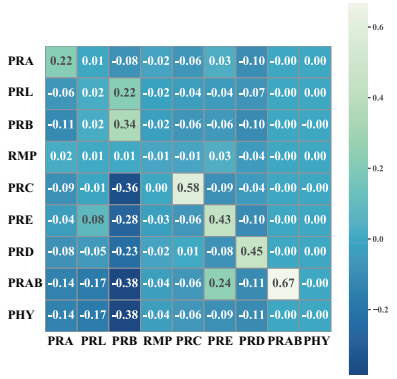


图 3 以期刊 RMP 为中间节点的转移概率差值矩阵示意图

(Figure 3 Transition probability difference matrix with RMP as middle node)

3.3 基于引用多样性信息熵的网络性能比较

本部分通过计算本文所提基于引用多样性信息熵的高阶网络信息增益指标来比较一阶与二阶引文网络在知识流动表示方面的性能差异. 首先, 基于本文收集的文献引用数据, 由式 (2.7)、(2.8)、(2.10) 和 (2.11) 计算可得一阶和二阶引文网络的引用多样性信息熵, 值分别为 2.32 和 0.92. 进一步通过式 (2.12) 计算得出二阶网络表示增益为 1.40, 这说明与一阶引文网络相比, 二阶网络表示结构的随机性减少, 增加了 1.40 个单位的有效高阶信息. 最后, 由式 (2.13) 计算每种期刊在两个网络中的引用多样性信息熵值和增益, 数值如表 3 所示. 其中, 期刊 *Rev Mod Phys* 和 *Phys Rev Lett* 的二阶网络表示增益均大于 0.3; 期刊 *Rev Mod Phys* 的增益最大, 值为 0.39, 说明其知识流动规律在两种网络中呈现出一定差异. 通过计算以不同期刊为中间节点的各概率差值矩阵的方差 (见表 3 第五列), 可以发现期刊 *Rev Mod Phys* 和 *Phys Rev Lett* 对应的值最大, 说明这两种期刊的知识流动规律在两种网络中呈现出较大差异, 同时反映出本文所提基于引用多样性信息熵的高阶网络信息增益指标可以准确地判定不同阶数的网络间的差异度.

在一阶引文网络中, 期刊 *Rev Mod Phys* 被 *Phys Rev Lett* 和 *Phys Rev B* 引用的概率比约为1:2; 然而在二阶引文网络中, 受 *Phys Rev Lett* 的影响, 存储在 *Rev Mod Phys* 中的知识

流向 *Phys Rev Lett* 和 *Phys Rev B* 的比例约为1:3. 这说明引用了 *Phys Rev Lett* 的文献更倾向于输出知识到 *Phys Rev B*, 反映出受 *Phys Rev Lett* 影响, 期刊 *Rev Mod Phys* 具备很强的向 *Phys Rev B* 传递知识的能力. 综上所述, 如果只使用一阶引文网络进行建模分析, 会忽略多种期刊间频繁而复杂的知识流动现象, 而仅得到两两期刊间过于简单的知识流动规律, 最终导致对现实引文复杂网络系统中知识流动整体认知的偏差.

表 3 高阶网络表示信息增益

(Table 3 High-order network representation information gain)

期刊名称	一阶引文网络下的引用多样性信息熵	二阶引文网络下的引用多样性信息熵	信息增益	概率差值矩阵方差
PRA	0.23	0.08	0.15	0.01
PRL	0.60	0.30	0.30	0.03
PRB	0.36	0.20	0.16	0.01
RMP	0.60	0.21	0.39	0.03
PRC	0.14	0.02	0.12	0.01
PRE	0.12	0.03	0.09	0.01
PRD	0.27	0.07	0.20	0.02
PRAB	0	0	0	0
PHY	0	0	0	0
总和	2.32	0.92	1.40	0.12

4 结束语

传统网络建模方式基于一阶马尔可夫假设, 认为网络中节点间的关联关系只与最近邻的两个节点有关, 导致难以捕捉节点间的间接依赖关系, 对现实系统的准确表征能力不足. 因此, 为提高网络分析结果的精度, 本文在传统的一阶引文网络中引入了二阶马尔可夫依赖, 构建高阶引文网络马尔可夫模型, 并从引文网络知识流动视角探究两种网络的差异. 首先, 本文构建了基于引文大数据的二阶引文网络马尔可夫模型, 将一阶引文网络中的边重构为二阶引文网络中的节点, 改变了传统网络模型以个体为节点的做法. 其次, 本文定义了基于引用多样性信息熵的高阶网络信息增益指标, 以此衡量期刊间知识流动效应在低阶和高阶引文网络中的差异, 实现了对低阶网络建模方式信息损失的全方位和多角度量化. 最后, 本文使用网络拓扑图和桑基图对一阶和二阶引文网络进行可视化, 直观地展示了期刊间的知识流动过程.

研究发现, 本文构造的基于引用多样性信息熵的高阶网络信息增益指标可以有效表征期刊在低阶引文网络中的高阶信息损失量. 如果期刊的信息增益较大, 则说明其在两种网络中的知识流动规律存在较大差异; 因此, 有必要在已有的引文数据基础上构建高阶网络来捕捉期刊间的间接关系, 实现对真实知识流动过程的整体认知. 以本文使用的引文数据为例, 期刊 *Rev Mod Phys* 的二阶网络表示增益最大, 说明其知识流动规律在两种网络中存在明显差异. 进一步分析发现, 受前序期刊 *Phys Rev Lett* 的影响, 存储在 *Rev Mod Phys* 中的知识更倾向于输出到 *Phys Rev B* 中, 说明以期刊 *Phys Rev Lett*、*Rev Mod Phys*、*Phys Rev B* 形成的知识社区的交流活动更为活跃. 期刊 *Rev Mod Phys* 的知识转移和渗透的能力较高, 其发展主要需要来自期刊 *Phys Rev B* 和 *Phys Rev Lett* 的知识交叉融合. 本文仅以引文网络为

例, 对二阶网络马尔可夫模型进行了研究, 未来工作可以考虑扩展研究领域, 根据不同领域的数据集构建网络进行对比分析, 以及针对三阶或四阶等复杂网络系统开展研究.

参 考 文 献

[1] Milgram S. The small world problem. *Psychology Today*, 1967, **2**(1): 60–67.

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

[3] Barabási A L, Albert R. Emergence of scaling in random networks. *Science*, 1999, **286**(5439): 509–512.

[4] Barabási A L, Albert R, Jeong H. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and Its Applications*, 1999, **272**(1–2): 173–187.

[5] Barrat A, Barthélemy M, Vespignani A. Weighted evolving networks: Coupling topology and weight dynamics. *Physical Review Letters*, 2004, **92**(22): 228701.

[6] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks. *Science*, 2002, **297**(5586): 1551–1555.

[7] Ravasz E, Barabási A L. Hierarchical organization in complex networks. *Physical Review E*, 2003, **67**(2): 026112.

[8] 周涛, 柏文洁, 汪秉宏, 等. 复杂网络研究概述. 物理, 2005, (1): 31–36.
(Zhou T, Bai W J, Wang B H, et al. A brief review of complex networks. *Physics*, 2005, (1): 31–36.)

[9] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用. 北京: 清华大学出版社, 2006.
(Wang X F, Li X, Chen G R. The Theory and Applications in Complex Networks. Beijing: Tsinghua University Press, 2006.)

[10] 何大韧, 刘宗华, 汪秉宏. 复杂系统与复杂网络. 北京: 高等教育出版社, 2009.
(He D R, Liu Z H, Wang B H. Complex Systems and Complex Networks. Beijing: Higher Education Press, 2009.)

[11] Xu J, Wickramarathne T L, Chawla N V. Representing higher-order dependencies in networks. *Science Advances*, 2016, **2**(5): e1600028.

[12] Lambiotte R, Rosvall M, Scholtes I. From networks to optimal higher-order models of complex systems. *Nature Physics*, 2019, **15**(4): 313–320.

[13] Scholtes I. When is a network a network? Multi-order graphical model selection in pathways and temporal networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, 1037–1046.

[14] Scholtes I, Wider N, Garas A. Higher-order aggregate networks in the analysis of temporal networks: Path structures and centralities. *The European Physical Journal B*, 2016, **89**(3): 1–15.

[15] Benson A R, Gleich D F, Leskovec J. Higher-order organization of complex networks. *Science*, 2016, **353**(6295): 163–166.

[16] Yin H, Benson A R, Leskovec J, et al. Local higher-order graph clustering. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, 555–564.

[17] Yin H, Benson A R, Leskovec J. Higher-order clustering in networks. *Physical Review E*, 2018, **97**(5): 052306.

[18] Scholtes I, Wider N, Pfitzner R, et al. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications*, 2014, **5**(1): 1–9.

[19] Saebi M, Xu J, Kaplan L M, et al. Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. *EPJ Data Science*, 2020, **9**(1): 15.

[20] 张艺蔓, 马秀峰, 程结晶. 融合引文内容和全文本引文分析的知识流动研究. 情报杂志, 2015, **34**(11): 49–54.
(Zhang Y M, Ma X F, Cheng J J. Research of knowledge flows based on citation content analysis. *Journal of Intelligence*, 2015, **34**(11): 49–54.)

[21] 王伟, 杨建林. 基于引文网络重叠社团发现的图书情报领域学科主题结构分析. 情报学报, 2020, **39**(10): 1021–1033.
(Wang W, Yang J L. Mapping the subject structure of library and information science through overlapping community detection in citation network. *Journal of the China Society for Scientific and Technical Information*, 2020, **39**(10): 1021–1033.)

[22] https://journals.aps.org/datasets.

[23] 吴海峰, 孙一鸣. 引文网络的研究现状及其发展综述. 计算机应用与软件, 2012, **29**(2): 164–168.
(Wu H F, Sun Y M. On status quo of citation network research and the overview on its development. *Computer Applications and Software*, 2012, **29**(2): 164–168.)

[24] 吴琼, 王如松, 李宏卿, 等. 土地利用/景观生态学研究中的马尔可夫链统计性质分析. 应用生态学报, 2006, (3): 3434–3437.
(Wu Q, Wang R S, Li H Q, et al. Statistical properties of Markov chain in land use and landscape study. *Chinese Journal of Applied Ecology*, 2006, (3): 3434–3437.)

[25] 刘臣, 单伟, 于晶. 中国学科知识网络的演化研究 —— 基于 1981–2010 年引文数据. 系统工程理论与实践, 2013, **33**(2): 430–436.
(Liu C, Shan W, Yu J. Evolution of discipline knowledge network in China — Empirical data from 1981 to 2010. *Systems Engineering — Theory & Practice*, 2013, **33**(2): 430–436.)

[26] 岳增慧, 许海云. 学科引证网络知识扩散特征研究. 情报学报, 2019, **38**(1): 1–12.
(Yue Z H, Xu H Y. Characteristics of knowledge diffusion in disciplinary citation network. *Journal of the China Society for Scientific and Technical Information*, 2019, **38**(1): 1–12.)

[27] Zhuge H. A knowledge flow model for peer-to-peer team knowledge sharing and management. *Expert Systems with Applications*, 2002, **23**(1): 23–30.

[28] Strelioff C C, Crutchfield J P, Hübler A W. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Physical Review E*, 2007, **76**(1): 011106.

[29] Shannon C E. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 2001, **5**(1): 3–55.

[30] 刘子琦, 郭炳晖, 程臻, 等. 基于熵值模糊层次分析法的科技战略评价. 计算机科学, 2020, **47**(S1): 1–5.
(Liu Z Q, Guo B H, Cheng Z, et al. Science and technology strategy evaluation based on entropy fuzzy AHP. *Computer Science*, 2020, **47**(S1): 1–5.)

[31] Baez J C, Fritz T, Leinster T. A characterization of entropy in terms of information loss. *Entropy*, 2011, **13**(11): 1945–1957.

[32] Lu X, Horn A L, Su J, et al. A universal measure for network traceability. *Omega*, 2019, **87**: 191–204.

[33] Dai J, Xu Q, Wang W, et al. Conditional entropy for incomplete decision systems and its application in data mining. *International Journal of General Systems*, 2012, **41**(7): 713–728.

[34] Jaffe A B, Trajtenberg M, Henderson R. Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 1993, **108**(3): 577–598.

[35] Van Leeuwen T, Tijssen R. Interdisciplinary dynamics of modern science: Analysis of cross-disciplinary citation flows. *Research Evaluation*, 2000, **9**(3): 183–187.

[36] Ma R, Yan E. Uncovering inter-specialty knowledge communication using author citation networks. *Scientometrics*, 2016, **109**(2): 839–854.

[37] Yan E. Disciplinary knowledge production and diffusion in science. *Journal of the Association for Information Science and Technology*, 2016, **67**(9): 2223–2245.

[38] Gates A J, Ke Q, Varol O, et al. Nature's reach: Narrow work has broad impact. *Nature*, 2019, **575**(7781): 32–34.