

科研大数据质量控制研究:角色定位、演化博弈与实施策略

佟泽华 陈兆娟 许海云 薛晓娜 韩春花

【摘要】[目的/意义]数据污染、数据风险、数据壁垒等数据质量隐忧,影响着科研工作者的科技创新步伐,科研大数据质量控制等问题的研究迫在眉睫。[方法/过程]文章在对科研大数据质量控制困境进行分析的基础上,对科研大数据质量控制主体进行了角色定位,进而从收益和成本角度出发,构建了科研大数据质量控制的“三方”演化博弈模型,对参与博弈的科研大数据提供者、科研大数据消费者、科研大数据监管者的行为策略进行了博弈分析,并利用 Matlab 对三方的策略选择进行仿真,最后提出了科研大数据质量控制策略。[结果/结论]科研大数据质量控制演化博弈模型由科研大数据提供者、科研大数据消费者、科研大数据监管者三方主体构成,通过模型分析与仿真得出了博弈三方的演化稳定策略,提出了“基-翼”相济式优化策略、“响应式-联盟化”管控策略、全过程技术融合策略、人才-平台双驱联动策略四位一体的科研大数据质量控制实施策略体系。

【关键词】科研大数据质量控制;角色定位;“三方”演化博弈;实施策略

【作者简介】佟泽华(1975-),男,山东理工大学信息管理学院教授,博士生导师;陈兆娟(1992-),女,山东理工大学信息管理学院硕士研究生(淄博 255000);许海云(通讯作者)(1982-),女,山东理工大学管理学院教授,博士生导师(淄博 255049);薛晓娜(1994-),女,山东理工大学信息管理学院硕士研究生(淄博 255000);韩春花(1979-),女,山东理工大学管理学院副教授(淄博 255049)。

【原文出处】《情报资料工作》(京),2024.3.18~30

【基金项目】本文系国家自然科学基金面上项目“基于弱信号时效网络演化分析的变革性科技创新主题早期识别方法研究”(批准号:72274113)、山东省泰山学者工程(项目编号:tsqn202103069)的研究成果之一。

科研大数据的质量是其价值实现的关键,高质量的科研大数据能够将数据资源转化为巨大的科学、经济和社会效益。2018年,我国发布了《科学数据管理办法》^[1],提出要“建立科学数据质量控制体系”。同年5月,中科院启动“十三五”信息化专项科研大数据工程,旨在为国家、科研人员提供更多高质量、高价值科研大数据^[2]。伴随着新兴技术与数字经济的融合发展,数据体量不断增加,数据冗余、数据标准缺失、数据更新缓滞、数据安全风险较高等问题不断涌现,阻碍了科研大数据开放与共享的进程。基于此,本文在对科研大数据质量控制困境分析的基础上,从收益和成本角度出发,构建了科研大数据质量控制的演化博弈模型,探讨了博弈主体不同策略选择对科研大数据质量的影响,并给出质量控制策略以促进科研大数据的应用与科研创新。

1 相关理论概述

1.1 科研大数据

科研大数据(Scientific Research Big Data)衍生

于科研工作者对“大数据”的研究。Demchenko等^[3-4]指出,科研大数据是与科学研究相关且反映复杂的自然和社会科学现象与关系的大数据类型。王胜开和马欣^[5]认为,科研大数据已经成为国家核心竞争力的重要组成部分以及“数据密集型”科学研究范式下不可缺少的工具。薛晓娜等^[6]认为,科研大数据以数据生产者、数据传播者、数据消费者、数据监管者为主要参与主体,其流转过程呈现出“链式结构”。佟泽华等^[7-8]认为,科研大数据的核心活动包括“初生”“再生”“共生”等,是科研创新活动的“养料”和“源泉”。

综上,本文认为科研大数据是指在大数据环境下进行自然科学、社会科学等科研活动所产生的数据集合,是科研创新活动的源泉和实现科技强国战略的新动能,其具有大数据本身的特征外,还表现出结构复杂性、高维度、高价值密度等显著特征。

1.2 科研大数据质量控制

国际标准化组织发布的《ISO/IEC 25012-2008》中将数据质量按属性划分为数据质量固有属

性、系统相关数据质量属性、固有与系统相关关联属性三大类^[9]。关于数据质量的研究出现了诸多代表性观点,Sukumar等^[10]在宏观层面上认为数据质量主要指数据资源的准确性,数据资源的准确性与数据价值息息相关,而数据价值又是数据质量的外在体现;丁小欧等^[11]指出数据的时效性、一致性、完整性、精确性是影响数据质量的关键特性;Wahyudi等^[12]认为数据质量已是当前数据资源应用的突出挑战,与数据消费者对数据资源的期望、满意程度相关。综上可看出,数据质量并非由某单一维度而定,而是多个维度相互作用、共同决定,其特性主要表现为准确性、可靠性、及时性、完整性、一致性、规范性等多个方面。

科研大数据作为当下新型战略性资源,其质量的优劣至关重要。为了提升科研大数据质量水平,不同的学者从不同角度进行了分析。李晓辉^[13]从技术、组织与服务方式、机构与人员配置等方面探讨了科研大数据管理模式;李斌等^[14]提出了建设科研大数据综合服务平台,整合科研大数据资源,实现其标准化管理;邹亚霏^[15]从认知、机制和管理模式等方面分析了科研大数据质量控制的问题,认为“增强管理认知”“加快体系构建”“升级管理模式”是实现科研大数据有效管理的可行性方案。

根据科研大数据质量特性及科研大数据质量问题,本文将科研大数据分为高品质科研大数据和低品质科研大数据两类,其中高品质科研大数据主要包括数据准确、安全可靠、更新及时、完整无差、表达规范等的数据;低品质科研大数据则是指数据缺失或重复、存在使用风险、更新滞后等无序、价值密度低的数据。科研大数据作为当下数据密集型科学的关键要素,其质量问题与科技创新率、数字经济建设等休戚相关,已有研究多从科研大数据质量控制过程中的平台建设、标准服务或科研大数据管理服务模式等方面展开,缺乏一个科学的、完整的科研大数据质量控制体系。

1.3 演化博弈理论

演化博弈理论是研究博弈主体动态化策略选择的基础理论。近年来,专家学者广泛应用该理论于数据质量控制研究中。刘文奇^[16]利用演化博弈方法,构建了数据质量变权综合评估模型;王娟和李玉海^[17]构建了数据监管者与数据消费者的演化博弈模型,推演出政府大数据质量控制的均衡策略;刘晓婷等^[18]以大数据时代为环境背景,研究了科研数据在信任机制视角下的数据共享行为。

关于科研大数据质量控制的演化博弈研究,相关学者只是初步探讨了数据提供者与数据消费者之间的博弈行为,而对于科研大数据质量控制的重要参

与者——数据监管者,并没有过多的研究。数据提供者、数据消费者、数据监管者是科研大数据质量控制体系建设中最主要的参与者:(1)科研大数据提供者主要负责生成、采集或发布科研大数据,是科研大数据资源的供给方,主要包括提供科研大数据的科研院所、数据平台、企业、个人等;(2)科研大数据消费者主要指通过数据平台、数据中介商获取并利用科研大数据实现某种目的的数据需求方,包括科研院所、企业、科研团队或个人等;(3)科研大数据监管者是指对科研大数据资源的发布、交易与使用等流程进行审核、管理的政府机构或第三方监管平台。参与博弈的数据提供者、数据消费者及数据监管者均为有限理性人,三方活动主要受自身利益影响。探讨三方在科研大数据质量控制中的动态博弈行为,有利于推导各要素对科研大数据质量控制效率的影响度,进而为构建科学合理的科研大数据质量控制体系提供可行性参考。

2 科研大数据质量控制困境

2.1 科研大数据质量问题多样化

科研大数据作为大数据资源的重要组成部分,其体量及更新速率都有着飞速转变,庞大的数据资源带来海量数据红利的同时,也滋生了数据污染、数据赘余或缺漏、数据烟囱、数据风险等质量问题,并呈现出多样化的特点:(1)数据污染:当科研大数据的准确性与完整性遭到破坏时,数据的价值就会降低,数据造假、数据赘余或缺漏都会导致数据污染问题^[19]。例如,论文数据造假导致的数据污染现象时有发生,在医学领域,数据污染的现象较多^[20],其带来的后果也甚为严重。据Nature报道称,生物医学SCI数据库PubMed(共计约3400万生物医学文献),至少有1%的论文可能出自论文工厂^[21]。(2)数据壁垒:也称为数据烟囱,是指科研机构或个人缺乏协同合作而导致的数据质量问题^[22],数据壁垒会导致科研大数据主体之间的数据互操作性严重缺失,进而影响科研效率。随着网络安全、数据安全问题的全球范围内的重要性凸显,数据在成为新生产要素的同时,数据竞争和数据保护制度也正在成为一种新的“壁垒”^[23],尤其是在高校及研究机构进行科研创新活动过程中,数据的开放严重不足,数据壁垒的现象显著^[24]。(3)数据风险:主要是指数据隐私、机密泄露等风险,监管缺失、个人数据保护意识不足、技术漏洞等导致数据盗用、篡改、泄露等问题频发,数据质量问题愈加严重^[25]。例如,2016年数据泄露水平指数报告中指出,2016年数据泄露事件导致近14亿条记录外泄,其中科技部门数据泄露高达3.9亿条^[26]。此外,数据权益模糊、数据更新之后数据来源不明等使得科研大数据

的准确性、及时性、可靠性等受到威胁。科研大数据质量问题层出不穷,迫切需要控制科研大数据质量。

2.2 科研大数据质量控制体系不完善

(1) 科研大数据质量控制意识不足。科研大数据质量控制体系建设初期,利益相关者的质量控制意识尚存在不足之处。对于科研大数据监管者来说,科研大数据更新速度快、结构复杂,传统监管模式并不适用于科研大数据的质量控制体系。对于科研大数据提供者而言,为追求利益最大化或由于人员、技术储备不足,使得相关人员缺乏对高品质科研大数据的认知与追求,从而从源头上影响了科研大数据质量。对于科研大数据消费者而言,不同渠道获得的科研大数据质量不一,相关人员缺乏对高品质科研大数据的鉴别能力,进而影响科研大数据质量。

(2) 缺乏统一的科研大数据质量标准体系。科研大数据质量是衡量科研大数据质量的硬性指标,目前,我国尚未形成统一的科研大数据质量标准。从科研大数据流通过程看,科研大数据的采集、处理、分析、应用、分解等缺乏统一的标准规范,由于数据流通过程中的技术不一致、人员的知识储备不同,由此产生的科研大数据结构不一、格式多样,进而影响了科研大数据的应用。从科研大数据类型来看,不同领域、不同学科之间尚存在数据鸿沟,缺乏统一的质量标准,影响了跨学科的数据交叉融合的进程。

(3) 相关政策落地具有局限性。科研大数据作为新兴的基础性科技资源,相关政策的提出与实现尚在起步阶段^[27]。一方面,我国科研大数据质量控制的相关政策不完善,不仅缺乏对科研大数据利益相关者质量控制意识的引导,而且对科研大数据质量风险的处罚政策也不完善。另一方面,科研大数据涉及的个人、企业、高校、科研院所等多种群体,相关政策实践落地困难且缺乏相应的监管体系。此外,科研大数据质量控制政策的细粒度不够^[28],针对不同领域、不同数据类型等的科研大数据质量控制政策缺乏,因而影响了科研大数据质量控制政策的实施。

(4) 基础设施建设不足。科研大数据质量控制的基础设施建设是实现科研大数据质量控制的前提条件^[29]。目前,我国科研大数据质量控制的基础设施建设不完善,尚未形成科学、高效的基础设施服务体系。从科研大数据平台建设来看,缺乏个性化的科研大数据服务,如科研大数据质量一览、科研大数据质量反馈、科研大数据质量咨询等。从科研大数据技术设施来看,人工智能、云计算、互联网+、区

块链等新兴技术与科研大数据质量控制的融合度不够,影响了科研大数据质量控制的效率。

3 科研大数据质量控制演化博弈模型构建

3.1 问题描述

有利益冲突的地方就有博弈,非公益性的科研大数据质量控制研究可以借助演化博弈理论展开。演化博弈理论在质量控制领域应用甚广,如国家审计质量控制^[30]、供应链质量控制^[31]等,本文参考、改进了以上质量控制模型,认为演化博弈角度下科研大数据质量控制影响因素主要是从利益相关者的角度出发,考虑利益驱动下的科研大数据质量优劣。经济学意义上的“利益驱动”主要由科研大数据利益相关者的收益-成本差来决定:

(1) 收益:是指利益相关者利用科研大数据在其流转与应用过程中所创造的直接或间接价值。收益因素分为直接收益和间接收益两个部分。直接收益:指利益相关者利用科研大数据所得到的价值产出。科研大数据的优劣程度是影响直接收益的主要因素,高品质科研大数据的价值转换率高,利益相关者所得到的收益就会高,相应的,低品质科研大数据的价值产出较低。但现实生活中,不排除数据提供者受利益驱动,为获得短期内的快速收益而将科研大数据以次充好的情况。数据消费者和数据监管者的收益则完全由其价值转化率决定。间接收益:与直接收益互为补充。数据消费者的间接收益包括反馈激励收益、罚金赔偿等;数据提供者的间接收益包括经营口碑、社会影响力、长期合作激励等;数据监管者的间接收益主要来源于罚款罚金所得、监管部门公信力、口碑等。间接收益对科研大数据质量的影响是潜移默化的。

(2) 成本:是指科研大数据活动中所要付出的总额,本文将科研大数据质量控制成本分为以下3个方面:时间成本:科研大数据体量庞大,结构复杂,科研大数据流通过程中数据提供者、数据消费者、数据监管者都需要付出时间来处理应用数据,即三方所付出的遴选、加工、应用、监管时间为时间成本。建设成本:科研大数据的建设成本主要指科研大数据质量控制体系的建设产出,如平台建设、技术优化投入、人员培养支出以及其他需要投入的人力物力财力等。风险成本:风险成本主要指科研大数据利益相关者在科研大数据活动中需要承担的风险。若风险成本过小,则会有更多的利益相关方“冒险”违规经营,不利于科研大数据质量优化。

3.2 科研大数据质量控制主体的角色定位

3.2.1 科研大数据监管者的角色定位

对于科研大数据监管者(简称为数据监管者或监管者)而言,作为科研大数据质量控制的参与者

及初期主导者,其监管行为及监管力度会间接影响科研大数据质量。监管的主体主要是不同层面的政府相关部门或第三方监督机构^[32]。在科研大数据质量控制体系建设初期,科研大数据监管者应发挥主流作用,积极引导构建科学的、健康的科研大数据质量控制体系。在此情况下,监管者应出台并宣传科研大数据质量控制政策、标准等,加大基础设施服务的完备性,搭建科研大数据质量控制主体间的沟通桥梁,加强数据质量立法,加大数据污染、数据造假等行为的惩罚力度等,通过提高数据提供者的风险成本来控制科研大数据质量。由此可以看出科研大数据监管者的角色有以下几种:

(1)初期主导者:在质量控制初期,负责制定与实施质量控制策略、规划,主导科研大数据的发展方向。“主导方向”可能是国家层面、区域层面或行业层面的,可以在不同层面引领科研大数据的发展,以保障科研大数据发展的方向性、准确性与可靠性。(2)审核者:科研大数据监管者需要对提供者的数据来源、数据可用性、数据可传播性等方面进行审核,以确保数据符合相关法律法规的要求。(3)管理者:科研大数据监管者为了保障整个控制流程的科学性,需要对数据收集、存储、处理等各环节负责,还要协调相关人员参与数据质量控制工作,持续监测数据质量。(4)监督者:科研大数据监管者需要对数据提供者提供的数据进行质量监督,确保数据真实可用。

3.2.2 科研大数据提供者的角色定位

对于科研大数据提供者(简称为数据提供者或提供者)而言,作为科研大数据资源的责任者、生成者与供给方,一方面,利益主导下的科研大数据流转具有较强的趋利性,即科研大数据提供者的行为受利益影响较大,故应该降低科研大数据的生产成本,如严格按照现有的政策、标准执行工作,充分利用监管者所提供的优惠政策,寻求技术扶持等。另一方面,可以通过自主改进设备、技术引进,建立联盟网络等提高生产、流通效率,进而提高收益,如扩大宣传,增加生产高品质科研大数据激励收益、声誉收益等。数据提供者的策略选择受风险成本影响,提供冗杂、造假、污染的科研大数据所受的声誉损失及罚金越大,数据提供者越倾向于提供高品质科研大数据。由此可以看出,科研大数据提供者的角色有以下几种:

(1)责任者:科研大数据提供者是保障数据质量的第一责任人,负责确保自己所提供数据的准确性、完整性、可靠性、一致性,一旦出现质量问题可追溯至科研大数据提供者。(2)生成者:科研大数据提供者需要具备一定的数据生产能力,能够产出

高质量、可靠的科研数据资源。科研大数据提供者通过科研大数据的初生(初次产生的新数据)、再生(旧数据衍生出的新数据)及共生(合作生成的数据)等核心活动,可以生成大量的科研数据。科研大数据提供者的数据核心活动是生成科研大数据的动力和源泉。(3)供给方:科研大数据提供者作为供给方需要为科研大数据消费者提供准确、完整、可靠的数据,不仅如此,还需要提供元数据文档,以便后续对数据进行修改与完善,并持续更新数据。

3.2.3 科研大数据消费者的角色定位

对于科研大数据消费者(简称为数据消费者或消费者)而言,其作为科研大数据红利的直接受益者与衡量者,科研大数据质量的高低与消费者所创造的价值休戚相关,科研大数据消费者的价值反馈行为对于提高科研大数据质量控制效率有着较强的影响力。通过加大基础设施建设力度,加强5G、大数据、人工智能、区块链等新兴技术在反馈中的应用与融合,增加科研大数据消费者反馈激励等措施,可以促进科研大数据价值的挖掘和发挥。科研大数据消费者的角色有以下几种:

(1)受益者:科研大数据消费者直接受益于科研大数据质量控制,高质量的科研大数据可以提高科研成果的准确性与可靠性,从而帮助科研机构、政府、企业、社会公众等在科研大数据的基础上得出更加客观、有效的结论,为社会的发展提供决策支撑。(2)衡量者:科研大数据消费者在数据分析和使用的过程中,可以对科研大数据质量进行合理的评价,为其他用户提供关于数据质量的反馈和建议,推动数据质量的不断提高。(3)再生者:科研大数据消费者在使用和分析数据的过程中,会在原来数据基础上产出新的数据资源,也可能会发现数据中存在的问题和不足,对数据进行修复和完善,实现数据再生。(4)需求方:科研大数据消费者是科研数据的需求方。为了确保产出科研成果的科学性与准确性,科研大数据消费者向提供者明确所需的数据类型、质量标准等要求,以确保获得高质量科研数据。

科研大数据质量控制主体的角色定位如下页图1所示。

3.3 演化博弈策略

科研大数据质量控制的博弈主体:科研大数据提供者、科研大数据消费者、科研大数据监管者,三方均为有限理性人。科研大数据提供者可能主动提供高品质科研大数据,也可能为节约成本以次充好提供低品质科研大数据,因此其行为策略集合为{高品质科研大数据,低品质科研大数据},主要依据

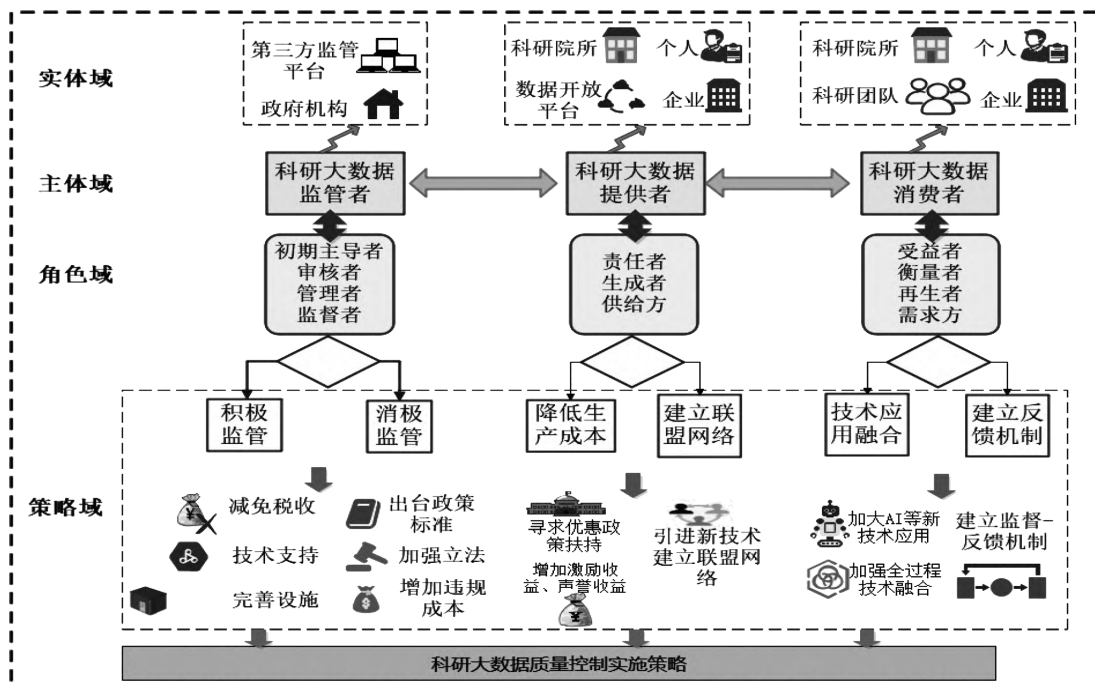


图1 科研大数据质量控制主体的角色定位

科研大数据带来的收益及违规操作的风险大小选择自身供给科研大数据的优劣性;科研大数据消费者是开发利用科研大数据的门户,可以依据自身利益得失选择是否参与反馈,故其策略集合为{反馈,不反馈};科研大数据监管者是科研大数据质量的管理者,主要对科研大数据的供给与需求行为进行监管,其数据监管行为需要投入一定的成本,因此数据监管者出于自身利益考虑可以选择公众参与且全面高效的积极监管策略或者经费投入较低的消极监管策略,则数据监管者的策略集合为{积极监管,消极监管}。

3.4 演化博弈模型的基本假设与参数设定

假设1:假设科研大数据提供者提供高品质科研大数据的概率为 x , $x \in [0,1]$,则提供低品质科研大数据的概率为 $1-x$;科研大数据消费者选择“反馈”策略的概率为 y , $y \in [0,1]$,则选择“不反馈”的概率为 $1-y$;科研大数据监管者选择“积极监管”的概率为 z , $z \in [0,1]$,则选择“消极监管”的概率为 $1-z$;三方均为有限理性人。

假设2:记科研大数据提供者提供低品质科研大数据的成本为 C_1 ,提供高品质科研大数据的成本记为 $C_1 + \Delta C$,通过流通科研大数据所获得的直接收入为 R_1 ; R_2 表示当数据提供者流通高品质科研大数据时,科研大数据消费者主动参与反馈可得的激励收益, C_2 表示数据消费者在相应平台、机构参与反馈的成本;科研大数据监管者选择消极监管的成本为 C_3 ,选择积极监管且消费者参与反馈时需额

外投入技术、场地、激励等的费用为 P ,监管机构固有的机构公信力、声誉、口碑等社会效益为 R_3 。

假设3:当科研大数据提供者选择流通高品质科研大数据且数据消费者选择给予正面反馈时,数据提供者会得到信誉增加、影响力提升等带来的额外收益 W ,当数据消费者选择“不反馈”时,其额外收益为0;当数据提供者选择流通低品质科研大数据时,数据消费者需承担损失 M ,若参与反馈且经数据监管者有效识别,则得到数据提供者给予的赔偿 K ,数据提供者需承担名誉损失 L 。

假设4:科研大数据监管者采取“积极监管”的策略下,若科研大数据消费者参与反馈则数据监管者对低品质科研大数据的识别率为100%,其余策略下数据监管者对低品质科研大数据的识别率为 β , $\beta \in [0,1]$;若数据监管者有效识别低品质科研大数据,数据提供者需承受数据监管者经济处罚 F 。

3.5 演化博弈模型构建

根据上述研究假设及参数设置,构建科研大数据质量控制演化博弈模型,收益分析如下。

(1)当博弈参与方的策略选择为{高品质科研大数据,反馈,积极监管}时,数据提供者的成本增加,由于数据消费者积极反馈,数据提供者声誉提高,其收益为 $R_1 - C_1 - \Delta C + W$,数据消费者的收益为 $R_2 - C_2$,数据监管者的收益为 $R_3 - C_3 - P$ 。

(2)当博弈参与方的策略选择为{高品质科研大数据,反馈,消极监管}时,数据提供者、数据消费者的收益与(1)中所述一致,数据监管者没有额外的

表1 演化博弈模型参数

符号	描述	符号	描述
C_1	供给方提供低品质科研大数据的成本	R_1	数据提供者的直接收益
ΔC	供给方提供高品质科研大数据的额外成本	R_2	高品质数据下数据消费者主动参与反馈的激励收益
C_2	数据消费者参与反馈的成本	R_3	监管机构固有的社会效益
C_3	数据监管者消极监管成本	P	消费者参与反馈时,数据监管者积极监管策略的额外支出
W	需求方参与正面反馈时数据提供者所获的声誉收益	M	数据消费者应用低品质资源所承担的损失
K	数据消费者得到供给方流通低品质资源的赔偿	L	数据提供者流通低品质资源的声誉损失
F	数据监管者针对数据提供者流通低品质资源的处罚	β	消极监管时及需求方不参与反馈时效率
x	数据方提供高品质科研大数据的概率	y	数据消费者选择“反馈”策略的概率
z	数据监管者选择“积极监管”的概率		

注:参数大小仅用于反应参与博弈的主体进行不同策略选择时的收益分析,以此来衡量各个参数对于科研大数据质量的影响程度,并不代表具体数值。

监管建设支出,其收益变为 $R_3 - C_3$ (消费者不反馈时数据监管者没有额外成本支出)。

(3) 当博弈参与方策略选择为{高品质科研大数据,不反馈,积极监管}时,数据提供者失去了其声誉增量,收益变为 $R_1 - C_1 - \Delta C$,数据消费者收益为 0,数据监管者收益为 $R_3 - C_3$ 。

(4) 当博弈参与方的策略选择为{高品质科研大数据,不反馈,消极监管}时,三方的收益与(3)所述一致。

(5) 当博弈参与方的策略选择为{低品质科研大数据,反馈,积极监管}时,数据提供者会因违规操作遭受监管部门处罚,给予数据消费者经济补偿,并承担声誉损失,其收益为 $R_1 - C_1 - L - F - K$,此时,数据消费者的收益为 $K - C_2 - M$,数据监管者的收益为 $R_3 - C_3 - P + F$ 。

(6) 当博弈参与方的策略选择为{低品质科研大数据,反馈,消极监管}时,数据提供者有 β 的概

率会被监管者识别,其收益为 $R_1 - C_1 - L - \beta F - \beta K$,数据消费者的收益为 $\beta K - C_2 - M$,数据监管者的收益为 $\beta R_3 - C_3 + \beta F$ 。

(7) 当博弈参与方策略选择为{低品质科研大数据,不反馈,积极监管}时,数据提供者收益为 $R_1 - C_1 - \beta F - \beta K$,此时,数据消费者收益为 $\beta K - M$,数据监管者收益为 $R_3 - C_3 + \beta F$ 。

(8) 当博弈参与方的策略选择为{低品质科研大数据,不反馈,消极监管}时,三方的收益与(7)所述一致。

综合以上分析,本文构建了三方收益矩阵,如表2所示。

表2 三方策略选择及收益矩阵

数据提供者	数据消费者	数据监管者	
		积极监管(z)	消极监管($1-z$)
高品质数据(x)	反馈(y)	$R_1 - C_1 - \Delta C + W$ $R_2 - C_2$ $R_3 - C_3 - P$	$R_1 - C_1 - \Delta C + W$ $R_2 - C_2$ $R_3 - C_3$
	不反馈($1-y$)	$R_1 - C_1 - \Delta C$ 0 $R_3 - C_3$	$R_1 - C_1 - \Delta C$ 0 $R_3 - C_3$
低品质数据($1-x$)	反馈(y)	$R_1 - C_1 - L - F - K$ $K - C_2 - M$ $R_3 - C_3 - P + F$	$R_1 - C_1 - L - \beta F - \beta K$ $\beta K - C_2 - M$ $\beta R_3 - C_3 + \beta F$
	不反馈($1-y$)	$R_1 - C_1 - \beta F - \beta K$ $\beta K - M$ $R_3 - C_3 - \beta F$	$R_1 - C_1 - \beta F - \beta K$ $\beta K - M$ $R_3 - C_3 + \beta F$

根据表2中的三方收益矩阵,可以求得科研大数据提供者提供高品质科研大数据、低品质科研大数据的期望收益 E_{x1} 和 E_{x2} 及科研大数据提供者的平均期望收益 $\bar{E}_x$ 分别为:

$$E_{x1} = yW + R_1 - C_1 - \Delta C$$

$$E_{x2} = zy(\beta F + \beta K - F - K) - \beta F - \beta K - yL + R_1 - C_1$$

$$\bar{E}_x = xy(W + L) - x\Delta C + (1-x)(F + K)(zy\beta - \beta - zy) - yL + R_1 - C_1$$

在求得科研大数据提供者的平均期望收入后进一步构造科研大数据提供者提供高品质科研大数据、科研大数据消费者在数据流转过程中参与反馈及科研大数据监管者采取积极监管策略时的复制动态方程:

$$F(x) = \frac{dx}{dt} = x(1-x)[y(W + L) + zy(F + K - \beta F -$$

$$\beta K) + \beta F + \beta K - \Delta C]$$

$$F(y) = \frac{dy}{dt} = y(1-y) [xR_2 - C_2 + (1-x)z(1-\beta)K]$$

$$F(z) = \frac{dz}{dt} = z(1-z) [(1-x)y(1-\beta)(R_3 + F) - yP]$$

4 科研大数据质量控制演化博弈模型解析

4.1 演化博弈模型分析

经过科研大数据提供者、消费者及监管者的博弈,最终何种策略为稳定策略不能直接进行判别,而科研大数据数据流转复制动态方程可以描述三者策略选择的变化速度与方向。当且仅当复制动态方程等于0,即科研大数据提供者、消费者及监管者的策略调整速度为0时,意味着博弈达到了一个相对稳定的状态。因此,令 $F(x) = 0$,即对 $F(x)$ 求偏导,求得:

$$\text{当 } z = \frac{y(L+W) - \Delta C + \beta F + \beta K}{y(\beta F + \beta K - F - K)} \text{ 时,所有水平处于}$$

稳定状态。通过均衡点稳定性分析,可知科研大数据提供者、消费者及监管者的演化博弈策略。由此,可以推论:

推论1:科研大数据提供者的策略选择不仅受提供优劣两种数据的成本差值、提供低品质科研大数据时的违规处罚与赔偿影响,还与科研大数据监管者的监管力度有关。

推论2:随着科研大数据监管者对经营低品质科研大数据的识别率与处罚增加,科研大数据提供者选择提供低品质科研大数据的比率下降。

推论3:当提供低品质科研大数据的风险低于其收益时,无论消极监管还是积极监管状态下,科研大数据提供者都会选择提供低品质科研大数据。

$$\text{对 } F(y) \text{ 求偏导,求得:当 } x = \frac{z(1-\beta)K - C_2}{z(1-\beta)K - R_2} \text{ 时,所有水平处于稳定状态。由此,可以推论:}$$

推论4:科研大数据消费者的策略选择与科研大数据监管效率有关,此外,消费者是否对科研大数据质量的优劣进行反馈与激励收益、反馈成本相关,科研大数据提供者提供低品质科研大数据被识别后基于消费者的经济赔偿也影响着消费者的策略选择。

推论5:当科研大数据提供者提供低品质科研大数据时,监管者对低品质科研大数据的识别率越高,消费者选择反馈策略的概率随着成本的增加而降低。

推论6:若科研大数据提供者选择提供低品质科研大数据,且监管者选择消极监管,消费者选择反馈策略的概率会随着反馈成本的提高而降低。

对 $F(z)$ 求偏导,求得:

$$\text{当 } x = 1 - \frac{1}{(R_3 + F)(1-\beta)} \text{ 时,所有水平趋于稳定。}$$

由此,可以推论:

推论7:科研大数据监管者的监管效率与其社会固有的影响力及对低品质科研大数据提供者的行政处罚有关。

推论8:随着消极监管对低品质科研大数据的识别率降低,监管者为了巩固其社会影响力及获得对低品质科研大数据提供者的行政处罚,监管者选择积极监管的概率会增加。

4.2 仿真分析

利用 Matlab 进一步研究科研大数据提供者、消费者及监管者采取不同策略时的相互影响关系。三方的初始策略演化仿真如图2,当 $x=0.6$ 时,提供者选择提供高品质科研大数据的收益最高,当 $y=0.7$ 时,科研大数据消费者参与反馈的收益最高,当 $z=0.5$ 时,科研大数据监管者选择积极监管的收益最高。

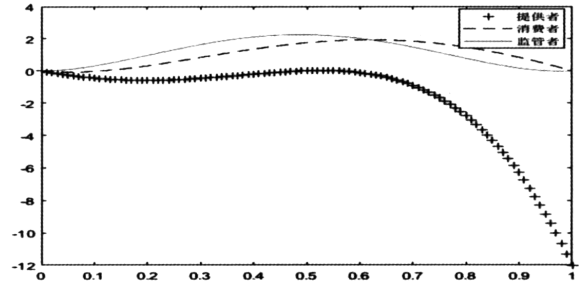


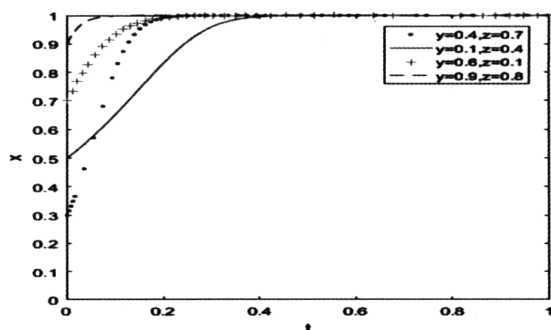
图2 数据提供者-数据消费者-数据监管者的三方演化趋势图

当参数变化时,参与博弈的三方演化策略如下:

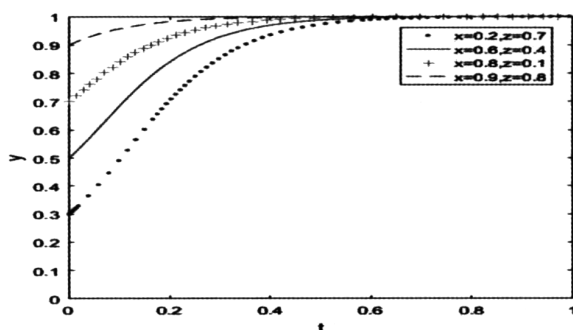
(1)当提供优、劣科研大数据的成本差小于提供低品质科研大数据所付出的风险成本 $(F+K)$ 时,令 $\Delta C=2, (F+K)=40, \beta=0.2$,如图3所示,在此条件下,提供者提供优、劣大数据的成本差值过小,科研大数据提供者会因想得到更高的声誉收益而选择提供高品质科研大数据,而科研大数据监管者节约监管成本会选择消极监管策略。

(2)当提供优、劣科研大数据的成本差大于提供低品质科研大数据的风险成本,即 $\Delta C > \beta(F+K)$,且 $R_2 > C_2$ 时,如图4所示,低品质科研大数据提供者会选择甘冒风险,生产低品质科研大数据资源。科研大数据消费者的演化稳定策略与其参与反馈的成本与收益有关,若参与反馈的成本过高,大于消费者参与反馈可得到的赔偿及损失之差,消费者就会选择不反馈。

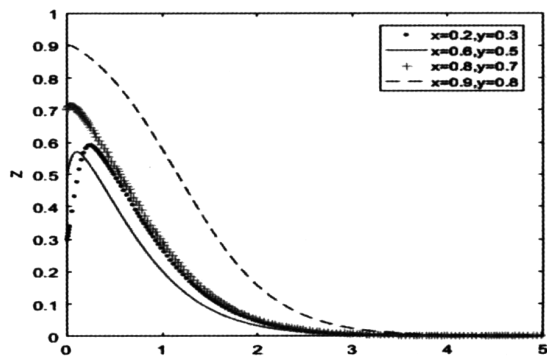
(3)当提供优、劣科研大数据的成本差大于提供



3(a) 提供者



3(b) 消费者



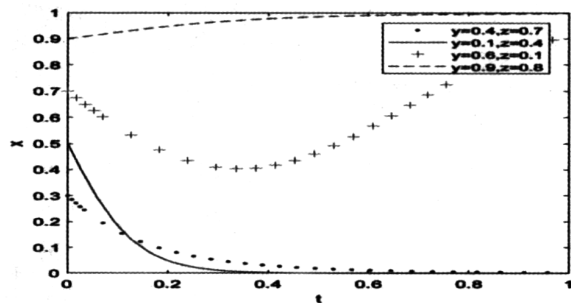
3(c) 监管者

图3 $\Delta C < \beta(F+K)$ 时,科研大数据演化策略仿真
低品质科研大数据的风险成本时,即 $\Delta C > \beta(F+K)$,
且 $R_2 < C_2$ 时,如下页图5所示,此时,科研大数据消
费者参与质量反馈的收益低于成本,因此,科研大
数据消费者会规避成本选择不参与反馈(与推论6
相呼应),此时,科研大数据提供者提供低品质科研
大数据的收益高于成本之和,故科研大数据提供者
会选择提供低品质科研大数据。此时,科研大数据
监管者的策略不稳定。

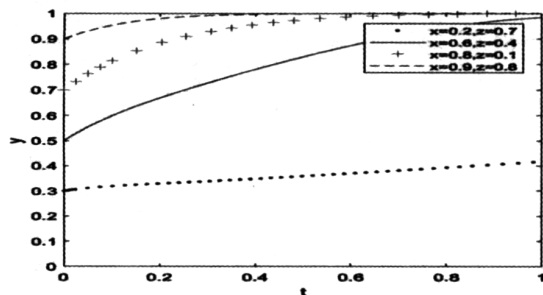
(4) 当 $\Delta C > F+K+W+L$ 时,如下页图6所示,当
科研大数据提供者生产优、劣两种科研大数据的成
本差远高于提供低品质科研大数据所需要付出的
风险成本时,此时,无论科研大数据监管者、消费
者采取何种策略,科研大数据提供者都会甘冒风险提

供低品质科研大数据(与推论3相呼应)。科研大
数据监管者为巩固其社会公信力选择积极监管策
略(与推论8相呼应),而科研大数据消费者因想挽
回一定的损失选择参与反馈。

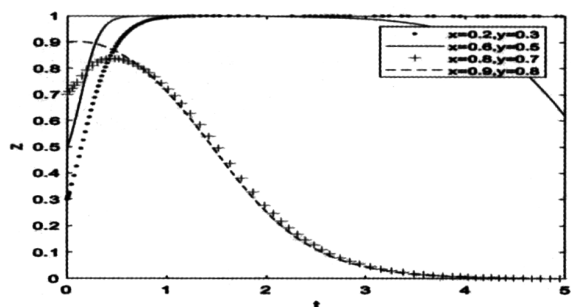
科研大数据质量受多方策略影响,对于科研大
数据提供者而言,其生产优、劣科研大数据的成本
差以及风险成本的增减会影响其策略选择,当风险
成本过高时,科研大数据提供者会因规避风险、求
得社会收益而选择提供高品质科研大数据。对于
科研大数据消费者而言,其策略选择与激励收益、
反馈成本、经济收益或损失相关,还与监管者的策
略选择相关。对于科研大数据监管者而言,其策略
选择与其社会公信力、监管效率相关,当科研大数
据提供者选择提供低品质科研大数据资源时,科研
大数据监管者的策略选择对其尤为重要。



4(a) 提供者



4(b) 消费者



4(c) 监管者

图4 $\Delta C > \beta(F+K)$ 且 $R_2 > C_2$ 时,
科研大数据演化策略仿真

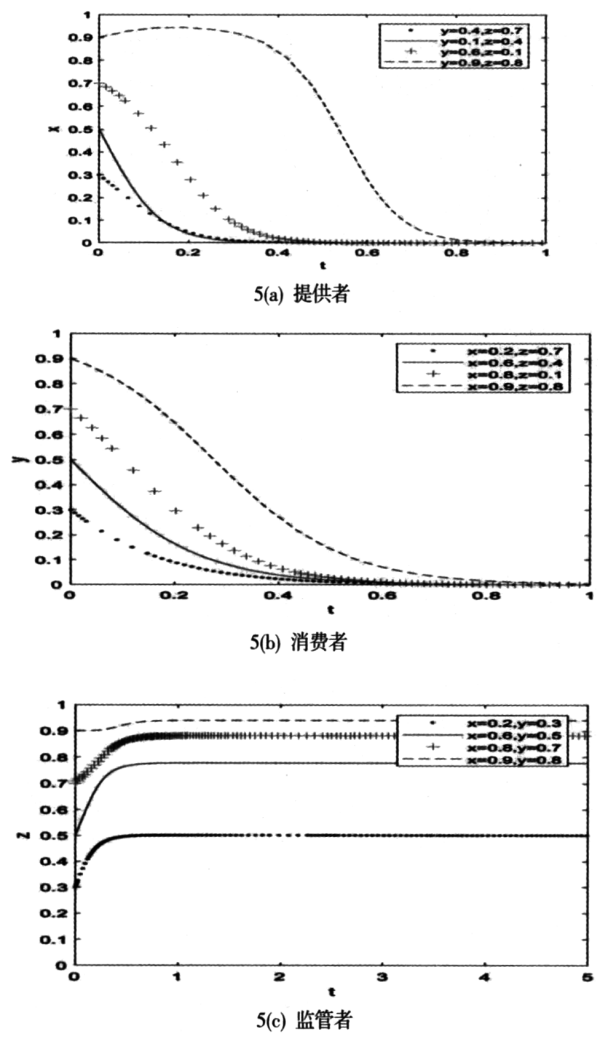


图5 $\Delta C > \beta(F+K)$ 且 $R_2 < C_2$ 时, 科研大数据演化策略仿真

4.3 演化博弈稳定策略

综合上述仿真分析,可以得出以下科研大数据质量控制的演化博弈稳定策略:

(1)对于科研大数据提供者而言,适当降低生产高品质科研大数据成本,提高对科研大数据提供者违规行为的处罚及赔偿力度是提升科研大数据质量的有效措施。一方面,在科研大数据质量控制演化博弈过程中,提供者的策略选择受成本影响较大,加大技术投入、建立科研大数据生产联盟网络等是降低优质科研大数据生产成本的有效举措。数据提供者提供高品质数据的目的是获取一定的收益,可通过增加生产高品质数据的激励收益、声誉收益等增加数据提供者的收益。另一方面,数据提供者的策略选择受风险成本影响,违规行为(如数据造假)所带来的声誉损失及罚金越大,则其更愿意提供高品质数据。

(2)降低科研大数据消费者的反馈成本是推进

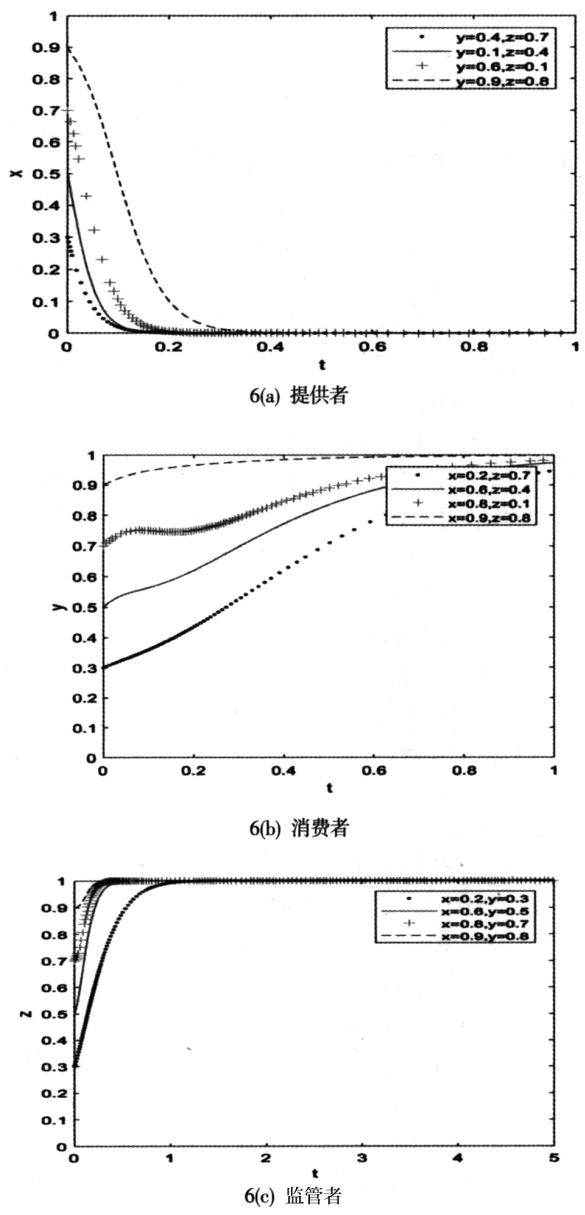


图6 $\Delta C > F+K+W+L$ 时,科研大数据演化策略仿真

监督-反馈机制建立的有效举措。科研大数据消费者作为科研大数据价值的直接创造者,科研大数据的质量与其利益有着较大的关联。科研大数据质量反馈有助于科研大数据质量控制体系的高效运行。反馈成本是影响科研大数据消费者策略选择的关键因素,若反馈成本过高,甚至高于科研大数据所创造的收入,消费者就会选择不反馈。

(3)科研大数据监管者应提升监管效率,杜绝或降低低品质科研大数据产生与流转的可能性。科研大数据监管者是数据质量控制主要参与者与核心主体,其监管效率与数据质量控制效率息息相关。政府部门应出台并宣传相关政策、标准等,加强数据质量立法,加大数据污染/造假等行为惩罚力

度等。第三方监管机构贯彻落实科研大数据质量控制政策、法律、标准等,加强宣传,构建数据提供者与消费者的沟通桥梁,辅助政府部门的科研大数据质量监管工作。

5 科研大数据质量控制实施策略

5.1 “基-翼”相济式质量优化策略

首先,要健全科研大数据质量控制的政策体系。一方面,加强国家层面的通用型政策的制定。我国已经制定《科学数据管理办法》,该办法对管理原则、管理职责、数据汇交/利用/安全等方面进行了引领性或指导性说明,为科研大数据质量控制和优化的发展指明了方向。但是,从国家层面还应进一步制定更为系统、细致的政策。另一方面,加强具体行业领域的政策制定。如针对工业大数据,工信部发布了《关于工业大数据发展的指导意见》^[33],还制定了《地球系统科学数据共享联盟章程》《地震科学数据共享管理办法》等。但存在领域政策覆盖不全面、独自一体的特点。在今后的实际中应坚持“国家顶层设计与行业领域实践相结合”“系统性与灵活性相结合”“通用性与专用性相结合”的实践策略,进行科研大数据质量政策体系的制定。

其次,要开展科研大数据质量标准化工作。完整、规范的科研大数据质量标准化体系是实现科研大数据质量提升的重要举措和方式。2017年,国家标准化管理委员会发布了《信息技术数据质量评价标准》^[34],涉及数据规范性、完整性、准确性、一致性、时效性、可访问性六大类评价指标。2018年,我国出台《大数据标准化白皮书》^[35],提出标准化工作的开展应围绕科研大数据流通过程逐步展开。我国积极开展数据标准化工作:国家标准化管理委员会发布了“关于征求《数据中心能效限定值及能效等级》强制性国家标准意见的通知”^[36],就数据建设平台相关问题征求意见。这些标准的实施为科研大数据质量的提升奠定了基础。

在优化科研大数据质量控制体系的过程中,应以科研大数据质量控制的政策体系为“基”,以科研大数据质量标准化为“翼”,二者相辅相成,更好地发挥科研大数据监管者的“初期主导者”“审核者”“管理者”等的角色价值,从而形成“基-翼”相济式优化策略,共同助力科研大数据质量的提升。

5.2 “响应式-联盟化”管控策略

(1)要建立监督-反馈-应急机制。要明确科研大数据权属,明确科研大数据提供者、消费者及监管者的主要权力与责任。此外,要建立行之有效的科研大数据质量问题风险应急机制,主要包括风

险预警及风险解决两个方面:一方面,相关部门应建立科研大数据质量应急机制,识别潜在的质量风险,另一方面,相关部门应着力构建科研大数据质量风险应急机制,以帮助数据监管者在科研大数据质量风险发生的第一时间给出科学的解决方案,提升数据质量监管效率。

(2)构建联盟化发展模式。科研个体开展协同合作研究,进而构建联盟化发展模式有利于提升科研大数据质量控制效率。一方面,可以促进领域内不同科研机构、企业、监管部门的协同合作,并缓解因缺乏合作导致的科研大数据质量控制成本过高的问题。另一方面,不同学科、不同领域的科研团队、企业、监管机构等加强协同合作,有利于碰撞出科研大数据质量控制的新方法、新范式,并实现多主体间的互利共赢。

“监督-反馈-应急机制”为科研大数据质量控制提供了良好的“响应机制”,而“联盟式发展模式”则为科研大数据的发展和提升提供了泛化的“数据网络基础”,二者相互促进,从而更好地发挥科研大数据消费者的“需求者”“衡量者”“再生者”等的角色作用,共同形成了“响应式-联盟化”管控策略。

5.3 全过程技术融合策略

(1)数据聚合与分析过程的技术融合。将大数据采集与分析技术、人工智能技术应用到科研大数据聚合与分析阶段的质量控制中,可以很好的满足相关技术需求。一方面,大数据采集与分析技术,有利于实现数据价值的深度挖掘,提升数据可用性。另一方面,充分利用人工智能技术可改变传统数据聚合与分析技术的低效、低质量等问题,实现“预测性”分析,提升数据可解释性。

(2)数据流转与应用过程的技术融合。通过运用具有高速率、大容量、低功耗、高可靠性、低时延、万物互联的5G技术,可满足科研大数据流转与应用过程的技术诉求。另外,科研大数据应用过程中尤其要确保数据的可用性与可解释性。人工智能技术可以更好的帮助科研人员改变传统的数据应用模式,重构科研大数据质量控制体系,使其更加智能、便捷,实现科研大数据的多样化应用。

(3)数据存储与分解过程的技术融合。云计算、云存储技术具有灵活度高、可靠性高、安全性强等特点,可应用于科研大数据存储。首先,云存储技术可容纳超大规模的数据存储,且存储容量动态可扩展,可保证科研大数据的全面性。其次,突破了时空界限,科研人员可实现跨时间、跨地域的数据存储。再者,支持用户动态访问,科研大数据的

提取、存储等过程简洁,不必困扰于繁杂的存储程序。最后,云存储技术具有较好的私密性与安全性,且数据分解过程操作便捷,有利于提升科研大数据质量控制效率。

(4)数据监测与管控过程的技术融合。人工智能、区块链等新兴技术的应用有助于解决重大的科研大数据质量问题。一方面,人工智能、卫星遥感技术的应用可丰富数据监测与管控方法,实现对科研大数据生产、组织、传播、消费、存储、分解等流程的实时监测与智能识别,减少或避免监测盲点,实现完全、无死角的科研大数据质量控制体系的建立。另一方面,将区块链技术应用到传统的科研大数据质量监测与管控过程中,可引导构建“公开透明”“全民监督”“公正严明”的科研大数据质量监管体系,确保数据质量。

数据聚合与分析过程技术融合一般是科研大数据生命周期中的成长期(前期)阶段,涉及主体主要是科研大数据提供者;数据流转与应用过程技术融合一般是在成熟期(中期)阶段,涉及主体主要是科研大数据消费者;数据存储与分解过程技术融合一般是在重生期或衰落期阶段(后期),该时期也常常是孕育数据“新生命”的时期,涉及的主体为科研大数据消费者与提供者;数据监测与管控过程技术融合则一般涉及生命周期的全过程等,涉及主体为监管者。可以看出,以上几方面的技术融合实现了数据生命周期全过程覆盖,由此共同构成了“全过程技术融合策略”,从而为“三方”主体角色作用的发挥提供有力支撑。

5.4 “人才-平台”双驱联动策略

(1)培养“跨行业-跨学科-多技能”复合型数据人才。复合型数据人才是科研大数据质量控制及科技创新的关键。复合型人才主要指在科研大数据应用与质量控制过程中掌握跨行业、跨学科的多种先进领域知识、关键技术与快速学习能力的人才。监管部门、科研院所、高校、企业等科研大数据利益相关者应加大复合型人才的培养力度,定期组织人员培训、学习,提高各科研人员的数据能力,使其快速适应人们对于科研大数据质量日益增长的要求,不断激发科技创新。

(2)加快“开放式-预警式”科研大数据平台建设。首先,科研数据平台的建设应优化布局,以满足科研大数据开放与应用需求。其次,要完善科研大数据平台的功能,建设集科研大数据质量评估、质量监测、质量反馈、风险预警、质量优化等于一体的科研大数据平台。再者,要丰富科研大数据平台的数据呈现方式以增加数据可用性。在平台建设

与完善过程中,多样化的科研大数据呈现方式有利于消费者对科研大数据的理解,提高科研产出效率。最后,平台建设应根据用户需求,智能化的提供相关领域、相关类型科研大数据平台的链接,以满足用户潜在的数据汇交需求。

综上所述,在科研大数据实践过程中,“跨行业-跨学科-多技能”复合型人才是“根”,是科研创新力量与科研大数据之“源”,而“开放式-预警式”数据平台是“支撑”,是“保障”,二者相互推动,共同形成了“人才-平台双驱联动策略”,以更好地为科研大数据提供者、消费者及监管者“三方”演化博弈主体角色作用发挥提供有力支撑。

6 结论与展望

6.1 研究结论与启示

(1)现阶段,科研大数据面临质量问题日趋严重、质量控制体系不完善等科研大数据质量控制困境。科研大数据的质量问题主要包括数据污染、数据烟囱、数据壁垒等。科研大数据质量控制体系存在质量控制意识不足、缺乏统一的科研大数据质量标准体系、相关政策落地具有局限性、基础设施建设不足等。

(2)科研大数据质量控制博弈的主要参与方有科研大数据提供者、科研大数据消费者、科研大数据监管者,其博弈行为主要受需要付出的成本(时间成本、建设成本、风险成本)和可以获得的收益(直接收益、间接收益)之间的差额决定。

当其生产高品质、低品质科研大数据的成本差值过小时,科研大数据提供者为避免风险成本、增加社会收益会选择提供高品质科研大数据。科研大数据监管者会因避免成本投入选择消极监管。此时,若科研大数据消费者参与反馈的成本高于可获得的激励,数据消费者的策略会趋向于不反馈科研大数据质量;若成本低于可获得的激励收益,数据消费者会选择反馈策略。

当提供高品质、低品质科研大数据的成本差大于提供低品质科研大数据的风险成本时,低品质科研大数据提供者会选择甘冒风险,此时,科研大数据消费者的策略选择与数据监管效率有关,若参与反馈的成本过高,大于数据消费者参与反馈可得到的赔偿及损失之差,数据消费者就会选择不反馈。

当提供高品质、低品质科研大数据的成本差大于提供低品质科研大数据的风险成本时,科研大数据消费者参与质量反馈的收益低于成本,因此,科研大数据消费者会规避成本选择不参与反馈,科研大数据提供者会选择提供低品质科研大数据。此

时,科研大数据监管者的策略不稳定。

当科研大数据生产高品质、低品质两种科研大数据的成本差远高于提供低品质科研大数据所需要付出的风险成本,无论科研大数据监管者、消费者采取何种策略,科研大数据提供者都会甘冒风险提供低品质科研大数据。科研大数据监管者为巩固其社会公信力选择积极监管策略,而科研大数据消费者因想挽回一定的损失选择参与反馈。

(3)本文提出的“四位一体”科研大数据质量控制策略体系有利于科研大数据质量的提升。主要涉及以下方面:“基-翼”相济式优化策略:加强政策法规研究,健全关于科研大数据质量的政策,积极开展科研大数据质量标准化工作,完善数据质量标准;“响应式-联盟化”管控策略:创新科研大数据质量控制模式,建立以“监督-反馈-应急”为形式的响应机制,促进协同合作,构建“联盟化”发展模式;全过程技术融合策略:相关人员应加快新兴技术的融合应用,促进科研大数据聚合与分析、流转与应用、存储与分解、监测与管控等多个阶段的技术融合,以实现在降低成本的条件下提升科研大数据质量控制效率;“人才-平台”双驱联动策略:培养复合型数据人才,加快科研大数据平台建设与完善,以实现科研大数据质量的高效控制。

6.2 研究展望

本文在已有研究的基础上,构建了三方演化博弈模型,分析了科研大数据提供者、科研大数据消费者、科研大数据监管者的质量控制策略选择,并推演出科研大数据质量的控制策略。但本文仍存在一定不足:本文只是对科研大数据质量的控制策略进行了初步探讨,许多问题待深入分析,例如,如何结合具体领域的科研大数据进行“领域科研大数据质量优化”,“领域科研大数据质量治理模型”如何构建等。在今后的研究中,将对这些问题进行深入剖析。本文希望通过对科研大数据质量控制的初步探索,达到抛砖引玉的效果,为国家的科技创新提供助力。

参考文献:

- [1] 国务院办公厅.《科学数据管理办法》[EB/OL]. [2018-03-17]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
- [2] 中国科学院.中科院“十三五”信息化专项科学大数据工程课题启动[EB/OL]. [2018-05-24]. http://www.cas.cn/yx/201805/t20180525_4647257.shtml.

[3] DEMCHENKO Y, LAAT C D, MEMBREY P. Defining architecture components of the big data ecosystem[C]. International Conference on Collaboration Technologies and Systems, IEEE, 2014:104-112.

[4] CHAO C, CRAGIN H, PALMER L. Data Practices and Curation Vocabulary(DPCVocab):an empirically derived framework of scientific data practices and curatorial processes[J]. Journal of the Association for Information Science and Technology, 2015, 66(3): 616-633.

[5] 王胜开, 马欣. 科研大数据发展状况探讨[J]. 互联网天地, 2015(3): 63-66.

[6] 薛晓娜, 佟泽华, 丰佰恒, 等. 5G 环境下数据生态链的优化策略研究[J]. 情报理论与实践, 2021, 44(3): 37-44.

[7] 佟泽华, 韩春花, 孙杰, 等. 科研大数据再生的内涵解析[J]. 情报理论与实践, 2020, 43(9): 39-46, 78.

[8] 佟泽华, 孙杰, 韩春花, 等. 科研大数据共生的机理研究[J]. 情报理论与实践, 2021, 44(3): 19-26, 52.

[9] ISO/IEC 25012 [EB/OL]. [2020-06-06]. <https://iso25000.com/index.php/en/iso-25000-standards/iso-25012>.

[10] SUKUMAR R, RAMACHANDRAN N, FERRELL R K. 'Big Data' in healthcare: how good is it? [J]. International Journal of Health Care Quality Assurance, 2015: 2-9.

[11] 丁小欧, 王宏志, 张笑影, 等. 数据质量多种性质的关联关系研究[J]. 软件学报, 2016, 27(7): 1626-1644.

[12] WAHYUDI A, KUK G, JANSSEN M. A process pattern model for tackling and improving big data quality[J]. Information Systems Frontiers, 2018, 20(3): 457-469.

[13] 李晓辉. 图书馆科研数据管理与服务模式探讨[J]. 中国图书馆学报, 2011, 37(5): 46-52.

[14] 李斌, 闫华, 顾静秋, 等. 基于“互联网+科技管理”的科研大数据综合服务平台构建与实践——以国家农业信息化工程技术研究中心为例[J]. 农业科技管理, 2018, 37(2): 37-41.

[15] 邹亚霏. 大数据背景下我国科研管理的问题与对策研究[D]. 武汉: 武汉科技大学, 2018.

[16] 刘文奇. 中国公共数据库数据质量控制模型体系及实证[J]. 中国科学: 信息科学, 2014, 44(7): 836-856.

[17] 王娟, 李玉海. 基于演化博弈论的政府开放数据质量控制机制研究[J]. 现代情报, 2019, 39(1): 93-102.

[18] 刘晓婷, 佟泽华, 师闻笛. 大数据时代科研人员数据共享演化博弈研究: 信任机制视角[J]. 情报理论与实践, 2019, 42(3): 92-100.

[19] 温亮明, 张丽丽, 黎建辉. 大数据时代科学数据共享伦理问题研究[J]. 情报资料工作, 2019, 40(2): 38-44.

[20] 彭传静, 叶婷婷, 魏璐, 等. 医学期刊面对论文数据造假的角色选择[J]. 编辑学报, 2023, 35(5): 478-482.

[21] 全球知名数据库中超过 34 万篇论文涉嫌造假[EB/OL]. [2023-04-27]. <https://baijiahao.baidu.com/s?id=1764318587589566732&wfr=spider&for=pc>.

[22] 胡海波, 姜策群. 数据开放环境下的政府数据治理: 理论逻辑与实践指向[J]. 情报理论与实践, 2019, 42(7): 41-47.

[23] 数据壁垒高墙渐起 [EB/OL]. [2021-07-27]. <https://baijiahao.baidu.com/s?id=1706414648167699914&wfr=spider&for=pc>.

[24] 释放数据资源价值亟需拔掉“数据烟囱” [EB/OL]. [2023-01-09]. http://www.cqj.gov.cn/bm/qdsjzfj/zwxw_58673/dt_58674/202301/t20230109_11471885.html.

[25] 夏义堃. 论政府数据开放风险与风险管理 [J]. 情报学报, 2017, 36(1): 18-27.

[26] First half findings from the 2016Breach Level Index [EB/OL]. [2021-12-25]. <https://kdocs.cn/l/sgIweOTx2Zio>.

[27] WAHYUDI A, KUK G, JANSSEN M. A process pattern model for tackling and improving big data quality [J]. Information Systems Frontiers, 2018, 20(3): 457-469.

[28] 夏义堃. 数据要素市场化配置与深化政府数据治理方式变革 [J]. 图书与情报, 2020(3): 14-16.

[29] 黄如花, 温芳芳. 我国政府数据开放共享的政策框架与内容: 国家层面政策文本的内容分析 [J]. 图书情报工作, 2017, 61(20): 12-25.

[30] 吕志明. 国家审计质量控制多阶段三方博弈分析

[J]. 审计与经济研究, 2012, 27(4): 19-25.

[31] 崔丽, 庞舒. 基于供应链质量控制的“农超对接”零供关系演化博弈分析 [J]. 北京工商大学学报 (社会科学版), 2013, 28(5): 48-54.

[32] 王芳, 卜昊昊. 科学数据管理政策发展比较研究 [J]. 中国图书馆学报, 2022, 48(6): 77-96.

[33] 工信部. 工业和信息化部关于工业大数据发展的指导意见 [EB/OL]. [2020-04-28]. http://www.gov.cn/zhengce/zhengceku/2020-05/15/content_5511867.htm.

[34] 国家标准化管理委员会. 信息技术科学数据引用 [EB/OL]. [2017-12-29]. <http://www.sac.gov.cn/gzfw/ggcx/gjbzgg/201732/>.

[35] 中国电子技术标准化研究院发布大数据标准化白皮书 (2018) [EB/OL]. [2018-03-29]. <http://www.cesi.cn/201803/3709.html>.

[36] 国家标准化管理委员会. 关于征求《数据中心能效限定值及能效等级》强制性国家标准 (报批稿) 意见的通知 [EB/OL]. [2021-01-02]. http://www.samr.gov.cn/bzjss/zqyj/202102/t20210202_325775.html.

An Analysis of Scientific Research Big Data Quality Control: Role Positioning, Evolution Game and Implementation Strategy

Tong Zehua

Chen Zhaojuan

Xu Haiyun

Xue Xiaona

Han Chunhua

Abstract: [Purpose/significance] Data pollution, data risks, data barriers and other data quality concerns affect the pace of scientific innovation of researchers, research on issues such as quality control of scientific research big data is imminent, the state, the community are actively exploring research on the quality control of scientific research big data. [Method/process] The study shows that the evolutionary game model of research big data quality control is composed of three parties: research big data providers, research big data consumers and research big data regulators, and the evolutionary and stable strategies of the three parties of the game are derived through model analysis and simulation, and then the roles of research big data quality control are analyzed, and the principles of quality control are given, and finally the four-in-one research big data quality control implementation strategy system of "base-wing" phase optimization strategy, "responsive-alliance" control strategy, whole process technology integration strategy, and talent-platform dual-drive linkage strategy is proposed. [Result/conclusion] The study shows that the evolutionary game model of scientific research data quality control is composed of three parties: scientific research data providers, scientific research data consumers and scientific research data regulators, and the evolutionary stabilization strategies of the three parties are derived through model analysis and simulation. Furthermore, the role positioning of quality control subject of scientific research big data was analyzed. Finally, we propose a four-in-one research big data quality control strategy system, including "base-wing" optimization strategy, "responsive-alliance" control strategy, whole process technology integration strategy, and talent-platform dual drive strategy.

Key words: Scientific research big data quality control; Role positioning; Tripartite evolutionary game; Implementation strategy