

数据管理交叉学科研究

破解数据垄断的几种治理模式研究

孟小峰

【摘要】随着数据的累积，不同科技企业在数据资源的储备量上的差异愈加明显，数据垄断逐渐形成，并催生了“堰塞湖”，导致各企业间的数据难以互通，用户隐私泄露问题随之凸显。因此，通过有效的数据治理来缓解数据垄断形势、促进数据安全与公平的共享流通刻不容缓。一方面应完善当前的数据治理模式，发挥现有治理手段的作用；另一方面要积极开拓透明化的数据治理框架，解决以数据垄断为主的数据伦理问题，构建健康有序的中国大数据生态。

【关键词】数据垄断 数据治理 数据透明 【中图分类号】F49 【文献标识码】A

大数据时代，海量数据的累积催生了数据挖掘、机器学习等新兴技术，同时也为这些技术预测未来、作出决策提供了基础，为社会创造了前所未有的价值。随着数据的累积，数据作为驱动人工智能等技术发展的重要资源，逐渐成为各科技公司争夺的主要对象，不同科技企业在数据资源的储备量上的差异也愈加明显，数据垄断逐渐形成，并催生了“堰塞湖”，各企业间的数据难以互通，并且由于数据本身与个人隐私的密切关系，用户隐私泄露问题亦随之凸显。笔者带领团队基于 3000 万真实用户数据和 30 万 APP 数据，对当前的数据收集情况进行了量化分析发现，当前数据垄断形势异常严峻，对数据进行有效治理迫在眉睫，而数据透明化应是未来数据治理的主题和必经之路。

当前移动应用软件市场的数据垄断现状

为量化当前移动应用市场的数据垄断情况，笔者基于 3000 万真实用户数据和 30 万 APP 数据，使用权限分析法对 2018 与 2019 两年大数据收集现状进行分析。分析的主要对象包括：数据生产者，即产生数据的个人或机构，在移动应用场景中通常指移动用户；数据收集者，即以主动或被动的方式收集数据的个人或机构，在移动应用场景中通常指 APP 开发商；数据使用者，即以任何形式处理或使用数据的个人或机构，在移动应用场景中它可以是数据收集者，也可以通过数据流通、共享等方式获取数据的第三方；数据监管者，即在数据收集、流通、使用过程中对数据进行合法监管的个人或机构，通常包括相关政府机构和可信第三方等。分析结果显示，当前移动应用市场数据垄断形势十分严峻，10% 的

数据收集者可获取 99% 的用户权限数据，数据收集的不平衡现象远甚于社会财富分配中的二八定律。

首先，从总体数据垄断现状来看，为详细阐明该数据收集现状，笔者根据获取权限数据的数量级对数据收集者进行划分，将获取 1 亿及以上权限数据的收集者定义为“亿级权限数据收集者”，获取 1 亿以下 1 千万以上权限数据的数据收集者定义为“千万级权限数据收集者”，并以此类推。主要结论如下：根据 2019 年总体数据收集状况，当前数据垄断形势严峻，极少数数据收集者垄断了绝大部分权限数据。2019 年度数据垄断的“主力军”是占据所有数据收集者数量 1% 的“百万级、千万级、亿级的权限数据收集者”，他们可获取约 92% 的权限数据。对比 2018 年度与 2019 年度数据垄断状况，前 10% 的权限数据收集者获取的权限数据量占比略有减少，但总体上数据垄断态势居高不下。具体而言，不同级别权限数据收集者的数量与获取数据量的对比分布如图 1 所示，“百万级、千万级、亿级的权限数据收集者”本身的数量极小，但权限数据获取量均在 10% 以上，而其余大量的数据收集者可获取的数据量不足 3%。该状况从不同比例数据收集者获取权限数据分布情况中体现得更为明显，如图 2 所示。表 1 给出 2018 年度与 2019 年度权限数据收集的对比情况，其变化量为负值说明这些权限数据收集者获取数据量占比有所减少，

图 1 2019 年权限数据收集者个数及数据获取分布图

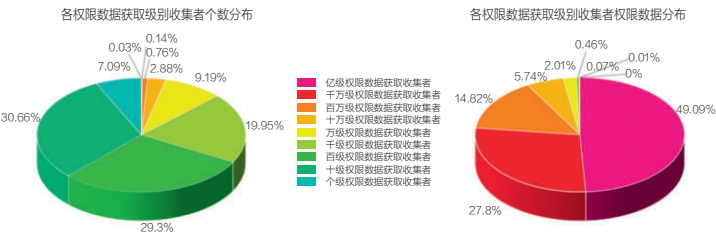


图 2 2019 年权限数据收集者数据收集分布图

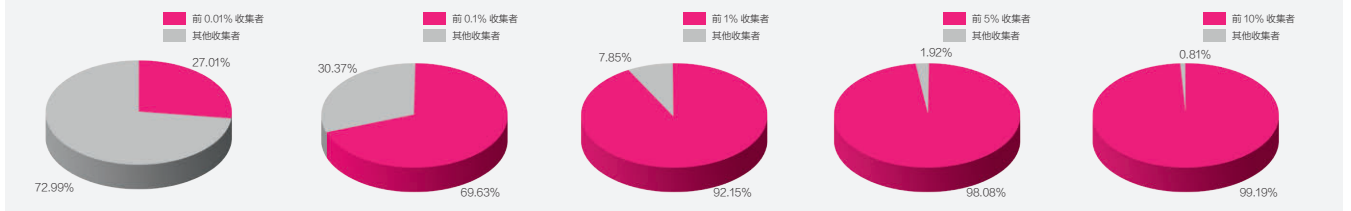


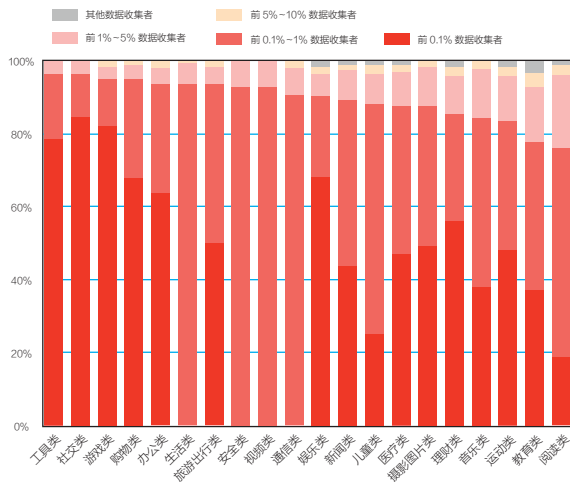
表 1 2018 与 2019 年度数据收集情况对比

权限数据收集量占比	权限数据收集者占比				
	0.01%	0.1%	1%	5%	10%
2018 年度	38.44%	75.48%	93.81%	98.57%	99.42%
2019 年度	27.01%	69.93%	92.15%	98.08%	99.19%
变化量	-11.43%	-5.55%	-1.66%	-0.49%	-0.23%

但权限数据收集者数量超过 5% 后，其获取数据量的变化微乎其微。可见，我国总体数据垄断形势依旧严峻。

其次，从分类数据垄断现状来看，笔者所在团队对 Google Play 及国内第三方应用网站中 APP 分类进行调研，将当前市场上的 APP 划分为 20 类，分别是安全类、生活类、社交类、办公类、理财类、购物类、教育类、儿童类、旅游出行类、摄影图片类、视频类、工具类、通信类、新闻类、医疗类、音乐类、游戏类、娱乐类、阅读类和运动类。基于该分类，得出如下结论：每类 APP 的数据垄断形势都十分严峻，前 10% 的数据收集者均收集了不少于 97% 的权限数据。各类 APP 中，工具类、社交类和游戏类为数据垄断的重灾区，教育类和阅读类的数据垄断状况较总体水平有所缓解。具体情况如图 3 所示，工具类、社交类和游戏类的前 0.1% 数据收

图 3 各类别权限数据收集分布情况



集者收集了约 80% 的权限数据，前 1% 的数据收集者收集了约 95% 的权限数据，而前 5% 的数据收集者就收集了约 99% 的权限数据。在形势较为缓和的教育类和阅读类，前 1% 的数据收集者收集了约 75% 的权限数据，低于该比例数据收集者对应的总体占比。

最后，从主要数据收集者垄断现状来看，笔者对数据获取量排名前 5 的数据收集者对比分析，以展示当前主要数据收集者的垄断现状。为保护数据收集者的个体隐私，该分析隐藏这 5 个数据收集者的名称，仅提供统计性结果。这 5 个数据收集者，最多的可获取 8% 的权限数据，最少者可获取 3% 的权限数据，累计可获取近 24% 的数据。也就是说，仅这 5 个数据收集者，就可获取约 1/4 的用户数据。其中，3 个数据收集者所开发 APP 涉及了 18 个以上的 APP 类别，其余 2 个数据收集者侧重于单个领域，其开发 APP 仅涉及了不足 5 个类别。这 5 个数据收集者的共同点是：其开发 APP 对应的用户量群体均十分庞大。

以当前数据收集者们的数据获取量为依据，分析数据垄断的成因

在严峻的数据垄断形势下，探究数据垄断成因十分关键。当前数据垄断的形成与数据自身的特点、数据收集者们的商业运营模式以及人工智能时代的网络效应密切相关。

第一，数据易聚集、难确权的特性，使得数据垄断易形成。大数据时代，海量数据通过移动设备、传感器网络等源源不断地自动产生，数据的生产成本较低，同时其本身的价值密度也较低，海量数据的价值需通过数据挖掘、机器学习等技术提取。而这些技术本质上是数据驱动型技术，需基于大量数据的输入才能获取高准确性、高可用性的输出结果，造成数据本身易聚集的特点。此外，数据本身的特殊性使其既不同于石油、矿藏类的自然产物，也不同于专利、作品等精神产物，

难以确定其所有权。在当前数据不能依据法律法规确权的环境下，数据收集的合理合规性得不到有效保证，易形成数据垄断。

第二，数据寡头多产品、跨领域、高用户量的商业运营特点，是数据垄断形成的重要因素。数据寡头即当前数据垄断的主要对象，对应的就是排名前 0.1% 的数据收集者。当前数据寡头们通过业务扩张、资本运作、并购等方式完成企业扩张，导致其具有多产品、跨领域的商业特点，并据此吸引或维系海量用户，从而具有海量数据收集的能力，形成数据垄断。分析结果表明，在移动应用市场，数据收集者们开发 APP 的数量越多、使用量越高、涉足的领域越多，其获取的权限数据量越大，越有可能成为数据寡头，形成数据垄断。显然，前 0.1% 的权限数据收集者的这三个因素比其他权限数据收集者明显高出数倍。

第三，人工智能时代的网络效应促进数据垄断形成。人工智能技术数据驱动的特点使其本身就具有网络效应。随着人工智能技术产品使用的用户量激增，该技术可获取更多用户的数据输入，从而可创建可用性更高的数据模型，增加其自身价值的同时吸引并服务于更多用户。当前移动应用市场上的数据寡头均为大型科技公司，他们均受益于人工智能等技术的支持。相应地，基于其海量的用户数据，他们可持续发展优化其产品与服务，进一步维持并吸引新用户。而本身处于弱势的数据收集者们则限于其产品或服务的升级能力，迫于数据寡头发展的压力逐渐流失用户，滚雪球效应产生，数据垄断现象随之加剧。

缓解数据垄断形势、促进数据安全与公平的共享流通，三种数据治理模式更为有效

严峻的数据垄断形势给当前移动互联网的发展带来了巨大的挑战。数据垄断使得寡头公司拥有大部分的用户数据，在数据驱动的发展模式下，压缩了该领域内其他公司的生存空间，不利于小型企业的发展。数据垄断一定程度上破坏了市场自由竞争的规则，数据寡头公司基于海量数据资本掌握市场主导权。对小型企业的打压，使得消费者失去同类服务的可替代选项。数据垄断有可能阻断小型企业的技术创新，而大型企业利用其丰富的数据可开发多领域的生产经营活动，技术壁垒进一步抑制了新技术的产生。数据垄断使得寡头企业一家独大，掌握对用户数据的控制权，易加剧数据滥用、隐私泄露、用户歧视等其他数据伦理问题的产生。因此，一

方面，应规范数据的收集、流通和使用，促进数据资源的合理配置；另一方面，应积极探索用户隐私保护的数据共享方式，促进数据共享流通。现有的数据治理模式包含以下三种：

一是局部模式。在数据流通过程中，从数据源头基于隐私保护技术对数据进行处理，一定程度上能够限制企业收集大规模数据的行为。当前应用的隐私保护技术主要包括基于扰动的匿名化、差分隐私技术和基于密码学的安全多方计算等，这些技术提供的隐私保护程度越高，收集数据的准确性越差，计算成本也就越高。数据收集者必须平衡隐私保护与数据有效价值之间的关系，从而缓解当前低成本的数据收集垄断局势。在该治理模式下，数据寡头仍持有大部分数据的控制权，数据垄断有所缓解但并未根除，并且需要权衡好数据治理与产业输出之间的关系。

二是中介模式。在数据流通过程中增加第三方中介平台，参与数据流通，促进数据共享。当前的中介平台主要包括数据交易平台、数据众包平台和数据共享平台三种模型，分别适用于不同情景。自 2015 年国务院印发《促进大数据发展行动纲要》以来，全国范围内涌现出多个数据交易平台，包括以数据包交易为主的政府类数据交易所，如贵州大数据交易所、上海数据交易中心、长江大数据交易中心等，以及以 API 接口模式为主的民营平台，如聚合数据、京东万象、数据堂等。数据众包平台为企业或个人提供有偿的数据供应及下载途径，目前有百度数据众包、有道众包、蚂蚁众包等平台。数据共享平台包括数据直接共享和数据间接共享两种方式。直接数据共享平台依据必要的设施规则，推动公共部门之间不对称信息的流通和企业之间数据的合理共享，较为典型的是英国人工智能实验室与开放数据研究所合作建立的“数据信托”实验点，其目的是促进多集团之间的数据共享。间接数据共享平台拒绝对源数据的直接共享，支持对本地数据训练得到的模型参数进行共享，而后由多方参与者共同训练效果较强的机器学习模型。该方法符合当前数据驱动的技术发展情景与用户隐私保护的需求，具代表性的是微众联邦学习项目与华为 NAIE 联邦学习平台。从总体发展现状来看，第三方中介的项目众多，但目前数据交易、共享的规模并不大，具有很大的发展空间。

三是全局模式。对数据产生、流通和使用的整个生命周期进行监管，弱化数据寡头对数据的掌控权，增强数据生成者（即用户）和数据监管者对数据的控制权。该模式主要分为中心化和去中心化两种形式。中心化全局模式是指建立统

一的数据监管平台，对数据进行统一管理，如库克提议美国联邦贸易委员会组建的“数据清算所”，通过监管数据流通状况来确保用户对数据的控制权。去中心化全局模式指借助区块链、智能合约等去中心化技术与平台，对数据收集、流通、共享、使用、结算等过程存证，构建可验证、可追踪、可溯源的数据共享与监管机制，目前已有众多政府机构与学术机构在此方面展开研究。全局模式相较其他两种治理模型成本更高，目前该数据治理体系正在构建中，其应用尚不成熟。

数据透明是解决数据垄断问题的根本途径，是未来数据治理的必经之路

上述数据治理模式以政府和IT企业为主要参与者，针对数据垄断、阻塞、不互通等问题提出局部或全局的治理方案，重点在于可监控的数据资产平衡分配。然而，当下的数据垄断问题不仅仅是数据资产的分配失衡问题，更是人工智能时代数据伦理的问题，数据垄断的加剧会导致数据隐私、数据歧视等其他伦理问题的发生。笔者认为，当下大数据的“堰塞湖”已然形成，数据垄断愈发严重，数据隐私与公平问题层出不穷，归根结底是数据收集、流通、共享、使用和决策过程中的不透明性所致。因此，数据透明是解决上述问题的根本途径，是未来数据治理的必经之路。

数据透明，并不表示数据对所有人公开可见，它指的是数据在其生命周期中对其从属主体透明化，即在数据收集、流通、共享、使用和决策过程中，保证数据对其拥有者、使用者和监管者显示部分或全部的透明性。在整个数据透明框架中，数据的隐私必须加以考虑并得到保证。对数据垄断而言，数据透明的应用可促进数据收集、流通和使用记录的生成，从而完成数据的审计、溯源与问责。该方式既可达到数据监管的目的，又可为数据共享方向与方式提供评估依据，结合数据访问控制技术可全方面监控并防止数据垄断的生成。

宏观上，基于数据透明的数据治理应聚焦于以下三个方面内容：第一方面，保证数据质量与价值。数据作为大数据时代科技企业的主要资源，在使用数据治理手段协调各个社会主体利益时，应基于数据透明机制保证数据的真实性、正确性，统一多源数据标准，评估有效数据价值，从而保证数据驱动决策的可靠性。第二方面，评估和监管个人隐私数据的使用。用户作为大数据生产者，极易在数据流通过程中丢失对自身数据的控制权。基于数据透明，可评估和监管个人

隐私数据的流向及用途，使用户重拾数据控制权，有效避免数据过度收集与聚积，预防个人隐私数据泄露。第三方面，监管并促进数据流通与共享。这也是阻断数据垄断的重要举措，但在实施时需兼顾数据隐私，考虑各参与主体间的信任模型，平衡各方利益。

具体而言，基于数据透明的数据治理可借助区块链技术实现。基于区块链公开透明、去中心化和不可篡改的特性，可在数据生命周期中的各阶段分别进行有效的数据治理。在数据存储阶段，基于区块链和智能合约存储数据，可达到支持审计的目的，防止该过程中数据伪造、数据篡改、数据标准不统一等问题的出现。在数据收集与共享阶段，可使用区块链保存数据的收集与共享日志，对数据流通过程进行追踪溯源；同时结合策略承诺、违法检测、隐私审计，可在隐私保护技术失效的情况下通过溯源问责保护隐私，并为实施数据监管、防止数据垄断提供技术支持。在数据使用与决策阶段，可基于区块链对数据计算节点进行验证，通过经济惩罚等手段防止恶意参与方的加入，同时验证决策结果的可靠性，确保数据的高效合理产出。

2020年4月6日，中共中央、国务院印发的《关于构建更加完善的要素市场化配置体制机制的意见》提出，要加快培育数据要素市场的概念，并强调了数据的开放与共享。这使得解决数据垄断问题、评估和监管数据的合理分配与使用，变得更加紧迫和必要。同时，它也对数据共享流通方式和数据质量等提出了更高的要求。将数据作为要素应该放在数据治理的框架下加以考量，需要综合考虑数据生命周期内相关参与主体的权利与义务。在未来数据治理的过程中，我们一方面要完善当前的数据治理模式，发挥现有治理手段的作用；另一方面要积极开拓透明化的数据治理框架，解决以数据垄断为主的数据伦理问题，构建健康有序的中国大数据生态，促进大数据产业合理规范发展。

（作者为中国人民大学信息学院教授、博导）

【参考文献】

- ①《中共中央 国务院关于构建更加完善的要素市场化配置体制机制的意见》，中国政府网，2020年4月6日。
- ②《国务院关于印发促进大数据发展行动纲要的通知》，中国政府网，2015年9月5日。
- ③《习近平：实施国家大数据战略，加快建设数字中国》，《人民日报》，2017年12月10日。

责编/韩拓 美编/陈媛媛

大数据与社会计算研究进展

——第五届全国大数据与社会计算学术会议

摘要 BDSC2020 的大会主题是“社会计算与社会智能”，旨在通过多学科交叉融合，以社会计算为方法论，以人工智能、大数据等信息技术为科学工具，构建“社会计算试验场”，深刻剖析社会计算与社会智能的内在机制，实现对新型社会现象的发现与机理揭示，促进社会计算与社会智能的发展。本文对 2020 年 BDSC 的 4 个特邀报告和 8 个专题进行总结。

1 引言

2020 年第五届全国大数据与社会计算学术会议（China Conference on Big Data & Social Computing, BDSC2020）于 2020 年 8 月 22 日至 23 日以在线方式（Zoom 会议视频+B 站直播）成功举办。本次会议由中国人工智能学会主办，社会计算与社会智能专委会具体落实，承办单位有人大、清华、北师大、电子科大，集智俱乐部与洛阳师范学院参与组织工作。孟小峰教授担任本次大会的共同主席。

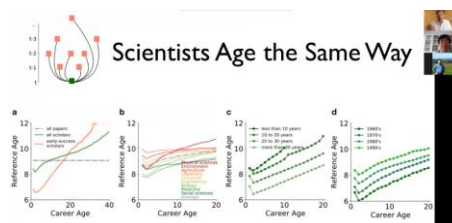


孟小峰教授代表会议组织方介绍了大会组织情况，对主办单位中国人工智能学会的大力支持表示感谢，并指出社会变革已然在技术变革的作用下悄然发生，本次会议直面社会变革所面临的真实问题（疫情应对、政府治理、大型公共活动、风投网、舆情、数据伦理、因果科学），在新的信息基础设施（Infrastructure）形成的基础上，探索建立新的社会计算研究方法，架构未来智能社会，是初心，也是责任，更是挑战，愿更多的学者参与其中！

2 大会特邀报告

2.1 多样性与发现

美国芝加哥大学社会学教授 James A. Evens 作了“多样性与发现”的报告。首先考察了科学家和发明家是如何在整个生命过程中进行科学研究的：这一过程的变化趋势往往一开始很快，然后越来越慢，直到后面慢慢僵化，最终形成科学的界限。然后，我将其与突破性的发现和发明如何跨越这些界限联系起来，这些突破性的发现和发明是通过内容的意外组合来实现的，这些内容包括问题、方法和自然实体，跨越不同的上下文，如期刊、子领域和会议。基于数以千万计的研究论文、专利和研究人员的数据库，我们构建了一个模型，并以此来预测下一年年的内容和上下文组合，基于高维随机块模型构建的嵌入，AUC 为 95%，其中新组合本身的不可能性-预测指数高达 50%的可能性他们将获得巨大的引用和重大奖项。

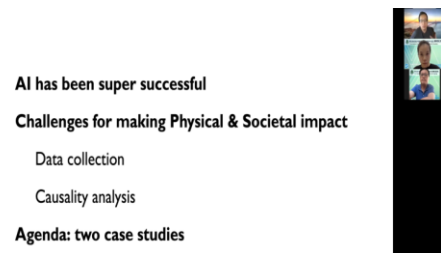


这些突破大多发生在一个领域的问题被来自另一个领域的研究人员意外地解决时。这些发现证明了提前出其不意的关键作用，并使科学机构的评估从教育和同行评议到支持它的奖项。我演示了新兴的人工智能知识发现方法，明确地结合了相关人员的分布，在预测未来发展方面的表现如何比那些只考虑思想分布的方法高出一倍。这项工作提出了一条创造最多，但也是最少的人类人工智能的道路，并优化设计了从根本上增强人类智能的集体认知多样性。

2.2 人工智能和它的物理与社会影响

IEEE Fellow、Instacart 副总裁和杰出科学家王海勋作了“人工智能和它的物理与社会影响”的报告。AI 发展到今天已经取得了许多的成果，其中用户接触较多，比较有代表性的就有实时新闻推送、广告投放和无人商超。而在不久的将来，我们还可能会迎来无人驾驶、智慧家庭等“科幻”应用。这说明科技已经不再只留于纸面，它已经切实的走进了我们的生活，AI 与其所代表的科学技术对物质和精神的影响才刚刚开始显现。但对应到复杂真实的物理社会，AI 技术可能还面临着两个巨大的问题。对这两个问题进行详细的研究，我们能够得到相当多的启示：（1）如何完整的收集正确可用的数据；（2）如何在 AI 中解决因果推断问题。今天应用到生活中的 AI 应用没有一个不是用大量的数据训练得到的。所以我们可以认为是海量数据的增长推动了 AI 的发展。但反观 AI 本身，如果使用不正确的数据进行训练，这个模型也无法得出正确的结果，所以研究数据是 AI 的关键点。

AI 技术的发展毫无疑问给我们物理社会的日常生活带来了便利。当然这也不仅仅是 AI 领域内的问题，不是任何单一计算过程所能够阐释的。作为技术本身 AI 的帮助显而易见，但我们整个社会是一个复杂的整体，AI 所给出的判断是否符合人类本身的情感逻辑还有待研究，机器人决策思考上的差异也必须被我们所考虑，所以在这一部分，还有非常长的路要走。未来，我们应该带着问题去研究。



AI has been super successful

Challenges for making Physical & Societal impact

Data collection

Causality analysis

Agenda: two case studies

2.3 社交媒体上的隐私悖论——用户能同时拥有隐私和效用吗？

ACM Fellow、AAAI Fellow、美国亚利桑那州立大学计算机科学与工程学院教授刘欢作了“社交媒体上的隐私悖论——用户能同时拥有隐私和效用吗？”的报告。社交媒体用户经常会遇到所谓的隐私悖论。在线服务的良好实用性似乎需要通过用户的

个人数据来了解其独特需求。因此，必然要以用户的隐私权换取效用。刘欢老师再讲座中，就用户是否仍然可以同时拥有隐私和实用性，以规避隐私悖论进行了相关的调研。由于用户对隐私可能有不同的看法，因此用户隐私是一个复杂的问题，需要受到法律，政策和技术的保护。自己所在的研究小组从技术角度提出一些讨论和观点。



首先，报告以数据引入问题。当前科学技术迅速发展，越来越多的人通过阿胶媒体来进行信息获取，工作互联、网络购物等活动，调查表明超过 68% 的美国人通过社交媒体获取消息资讯。用户进行的各种各样的活动，生成海量的数据。这些数据的价值在于，一方面可以提供给用户个性化的服务，另一方面可以让商业合作者的信息和数据实现共享。

然而，用户产生的行为数据中可能包含敏感和隐私的信息。例如：Dr.appt Friday morning told I need to loose 50 pounds by X-Mas, have a fatty liver, high cholesterol and high blood pressure. 隐私者强烈需要数据的脱敏和匿名化。相反，数据实用性对商家提供个性化服务是非常重要的。

正如刘欢老师再报告中讲述的那样，用户声明他们关注隐私，但在社交媒体上的行为举止似乎并非如此。因此，作为计算机和数据科学从事者如何提供帮助？

针对上面的问题，刘欢老师研究小组主要从两点解决问题：

(1) 识别风险；要明确如何识别风险，社交媒体数据的隐私风险是什么，其中可以包含医疗、财务、家庭信息，在线购买、媒体使用行为等。

(2) 风险消减方法优化。如何采用计算机和数据科学的方法，让我们同时拥有隐私权和数据实用性吗？而针对这些问题，现有研究已有例如 K-anonymity, 差异化隐私等方法，为了更好的解决多样化社交数据的风险，刘欢老师的研究小组提出 De-anonymization attacks 的方法。

2.4 大规模移动购物数据的跨平台消费者行为分析

欧洲科学院院士、德国哥廷根大学数学与计算机科学学院终身教授傅晓明作了“大规模移动购物数据的跨平台消费者行为分析”的报告。移动设备（尤其是智能手机）的普及为行业和学术界带来了巨大的机遇。特别是，从用户使用日志中生成的海量数据为利益相关者提供了借助数据挖掘来了解消费者行为的可能性。



此次报告中，傅晓明老师的研究小组基于电信运营商的大规模移动 Internet 数据集，研究了跨多个电商平台的消费者行为。该数据集涵盖了来自两个地区的 970 万用户，其中 140 万在小组研究的一周内访问了电子商务平台。

傅晓明老师在报告中介绍，研究小组尝试回答几个问题：

- (1) 用户的位置（以及更普遍的时空因素）如何影响他们的购物行为；
- (2) 面对来自多个平台的许多替代选择时，用户决定购买产品需要多少时间；
- (3) 用户是否表现出对首选购物平台的忠诚度迹象？

首先，不同社会经济地位的人群在浏览、购买、成功购买的比例是不一样的，低端、中产、高端人群的购物的各个阶段的决策是不同的。例如上图所示，中产人群浏览和购买的比例较多，相反成功购买的比例却比较小，这其中考虑包含这类人群的家庭环境、经济收入等

方面的影响因素。其次，不同的购买者的花费时间各有不同，61.2%的用户是快速购买者，21.93%的用户是犹豫购买者，还有小部分的短期和长期购物决定者。最后，正如傅老师在报告中说到，经济较发达地区的用户比不发达地区的用户更快地做出购物决定。同样，在可用的多个电子商务平台中，大多数移动用户忠于其有利的平台。此外，人们倾向于在不到 30 分钟的时间内做出快速的在线购买决定。所以针对不同的用户，分析不同的购买行为特征，在策略营销时，采用不同的手段，这对运营推广具有一定的借鉴意义。这些新见解可以为电子商务未来战略规划提供有用的信息。

通过研究分析用户的移动购物行为数据，获得用户的位置、社会经济状况、使用行为、时间特征等因素指标，研究通过因素分析，引入用户忠诚度指标因子，结合有监督的机器学习模型方法，以实现用户的消费行为的预测。研究小组通过获取的大规模跨平台移动购物数据开展验证测试，测试实验的有效性。在报告结尾，傅老师对未来的相关研究做出了进一步展望，未来在大规模移动用户行为分析研究方面，将长期学习与社会关系相结合，从时序和社交等视角进一步研究分析用户的购物行为，以期在大规模跨平台移动购物行为研究方面深入探索。

3 社会计算与 AI（因果科学）专题

3.1 大数据环境下的因果推理

清华大学崔鹏副教授作了“大数据环境下的因果推理”的报告。近年来人工智能技术的发展，在人脸识别、棋类游戏若干垂直领域取得了性能突破。但当我们重新审视人工智能技术的“能”与“不能”时，我们发现其在理论方法层面存在的两个短板：（1）缺乏“举一反三”的能力，在一个环境或场景中学习的模型难以泛化到其他环境和场景；（2）推理结果和推理过程难以解释，限制了人工智能技术在智慧医疗、金融科技等风险敏感领域的应用。究其深层次原因，均是因为今天的人工智能技术“只知其然，不知其所以然”。针对这一问题，介绍了大数据环境下的因果推理，并将因果推理引入预测性问题，介绍稳定预测模型方面的研究进展。

3.2 复杂系统自动建模与因果发现

北京师范大学张江教授作了“复杂系统自动建模与因果发现”的报告。在人工智能、深度学习的助力下，复杂系统建模已经步入了自动化的阶段。根据复杂系统的运行数据（时间序列），深度学习系统即可以模拟系统的运行、预测系统的未来状态。随着图网络、神经微分方程（Neural ODE）以及标准化流（Normalization Flow）等技术的发展，这方面的研究如今呈现出了井喷的模式。人们不仅能够精准地构建复杂系统的动力学模型，而且还能在带有噪声、带有隐含节点、隐含变量，以及小数据的系统上自动构建模型。另一方面，随着因果推断技术的发展成熟，越来越多的能够自动从数据中提炼出因果关系，并具备一定可解释性能力的深度学习模型逐渐被提出。而近来这两大方向的发展正在逐渐呈现新的交叉、合并

之势。通过引入图结构学习技术，深度学习算法不仅可以精准地预测系统动力学，还能够自动提炼因果结构，甚至能够与系统进行互动和干预，还能逐渐攀爬 Judea Pearl 所说的三阶因果之梯。张江老师站在一种较宏观的视角对这些技术进行概述，内容将涉及但不限于：运用 Reservoir 计算预测混沌、基于图网络的自动建模与控制、基于最优控制的可微分 ODE 求解技术、基于自注意力机制的人工智能统计物理学家、基于 Gumbel softmax 技术的网络重构、基于神经网络的格兰杰因果检验、基于强化学习的干预因果模型等。

3.3 识别数字平台亲社会行为因果关系

来自麻省理工学院系统和社会研究所的博士生袁源，则带来了基于数据平台识别亲社会行为的影响相关研究，其题目是《Identifying the impact of prosocial behavior on digital platforms》。他目前也是麻省理工学院媒体实验室（MIT Media Lab）人类动力学小组的研究助理。现在社交网络兴起，存在许多群组，无论是家庭群、朋友群还是工作群，在其中朋友之间礼物可以起到很重要的社会连接作用，而雇主对通过分发礼物给他们的员工，也可以帮助提高员工的工作的效率。但在发群礼物、群红包时，一般不会直接发给某个接收者，而是直接发在群里。因此这种发礼物行为就可能具有一种社会传染性。即是说，当一个人收到了礼物，会促使他们去发更多的礼物。这样的关系其实就是一种因果的关系。因此就可以通过一些因果推断手段去分析传递性的因果机制。袁源研究工作从红包机制构建数据模型开始。首先是对使用的观察数据寻找社会传染性。这需要时间上的一种聚集性，称之为 temporary classroom；其次是社会学一个概念叫红包分裂，意思大概就是说人们喜欢跟他相近的人，接触或者是成为朋友。这就是说在微信或者在其他社交网络平台的场景下，有钱的人或者是更喜欢发红包的人，这些相似的人更可能跟他相似的人聚集在同一个群里。

3.4 计量经济学因果分析工具在快手中的应用

快手经济学家、华盛顿大学经济学博士杨淼钰作了“计量经济学因果分析工具在快手中的应用”的报告。在产品迭代和公司决策中，一个很重要问题是：一个动作 A 是如何影响 B 的？例如设置上下滑的功能是如何影响用户的视消费、直播消费、生产等体验，甚至影响用户对于平台的长期留存的？传统解决以上问题的方法是使用 A/B 实验。但如果没有条件、在不方便传统方法情况下，那就需要使用因果分析的方法，结合观测数据来回答这个问题。在快手则是基于因果分析的计量经济学和机器学习方法来解决这些归因问题。快手经济学家杨淼钰分别介绍了这两种因果分析方法是如何在产品具体业务中与实践相结合的。

4 社会计算与数据伦理专题

4.1 大数据与人工智能的伦理挑战

电子科技大学周涛教授作了“大数据与人工智能的伦理挑战”报告。一个以大数据为原材料，以人工智能为引擎的新科技时代的到来不可阻挡。大数据和人工智能在给人类社会带

来巨大利益的同时，也带来了诸如个人隐私、数据独裁、新型智能生命等让人担忧的问题。回顾了大数据和人工智能伦理研究的背景、意义和现状，着重从中立性、时效性、导向性、边界问题、隐私问题和责权问题六个方面（根据报告时间，或有详略不同）介绍大数据于人工智能发展带来的具体伦理挑战，最后开放式讨论可能的有效应对策略以及相关的政策和技术问题。

4.2 实现个人的数据控制权利：无用成本 vs. 效率投资？

北京理工大学法学院洪延青研究员作了“实现个人的数据控制权利：无用成本 vs. 效率投资？”的报告。国外成熟的个人信息保护法律，都赋予个人对其个人信息的一些控制性的权利，例如查询、更正、删除。但不同法域在不同场景下对个人权利做了不同的配置。中国正处于立法过程之中，对个人信息权利的边界，还有很多争议。许多企业认为实现个人权利对其运营是一种无用的负担，立法不应当过度增加组织的成本。但实际上，允许个人一定程度上“参与到”组织的数据伦理，能够倒逼组织提升自己的数据伦理能力。

5 社会计算与疫情应对专题

5.1 新型冠状病毒传播示踪与预警系统（TWS）

国防科技大学郭得科教授作了“新型冠状病毒传播示踪与预警系统（TWS）”的报告。在新型冠状病毒等重大传染病爆发期间，现有的疾控技术无法全时空细粒度地感知大规模人群的密切接触行为，严重制约了重大传染病在大时空尺度上的准确认知、预警和应对。为此，鹏城实验室刘韵洁院士牵头，联合南方科大、国防科技大学于2月1日紧急启动“新型冠状病毒传播示踪与预警系统(TWS)”项目，利用蓝牙通信技术提供轻量级的常态化防疫技术手段，期望准确定位病毒感染者密切接触人群，示踪病毒传播路径和细粒度预警，阻断病毒二次传播。TWS系统的技术体制在国际上走在了前列，比新加坡、德国、美国、英国、澳大利亚等国的同类应用早上线和推广应用，系统得到国家发改委官网、人民网、中国计算机学会官网、科技日报等重要网站和媒体的报道。本次报告将从项目背景、项目简介、工作机制、工作模式、适用场景、当前进展、社会评价等方面系统性介绍新型冠状病毒传播示踪与预警系统（TWS）。

5.2 新冠肺炎疫情大规模公开病例数据的分析与建模

大连民族大学许小可教授作了“新冠肺炎疫情大规模公开病例数据的分析与建模”的报告。新冠肺炎疫情暴发以来，各级政府按照《政府信息公开条例》要求，纷纷启动了疫情数据的信息公开工作。疫情数据特别是病例信息的公开，对于满足公众知情权，加强公众自我防护意识和抗击新冠肺炎疫情起到了重要作用。基于这些数据，社会上的商业公司开发出各种深受公众好评的应用，科学家完成了新冠肺炎等新发重大传染病的监测和防治等诸多研究工作，为全世界人民抗击疫情做出了贡献。报告中介绍了基于公开疫情数据驱动的多学科研

究进展，基于收集和整理的全国 10005 条公开病例数据，分析网络科学在抗击新冠肺炎疫情中的具体应用以及潜力，探索基于实证数据驱动和网络科学理论相结合进行新冠肺炎传播的研究范式，为预防可能的新冠肺炎秋冬季节第二波大暴发提供一定参考。

5.3 全媒体背景下的突发事件感知与管理决策

上海财经大学刘建国教授作了“全媒体背景下的突发事件感知与管理决策”的报告。智能手机、微博、微信等现代通讯工具的快速发展为网络信息的收集、发布和分析带来了便利，其去中心化，多源异构的特征也为突发事件感知与应急管理带来了的机遇和挑选。如何从全媒体角度感知热点和突发事件，进行情绪感知和决策应对，进而服务于政府的重大决策分析，提升企业信息感知和管理水平等工作又具有重大的需求。报告从全媒体背景下的突发事件感知、评估、应对和评估等角度对全媒体背景下的信息感知与应对工作进行系统介绍，抛砖引玉。

5.4 基于人口流动大数据定量评估新冠肺炎的传播风险和防控效果

英国南安普顿大学高级研究员赖圣杰“基于人口流动大数据定量评估新冠肺炎的传播风险和防控效果”的报告。现代交通方式的发展促进了人类在全球范围的流动，同时也加速了传染病跨地域的传播。随着信息和通讯技术的快速发展，手机定位、社交媒体、火车/飞机订票等大数据，可以用于实时测量人群流行的时空模式，也为及时理解传染病的传播规律、扩散风险和防控效果评估提供重要的数据源。新型冠状病毒肺炎疫情发生以来，这些数据被广泛应用于病例和密切接触者的追踪、流行病学调查、测量和预测疫情传播风险、评估干预措施的效果等。报告主要介绍了英国南安普顿大学全球人口与健康（WorldPop）团队，如何利用手机、民航客运、火车客流等大数据和流行病学数据，开展的一系列新冠肺炎研究，包括：疫情早期病毒在国内和国际传播扩散风险和趋势；乘坐不同交通工具的感染风险；中国和全球不同国家采取的非药物干预措施效果；各国应如何采取协同、有效的方式开展疫情防控和实施解禁策略，以避免疫情的反弹等。

6 社会计算与舆情治理专题

6.1 突发公共卫生事件中媒体信息透明度与社会情绪研究——以新冠肺炎疫情为例

北京师范大学吴晔教授报告了“突发公共卫生事件中媒体信息透明度与社会情绪研究——以新冠肺炎疫情为例”的报告。以新冠肺炎疫情为案例，探讨了突发公共卫生事件中媒体信息透明度的变化对社会情绪的影响。选择疫情爆发期样本地市官方微信公众号和微博账号为研究对象，深入分析疫情通告透明度及用户评论社会情绪，探讨媒体信息供给透明度能否以及如何实现对社会情绪的调节作用，建议突发公共卫生事件中媒体信息传播时扩大信息供给覆盖面、从受众出发实现信息供需平衡、讲求信息供给的时效性、重视舆情监测与情绪的

反馈作用。

6.2 ABM 仿真与大数据双轮驱动的网络舆情演化全周期模型

中南大学吕鹏教授作了“ABM 仿真与大数据双轮驱动的网络舆情演化全周期模型”的报告。网络舆情事件呈现稳健的“生命周期”宏观涌现规律。综合采用大数据分析 with ABM 仿真两种方法予以双向考察。大数据分析真实网络事件的宏观涌现特征,构建目标拟合函数。微观行为研究方面,设置事件(Hots)与网民(Netizens)两类智能体,进行 ABM 多主体仿真。通过参数遍历,求解拟合度最高的最优参数组合。拟合度考察指标包括寿命、峰值、差值、比值、分布等。最优解的存在,表明模型较好地刻画了微观行为机制,进而才能实现宏观演化过程的高精度复现。基于此客观模型的精准研判、轨迹预测、双向干预等实践操作,其精确性、科学性、自适应性将大为增强。

6.3 社会计算背景下的虚假信息治理

北京师范大学徐敬宏教授作了“社会计算背景下的虚假信息治理”的报告。虚假信息的传播与治理,是一个传统的研究话题。社交媒体等新媒体技术和应用,助长了虚假信息的传播速度和范围。另一方面,大数据的出现和计算机技术的发展,为作为社会科学和计算科学交叉的社会计算带来了新的研究方法。社会计算可以在虚假信息的监测与识别、预防与治理等方面大有作为。采用比较法,梳理各主要国家和地区在社会计算背景下虚假信息治理的主要政策与措施、成功的经验和失败的教训,探讨未来可能的发展方向以及适合我国国情的治理建议。

6.4 基于超网络的突发事件舆情风险多元主体治理“三力”研究

中国科学院科技战略咨询研究院马宁助理研究员作了“基于超网络的突发事件舆情风险多元主体治理“三力”研究”的报告。互联网时代,重大突发事件发生后所产生的各种舆情风险冲突加剧,给公共生活秩序和网络安全带来新的挑战。报告以构建舆情风险治理体系为切入点,建立决策层、施政层、受众层多元主体协同共治的“三位一体”舆情风险治理体系,针对各主体分别构建不同的多层、多级、多维超网络模型,综合链路预测算法、超边排序算法、特征相似度算法等,预测舆情风险预控关键点、识别舆情风险敏感人物、仿真舆情风险演化态势,以提升决策层关口前移洞察力、施政层风险应对执行力和受众层舆情风险偏好辨别力。通过案例反演,完善不同治理主体之间的反馈-优化机制,提升突发事件舆情风险综合治理水平。

7 社会计算与大型公共活动治理专题

7.1 全球疫情及城市大型活动

清华大学张辉教授作了“全球疫情及城市大型活动”的报告。当前，正面临着新冠疫情与严峻国际形势的双重挑战，如何总结我国抗疫经验，并阻断事件在全球进一步升级/恶化，如何重新恢复城市活力，如举办奥运会等大型活动，及如何实现后疫情时代的科技创新是我们急需进行的研究。我们通过国际/国内调研，为国际/国内标准的研制奠定基础，为公共卫生事件的预测、防控、处置和管理提供支持，综合分析各国的国情、体制机制、政府执行力、文化背景与民众接受度等。2022 年的冬季奥运会很快就要来到，届时又逢冬季呼吸系统传染病高发，如何防控新发、突发传染病，是迫在眉睫的问题，如何在疫情下成功举办奥运会测试赛及正式比赛，将需要构建很好的数字城市模型，用社会计算方法进行奥运比赛风险评估及政策效果评估。同时兼顾增进各国交流合作的奥林匹克精神与抑制城市疫情传播的风险，构建风险可控的奥运比赛流程，国内外合作机制，动态监测及闭环控制策略。

7.2 疫情防控常态化下如何稳步有序开放国门？

中国科学院自动化研究所曹志冬副研究员作了“疫情防控常态化下如何稳步有序开放国门？”的报告。新冠肺炎疫情全球大流行愈演愈烈，新冠病毒将在未来很长一段时间内与人类共存，疫情防控常态化已成为世界各国不得不痛苦面对的巨大挑战。各个国家都在积极探索一条适应于本国国情的与新冠病毒共存的可持续发展道路。报告将就这一命题开展讨论，一方面分析未来 1-3 年全球新冠疫情演化后可能形成的格局，另一方面，探讨我国如何根据高度分化的各国疫情风险态势，在确保我国疫情不会失控和沦陷的安全前提下，科学、稳步、有序开放国门的可行之策。

7.3 后疫情与大数据时代的社区风险防范与治理

清华大学刘奕副研究员作了“后疫情与大数据时代的社区风险防范与治理”的报告。社区是社会治理的基本单元，社区是基层基础，只有基础坚固，国家大厦才能稳固。社区具有风险因素高度汇聚、耦合关联和动态不确定的特性，随着大数据等信息技术的发展，海量数据信息给社区风险治理提出新的机遇和挑战。爆发于 2019 年末、持续到 2020 年尚未结束且已席卷全球的新冠肺炎疫情，对全社会的生产、运行乃至人们的生活方式，都带来巨大的冲击，社区成为抗击疫情的前沿阵地和基础防线，数据的采集和使用也达到前所未有的规模。面对后疫情和大数据时代，社区风险防范与治理，能交出怎样的答卷，是科学研究和社会治理面临的共同问题。

7.4 地理人工智能如何助力公共卫生管理

中国科学院计算机网络信息中心郭旦怀副研究员作了“地理人工智能如何助力公共卫生管理”的报告。随着人工智能时代的来临，以深度学习为代表的人工智能技术正与不同的学科产生交叉碰撞。在全球面临新冠疫情等重大公共卫生事件的背景，地理人工智能将如何助力公共卫生管理是一个值得关注的话题。本报告将从地理人工智能的内涵和外延开始，探讨地理人工智能在疾病监测、疫情分析、地理模拟、仿真对抗、医疗资源配置、政策分析与模

拟等方面的研究和应用进展。

8 社会计算与风投网结构治理专题

8.1 使用加权 k 均值结合集中度度量来识别中国领先的风险投资企业

清华大学罗家德教授作了“使用加权 k 均值结合集中度度量来识别中国领先的风险投资企业”的报告。尽管确定领先的风险投资公司（VCs）在分析中国投资市场方面是一个有意义的挑战，但相关文献很少提及该研究主题。给定风险投资的共同投资网络，确定领先的风险投资就等于确定复杂网络分析领域中的影响节点。由于使用单一集中度度量和多准则决策分析（MCDA）方法来识别领先的风险投资存在一些弊端和局限性，因此罗老师结合了几种不同的风险投资共同投资网络的集中度度量，然后提出了一种新的方法。用加权 k 均值对 VC 进行分组和个人排名，并确定领先的 VC。所提出的方法不仅显示基于多个评估标准的替代分组，而且根据其综合得分对这些分组进行排序，该综合得分是这些标准的加权总和。实证分析表明，该方法可用于识别中国领先的风投公司。

8.2 大数据风险投资领袖识别下的联合投资绩效研究

中央财经大学信息学院杨虎副教授作了“大数据风险投资领袖识别下的联合投资绩效研究”的报告。在中国风险投资市场，风险投资机构为获取更多投资机会和资源、分散投资风险、产生协同效应等原因，常常与其他投资机构进行联合投资，以提升自身投资绩效。本文通过对中国风险投资机构的联合投资事件的实证研究，基于综合评价方法整合多个中心度指标评价风险投资机构在联合投资网络中的排名和投资领袖，并回答投资机构的排名和投资领袖是否能够反映风险投资机构的投资绩效的问题。报告的创新点在于：（1）借助综合评价方法来评估风险投资机构在联合投资网络中的排名，并论证投资机构的排名与其投资绩效之间的关系：风险投资机构的排名越靠前，其投资绩效越高；（2）基于投资机构在联合投资网络中的综合排名定义投资领袖，并证实投资领袖的投资绩效显著优于其他投资机构；（3）验证其他机构与投资领袖联合投资能够促进其投资绩效的提升。本文的研究结果为评价投资机构的排名、识别投资领袖，选择联合投资伙伴提供参考依据，并具有较强的理论和实践指导意义。

8.3 Research on semi-supervised community detection of industrial network: Taking Chinese venture capital industry as an example

北京师范大学樊瑛教授作了“Research on semi-supervised community detection of industrial network: Taking Chinese venture capital industry as an example”的报告。在风险投资（VC）领域，VC 公司更倾向于和其他公司联合投资，它们之间构成连边便形成 VC 网络。本文借助联合投资数据和 VC 领袖名单，来实现一种对 VC 网的半监督社群划分。研究方法的核心是根据 VC 领袖的连通集团的演变情况来设计社群划分的初始标签。结果显示 VC 网

的社群结构具有较明显的区分特征。报告利用 VC 网（Venture capital networks）中隐含的先验信息，设计一种半监督社群划分算法，应用于 VC 网，检验该算法的有效性，以及该算法在社群划分准确性及合理性上优于传统的无监督的 EO 社群划分算法的特征。同时，对 VC 网上的半监督社群划分结果进行比较分析，以发掘 VC 网的社群结构及性质。

8.4 The Methodology of Social Computing—Taking VC Network Dynamics as an Example

北京化工大学谷伟伟老师作了“The Methodology of Social Computing—Taking VC Network Dynamics as an Example”的报告。风险投资是一项高风险和高回报并存的商业活动，也是帮助潜力企业快速成长的重要力量。中国的风险投资，通常是由大型投资机构领投，其他小型机构跟投，机构之间通常具有 leader 和 follower 的角色关系。大型投资机构为了分散风险，今后方便获得更好的投资机会通常会和其他机构合作投资，这种合作投资的关系构成了风险投资合作网。中国有成千上万家风险投资机构，它们之间的交互关系会形成什么特性的网络？演化机制又是如何？我们应该如何对其动态建模？又能够得出中国风险投资怎样的特点？本报告基于中国风险投资的实证数据，建模了风险投资网络的演化过程，揭示了中国风险投资界的小世界及精英俱乐部的性质。

9 社会计算与社会发展专题

9.1 算法的价值取向与社会后果：以抖音和快手为例

清华大学社会科学学院社会与金融研究中心主任郑路教授作了“算法的价值取向与社会后果：以抖音和快手为例”的报告。算法作为一种技术，是否是中立的？互联网企业在算法上的选择将会带来什么样的社会后果？本研究以抖音和快手为比较案例，围绕二者在推荐算法上的差异，考察这两个平台在运营思路、内容风格、平台治理等方面的异同，以及所产生的社会和经济后果。本研究试图表明，在这个算法为王的时代，企业的价值取向影响算法的设计，制造出特定的平台生态和治理模式，不仅影响着企业的长远发展，而且产生了深远的社会影响。

9.2 中国计算社会学的崛起：宏观定量社会学和社会预测

南京大学陈云松教授作了“中国计算社会学的崛起：宏观定量社会学和社会预测”的报告。计算社会学的发展方兴未艾，中国计算社会学也正在快速崛起。其中，宏观定量社会学从大数据中构建出可以用传统计量模型处理分析的社会指标，拓展了传统社会定量分析的领域；社会预测则借助机器学习方法形成了社会定量研究的全新路径。

9.3 科技“抗疫”中的平台治理：企业自主性与社会学干预

中国社会科学院社会学研究所吕鹏研究员作了“科技“抗疫”中的平台治理：企业自主性与社会学干预”的报告。新冠病毒疫情在世界范围内为数字平台企业参与治理提供了一次外生的刺激。理解数字平台企业在疫情期间与国家、市场、社会的四角关系，离不开对疫情前“数字平台治理场域”的分析，并着眼于疫情之后社会治理的常态。本文从自我治理、外部治理、共同治理三个“亚场域”出发，认为我国在疫情之前就已经形成了一个由国家和数字平台企业非对称共栖的治理生态；而疫情则进一步强化了平台企业参与治理的合法性。理解这一变化，不能仅仅把企业视为被动的国家治理工具，而是要从“企业自主性”出发理解企业作为行动主体的逻辑与策略。与批判资本与技术作恶的研究范式不同，本文强调辩证看待由此带来的治理效能，数字平台企业亦有可能成为善治的主体。防止这一目标偏离的关键之一，是通过源头参与、内容治理、赋能社群等途径进一步积极地促成算法审计，搭建社会平台。计算社会学的使命不仅包括预测，更需要带入社会学干预的行动方案。

9.4 基于多源大数据的城市犯罪时空结构与影响因素研究

西安交通大学贺力副教授作了“基于多源大数据的城市犯罪时空结构与影响因素研究”的报告。城市建成环境和社会环境是刻画城市犯罪时空分异的两大重要维度，是决定罪犯、目标、监视时空趋同频率的关键因素，掌握两者对犯罪时空分布的影响规律是犯罪地理学、犯罪社会学的核心内容。以暴力犯罪和盗窃犯罪统计数据为基础，借助手机信令大数据和街景图像大数据，可实现对城市风险人口的准确刻画以及对微观尺度建成环境的精细量化。进而，通过剖析犯罪时空格局的驱动因素，可揭示建成和社会环境对暴力和盗窃犯罪的促进和抑制机理，丰富犯罪社会学和犯罪地理学的理论和方法，为城市犯罪的事前防控提供切实支撑。

10 社会计算与数字空间政府治理专题

10.1 数字化国家能力与治理现代化

清华大学孟天广副教授作了“数字化国家能力与治理现代化”的报告。人类社会正在经历第四次工业革命，新一轮工业革命驱动着人类社会的快速化、大尺度和深层次数字化转型，政府数字化转型成为未来政府建设的必由之路。讲座将结合全球趋势和中国经验，阐述数字与智能技术如何通过技术赋能和技术赋权双重逻辑，信息、平台、信任、普惠和协同五大机制，成为现阶段国家治理现代化的技术引擎。伴随着政府数字化转型，数字化国家能力和数字治理生态成为治理现代化的两大抓手，驱动着数字政府和智慧社会的同步演进，为治理现代化贡献重要力量。

10.2 基于大数据和人工智能的社会治理现代化

哈尔滨工业大学刘鲁宁教授作了“基于大数据和人工智能的社会治理现代化”的报告。随着信息技术不断发展，大数据和人工智能等新一代信息技术逐渐开始被政府应用于社会治

理。本报告将从理论研究到应用研究、从已完成研究到未来研究方向，系统地讨论大数据和人工智能对社会治理现代化的推动作用。

10.3 数字空间的基本特征与治理何以可能

桂林理工大学章昌平教授作了“数字空间的基本特征与治理何以可能”的报告。报告主要介绍了四个部分的内容：一是数字空间概念的基本界定，二是数字空间呈现出的基本特征，三是第四次工业革命作用下，物理空间-社会空间-数字空间构成的三元空间高度互动和互嵌背景下对政府治理提出的挑战，四是数字空间治理的驱动力量与实现逻辑。

10.4 问题与延伸：进一步推进数字空间政府治理的思考

北京理工大学徐磊教授作了“问题与延伸：进一步推进数字空间政府治理的思考”的报告。数字空间的政府治理，是在网络链接、数字处理以及智能计算等新一代技术基础上发展起来全新政府治理模式。目前，这一模式在中国各地所呈现出的形态各异的实践特性，表明它不仅是一种技术或程序层面的变化，实际上它也引发了制度安排与治理理念的演变，这似乎是一种塑造新型文明的行为。报告从当下技术、组织及观念的等方面存在的问题，展望进一步推进数字空间政府治理的问题与路径，以期梳理形态各异的经验素材，发现其中可用于推演和展望未来的理论脉络。

11 总结

本次会议以大数据和社会计算交叉学科研究为背景，探索大数据与社会计算在国家发展重大战略中的应用，展现跨学科领域最新的突破性研究进展，交流新的学术思想和方法。由于疫情影响，本次会议全程通过 bilibili 网络直播，引起了广泛的社会反响。持续观看人数 8-9 千，最高峰观看人数达到 1.4 万人为历史之最。同时来自各个不同领域的与会者在各个专题汇报与论坛展开了讨论，有效的推动了大数据与社会计算间跨学科知识的融合。本文对 4 场主题报告和 8 场专题报告进行了总结。

· 专题一：区块链技术及应用 ·

区块链与数据治理

孟小峰* 刘立新

(中国人民大学 信息学院, 北京 100872)

[摘要] 当下,大数据的“堰塞湖”已经形成,数据治理问题迫在眉睫。传统的治理概念来自政府、企业、IT 领域,数据治理既有其一般性,也有其特殊性。本文提出数据治理的根本保障在于增加大数据价值实现过程的透明性。区块链凭借去中心、公开透明和不可篡改的特性与大数据价值实现的透明性需求相契合,能够克服当前数据治理存在的问题,为数据治理提供了新的解决思路。同时,基于区块链实现数据治理也面临诸多挑战。

[关键词] 数据治理;区块链;隐私保护;溯源问责;决策可信

DOI:10.16262/j.cnki.1000-8217.20200313.012

大数据时代,数据源源不断产生并自主汇聚至多方数据收集者,数据已经成为企业间竞争的关键和影响国家竞争力的重要因素,由此数据治理成为企业治理和国家治理的重点领域和重要方式^[1, 2]。然而,大规模数据收集也带来严峻的隐私泄露、数据滥用和数据决策不可信等问题,对传统的数据治理提出了新的挑战。例如,“Facebook-剑桥分析”事件^[3]就是大规模数据收集导致的隐私泄露、数据滥用和决策不可信的典型案例。进一步,大规模数据自主汇聚还导致数据垄断困境的出现,使数据被不合理的分配与享用^[4]。大数据的“堰塞湖”已经产生,如何使这些问题得到有效解决,并使数据得到正确和规范的使用是决定大数据继续发挥价值的关键,也是目前数据治理亟待解决的问题。

上述问题产生的主要原因是大数据价值实现过程的不透明。大数据收集和共享流通过程不透明导致隐私泄露和数据滥用等问题追踪问责困难,并且致使数据垄断问题悄然形成却缺乏评估和解决依据;大数据存储、处理和共享流通过程中缺乏透明导致数据被篡改等问题难以被发现,影响决策数据质量并最终导致数据决策不可信。由此可以得出,当前数据治理的根本保障在于增加大数据价值实现过程的透明性。数据收集和共享流通过程透明地对数据流向进行记录,以溯源问责的方式进行隐私保



孟小峰 博士,中国人民大学教授,博士生导师,CCF 会士,主要研究方向为数据库理论与系统、大数据管理系统、大数据隐私保护、大数据融合与智能、大数据实时分析、社会计算等。

护^[5]和为解决数据垄断提供依据;数据存储、处理和共享流通等过程透明使决策数据可审计和促进数据决策可信。数据治理实现途径有多种方式,除了法律法规和政策标准,还需要技术方法的保驾护航。区块链起源于数字货币,具有公开透明、去中心和不可篡改的特性。该技术的进步发展为解决当前数据治理面临的问题带来新的机遇^[6-10]。

本文提出了数据治理的根本保障在于增加大数据价值实现过程中的透明性,总结了数据治理的发展历程和技术上实现数据治理的关键内容,并对基于区块链实现数据治理的研究现状进行分析和总结,最后提出目前数据治理面临的挑战。

1 数据治理概述

“治理”(Governance)一词起源于拉丁文“掌舵”(Steering),最初用于“政府治理”,目标是协调政府与其他社会主体之间的利益。后来逐渐受到企业的

收稿日期:2019-12-30;修回日期:2020-02-13

* 通信作者,Email:xfmeng@ruc.edu.cn

本文受到国家自然科学基金项目(91646203,61532010,91846204,61532016)的资助。

认同和重视,出现了“企业治理”,目标是协调企业内部利益相关者的利益。伴随着IT资源和数据资源的日益丰富,又出现了“IT治理”和“数据治理”^[1,2]。后来,由于大数据的流通性、多源数据融合和涉及多方参与主体等应用特性,“数据治理”又进一步延伸,出现了“大数据治理”。“大数据治理”关注大数据生命周期中数据生产者、数据收集者、数据使用者、数据处理者和数据监管者^①等各方参与主体,其目标是在兼顾各方参与主体的权利、责任和利益的前提下发挥数据价值,即大数据价值实现和风险规避。

由于“大数据治理”是“数据治理”的延伸,为避免混淆,本文后续内容采用“数据治理”的概念来探讨大数据时代的数据治理。数据治理的发展过程和涉及的参与主体如图1所示。

大数据的应用特性与数据治理的目标决定了当下数据治理的关键内容。目前,数据治理的关键内容和挑战聚焦在以下3个方面:

(1) 提高决策数据质量。大数据价值实现需要多源数据的融合,然而大数据来源广泛且生命周期内涉及多方参与主体,数据是否真实产生、数据被篡改和多源数据的标准和类型不一致等问题都会影响决策数据质量,进而影响数据使用者的数据决策结果。所以,数据治理需要支持大数据在其全生命周期内的溯源。

(2) 评估与监管个人隐私数据的使用。大数据应用的流通特征使数据生产者对数据获取和共享缺乏知情权和控制权。作为数据生产者,用户不知道哪些数据被收集、被谁收集、收集之后流向哪里和作何使用。同时,数据的收集汇聚导致数据垄断现象

出现。数据垄断可能会阻碍市场竞争、使消费者福利受损、阻碍行业技术创新和带来更严重的个人隐私泄露风险等问题,但数据监管者却无法对数据应用进行评估和监管;此外,大数据应用的多源数据融合特征还可能会引发更严峻的隐私泄露问题。所以,数据治理需要对个人隐私数据使用进行评估与监管。

(3) 促进数据共享。数据共享可以促进大数据价值实现和缓解数据垄断,但同时也需要解决隐私保护等问题。一方面,数据共享双方之间发生数据共享流通时,考虑到隐私问题,需要以有效的方式保护数据生产者的个人隐私。另一方面,限于法律和实际应用中的一些因素,需要在不直接传输原始数据情况下,依据多方数据持有者的数据实现分布式数据集进行统计分析和分布式机器学习。由于多方参与者之间不存在完全的可信性,此时应该能够保护数据使用者对其共享过程进行验证。所以,数据治理需要在权衡数据生产者和数据使用者等参与主体利益的前提下促进数据共享。

数据治理需要综合法律法规、政策标准和技术方法等多种途径实现。一方面,国际组织和国家相关部门出台相应的法律法规和政策标准。例如,国际数据治理研究所从组织、规则和过程三方面总结数据治理的要素^[11];以及,国际标准ISO/IEC 38505-1:《信息技术—IT治理—数据治理》为数据治理参与主体提供原则、定义以及模型,帮助数据治理参与主体评估、指导和监督其数据利用的过程^[12]。另一方面,数据治理亟需安全、可靠的技术方法,为大数据应用过程中数据隐私保护、提高决策

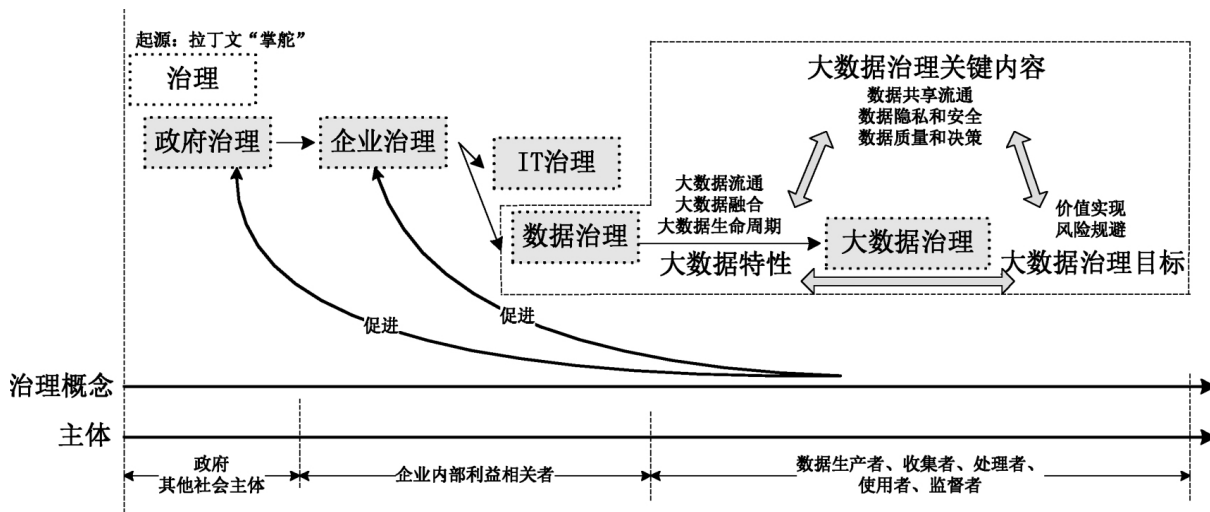


图1 数据治理发展过程和涉及的参与主体

① 各参与主体之间可能存在重合,例如当数据收集者自己使用数据并且具有处理能力时,数据收集者也充当数据处理者和数据使用者。

数据质量、促进数据共享和评估监管数据应用的合规性等问题提供技术支持。

2 基于区块链实现数据治理

区块链本质上是一种去中心化的分布式数据库,在增加大数据价值实现过程的透明性方面具有天然的优势,为解决当前数据治理的关键问题提供了可行性。

2.1 支持审计的数据存储和处理

数据决策渗透在人们生产、生活的方方面面,由于涉及多方利益相关者,数据在存储、处理和共享流通等过程中存在数据被篡改、数据伪造,以及不同来源数据的类型和标准规则差异等问题,这些问题都会影响决策数据质量。所以,数据使用者需要对决策数据进行审计。区块链作为去中心化的分布式数据库,可以实现支持审计的数据存储和处理。此外,基于区块链在不同利益主体之间构建去中心分布式数据库系统,数据通过全网快速广播至各个利益主体,也能够保证数据共享流通的真实性和及时性。

区块链网络内各节点都存储数据,数据一旦存入区块链就不会被篡改或者丢失,即使存在通信故障和蓄意攻击等问题,也仍然能保证数据存储的正确性,数据使用者可以对其进行审计。此外,将数据存入区块链还支持数据处理过程和处理结果的可审计性。对于传统的数据库管理系统,数据库中存储和维护当前数据状态,仅将数据处理过程等信息存在数据库日志,用于故障恢复,并不支持数据的历史状态查询。然而,区块链作为去中心分布式数据库,支持数据的历史状态查询,用以确认当前数据状态是否正确。基于区块链进行数据存储和处理,在保险^[13]、医疗^[14-17]和供应链^[18-21]等数据完整性要求较高领域是有重要意义的。由此,数据使用者可以对决策数据进行审计并在可信数据上执行分析和进行决策^[22-25]。

针对不同来源数据的类型和标准规则不一致等问题,可以基于区块链和智能合约制定统一的数据类型和标准规则。智能合约会被存储和同步在区块链各个节点,区块链会根据智能合约上的代码自动执行验证。由于智能合约的执行过程公开透明,使其执行过程和执行结果是可审计的,能提高多源数据共享效率且不存在单点失败。

2.2 支持溯源问责的数据获取和共享

在传统的数据获取和数据共享过程,由数据收集者制定数据使用协议并据此告知用户数据收集、

共享和使用等信息。用户作为数据生产者,对数据的知情权和可控权仍然限于法律约束和第三方信用背书。然而,由于数据获取和共享等过程对外不可见,其契约履行情况也无从考证。2014年皮尤研究中心关于美国隐私状况的报告指出,91%的受访者认为他们已经失去对数据收集者收集和使用个人数据的控制,61%的受访者对不了解数据收集者如何使用个人数据感到沮丧^[26];2016年《中国网民权益保护调查报告》显示,84%的网民对个人隐私泄露带来的不良影响有深切的感受^[27]。数据获取和数据共享不透明导致隐私泄露问题更为严峻。传统的加密、差分等隐私保护技术虽然对数据隐私具有一定的保护作用,但是目前还不足以应对大规模数据收集带来的隐私泄露风险。应用区块链的去中心性和不可篡改性,可以记录数据的获取和共享情况,进一步实施追踪溯源,并结合策略承诺(Policy Compliance)、违反检测(Violation Detection)和隐私审计(Privacy Audit),可以在隐私保护技术无效的情况下以溯源问责的方式保护隐私,也可以为评估监管数据和解决数据垄断问题提供技术支持。

目前,已有研究利用区块链增加移动应用^[28]、医疗^[29, 30]和物联网^[31-33]等领域的数据获取和共享流通的透明性。基于区块链实现数据获取和共享的框架可以分为四层:数据获取层—存储层—区块链层—共享层。在数据获取层,数据生产者对数据收集内容、形式和目的等具有知情权;在存储层,采用传统数据库管理系统、云存储和分布式存储系统等方式存储数据,并采用加密技术对数据进行加密来保护数据安全和隐私;在区块链层,由区块链执行去中心化的访问控制,使任何数据访问情况都通过区块链的交易被记录在区块链;在共享层,实现数据共享并对共享关系进行保护。正是通过上述四层,区块链增加数据获取和共享流通的透明性。

2.3 支持验证的分布式数据统计分析和机器学习

在医学研究、公共安全和商业合作等一些应用领域,需要在大规模分布式数据集上执行统计分析^[34-36]和机器学习任务^[38-41],但考虑法律法规等因素的限制,需要在不泄露隐私数据前提下进行分布式数据统计分析和机器学习。针对分布式数据集统计分析,现有方案基于安全多方计算、秘密共享、本地化差分隐私和同态加密等技术实现。然而,安全多方计算方法不适用于大规模数据提供者参与;秘密共享使数据提供者失去数据控制权;本地化差分隐私需要平衡数据的可用性和隐私损失;同态加密

能够保证数据提供者不失去数据控制权,而且不需要考虑隐私损失,但是实现的前提是数据提供者提供真实数据和计算节点的可信计算。针对分布式机器学习,由于数据提供者和数据需求者之间不存在完全的信任,各个数据提供者也可能会提供不可靠的数据或参数扰乱最终结果,以及由于经济利益等因素提前退出。所以,数据使用者需要对分布式数据集统计分析和分布式机器学习进行验证,以及需要合理的经济激励促进其顺利执行。

基于区块链实现可验证的分布式数据集统计分析常包括数据提供者、多个计算节点、多个验证节点和数据查询者。其中,数据提供者提供加密数据,多个结算节点执行密文计算,由区块链组成多个验证节点并对计算节点的计算进行验证。除此之外,分布式数据集统计分析需要考虑数据机密性、数据提供者和数据之间不可连接性、查询结果机密性和计算结果的鲁棒性等安全和隐私问题。为此通常采用洗牌和同态加密等技术进行保护。

基于区块链实现可验证的和公平的分布式机器学习,数据提供者将本地机器学习参数上传和存储至区块链,由区块链执行交叉验证,将分布式机器学习过程的每一步都记录在区块链。同时,还可以结合零知识证明和密码学承诺对恶意的参与方进行经济惩罚,通过经济激励促进公平。除此以外,分布式机器学习需要考虑数据提供者本地参数的安全性,因为本地参数也可能会泄露数据或者机器学习模型。为此通常采用差分隐私、秘密共享和同态加密等技术对其进行保护。

3 挑战与问题

区块链为数据治理提供了新的思路,但数据治理具体实现过程中也将面临诸多挑战,同时对区块链自身技术有了更高的要求。此外,基于区块链实现数据治理会导致政府和企业的管控机制和业务流程发生重大变革,这将对政府管理和企业管理提出新挑战。目前,数据治理实现过程面临的挑战与问题主要包括以下3个方面:

(1) 数据治理实现过程中面临的挑战。一方面,虽然将数据共享流通信息记录在区块链可以实现溯源问责,但是在大规模数据收集和数据共享流通错综复杂背景下,如何实现跨平台和跨领域的溯源问责是具有挑战性的问题。同时,溯源问责也可能带来隐私泄露问题,所以溯源问责过程的隐私保护也至关重要。另一方面,虽然将数据存入区块

链,可以一定程度上防止数据篡改和保证数据可以进行追踪溯源,但是保证数据存入区块链之前的真实性和可靠性仍存在挑战。

(2) 对区块链自身技术提出的新挑战。区块链自身的存储需求限制、隐私与安全、可扩展性和互操作性等方面还存在大量待解决的问题,现有比特币、以太坊和超级账本等主流的区块链还不能满足数据治理的需求。为此应该考虑设计轻量级的、高可扩展的、互联互通性较强的适用于数据治理需求的区块链。同时,伴随着各类区块链系统的出现,区块链系统评价标准与评估规范也成为亟待解决的问题。

(3) 对政府管理和企业管理提出的挑战。区块链的去中心化特性将打破传统的中心化管理方式,对政府和管理权威带来挑战;同时,去中心化特性还会使数据安全和保密的责任置于多方,对政府和企业的数据管理等方面带来新的挑战。此外,基于区块链实现数据治理并据此对数据执行相应的监管措施需要一个过程,而且随着区块链技术的迅猛发展,将会对传统的监管制度和法律法规政策提出新的要求。

4 结 语

数据治理已经成为国家治理和企业治理的重点领域和重要因素。随着各个领域数据的不断开放共享,数据治理对数据共享、数据监管和隐私保护等方面都提出了更高的要求。这些问题通过与区块链相结合可以提升数据治理的效率和透明度,将会有利于构建一个全新的数据信息时代。与此同时也会带来诸多新的挑战,需要多学科、多领域和多部门共同的努力去实现数据治理的新篇章。

参 考 文 献

- [1] 吴信东,董丙冰,堵新政,等. 数据治理技术. 软件学报, 2019, 30(9): 2830—2856.
- [2] 安小米,郭明军,魏玮,等. 大数据治理体系:核心概念、动议及其实现路径分析. 情报资料工作, 2018, (1): 5—11.
- [3] Jennifer Zhu Scott. Facebook and Cambridge Analytica: what you need to know as fallout widens. <https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html>. [2018-03-19]/[2020-01-01].
- [4] 孟小峰,朱敏杰. 数据垄断与其治理模式研究. 信息安全研究, 2019, 1(9): 789—797.
- [5] 孟小峰,张啸剑. 大数据隐私管理. 计算机研究与发展, 2015, 52(2): 265—281.

- [6] 祝烈煌, 高峰, 沈孟, 等. 区块链隐私保护研究综述. 计算机研究与发展, 2017, 54(10): 2170—2185.
- [7] 袁勇, 倪晓春, 曾帅, 等. 区块链共识算法的发展现状与展望. 自动化学报, 2018, 44(11): 93—104.
- [8] 邵奇峰, 金澈清, 张召, 等. 区块链技术: 架构及进展. 计算机学报, 2018, 41(5): 3—22.
- [9] 韩璇, 袁勇, 王飞跃. 区块链安全问题: 研究现状与展望. 自动化学报, 2019, 45(1): 208—227.
- [10] 李芳, 李卓然, 赵赫. 区块链跨链技术进展研究. 软件学报, 2019, (6): 1649—1660.
- [11] The Data Governance Institute. data governance institute framework. http://www.datagovernance.com/wp-content/uploads/2014/11/dgi_framework.pdf. [2014-11-15]/[2020-02-13].
- [12] 国家标准化管理委员会. 《信息技术—IT 治理—数据治理—第 1 部分: ISO/IEC 38500 在数据治理中的应用》. http://www.sac.gov.cn/sgzybzb/gzdt_2132/201705/t20170515_238441.htm. [2017-05-15]/[2020-02-13].
- [13] Vo H. Blockchain-based data management and analytics for micro-insurance applications//Proc of the ACM Int Conf on Information and Knowledge Management. New York: ACM, 2017: 2539—2542.
- [14] Vo, H. Research directions in blockchain data management and Analytics//Proc of Int Conf on Extending Database Technology. Bordeaux: Springer LNCS, 2018: 445—448.
- [15] Vo H. Blockchain-Powered big data analytics platform//Proc of the Int Conf on Big Data Analytics. Berlin: Springer, 2018: 15—32.
- [16] Shae Z, Tsai J P. On the design of a blockchain platform for clinical trial and precision medicine//Proc of the Int Conf on Distributed Computing Systems. Washington: IEEE, 2017: 1972—1980.
- [17] Tsai J. Transform blockchain into distributed parallel computing architecture for precision medicine//Proc of the Int Conf on Distributed Computing Systems. Washington: IEEE, 2018: 1290—1299.
- [18] Xu XW, Lu QH, Liu Y. Designing blockchain-based applications a case study for imported product traceability. Future Generation Computer Systems 2019, 92: 399—406.
- [19] Swan M. Blockchain: Blueprint for a new economy // O'Reilly Media Inc, 2015: 1—18.
- [20] Vasco L, Luis A. An overview of blockchain integration with robotics and artificial intelligence [EB/OL]. arXiv preprint, arXiv: 1810. 00329, 2018[2018-09-30]. <https://arxiv.org/abs/1810.00329>
- [21] Salah K, Rehman MHU, Nizamuddin N, et al. Blockchain for AI: review and open research challenges. IEEE Access, 2019, 7: 10127—10149.
- [22] Li Y, Zheng K, Yan Y. EtherQL: A query layer for blockchain system//Proc of the Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2017: 556—567.
- [23] Xu C, Zhang C, Xu J. vChain: Enabling verifiable boolean range queries over blockchain databases [EB/OL]. arXiv preprint, arXiv: 1812. 02386, 2018[2018-12-06]. <https://arxiv.org/abs/1812.02386>.
- [24] Zhang C, Xu C, Xu J, et al. GEM-2-Tree: A gas-efficient structure for authenticated range queries in blockchain//Proc of the 35th Int Conf on Data Engineering. Washington: IEEE, 2019: 842—853.
- [25] P Ruan, Chen G, TTA Dinh. Fine-grained, secure and efficient data provenance on blockchain systems//Proceeding of the Very Large DataBase. California: ACM, 2019: 975—988
- [26] Explainable artificial intelligence: A survey. Pew Research Center. Public perceptions of privacy and security in the post-Snowden era. <https://www.pewinternet.org/2014/11/12/public-privacy-perceptions/>. [2019-01-30]/[2020-01-01].
- [27] 中国互联网协会. 《中国网民权益保护调查报告 2016》. <http://www.isc.org.cn/zxzx/xhdt/listinfo-33759.html>. [2016-06-26]/[2020-01-01].
- [28] Zyskind G, Nathan O. Decentralizing privacy: using blockchain to protect personal data//Proc of IEEE Security and Privacy Workshops. Washington: IEEE, 2015: 180—184.
- [29] Azaria A, Ekblaw A, Vieira T. MedRec: using blockchain for medical data access and permission management//Proc of the Int Conf on Open & Big Data. Washington: IEEE, 2016: 25—30.
- [30] Dubovitskaya A, Xu Z, Ryu S. Secure and trustable electronic medical records sharing using blockchain. American Medical Informatics Association., 2017, 650—659.
- [31] Ouaddah A, Abou Elkalam A, Ait Ouahman A. FairAccess: a new blockchain-based access control framework for the Internet of Things. Security and Communication Networks, 2016, 9(18): 5943—5964.
- [32] Hossein S, Lukas B. Droplet: Decentralized authorization for IoT Data Streams [EB/OL]. arXiv preprint, arXiv: 1806. 02057, 2018[2018-11-14]. <https://arxiv.org/abs/1806.02057>.

- [33] Li R, Song T, Mei B. Blockchain for large-scale internet of things data storage and protection. *IEEE Transactions on Services Computing*, 2018; 1—8.
- [34] Henry C, Dan B. Prio: Private, robust, and scalable computation of aggregate statistics// *Proc of the 14th USENIX Symposium on Networked Systems Design and Implementation*, Berkeley CA: USENIX, 2017; 259—282.
- [35] Froelicher D, Egger P. UnLynx: a decentralized system for privacy-conscious data sharing// *Proc on Privacy Enhancing Technologies*. NJ: IEEE, 2017; 232—250.
- [36] Froelicher D, Juan R. Drynx: Decentralized, secure, verifiable system for statistical queries and machine learning on distributed datasets [EB/OL]. arXiv preprint, arXiv: 1902. 03785, 2019 [2019-02-11]. <https://arxiv.org/abs/1902.03785>.
- [37] Nelson Kibichi Bore, Ravi Kiran Raman. Promoting distributed trust in machine learning and computational simulation via a blockchain network. <http://arxiv.org/abs/1810.11126>.
- [38] Ravi K, Roman V, Michael H. Trusted multi-party computation and verifiable simulations: a scalable blockchain approach [EB/OL]. arXiv preprint, arXiv: 1809. 08438, 2018 [2018-09-22]. <https://arxiv.org/abs/1809.08438>.
- [39] Tsung T, Lucila O. ModelChain: decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks [EB/OL]. arXiv preprint, arXiv: 1802. 01746, 2018 [2018-02-06]. <https://arxiv.org/abs/1802.01746>.
- [40] Weng J, Zhang J. Deepchain: auditable and privacy-preserving deep learning with blockchain-based incentive. *Cryptology ePrint Archive*, Report 2018/679.
- [41] KUO, Tsung-Ting; GABRIEL, Rodney A, et al. Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. *Journal of the American Medical Informatics Association*, 2019, 26(5): 392—403.

Blockchain and Data Governance

Meng Xiaofeng Liu Lixin

(School of Information, Renmin University of China, Beijing 100872)

Abstract The “Quake lake” of big data has been formed, and the data governance is one of the most pressing problems that we face. The concept of “governance” comes from the “government governance”, “enterprise governance” and “IT governance”. This paper proposed that the fundamental guarantee of data governance lies in increasing the transparency of the data lifecycle. Blockchain provides a new solution for data governance by virtue of its characteristics of decentralization, transparency and tamper-proof, which can overcome the existing problems of data governance. At the same time, implementing data governance based on blockchain also faces many challenges.

Keywords data governance; blockchain; privacy protection; traceability; accountability

(责任编辑 齐昆鹏)