

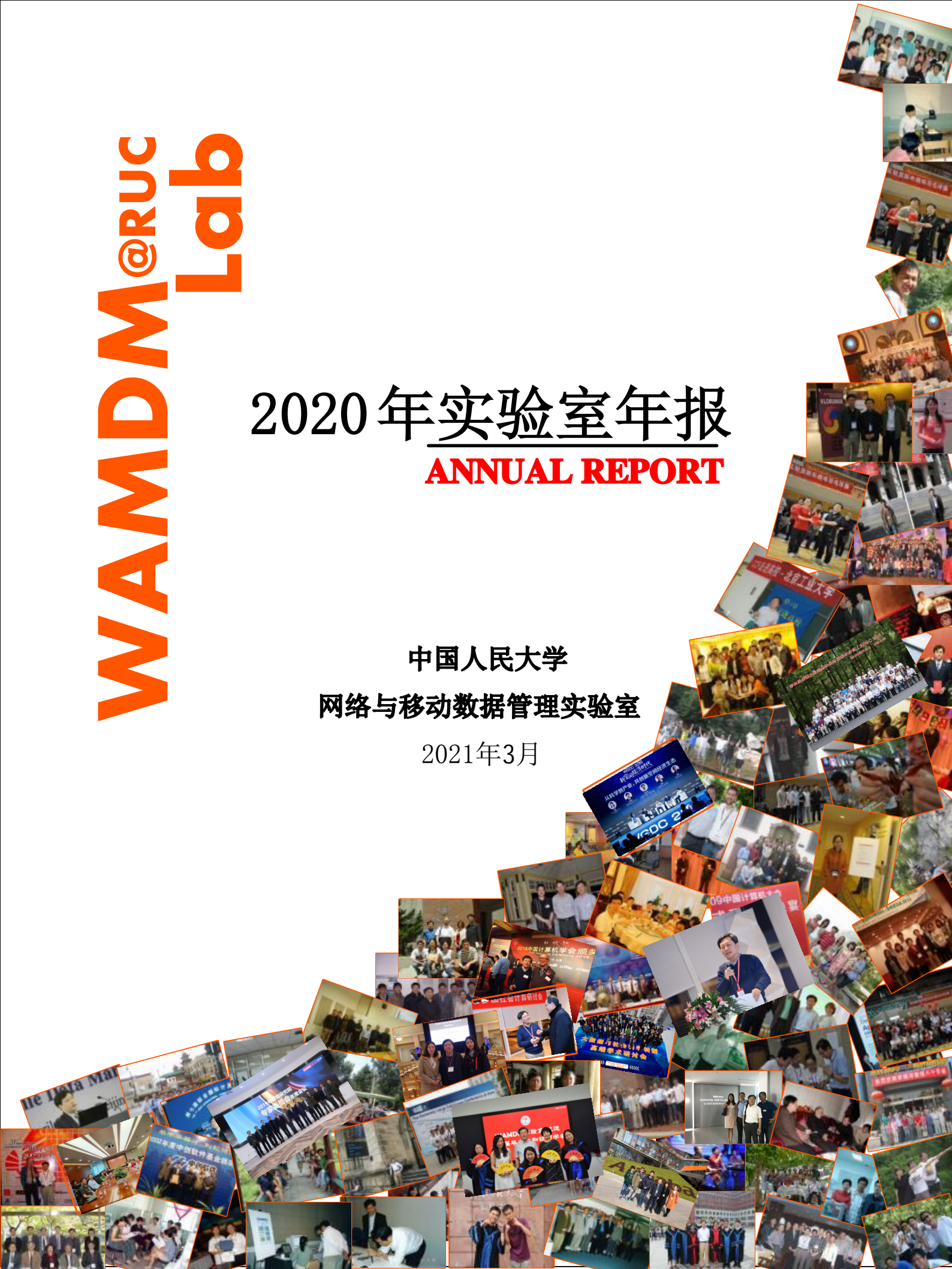
WAMDM@RUC
Lab

2020 年实验室年报

ANNUAL REPORT

中国人民大学
网络与移动数据管理实验室

2021年3月



序

2020 年，注定是不平凡的！居家度日，读些闲书是自然的，《沽酌集》便是偶读。“沽，买酒也；酌，饮酒也。我取这个题目，好像做了酒鬼似的，其实不然。打个比方罢了。平生兴趣甚少，烟酒茶均不沾，也不喜欢什么运动，只买些书来读；但我觉得就中意味，与沽酒自酌约略相近。若说不足与外人道未免夸张，总之是自得其乐。至于偶尔写写文章，到底还是余绪，好比闲记酒帐而已。”书的作者为止庵，本名王进文，学科学出身，没受过文科教育，喜欢谈书论语，不知章法，但不妄言，因此没有了文人的作态。其一句“闲记酒帐”很是能引起同感，想来这年报一记也有十六七年了，似乎有点沽酌的意味，算作闲记酒帐吧。

2020 年，注定是不平凡的！疫情使我们的社会变得更有韧性，而这一次源自我们特有的数据力量，但同时我们也将面对更多数据风险，因此我们急需要恰如其分数据治理！

2020 年，中国迎来了数据治理元年！实验室迎难而上，基于 2016 年开始的研究基础，在数据智能和数据治理方面取得了一些阶段性成果。

数据智能：信息技术的发展先后呈现三条主线，即“计算主线”、“应用主线”、“数据主线”。当下是数据主线主导的时代，即有什么数据便催生什么应用，进而促使算力的提升。记得 2016 年暑期在南开大学一次报告中首次提及数据智能的思考，之后在国家重点研发项目和基金委重大研究计划项目支持下，先后探索科学数据智能方法、空间数据智能方法、以及学术数据智能方法，逐渐积累了一套基于“层叠”历史数据辨规律（“算旧账”）、基于实时数据找异常（“抓现行”）、基于规律异常做决策（“拿主意”）的数据智能方法论。在此背景下，实验室联合国家天文台等单位以地基广角相机阵（GWAC）获得的真实巡天数据为基础，举办了首届科学数据智能发现大赛（SciDI-Cup）。从竞赛中我们也发现了一些全新的科学挑战，包括短时标暂现源的发现与验证、大规模数据的实时标注以及数据的长时存储等。为了解决这些问题，我们形成了题为“科学数据智能发现与管理”的相关研究成果，得到了广泛关注。未来，实验室会继续聚焦数据智能相关问题，致力于解决其中面临的综合挑战，并与社会计算相结合，形成新的科研范式。

数据治理：大数据时代，海量数据源源不断地产生与积累，数据垄断、隐私泄露等问题随之凸显。在此背景下，实验室连续三年发布了《中国隐私风险指数分析报告》。2020 年度相比前两年获取了更大规模的分析数据，同时导出了中国隐私风险指数三年的变化情况。三年的对比分析显示出用户与 APP 的隐私风险状况和数据垄断状况有轻微缓解的趋势，但整体形势依然严峻。为破解当下的数据治理难题，实验室在中国数据治理理论与实践方面进行了深入研究，本人应邀在第十七届中国信息系统及应用大会、第五届中国数据安全与隐私保护大会、首届数字经济与人口发展研讨会等学术会议上对相关成果作大会报告，并在人民论坛发布了有关数据垄断及其治理模式的研究成果，引发广泛关注。

在学术活动方面，实验室于疫情期间积极探索绿色高效的交流方式，为学术同行提供力所能及的服务和共享。恰逢 2020 年我国决定新增交叉学科作为新的学科门类，国家自然科学基金委新成立交叉科学部，在国家发力交叉学科的背景下，实验室参与举办首届中国空间数据智能学术会议（SpatialDI2020）、第五届全国大数据与社会计算学术会议（BDSC2020）、第五届中国数据安全和隐私保护学术会议（ChinaPrivacy2020）等多场线上的交叉领域学术会议，促进计算机领域与其他领域的交叉融合，取得圆满成功。同时，

实验室为空间数据管理和 GIS 领域多学科交叉的顶级学术会议 ACM SIGSPATIAL 2021 首次来到中国北京做了充足准备，欢迎大家关注参与！

在人才培养方面，实验室本年度成果显著。本年度博士毕业 3 人，硕士毕业 7 人。其中，博士毕业生叶青青在基于本地化差分隐私的隐私保护数据收集、对抗机器学习等研究方向上取得高水平的学术成果，入选香港理工大学 **Strategic Hiring Scheme**，入职拟聘研究助理教授。

一分耕耘，一份收获，同时也感谢多年来支持和帮助过我们的人们，谨以此致以诚挚的谢意！

孟小峰

2021 年 3 月 4 日于人大

目 录

实验室年度亮点

1. 实验室连续三年发布《中国隐私风险指数分析报告》	3
2. 实验室在中国数据治理理论与实践上取得重要成果	4
3. 孟小峰教授在人民论坛发表有关数据垄断及其治理模式的研究成果	4
4. 举办首届科学数据智能发现大赛 (SciDI Cup), 开创研究型大赛先河	5
5. 共同创办中国空间数据智能学术会议 (SpatialDI 2020)	6
6. 孟小峰教授应邀担任 ACM SIGSPATIAL 2021 大会主席	6
7. 实验室促进社会计算交叉学科的发展, 举办多场线上学术会议	7
8. 博士毕业生叶青青入选香港理工大学人才引进计划 (Strategic Hiring Scheme)	8

数据管理交叉学科研究

1. 破解数据垄断的几种治理模式研究 孟小峰	11
2. 大数据与社会计算研究进展 全国大数据与社会计算学术会议 (BDSC 2020)	15
3. 区块链与数据治理 孟小峰, 刘立新	28

中国隐私风险指数分析报告

中国隐私风险指数分析报告	37
--------------	----

发表论文精选

1. LF-GDPR: Graph Metric Estimation with Local Differential Privacy
Q. Ye, H. Hu, M. H. Au, X. Meng, X. Xiao.
IEEE Transactions on Knowledge and Data Engineering (TKDE), 2020.
.....49
2. Towards Locally Differentially Private Generic Graph Metric Estimation
Q. Ye, H. Hu, M. H. Au, X. Meng, X. Xiao.
IEEE International Conference on Data Engineering (ICDE), 2020, pp 1922-1925.
.....65
3. A Unified Adversarial Learning Framework for Semi-supervised Multi-target Domain Adaptation
X. Wu, L. Wang, S. Wang, X. Meng, L. Li, H. Huang, X. Zhang, J. Yan.
Database Systems for Advanced Applications (DASFAA), 2020.
.....69

科研成果

1. 论文列表.....87
2. 学位论文.....88
3. 已授权专利.....91

活动专题

1. SciDI Cup: 科学数据智能发现大赛.....95

学术交流

1. 学术活动任职.....107
2. 学术交流.....108

附录

1. 实验室研讨会.....119
2. 实验室网站.....125
3. 实验室成员.....126
4. 实验室新生感言.....127
5. 实验室毕业生寄语.....128
6. 历年年报回顾.....133

实验室年度亮点



实验室连续三年发布《中国隐私风险指数分析报告》

《中国隐私风险指数分析报告》由孟小峰教授团队（中国人民大学网络与移动数据管理实验室）发布，基于大规模真实用户数据与 APP 数据，对当前移动应用场景下用户隐私数据被收集的情况进行调研分析。目前该报告从 2018 年至 2020 年已连续编纂 3 年，2020 年度相比前两年获取了更大规模的分析数据，包括 3670 万真实用户数据（2018-2019 年度约 3000 万）和约 40 万 App 数据（2018-2019 年度约 30 万），能对当前移动应用场景下用户隐私数据被收集的情况进行更准确的调研分析，同时导出了中国隐私风险指数 3 年的变化情况。



在内容上，该报告继续沿用先前提出的中国隐私风险指数体系，旨在从用户、移动应用程序（Mobile Application，简称 App）和数据收集者（即 App 开发者）三个角度揭示当前用户隐私数据被收集的现状，及其产生的隐私风险状况。

2018-2020 三年分析对比的亮点结论如下：

- **总体风险指数：**2020 年度相比去年总体下降 15.8%，但数据垄断趋势并未减弱；
- **区域隐私风险：**区域差异化加大；
- **APP 隐私风险：**出现高等级 APP 向低等级转移的现象；
- **行为隐私风险：**“学生贷”、“团购”等行为隐私风险凸显。

报告揭示了当前严峻的数据垄断形势和用户隐私风险问题，即近两年来，用户个人隐私风险持续增高，移动应用市场数据垄断形势居高不下，旅游省份和经济发达省份与其他省份间的隐私风险差异增大。

与此同时，2018-2020 三年分析对比也显示出，用户与 APP 的隐私风险状况和数据垄断状况均有轻微缓解的趋势，这一现象得益于我国对 APP 治理的各项举措。然而，目前这些举措仍旧不足以扭转当前较高的数据隐私风险与数据垄断局势，发掘有效的数据治理技术势在必行，只有这样才能更好地响应中共中央国务院发布于 2020 年 4 月发布的《关于构建更加完善的要素市场化配置体制机制的意见》，加快培育数据要素市场。



实验室在中国数据治理理论与实践上取得重要成果

数据治理是当前大数据时代的核心问题之一，本年度实验室在中国数据治理理论与实践上取得重要成果。实验室提出主动式数据治理的概念理论，将其方法体系总结为数据治理炼金术框架，包括数据风险评估、数据风险操纵及数据风险问责三方面。最终将数据治理理论落地于以 App 治理为例的大规模数据收集治理与以雄安新区治理为例的大城市群数据治理两类实践。

在第十七届中国信息系统及应用大会、第五届中国数据安全与隐私保护大会、首届数字经济与人口发展研讨会等学术会议上，孟小峰教授应邀对该成果作大会报告，得到广泛关注。



孟小峰教授在人民论坛发表有关数据垄断及其治理模式的研究成果

近日，人民论坛发布了信息学院孟小峰教授有关数据垄断及其治理模式的研究成果。孟小峰教授团队基于 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。孟小峰教授首先以当前数据收集者们的数据获取量为依据，分析了数据垄断的成因。然后提出了三种数据治理模式，以缓解数据垄断形势、促进数据安全与公平的共享流通。最后，孟小峰教授指出：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。

该篇文章于 9 月下旬在《人民论坛》“精品力作”栏目刊发，在人民论坛网“思想理论/深度原创”栏目推广，引发广泛关注。



破解数据垄断的几种治理模式研究

孟小峰

【摘要】随着信息技术的飞速发展，数据已成为继土地、劳动力、资本、技术之后的第五大生产要素。数据垄断已成为当前数字经济发展的主要障碍。本文基于 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。文章从数据垄断的成因、危害及治理模式三个方面进行了探讨。首先，分析了数据垄断的成因，包括数据收集的不对称性、数据流动的封闭性、数据价值的独占性等。其次，阐述了数据垄断的危害，包括抑制创新、损害消费者权益、阻碍数据要素的流通等。最后，提出了三种数据治理模式：数据透明化治理、数据共享治理、数据竞争治理。文章认为，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。



图1 2014-2018年中国数字经济规模及增速

举办首届科学数据智能发现大赛，开创研究型大赛先河

为了寻找广袤银河中的“流浪地球”，实验室承办了首届科学数据智能发现大赛(SciDI Cup)。该竞赛是阿里云天池平台上首场研究型竞赛，吸引了全国高校和科研院所的广泛参与。

本次竞赛的主要目标是从时域天文大数据中发现微引力透镜和恒星耀发候选体这两种短时标稀有天体光变事件，完成从光变曲线(时序数据)中发现稀有异常子序列模式的计算任务。竞赛的数据来源于中国科学院国家天文台地基广角相机阵GWAC所采集的真实时域天文数据，数据总量近170万条(其中初赛约76万，复赛增加约93万)，观察时间跨度为6个月，能够充分支持竞赛环境。竞赛由ACM SIGSPATIAL中国分会主办，中国人民大学、中国科学院国家天文台、中国科学院计算机网络信息中心承办，国家天文科学数据中心、中国科技云协办。中国人民大学的孟小峰教授、国家天文台的魏建彦研究员、中科院计算机网络信息中心的廖方宇研究员等专家和学者担任大赛的评委和指导专家。

2020年7月4日，孟小峰教授主持了题为“天文发现面临的大数据挑战——以GWAC项目为例”的主题报告，讲座人为国家天文台的魏建彦研究员。本次讲座主要以GWAC项目为例，具体分析从天文大数据中获得科学发现的挑战所在，并解析本次比赛的科学意义和国际背景。帮助大家理解科学数据发现的真正意义。

比赛自2020年8月1日正式启动以来，共有来自全国317支队伍的362人报名参赛。参赛队伍分别来自中国人民大学、清华大学、北京大学、中科院、浙江大学、武汉大学、北京师范大学、南京大学、太原理工、陆军工程大学等40余所高校和科研院所。经过初赛比拼，共有45只队伍进入复赛。复赛的任务较初赛更具挑战性，采取机器和人工评判相结合的方式给出相应的得分及排名，最终排名靠前的6支队伍通过现场答辩决出冠亚季军。



★ 共同创办中国空间数据智能学术会议（SpatialDI 2020）

孟小峰教授作为 ACM SIGSPATIAL 中国分会主席，共同创办了 2020 ACM SIGSPATIAL 中国空间数据智能学术会议（SpatialDI 2020）。

该会议由 ACM SIGSPATIAL 中国分会主办，深圳大学地理空间信息研究团队承办，于 2020 年 5 月 8-9 日成功举办。会议以“空间数据智能：汇聚、融合及产能”为主题，特邀国内外空间数据研究领域的知名华人学者以及来自百度、阿里、滴滴、华为、京东等互联网企业的代表在空间数据智能获取、管理、分析、应用等方面进行了专题研讨。各位专家学者就高精度定位、城市众包多源感知、海量数据存储等前沿问题展开了深入的探讨。



本次会议采用腾讯会议与 B 站直播的形式同步进行，B 站同时在线观看人数峰值超过 2.4 万人，弹幕总数超过 1900 条，并在评论区产生了热烈的讨论交流。这种会议形式使更多人有机会了解空间数据智能，同时也探索了学术会议的新方式，不仅绿色高效，还促进了更多的交流和碰撞。

★ 孟小峰教授应邀担任 ACM SIGSPATIAL 2021 大会主席

2020 年 11 月 6 日，第 28 届 ACM SIGSPATIAL 2020 国际会议以在线形式在美国西雅图落下帷幕。会议宣布下届第 29 届 ACM SIGSPATIAL 2021 国际会议将于 2021 年 11 月 2-5 日在中国北京举办。ACM SIGSPATIAL 会议被公认为空间数据管理和 GIS 领域



多学科交叉的顶级学术会议。会议在北美已成功举办了二十八届（1993-2020 年），旨在聚集空间数据和地理信息领域的研究人员、开发人员、用户和从业者，促进 GIS 全方位的跨学科讨论和研究。ACM SIGSPATIAL 2021 将是该会议历史上首次在北美以外地区举办。

孟小峰教授应邀担任 ACM SIGSPATIAL 2021 大会主席，并与美国石溪大学汪富生教授，美国弗吉尼亚理工大学吕昌田教授在星期五的“Statistics and Awards”环节上进行了 ACM SIGSPATIAL 2021 会前汇报，为正式会议的举办开展积极的宣传并做足了充分准备。ACM SIGSPATIAL 2021 会议主题涵盖了空间智能、空间大数据、GIS、普适计算、空间搜索、空间系统等研究方向，具有重要的学术价值与应用价值，受到广泛关注和期待。



实验室促进社会计算交叉学科的发展，举办多场线上学术会议

实验室在社会计算领域的研究起步于 2012 年，历经近十年的深耕细作，为社会计算领域的发展做出了重要贡献。2020 年，恰逢我国决定新增交叉学科作为新的学科门类，国家自然科学基金委新成立交叉科学部，在国家发力交叉学科的背景下，实验室举办多场线上的社会计算学术会议，进一步促进社会科学与自然科学的交叉融合。

2020 年 8 月 22-23 日，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）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。会议以“社会计算与社会智能”为主题，共包含 4 个大会报



告，8 个专题，由来自社会学、管理学、经济学、复杂性科学、传播学、数字人文、计算机科学等多个学科专家学者，以及政府、企业等领域相关研究人员共计 44 位与大家分享了最新成果。会议得到 B 站观众的积极参与：自 22 日会议开始，B 站同时在线观看人数迅速攀升，峰值达到 1.4 万人。23 日，B 站同时在线观看人数均值超 6 千人。

孟小峰教授担任本次大会的共同主席，在闭幕致辞中提出三个期望：交叉学科是未来的趋势，社会计算应当领头雁，成为交叉典范；社会智能是未来的趋势，中国是社会智能的最大试验场，我们要勇于解决社会智能的真问题，新问题，不固守一城一池，敢于突破！教育变革是未来的趋势，单一学科专业的人才培养模式不足以支撑未来社会变革的需要，大数据与社会计算有足够多事情要做！我们的会议任重道远！

2020 年 12 月 14-15 日，第二届社会计算国际会议（The 2nd International Conference of Social Computing）以 Zoom 会议+B 站直播的形式线上成功召开。本次会议由清华大学社会网络研究中心，中国人民大学信息学院共同组织承办。会议旨在促进信息科学、社会学、管理学、经济学、金融学、传播学、政治学、地理学等多学科的对话与创新。大会邀请了来自清华大学、中国人民大学、南京大学、雪城大学等中外名校的 20 多位知名专家学者



者分别在大数据与分析、金融科技、公共卫生与社会计算等交叉领域作了主题报告。

孟小峰教授担任大会共同主席并在开场致辞中阐述了由理论驱动的社会科学向由数据驱动的社会科学的转型相关背景和问题，简述了当前国际社会科学转型的方向、方法和路径等相关研究状况，并对未来人工智能与社会计算在社会发展与治理的应用、跨学科领域最新的突破性研究发展、新的学术思想和方法交流等作了前景展望。

★ 博士毕业生叶青青入选香港理工大学 Strategic Hiring Scheme

2020 年,实验室在人才培养方面成果显著,本年度博士毕业 3 人,在读 8 人;硕士毕业 7 人,在读 2 人。

其中,博士毕业生叶青青在基于本地化差分隐私的隐私保护数据收集、对抗机器学习等研究方向上取得高水平的学术成果,在 S&P、ICDE、TKDE、INFOCOM 等顶级学术会议发表多篇论文。目前,叶青青博士入选香港理工大学人才引进计划 (Strategic Hiring Scheme),入职拟聘研究助理教授。



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

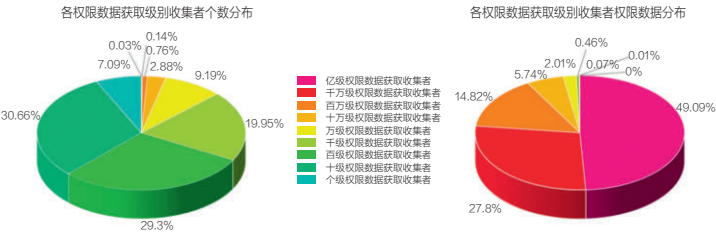


图 2 2019 年权限数据收集者数据收集分布图

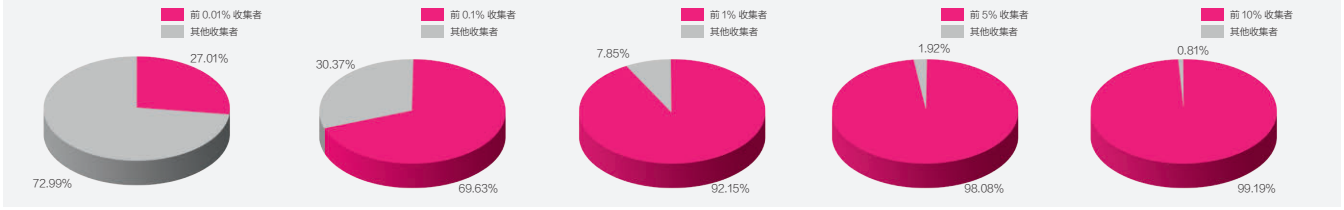


表 1 2018 与 2019 年度数据收集情况对比

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

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。

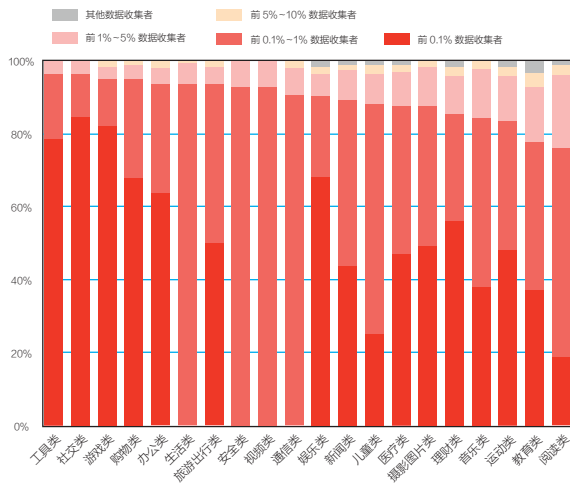
最后，从主要数据收集者垄断现状来看，笔者对数据获取量排名前 5 的数据收集者对比分析，以展示当前主要数据收集者的垄断现状。为保护数据收集者的个体隐私，该分析隐藏这 5 个数据收集者的名称，仅提供统计性结果。这 5 个数据收集者，最多的可获取 8% 的权限数据，最少者可获取 3% 的权限数据，累计可获取近 24% 的数据。也就是说，仅这 5 个数据收集者，就可获取约 1/4 的用户数据。其中，3 个数据收集者所开发 APP 涉及了 18 个以上的 APP 类别，其余 2 个数据收集者侧重于单个领域，其开发 APP 仅涉及了不足 5 个类别。这 5 个数据收集者的共同点是：其开发 APP 对应的用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

图 3 各类别权限数据收集分布情况



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。

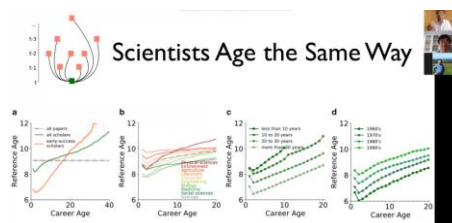


孟小峰教授代表会议组织方介绍了大会组织情况，对主办单位中国人工智能学会的大力支持表示感谢，并指出社会变革已然在技术变革的作用下悄然发生，本次会议直面社会变革所面临的真实问题（疫情应对、政府治理、大型公共活动、风投网、舆情、数据伦理、因果科学），在新的信息基础设施（Infrastructure）形成的基础上，探索建立新的社会计算研究方法，架构未来智能社会，是初心，也是责任，更是挑战，愿更多的学者参与其中！

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。

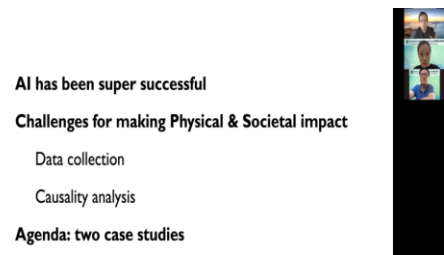


这些突破大多发生在一个领域的问题被来自另一个领域的研究人员意外地解决时。这些发现证明了提前出其不意的关键作用，并使科学机构的评估从教育和同行评议到支持它的奖项。我演示了新兴的人工智能知识发现方法，明确地结合了相关人员的分布，在预测未来发展方面的表现如何比那些只考虑思想分布的方法高出一倍。这项工作提出了一条创造最多，但也是最少的人类人工智能的道路，并优化设计了从根本上增强人类智能的集体认知多样性。

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

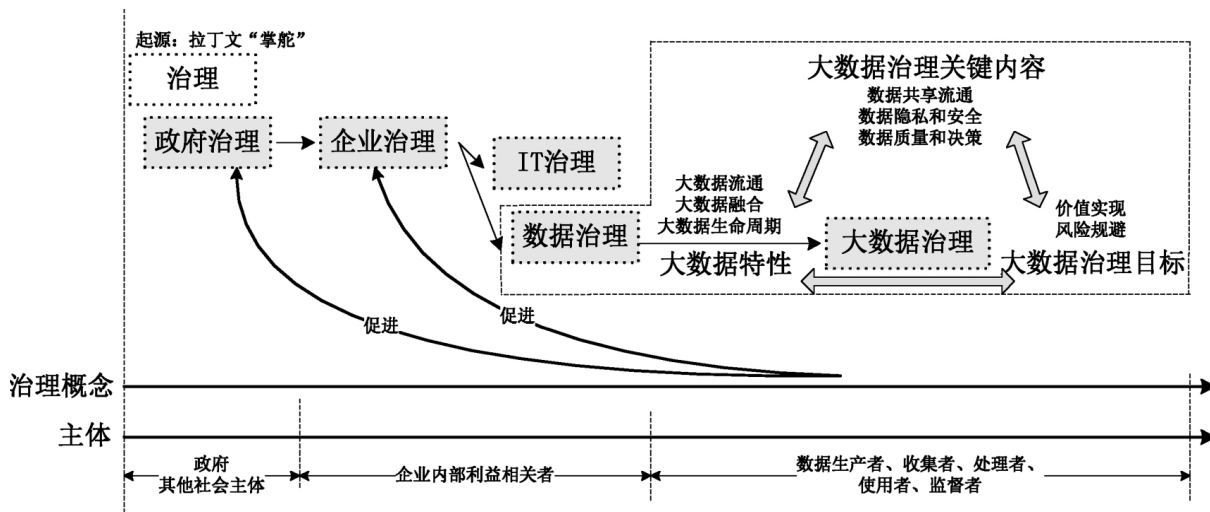


图1 数据治理发展过程和涉及的参与主体

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/sgybz/bz/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)

中国隐私风险指数 分析报告

2020 年度中国隐私风险指数分析报告

《中国隐私风险指数报告》已连续编纂 3 年，**2020 年度相比前两年获取了更多的分析数据**，包括 3670 万真实用户数据¹（2018-2019 年度约 3000 万）和约 40 万 App 数据²（2018-2019 年度约 30 万），能对当前移动应用场景下的用户隐私数据被收集的情况进行更准确的调研分析，同时**导出了中国隐私风险指数 3 年的变化情况**。在内容上，该报告继续沿用先前提出的中国隐私风险指数体系，旨在从用户、移动应用程序（Mobile Application，简称 App）和数据收集者（即 App 开发者）三个角度揭示当前用户隐私数据被收集的现状，及其产生的隐私风险状况。

基于数据拥有者（用户）量化模型，本报告对 2018-2020 年度中国总体隐私风险情况进行对比，并分析其原因。结论显示，**2020 年度中国总体隐私风险指数相比 2019 年度下降 15.8%，接近于 2018 年度的中国总体隐私风险**。而该隐私风险指数的下降伴随着用户平均安装 APP 数量的上升，该现象主要是 APP 平均请求权限敏感度下降所致，该平均敏感度相比 2019 年度同比下降 2.25%。此结论也揭示出各 APP 在 2020 年度国家各部门强力的 APP 治理政策的压力下，对敏感权限的收集有所收敛，说明了当前法律政策对 App 敏感数据收集的遏制作用。

表 1 2018-2020 中国总体隐私风险情况对比

	用户平均 App 安装量（个）	用户平均权限数据泄露量（份）	App 平均权限敏感度	总体隐私风险指数	增长率
2018 年	27	291	2.17	0.45	—
2019 年	31	336	2.22	0.57	26.7%
2020 年	56	532	2.17	0.48	-15.8%

2018-2020 中国总体隐私风险详细情况如表 1 所示。在该表中，总体隐私风险指数，是指基于数据拥有者量化模型中单个用户隐私风险指数，计算得到的所有用户样本隐私风险指数的均值。1 份权限数据指 1 个用户通过 1 次权限请求泄露的数据。对任一用户而言，1 个 App 通过其多个权限请求可获取多份用户权限数据，用户手机上的 1 个 API 可向多个 App 泄露多份用户权限数据。

在数据拥有者分析中，本报告继续基于移动用户的隐私风险指数，从区域、人群、行为三个角度对隐私风险进行评估，并进行 2018-2020 的三年对比。

2018-2020 年度区域隐私风险对比显示，**旅游省份与经济发达省份与其他区域的隐私风险指数差异增大，前者的隐私风险愈加明显高于后者**。从每年度的隐私风险排名变化上可发现，贵州省、云南省与吉林省的隐私风险排名连年上涨，吉林省三年间全国隐私排名更是提高了 14 个位次；而湖南省和江苏省的隐私风险排名连年下降，江苏省三年间累计下降了 11 个位次。从各省份的隐私风险值上，其总体的区域与中国总体隐私风险趋势保存一致，即 2019 年度的隐私风险值明显偏大，2020 年度相比之前有明显下降。

2018-2020 年度人群与行为隐私风险总体上无大变化。从人群隐私风险来看，**男性人群的隐私风险指数首年超越女性人群**。在其他属性上，**经济能力较高人群的隐私风险亦较高，该结论愈加凸显**；同时，已婚人群、青年人群（26-35 岁人群）的隐私风险相比同类别其他属性人群依旧保持着较低水平。从行为隐私风险来看，三年来出行方式与贷款方法对不同行

¹ 3670 万真实用户数据是由 AURORA 极光大数据公司提供的脱敏数据集，分析结果仅涉及统计信息，不涉及个体隐私。

² 40 万 App 数据由 WAMDM 实验室该项目参与者从应用宝、豌豆荚等应用市场爬取得到。

为人群的隐私风险影响一直较高，2020 年度“学生贷”和“团购”行为首次成为对隐私风险影响较大的偏好行为。

在数据收集者分析中，本报告对 2018-2020 年度中国的数据垄断状况进行分析对比，结论显示，10%的数据收集者依旧可收集 99%的权限数据，数据垄断的严峻形势仍居高不下。但从具体的数据来看，2018 至 2020 年，数据垄断的形势一直有着极其轻微的缓解趋势，数值上该缓解对应收集数据比例的变化幅度不足 0.1%。

在 2019 年新增的移动应用程序分析中，本报告进一步完成了 2019 与 2020 年度的结果对比。从 APP 的请求权限数量上，2020 年度各类 APP 请求权限的数量相比之前均有减少。在 APP 的权限设置隐私风险 P1 上，各类 APP 的 P1 隐私风险指数均有降低，且最高隐私风险级别对应的 APP 数量与用户量均有减少，相应地，这些 APP 与用户相对均匀地转移至其他较低的隐私风险级别中去；在 APP 的使用量隐私风险 P2 上，由于疫情期间用户对多种 APP 使用量的增加，各类别 APP 的 P2 隐私风险指数有明显上升，不同级别对应的 APP 数量与用户数量与之前的分布趋势相同，即使用量隐私风险越高，APP 数量与用户量越大。

综上，本报告展示了 2018 至 2020 三年来中国隐私风险与数据垄断局势居高不下的总体特征，从区域差异、用户差异、行为差异、APP 差异等各个角度展示了中国隐私风险的细微变化，亦证明了国家各部门对 APP 治理的阶段性成效，即各 APP 对用户敏感权限的请求均有所收敛。但目前这些举措仍旧不足以扭转当前较高的数据隐私风险与数据垄断局势，发掘有效的数据治理技术与体系势在必行，只有这样才能更好的响应中共中央国务院发布于 2020 年 4 月发布的《关于构建更加完善的要素市场化配置体制机制的意见》，加快培育数据要素市场！

一、中国隐私风险指数体系

中国隐私风险指数是一个反映我国在特定时段内数据拥有者（移动用户）因个人数据被收集者（App 开发者）获取而面临的隐私风险及数据收集者造成的隐私风险相对数值的宏观指标，用来反映不同移动用户个体或群体面临隐私风险的差异。该指数基于中国现有 334 个地级市分层抽样的 3670 万（36,722,417）真实用户的 App 使用数据、162 个维度的用户属性画像数据、通过爬取第三方应用网站得到 40 万（406,054）个 App 相关信息构成的数据集计算得到。

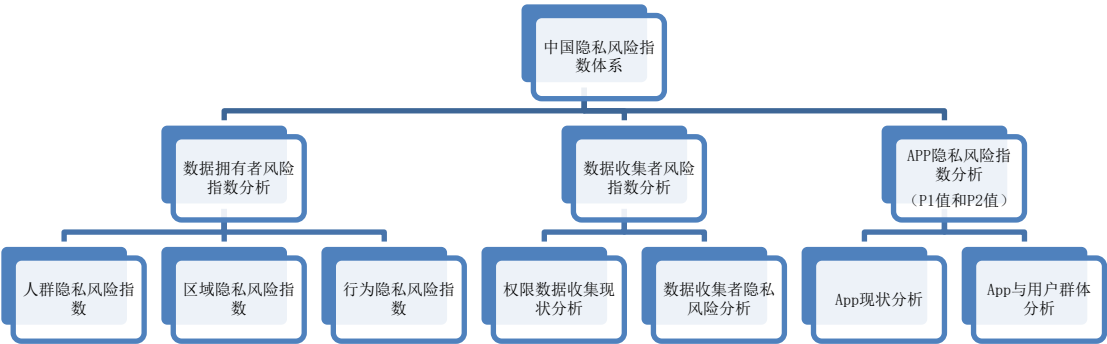


图 1-1 中国隐私风险指数体系

基于数据拥有者（移动用户）、数据收集者（App 开发者）、以及移动应用程序（APP）的隐私风险量化结果，本报告进一步制定中国隐私风险指数体系，如图 1-1 所示，从数据收集者角度揭示移动用户数据的流向，并从自然属性、社会属性、行为属性等维度揭示不同数据拥有者群体的隐私风险特征，最终分析汇总成《中国隐私风险指数分析报告》。

二、大规模数据收集现状分析

图 2-1 与图 2-2 分别展示了 2018-2020 年度不同比例的数据收集者收集数量及隐私风险占比变化情况。连续三年的分析显示：**10%的数据收集者可收集 99%的权限数据，数据垄断的严峻形势近两年只有极其轻微的缓解，总体上仍居高不下。**

如图 2-1 所示，2018-2019 年度，前 10%、5%、1%的数据收集者收集数量比例基本接近；2020 年，前 0.1%、0.01%的数据收集者收集数量比例相比 2019 年有所上涨，但仍低于 2018 年的收集数量比例。图 2-2 所示的不同比例的数据收集者的隐私风险占比变化趋势与之类似。因此可得出以下结论：

- 2019 年相比 2018 年数据垄断形势略有缓解；
- 2020 年相比 2019 年数据垄断形势有极其轻微的上涨，但相比 2018 年依旧有缓解的趋势。

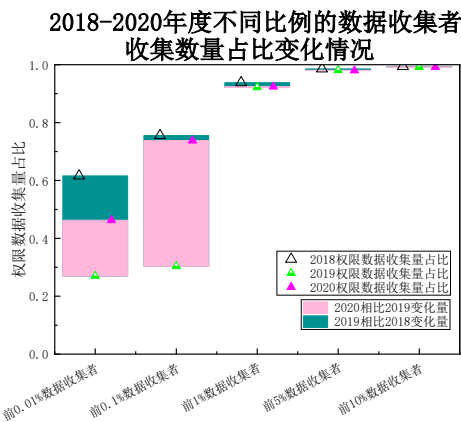


图 2-1 2018-2020 年度不同比例的数据收集者收集数量占比变化情况

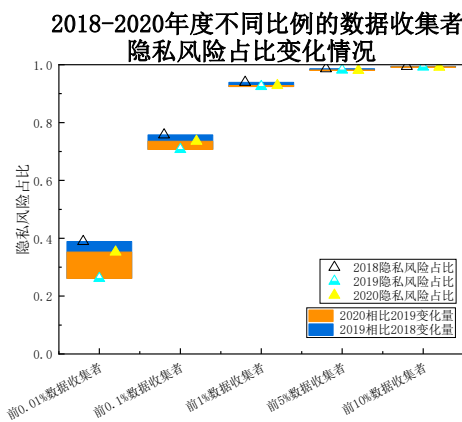


图 2-2 2018-2020 年度不同比例的数据收集者隐私风险占比变化情况

三、移动应用程序分析

本节对 2018-2020 年度各类别 App 的权限请求数量及隐私风险状况进行对比分析。

图 3-1 展示了 2018-2020 年度各类别 App 平均请求权限数量。由此可见：2020 年各类别 App 的权限请求数量均少于 2018 年、2019 年，说明**各类别 App 在 2020 年对用户权限的请求均有所收敛。**

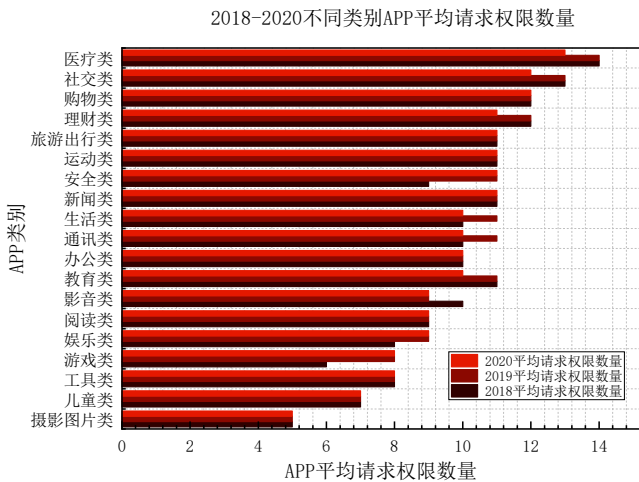


图 3-1 2018-2020 年度不同类别 App 平均请求权限数量

图 3-2、图 3-3 分别展示了 2019-2020 年度各类别 App 的 P1 隐私风险值和 P2 隐私风险值。可以看出：2020 年大部分类别 App 的 P1 隐私风险值相对 2019 年有所降低，P2 隐私风险值相对 2019 年提高较为明显。这与 2020 年“用户安装 APP 数量增加、但 APP 请求用户权限更为谨慎”结论一致。

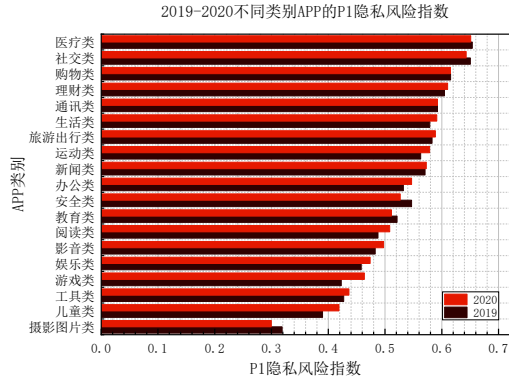


图 3-2 2019-2020 年度不同类别 App 的 P1 隐私风险指数

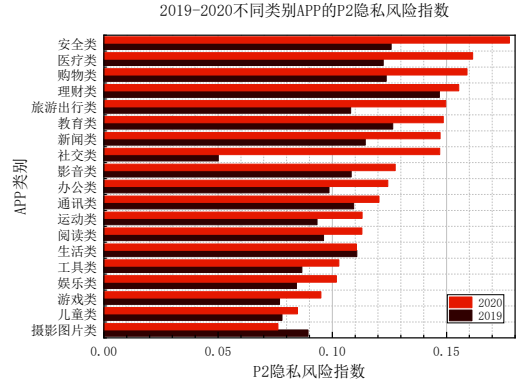


图 3-3 2019-2020 年度不同类别 App 的 P2 隐私风险指数

四、移动应用程序分级体系分析

本节对 2019-2020 年度不同隐私风险级别的 App 数量及用户量进行对比分析。由图 4-1 可得：

App 对高敏感度权限的请求有所收敛。2020 年，最高 P1-L10 隐私风险级别 App 的用户量减少，其他隐私风险级别 App 的用户量有所提高，而各隐私风险级别 App 数量主要呈现增长趋势。这与上节中“各类别 App 在 2020 年对用户权限的请求均有所收敛”的结论一致。

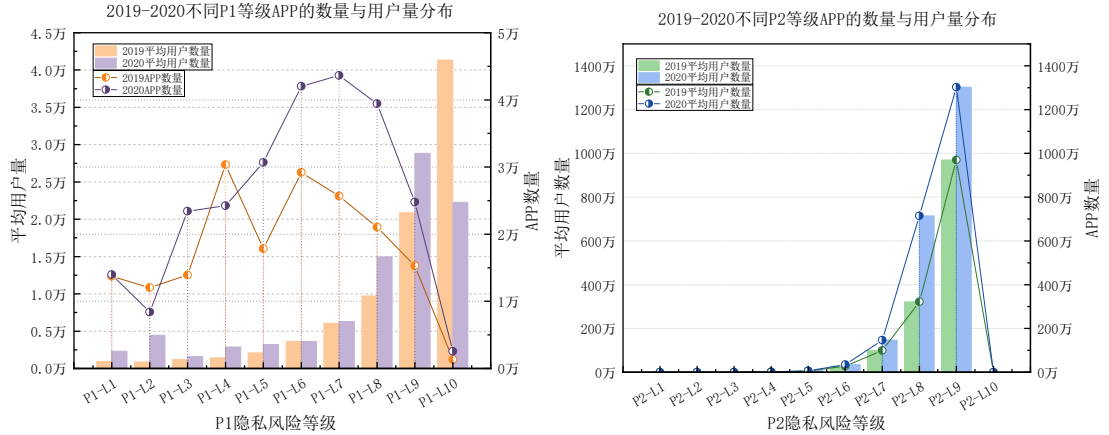


图 4-1 2019-2020 年度不同隐私风险级别的 App 数量与用户量分布

五、区域隐私风险指数

本节对 2018-2020 年度省级区域的隐私风险指数进行对比分析。由图 5-1 可以看出，省级区域隐私风险差异化趋势越来越明显：经济发达及旅游省份隐私风险较高，其他省份隐私风险较低。



图 5-1 2018-2020 年度中国隐私风险指数地图

六、人群隐私风险指数

本节将分别对职业、消费水平、收入能力、婚姻状况、年龄及性别与人群隐私风险指数的关系进行 2018-2020 年度对比分析。总体来看，除在性别基本属性上，男性人群的隐私风险首次略高于女性人群，其他基础属性对应的隐私风险分布并无明显变化。

6.1 职业

由图 6-1 可知：2018-2020 年，工程施工人员、旅游及健身娱乐场所服务人员、社会服务和居民生活服务人员的隐私风险指数相对较高，运输服务人员、安全保卫和消防人员、科学研究人员的隐私风险指数相对较低。

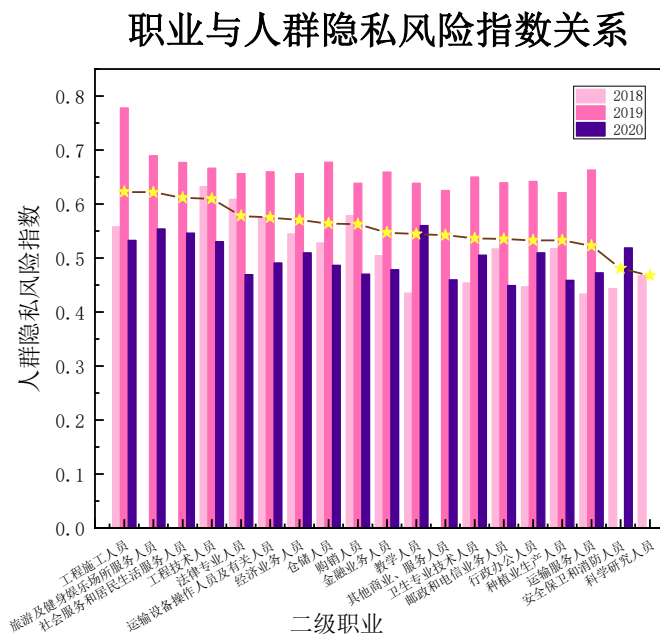


图 6-1 2018-2020 年度职业与人群隐私风险指数关系

6.2 消费水平

由图 6-2 可知：2018-2020 年，人群隐私风险指数与消费水平基本呈正比。

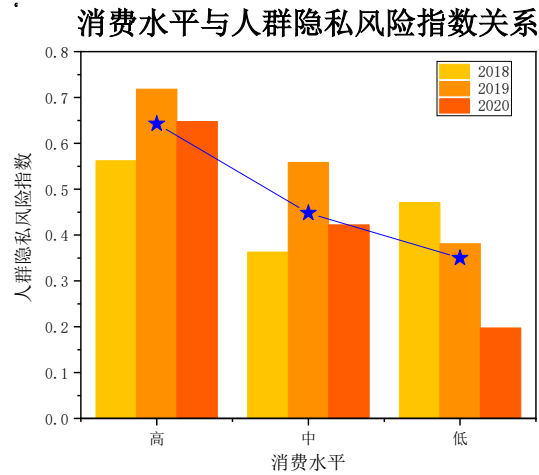


图 6-2 2018-2020 年度消费水平与人群隐私风险指数关系

6.3 收入能力

由图 6-3 可知：2018-2020 年，人群隐私风险指数与收入能力基本呈正比。

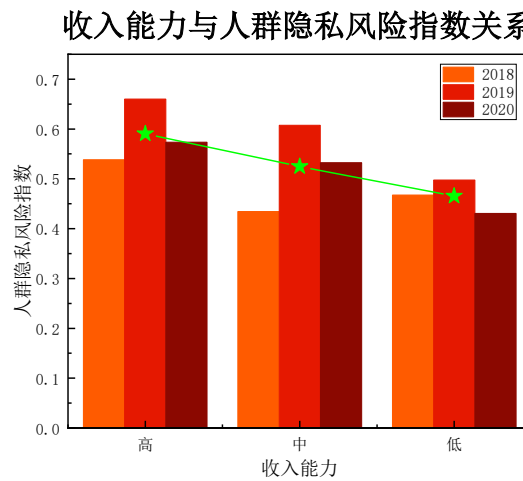


图 6-3 2018-2020 年度收入能力与人群隐私风险指数关系

6.4 婚姻状况

由图 6-4 可知：2018-2020 年，未婚人群相较已婚人群隐私风险指数高。

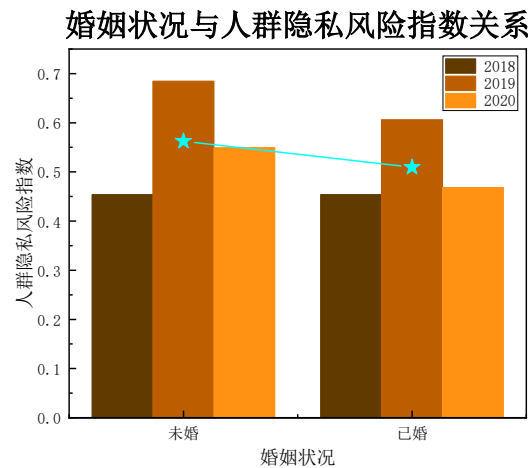


图 6-4 2018-2020 年度婚姻状况与人群隐私风险指数关系

6.5 年龄

由图 6-5 可知：2018-2020 年，26-35 岁人群相对其他年龄段人群隐私风险最低。

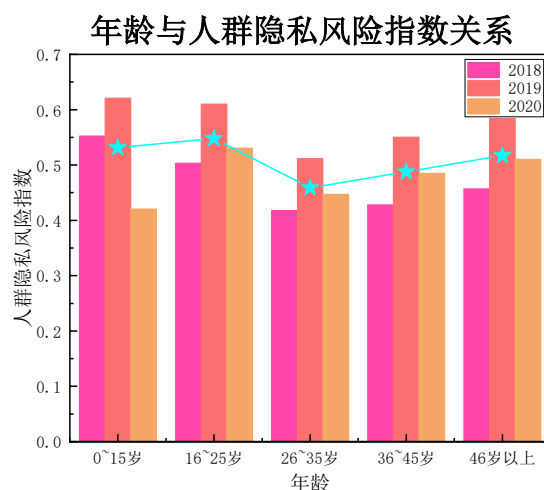


图 6-5 2018-2020 年度年龄与人群隐私风险指数关系

6.6 性别

由图 6-6 可知：2019 年之前，女性人群隐私风险偏高；但 2020 年，男性人群隐私风险略高。

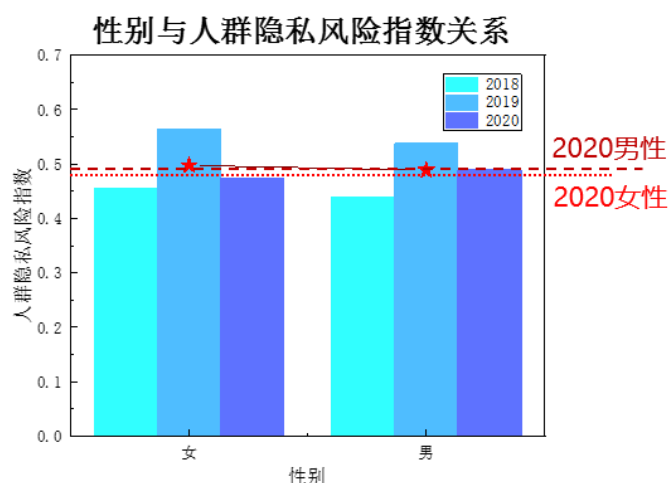


图 6-6 2018-2020 年度性别与人群隐私风险指数关系

七、行为隐私风险指数

相比 2018 年度数据集中各行为标签在全部用户中的分布较为均匀，2020 年同 2019 年的数据一样，部分标签的用户数呈现两级分化的特点，如教育行为中，超 38 万用户被标记为“儿童学习”，而“胎儿教育”的用户仅为 1 人。本报告仅对现有数据集中所体现的样本特征进行分析，并由此估计总体特征；同时，对基于各自的数据集以及分析模型得到的 2019、2020 年行为隐私风险指数的变化进行说明，并忽略数据集本身导致的系统误差。

对比 2019 年的分析结果，2020 年上述 12 类行为属性的隐私风险总体指数方差增大，隐私风险排名基本保持不变。在各类行为属性的具体分析中，不同倾向人群的人均 App 使用个数与隐私风险的变化呈现相同趋势，即该人群人均 App 使用个数越多，隐私风险指数越高。

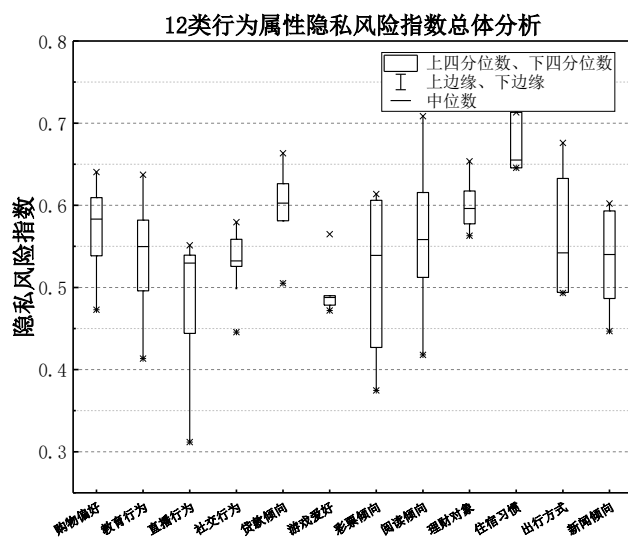


图 7-1 12 类行为属性隐私风险指数总体分析

如图 7-1 所示，2020 年度不同行为人群隐私风险分析得到如下结论：

- 不同购物偏好人群中，从事汽车专卖隐私风险由 2019 年的第七上升至第三。
- 不同教育行为中，从事 IT 教育隐私风险上升至最高，从事胎儿教育隐私风险降至最低。
- 不同直播行为中，喜欢收看美妆直播节目的人群隐私风险指数又降低至最低，收看明星直播节目的人群隐私风险指数提升至最高。
- 不同社交行为中，喜欢陌生人社交的人群隐私风险指数降低至最低。
- 不同游戏爱好行为中，偏好神话修真的人群隐私风险指数提升至最高，且明显高于其他类型。
- 不同阅读行为中，喜欢有声小说的人群隐私风险指数从 2019 年的第二降低至最低，喜欢金融知识的人群隐私风险仍然保持最高。
- 不同新闻倾向中，喜欢财经新闻的人群隐私风险仍然保持最高，喜欢综合新闻的人群隐私风险指数从 2019 年的第二降低至最低。

图 7-2 展示了 2018-2020 年度不同行为属性的隐私风险指数。分析可得，3 年来出行方式和贷款倾向对隐私风险的影响较大。

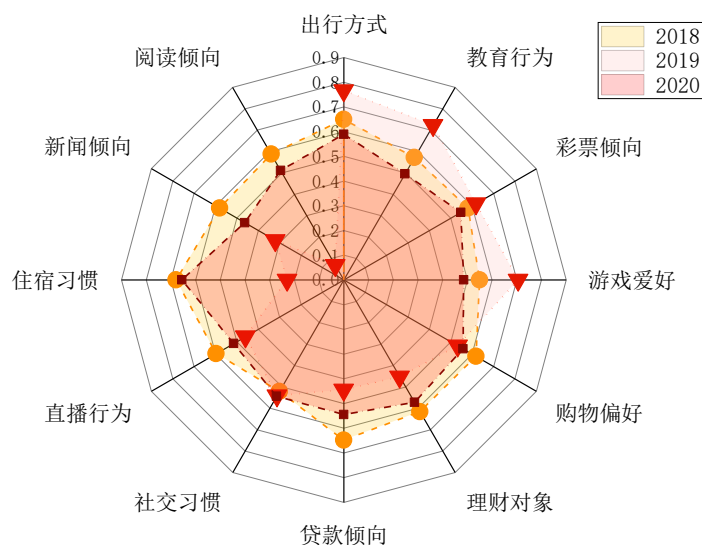


图 7-2 2018-2020 年度不同行为属性的隐私风险指数

图 7-3 展示了 2018-2020 年度的行为属性词云，字体越大表示隐私风险指数越高。分析可得以下结论：

- 2018 年，偏好阅读神话修真小说、喜欢民宿、大巴出行等行为人群隐私风险偏高；
- 2019 年，偏好星级酒店、飞机出行、喜欢返利购物、有车贷等行为人群隐私风险偏高；
- 2020 年，偏好星级酒店与民宿、大巴出行、阅读金融知识、有学生贷等行为的人群隐私风险偏高。



图 7-3 2018-2020 年度行为属性词云

八、总结

中国人民大学 WAMDM 实验室 2018、2019 年连续两年发布《中国隐私风险指数分析报告》，从数据拥有者（移动用户）、数据收集者（App 开发者）和移动应用（App）三个角度，分析了数据收集者获取的权限数据和造成的隐私风险，并从自然属性、社会属性、地域信息及行为属性多维度地分析数据拥有者（移动用户）的各群体隐私风险特征，从权限设置隐私 P1 分级和使用量隐私 P2 分级对应的不同级别 APP 分析其用户分布与隐私风险，从而构建中国隐私风险指数体系。本报告基于上述研究成果，基于 334 个地级市分层抽样的 3700 万用户和通过第三方网站爬取的 40 万 APP，进一步分析了 2020 年度中国隐私风险指数。

本报告已有成果主要如下:

- **大规模真实数据集：**约 3670 万用户集上的 179 个属性标签、40 万余 App 数据，相比前两年增加了约 670 万用户数据和 10 万 APP 数据。
- **四个主要贡献：**分析了 2018 至 2020 三年间中国隐私总体风险指数变化趋势、数据拥有者隐私风险变化趋势、数据垄断变化形势与 APP 发展趋势，揭示了当前中国居高难下的隐私风险状况与数据垄断形势。
- **三类分析对象：**数据拥有者（移动用户）、数据收集者（App 开发者）及移动应用程序（App）
- **六组分析结论：**大规模数据收集现状、移动应用程序分析、移动应用程序分级体系分析、区域隐私风险指数分析、人群隐私风险指数分析、行为隐私分析指数分析

后续将基于现有工作，进一步展开新一年度的分析，同时针对当前中国隐私风险指数体系中的不足加以改进，包括以下三个方面：

- 进一步开展 App 分级机制研究，扩宽分析维度，形成一套完整机制，促进其在实际应用程序市场落地；
- App 隐私风险预警系统构建；
- 和国家有关职能部门进一步沟通，为隐私相关政策的制定提供研究基础。

参考文献

- [1] Q Ye, H Hu, X Meng. PrivKV: Key-Value Data Collection with Local Differential Privacy[C]. IEEE Symposium on Security and Privacy (S&P). 2019:317-331.
- [2] M Zhu, Q Ye, X Yang, et al. Poster: AppPrivacy: Analyzing Data Collection and Privacy Leakage from Mobile Apps[C]. IEEE Symposium on Security and Privacy (S&P). 2019:41-42.
- [3] 孟小峰, 朱敏杰, 刘俊旭. 大规模用户隐私风险量化研究[J]. 信息安全研究, Vol5(9):778-788, 2019.
- [4] 孟小峰, 朱敏杰, 刘立新, 等. 数据垄断与其治理模式研究[J]. 信息安全研究, Vol5(9):779-797, 2019.
- [5] 孟小峰, 王雷霞, 刘俊旭. 人工智能时代的数据隐私、垄断与公平[J]. 大数据, Vol6(1): 35-46, 2020.
- [6] 中国消费者协会. 《100 款 App 个人信息收集与隐私政策测评报告》, 2018-11-28
- [7] 朱敏杰, 叶青青, 孟小峰, 等. 基于权限的移动应用程序隐私风险量化[J]. 中国科学: 信息科学, 2020. (Online)
- [8] J. Zang, K. Dummit, J. Graves, P. Lisker, and L. Sweeney, “Who Knows what About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps,” Technology Science, 2015.
- [9] Hamed A, Ayed H K B. Privacy risk assessment and users' awareness for mobile Apps permissions[C]// Computer Systems and Applications. IEEE, 2017:1-8.
- [10] Hamed A, Ayed H K B. Privacy risk assessment and users' awareness for mobile Apps permissions[C]//Computer Systems and Applications (AICCSA), 2016 IEEE/ACS 13th International Conference of. IEEE, 2016: 1-8.

发表论文精选

LF-GDPR: A Framework for Estimating Graph Metrics with Local Differential Privacy

Qingqing Ye, *Member, IEEE*, Haibo Hu, *Senior Member, IEEE*, Man Ho Au, *Member, IEEE*,
Xiaofeng Meng, *Member, IEEE*, Xiaokui Xiao, *Member, IEEE*

Abstract—Local differential privacy (LDP) is an emerging technique for privacy-preserving data collection without a trusted collector. Despite its strong privacy guarantee, LDP cannot be easily applied to real-world graph analysis tasks such as community detection and centrality analysis due to its high implementation complexity and low data utility. In this paper, we address these two issues by presenting LF-GDPR, the first LDP-enabled graph metric estimation framework for graph analysis. It collects two atomic graph metrics — the adjacency bit vector and node degree — from each node locally. LF-GDPR simplifies the job of implementing LDP-related steps (e.g., local perturbation, aggregation and calibration) for a graph metric estimation task by providing either a complete or a parameterized algorithm for each step. To address low data utility of LDP, it optimally allocates privacy budget between the two atomic metrics during data collection. To demonstrate the usage of LF-GDPR, we show use cases on two common graph analysis tasks, namely, clustering coefficient estimation and community detection. The privacy and utility achieved by LF-GDPR are verified through theoretical analysis and extensive experimental results.

Index Terms—Local differential privacy; Graph metric; Privacy-preserving graph analysis.

1 INTRODUCTION

With the prevalence of big data and machine learning, graph analytics has received great attention and nurtured numerous applications in web, social network, transportation, and knowledge base. However, recent privacy incidents, particularly the Facebook privacy scandal, pose real-life threats to any **centralized** party who needs to safeguard graph data of individuals while providing graph analysis service to third parties. In that scandal, a third-party developer Cambridge Analytica retrieves the personal profiles of 87 million Facebook users through the Facebook Graph API for third-party apps [1], [2]. The main cause is that this API allows these apps to access the **friends list** of a user by a simple authorization, through which these apps propagate like virus in the social network. Unfortunately, most existing privacy models on graph assume a centralized trusted party to release the graph data that satisfies certain privacy metrics, for example, the k -neighborhood anonymity [3], k -degree anonymity [4], k -automorphism [5], k -isomorphism [6], and differential privacy [7], [8]. However, in practice even Facebook cannot be fully trusted or is in the centralized position to release graph data on behalf of each user. For decentralized graphs in which each user or party locally maintains a limited view of the graph,

there is even no such a central party. These graphs, such as the World Wide Web, federated knowledge graphs, peer-to-peer (e.g., vehicular and mobile ad-hoc) and blockchain networks, and contact tracing graph for COVID-19, are in a more compelling need to find alternative privacy models without a trusted party [9].

A promising model is local differential privacy (LDP) [10], where each individual user **locally perturbs her share of graph metrics** (e.g., node degree and adjacency list, depending on the graph analysis task) before sending them to the data collector for analysis. As such, the data collector does not need to be trusted. A recent work *LDPGen* [11] has also shown the potential of LDP for graph analytics. In that work, LDP is used to collect node degree for synthetic graph generation. However, such solution is usually task specific — for different tasks, such as centrality analysis and community detection, dedicated LDP solutions must be designed from scratch. To show how complicated it is, an LDP solution usually takes four steps: (1) selecting graph metrics to collect from users for the target metric (e.g., clustering coefficient, modularity, or centrality) of this task, (2) designing a local perturbation algorithm for users to report these metrics under LDP, (3) designing a collector-side aggregation algorithm to estimate the target metric based on the perturbed data, (4) designing an optional calibration algorithm for the target metric if the estimation is biased. Step (4) is important as locally perturbed data often causes bias (i.e., deviation from the true mean) in the collector-side statistics. Obviously, working out such a solution **requires in-depth knowledge of LDP**, which hinders the embrace of LDP by more graph applications.

In this paper, we address this challenge by presenting LF-GDPR (Local Framework for Graph with Differentially Private Release), the first LDP-enabled graph metric estimation framework for general graph analysis. It simplifies

- Qingqing Ye is with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, and the School of Information, Renmin University of China. E-mail: qqing.ye@polyu.edu.hk
- Haibo Hu is with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, and Polyu Shenzhen Research Institute. E-mail: haibo.hu@polyu.edu.hk
- Man Ho Au is with the Department of Computer Science, The University of Hong Kong. E-mail: allenau@cs.hku.hk
- Xiaofeng Meng is with the School of Information, Renmin University of China. E-mail: xfmeng@ruc.edu.cn
- Xiaokui Xiao is with the School of Computing, National University of Singapore. E-mail: xkxiao@nus.edu.sg

Manuscript received April 19, 2020; revised August 26, 2020.

the job of a graph application to design an LDP solution for a graph metric estimation task by providing complete or parameterized algorithms for steps (2)-(4) as above. As long as the target graph metric can be derived from the two atomic metrics, namely, the adjacency bit vector and node degree, the parameterized algorithms in steps (2)-(4) can be completed with ease. Furthermore, LF-GDPR features an optimal allocation of privacy budget between the two atomic metrics. To illustrate the usage of LF-GDPR, we will also show use cases on two common graph analysis tasks, namely, clustering coefficient estimation and community detection. To summarize, our main contributions in this paper are as follows.

- 1) This is the first LDP-enabled graph metric estimation framework for a variety of graph analysis tasks.
- 2) We provide complete or parameterized algorithms for local perturbation, collector-side aggregation, and calibration.
- 3) We present an optimal solution to allocate the privacy budget between adjacency bit vector and node degree.
- 4) We show two use cases of LF-GDPR and compare their performance with existing methods on real datasets.

The rest of the paper is organized as follows. Section 2 introduces preliminaries on local differential privacy and its application in graph analytics. Section 3 presents an overview of LF-GDPR. Section 4 describes the implementation details of this framework. Sections 5 and 6 show the detailed usage of LF-GDPR in two use cases. Section 7 presents the experimental results, followed by Section 8 which reviews related work. Section 9 draws a conclusion with future work.

2 PRELIMINARIES

2.1 Local Differential Privacy

Differential privacy [12] (DP) is defined on a randomized algorithm \mathcal{A} of a sensitive database. \mathcal{A} is said to satisfy ϵ -differential privacy, if for any two neighboring databases D and D' that differ only in one tuple, and for any possible output s of \mathcal{A} , we have $\frac{\Pr[\mathcal{A}(D)=s]}{\Pr[\mathcal{A}(D')=s]} \leq e^\epsilon$. In essence, DP guarantees that after observing any output of \mathcal{A} , an adversary cannot infer with high confidence whether the input database is D or D' , thus hiding the existence or non-existence of any individual tuple.

Centralized DP requires the real database stored in a trusted server where the randomized algorithm \mathcal{A} can execute. However, this assumption does not hold in many real-world applications. **Local differential privacy (LDP)** [10], [13] is proposed to assume each individual is responsible for her own tuple in the database. In LDP, each user locally perturbs her tuple using a randomized algorithm before sending it to the untrusted data collector. Formally, a randomized algorithm \mathcal{A} satisfies ϵ -local differential privacy, if for any two input tuples t and t' and for any output t^* , $\frac{\Pr[\mathcal{A}(t)=t^*]}{\Pr[\mathcal{A}(t')=t^*]} \leq e^\epsilon$ holds. In essence, LDP guarantees that after observing any output tuple t^* , the untrusted data collector cannot infer with high confidence whether the input tuple is t or t' .

2.2 Local Differential Privacy on Graphs

In this paper, a graph G is defined as $G = (V, E)$, where $V = \{1, 2, \dots, n\}$ is the set of nodes, and $E \subseteq V \times V$ is the set of edges. For the node i , d_i denotes its degree and $\mathbf{B}_i = \{b_1, b_2, \dots, b_n\}$ denotes its *adjacency bit vector*, where $b_j = 1$ if and only if edge $(i, j) \in E$, and otherwise $b_j = 0$. The adjacency bit vectors of all nodes constitute the *adjacency matrix* of graph G , or formally, $M_{n \times n} = \{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_n\}$.

As with existing LDP works, we concern attacks where an adversary can infer with high confidence whether an edge exists or not, which compromises a user's relation anonymity in a social network. As a graph has both nodes and edges, LDP can be applied to either of them, which leads to *node local differential privacy* [14] and *edge local differential privacy* [15]. Node LDP (resp. edge LDP) guarantees the output of a randomized algorithm does not reveal whether any individual node (resp. edge) exists in G .

Definition 2.1. (*Node local differential privacy*). A randomized algorithm \mathcal{A} satisfies ϵ -node local differential privacy (a.k.a., ϵ -node LDP), if and only if for any two adjacency bit vectors \mathbf{B}, \mathbf{B}' and any output $s \in \text{range}(\mathcal{A})$, $\frac{\Pr[\mathcal{A}(\mathbf{B})=s]}{\Pr[\mathcal{A}(\mathbf{B}')=s]} \leq e^\epsilon$ holds.

Definition 2.2. (*Edge local differential privacy*). A randomized algorithm \mathcal{A} satisfies ϵ -edge local differential privacy (a.k.a., ϵ -edge LDP), if and only if for any two adjacency bit vectors \mathbf{B} and \mathbf{B}' that differ only in one bit, and any output $s \in \text{range}(\mathcal{A})$, $\frac{\Pr[\mathcal{A}(\mathbf{B})=s]}{\Pr[\mathcal{A}(\mathbf{B}')=s]} \leq e^\epsilon$ holds.

Both node and edge LDP satisfy sequential composition.

Theorem 2.3. (*Sequential Composition*) [11]. Given c randomized algorithms $\mathcal{A}_i (1 \leq i \leq c)$, each satisfying ϵ_i -node (resp. edge) LDP, the collection of these algorithms $\mathcal{A}_i (1 \leq i \leq c)$ satisfies $(\sum \epsilon_i)$ -node (resp. edge) LDP.

Edge-LDP is a relaxation of node-LDP, which limits the definition of neighbors from any two adjacency bit vectors to those that differ only in one bit (i.e., one edge). Nonetheless, edge-LDP can still achieve strong indistinguishability of each edge's existence, which suffices for many graph applications such as social networks while preserving high utility [14]. As such, in this paper we assume edge-LDP as with all existing graph LDP works.

3 LF-GDPR: FRAMEWORK OVERVIEW

In this section, we first introduce the rationale behind LF-GDPR for privacy-preserving graph analytics and then overview its workflow. Finally, we introduce two use cases of LF-GDPR.

3.1 Design Principle

The core of privacy-preserving graph analytics often involves **estimating some target graph metric** without accessing the original graph. Under the DP/LDP privacy model, there are two solution paradigms, namely, generating a synthetic graph to calculate this metric [11], [16], [17], [18], [19] and designing a dedicated DP/LDP solution for such metric [7], [14], [20], [21], [22]. The former provides a general solution but suffers from low estimation accuracy as **the neighborhood information in the original graph is**

TABLE 1
Popular graph analysis tasks and metrics

Graph Analysis Task	Graph Metric Concerned	Derivation from B , M , and D
synthetic graph generation	clustering coefficient	$cc_i = \frac{M_{ii}^3}{d_i(d_i-1)}$
community detection, graph clustering	modularity	$Q_c = \frac{\ M_c\ - \ D_c\ ^2}{\ D\ - \ D\ ^2}$
node role, page rank	degree centrality	$c_i = d_i$
	eigenvector centrality	$c_i = B_i M^k$
connectivity analysis (clique / hub)	structural similarity	$\tau(i, j) = \frac{\ B_i \cap B_j\ }{\sqrt{d_i d_j}}$
node similarity search	cosine similarity	$\tau(i, j) = \frac{B_i B_j'}{\sqrt{d_i d_j}}$

missing from the synthetic graph. The latter can achieve higher estimation accuracy but cannot generalize such a dedicated solution to other problems — it works poorly or even no longer works if the target graph metric or graph type (e.g., undirected graph, attributed graph, and DAG) is changed [8], [18].

LF-GDPR is our answer to both solution generality and estimation accuracy under the LDP model. It collects from each node i two atomic graph metrics that can derive a wide range of common metrics. The first is the **adjacency bit vector** B , where each element j is 1 only if j is a neighbor of i . B of all nodes collectively constitutes the adjacency matrix M of the graph. The second metric is **node degree vector** $D = \{d_1, d_2, \dots, d_n\}$, which is frequently used in graph analytics to measure the density of connectivity [21]. Table 1 lists some of the most popular graph analysis tasks in the literature [23], [24], [25] and their graph metrics, all of which can be derived from B , M , and D .

Intuitively, for each node, d can be estimated from B . However, given a large graph and limited privacy budget, the estimation accuracy could be too noisy to be meaningful. To illustrate this, let us assume each bit of the adjacency bit vector B is perturbed independently by the classic Randomized Response (RR) [26] algorithm with privacy budget ϵ . As stated in [26], the variance of the estimated node degree \tilde{d} is

$$Var[\tilde{d}] = n \cdot \left[\frac{1}{16(\frac{e^\epsilon - 1}{e^\epsilon + 1} - \frac{1}{2})^2} - (\frac{d}{n} - \frac{1}{2})^2 \right] \quad (1)$$

Even for a moderate social graph with extremely large privacy budget, for example, $d = 100$, $n = 1M$, and $\epsilon = 8$ (the largest ϵ used in [11] is 7), $Var[\tilde{d}] \approx 435 > 4d$, which means the variance of the estimated degree is over 4 times that of the degree itself. As such, we choose to spend some privacy budget on an independently perturbed degree. This further motivates us to design an optimal privacy budget allocation between adjacency bit vector B and node degree d , to minimize the distance between the target graph metric and the estimated one.

To summarize, in LF-GDPR each node sends two perturbed atomic metrics, namely, the adjacency bit vector \tilde{B} (perturbed from B) and node degree \tilde{d} (perturbed from d), to the data collector, who then aggregates them to estimate the target graph metric.

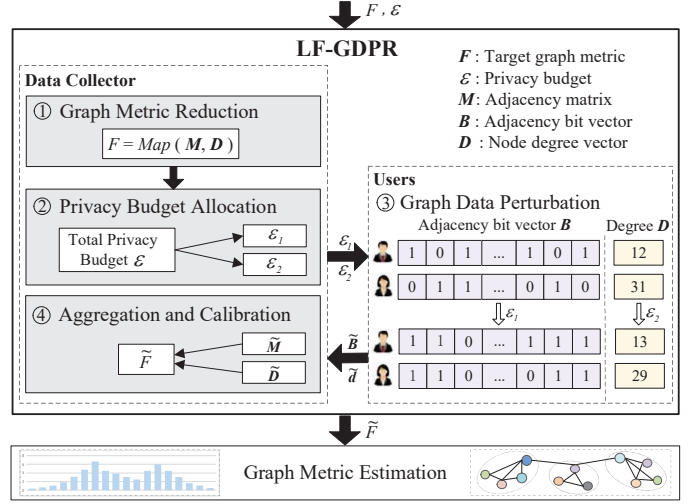


Fig. 1. An overview of LF-GDPR

3.2 LF-GDPR Overview

LF-GDPR works as shown in Fig. 1. A data collector who wishes to estimate a target graph metric F first reduces it from the adjacency matrix M and node degree vector D of all nodes by deriving a mapping function $F = \text{Map}(M, D)$ (step ①). Based on this reduction, LF-GDPR optimally allocates the total privacy budget ϵ between M and D , denoted by ϵ_1 and ϵ_2 , respectively (step ②). Then each node locally perturbs its adjacency bit vector B into \tilde{B} to satisfy ϵ_1 -edge LDP, and perturbs its node degree d into \tilde{d} to satisfy ϵ_2 -edge LDP (step ③). According to the composability of LDP, each node then satisfies ϵ -edge LDP. Note that this step is challenging as both B and d are correlated among nodes. For B , the j -th bit of node i 's adjacency bit vector is the same as the i -th bit of node j 's adjacency bit vector. For d , whether i and j has an edge affects both degrees of i and j . Sections 4.2 and 4.3 solve this issue and send out the perturbed B and d , i.e., \tilde{B} and \tilde{d} . The data collector receives them from all nodes, aggregates them according to the mapping function $\text{Map}(\cdot)$ to obtain the estimated target metric \tilde{F} , and further calibrates it to suppress estimation bias and improve accuracy (step ④). The resulted \tilde{F} is then used for graph analysis. The detailed implementation of LF-GDPR for steps ①②③④ will be presented in Section 4. Note that the algorithms in steps ①②④ are parameterized, which can only be determined when the target graph metric F is specified.

Example 3.2. LF-GDPR against Facebook Privacy Scandal. Facebook API essentially controls how a third-party app accesses the data of each individual user. To limit the access right of an average app (e.g., the one developed by Cambridge Analytica) while still supporting graph analytics, Facebook API should have a new permission rule that only allows such app to access the perturbed adjacency bit vector and degree of a user's friends list under ϵ_1 and ϵ_2 -edge LDP, respectively. In the Cambridge Analytica case, the app is a personality test, so the app developer may choose structural similarity as the target graph metric and use the estimated value for the personality test. To estimate structural similarity, the app then implements steps ①②④

of LF-GDPR. On the user side, each user u has a privacy budget ϵ_u for her friends list. If $\epsilon_u \geq \epsilon_1 + \epsilon_2$, the user can grant access to this app for perturbed adjacency bit vector and degree; otherwise, the user simply ignores this access request.

3.3 Two Cases of Graph Analytics Using LF-GDPR

To illustrate LF-GDPR, we show two use cases throughout this paper. In this subsection, we introduce their background and target graph metrics F . Their usage details, including the reduction of F (step ①), the optimal privacy budget allocation (step ②), and the aggregation and calibration (step ④), are presented in Sections 5 and 6 respectively.

3.3.1 Clustering Coefficient Estimation

The clustering coefficient of a node measures the connectivity in its *neighborhood*, i.e., the subgraph of its neighbors. Formally, the clustering coefficient cc_i of node i is defined as

$$cc_i = \frac{2t_i}{d_i(d_i - 1)},$$

where t_i denotes the number of edges in the neighborhood of node i , or equivalently, the number of triangles incident to node i . A clustering coefficient is in the range of $[0, 1]$, and a high value indicates its neighbors tend to directly connect to each other. It is an important measure of graph structure, and is widely used in graph analytics. For example, the graph model BTER [11], [27] needs clustering coefficient (as well as node degree) to generate a synthetic graph. As it depends on the neighborhood information and thus cannot be calculated locally in each node, existing LDP techniques for values, such as [28], [29], [30], cannot work. The detailed solution by LF-GDPR will be shown in Section 5.

3.3.2 Modularity Estimation and Community Detection

Communities (i.e., densely connected subgraphs) are commonly used in graph analytics to understand the underlying structure of a graph. The criterion of a good community is similar to a graph partition — with many intra-community edges and only a few inter-community edges. Many popular community detection methods are based on modularity maximization [31], which iteratively improves modularity, a widely-adopted metric to measure the quality of detected communities. Formally, the modularity Q of a graph is defined as the sum of individual modularities q_c of all communities \mathcal{C} :

$$Q = \sum_{c=1}^r q_c = \sum_{c=1}^r \left[\frac{L_c}{L} - \left(\frac{K_c}{2L} \right)^2 \right], \quad (2)$$

where r is the number of communities in the graph, L is the total number of edges, L_c is the total number of edges in community \mathcal{C} , and K_c is the total degree of all nodes in \mathcal{C} . Q is in the range of $[-1, 1]$, where a higher value is more desirable. As with clustering coefficient, neither individual nor overall modularity can be estimated by dedicated LDP techniques which do not send the adjacency bit vectors. Section 6 will elaborate on how to use LF-GDPR to estimate it.

4 LF-GDPR: IMPLEMENTATION

In this section, we present the implementation details of LF-GDPR. We first discuss graph metric reduction (step ①), followed by the perturbation protocols for adjacency bit vector and node degree, respectively (step ③). Then we elaborate on the aggregation and calibration algorithm (step ④). Finally, we present the optimal allocation of privacy budget between adjacency bit vector and node degree (step ②).

4.1 Graph Metric Reduction

The reduction outputs a polynomial mapping function $Map(\cdot)$ from the target graph metric F to the adjacency matrix $\mathbf{M} = \{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_n\}$ and degree vector $\mathbf{D} = \{d_1, d_2, \dots, d_n\}$, i.e., $F = Map(\mathbf{M}, \mathbf{D})$. Without loss of generality, we assume F is a polynomial of \mathbf{M} and \mathbf{D} . That is, F is a sum of terms F_l , each of which is a multiple of \mathbf{M} and \mathbf{D} of some exponents. Since F and F_l are scalars, in each term F_l , we need functions f and g to transform \mathbf{M} and \mathbf{D} with exponents to scalars, respectively. Formally,

$$F = \sum_l F_l = \sum_l f_{\phi_l}(\mathbf{M}^{k_l}) \cdot g_{\psi_l}(\mathbf{D}), \quad (3)$$

where \mathbf{M}^{k_l} is the k_l -th power of adjacency matrix \mathbf{M} whose cell (i, j) denotes the number of paths between node i and j of length k_l , ϕ_l projects a matrix to a cell, a row, a column or a sub-matrix, and $f_{\phi_l}(\cdot)$ denotes an aggregation function f (e.g., sum) after projection ϕ_l . Likewise, ψ_l projects a vector to a scalar or a sub-vector, and $g_{\psi_l}(\cdot)$ denotes an aggregation function g after ψ_l .

As such, the metric reduction step is to determine k_l , $f_{\phi_l}(\cdot)$, and $g_{\psi_l}(\cdot)$ for each term F_l in Eq. 3.

4.2 Adjacency Bit Vector Perturbation

An intuitive approach, known as *Randomized Neighbor List* (RNL) [11], perturbs each bit of the vector independently by the classic Randomized Response (RR) [26]. Formally, given an adjacency bit vector $\mathbf{B} = \{b_1, b_2, \dots, b_n\}$, and privacy budget ϵ_1 , the perturbed vector $\tilde{\mathbf{B}} = \{\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_n\}$ is obtained as follows:

$$\tilde{b}_i = \begin{cases} b_i & \text{w.p. } \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}} \\ 1 - b_i & \text{w.p. } \frac{1}{1+e^{\epsilon_1}} \end{cases} \quad (4)$$

Note that here basic RR rather than OUE [32] is adopted. This is because adjacency bit vector is a binary vector, and according to [33], RR can achieve better accuracy than OUE.

Note that in Eq. 4, the probability of preserving an edge (bit '1') or non-edge (bit '0'), i.e., $p = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}$, is not proportional to the amount of edge information disclosed to the collector. In fact, the success rate of the collector inferring an observed edge is a true edge is $\frac{\gamma p}{\gamma p + (1-\gamma)(1-p)}$, where γ is the edge density in a graph. Although the edge density γ is not considered in the definition of edge LDP, but it contributes to the posterior probability for the collector to infer the truth from an observed edge or non-edge. As such, a high edge density γ also plays an important role in raising the success rate. But it is normally very small in social networks, and furthermore, such statics are generally not precisely owned by the collector.

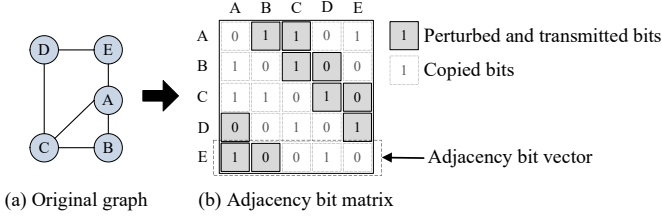


Fig. 2. Illustration of RABV protocol

RNL is proved to satisfy ϵ_1 -edge LDP for each user. However, **for undirected graphs, RNL can only achieve $2\epsilon_1$ -edge LDP for the collector**, because the data collector witnesses the same edge perturbed twice and independently. Let $\tilde{M} = \{\tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_n\}$ denote the perturbed adjacency matrix. The edge between nodes i and j appears in both \tilde{M}_{ij} and \tilde{M}_{ji} , each perturbed with privacy budget ϵ_1 . Then according to the theorem of composability, RNL becomes a $2\epsilon_1$ -edge LDP algorithm for an undirected graph, which is less private. A formal proof is as follows.

For the original adjacency matrix M of an undirected graph, $M_{ij} = M_{ji}$ always holds for any two nodes i and j . By observing two perturbed bits \tilde{M}_{ij} and \tilde{M}_{ji} in the perturbed adjacency matrix \tilde{M} , the posterior probability that there exists an edge between nodes i and j can be denoted by $\Pr[M_{ij} = M_{ji} = 1 \mid \tilde{M}_{ij}, \tilde{M}_{ji}]$. Further, we have

$$\begin{aligned}
 & \frac{\Pr[M_{ij} = M_{ji} = 1 \mid \tilde{M}_{ij}, \tilde{M}_{ji}]}{\Pr[M_{ij} = M_{ji} = 0 \mid \tilde{M}_{ij}, \tilde{M}_{ji}]} \\
 & \leq \frac{\Pr[M_{ij} = M_{ji} = 1 \mid \tilde{M}_{ij} = \tilde{M}_{ji} = 1]}{\Pr[M_{ij} = M_{ji} = 0 \mid \tilde{M}_{ij} = \tilde{M}_{ji} = 1]} \\
 & = \frac{\Pr[M_{ij} = 1 \mid \tilde{M}_{ij} = 1] \cdot \Pr[M_{ji} = 1 \mid \tilde{M}_{ji} = 1]}{\Pr[M_{ij} = 0 \mid \tilde{M}_{ij} = 1] \cdot \Pr[M_{ji} = 0 \mid \tilde{M}_{ji} = 1]} \\
 & = \frac{\frac{e^{\epsilon_1}}{1+e^{\epsilon_1}} \cdot \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}}{\frac{1}{1+e^{\epsilon_1}} \cdot \frac{1}{1+e^{\epsilon_1}}} = e^{2\epsilon_1},
 \end{aligned}$$

which proves that RNL only provides $2\epsilon_1$ -edge LDP.

Furthermore, RNL requires each user to perturb and send all n bits in the adjacency bit vector to data collector, which incurs a high computation and communication cost.

To address the problems of RNL, we propose a more private and efficient protocol *Randomized Adjacency Bit Vector (RABV)* to perturb edges in undirected graphs. As shown in Fig. 2(b), the adjacency matrix is composed of n rows, each corresponding to the adjacency bit vector of a node. For the first $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$ nodes, RABV uses RR as in Eq.4 to perturb and transmit $t = \lfloor \frac{n}{2} \rfloor$ bits (i.e., bits in grey) — from the $(i+1)$ -th bit to the $(i+1+t \bmod n)$ -th bit; for the rest nodes, RABV uses RR to perturb and transmit $t = \lfloor \frac{n-1}{2} \rfloor$ bits in the same way. In essence, **RABV perturbs one and only one bit** for each pair of symmetric bits in the adjacency matrix. The data collector can then obtain the whole matrix by copying bits in grey to their symmetric positions.

Following the same proof of RNL, RABV is guaranteed to satisfy ϵ_1 -edge LDP for the collector. Further, since each node only perturbs and transmits about half of the bits in an adjacency bit vector, RABV significantly reduces computation and communication cost of RNL.

4.3 Node Degree Perturbation

Releasing the degree of a node while satisfying edge ϵ -LDP is essentially a centralized DP problem because all edges incident to this node, or equivalently, all bits in its adjacency bit vector, form a database and the degree is a count function. In the literature, *Laplace Mechanism* [12] is the predominant technique to perturb numerical function values such as counts. As such, LF-GDPR adopts it to perturb the degree d_i of each node i . According to the definition of edge LDP, two adjacency bit vectors B and B' are two neighboring databases if they differ in only one bit. As such, the sensitivity of degree (i.e., count function) is 1, and therefore adding Laplace noise $Lap(\frac{1}{\epsilon_2})$ to the node degree can satisfy ϵ_2 -LDP. That is, $\tilde{d}_i = d_i + Lap(\frac{1}{\epsilon_2})$.

Similar to perturbing adjacency bit vector, however, in the above naive approach the data collector witnesses two node degrees d_i and d_j perturbed independently, but they share the same edge between i and j . As such, whether this edge exists or not contributes to both d_i and d_j . In the most extreme case where there are only two nodes and one edge in the graph, $d_1 = 1$ and $d_2 = 1$, both of which indicate the existence of this edge. If it is removed, both d_1 and d_2 will decrease by 1, causing the sensitivity of node degree perturbation to be 2. As DP or LDP does not refrain an adversary from possessing any background knowledge, in the worst case the collector already knows all edges except for this one. As such, witnessing the two node degrees d_i and d_j is degenerated to witnessing the edge between i and j twice and independently.

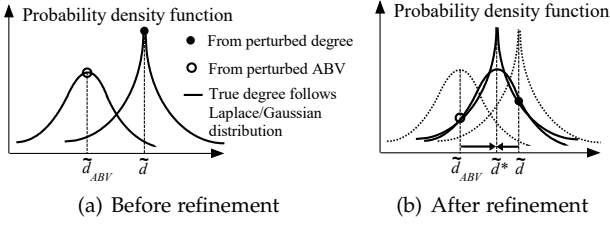
Unfortunately, the remedy that works for perturbing adjacency bit vector cannot be adopted here, as direct bit copy is not feasible for degree. As such, we take an alternative approach to increase the Laplace noise. The following theorem proves that if we add Laplace noise $Lap(\frac{2}{\epsilon_2})$ to every node degree, ϵ_2 -LDP can be satisfied for the collector.

Theorem 4.1. A perturbation algorithm \mathcal{A} satisfies ϵ_2 -LDP for the collector if it adds Laplace noise $Lap(\frac{2}{\epsilon_2})$ to every node degree d_i , i.e., $\tilde{d}_i = \mathcal{A}(d_i) = d_i + Lap(\frac{2}{\epsilon_2})$.

PROOF. By adding Laplace noise $Lap(\frac{2}{\epsilon_2})$ to any node degree d_i , i.e., $\tilde{d}_i = d_i + Lap(\frac{2}{\epsilon_2})$, the perturbation algorithm \mathcal{A} satisfies $\frac{\epsilon_2}{2}$ -LDP for node i . For the collector, whether there is an edge between any two nodes i and j can be derived from both perturbed degrees \tilde{d}_i and \tilde{d}_j . Then according to the composability property of Theorem 2.3, the perturbation algorithm \mathcal{A} satisfies ϵ_2 -LDP for the collector. \square

The perturbed degree \tilde{d} is a coarse estimation of the true degree. Now that we have both \tilde{d} and \tilde{d}_{ABV} , the degree estimated from the perturbed adjacency bit vector \tilde{B} ,¹ we can use *Maximum Likelihood Estimation* (MLE) [34] to obtain a refined estimation \tilde{d}^* . The rationale of this refinement is illustrated in Fig. 3. Before refinement (Fig. 3(a)), as each bit of B follows Bernoulli distribution, according to De Moivre-Laplace Central Limit Theorem, the probability density function of \tilde{d}_{ABV} can be approximated by a Gaussian

1. A naive and biased estimation is $\tilde{d}_{ABV} = \sum_{j=1}^n \tilde{b}_j$. In Example 4.4, we show a calibrated and unbiased estimation $\tilde{d}_{ABV} = \frac{\sum_{j=1}^n \tilde{b}_j}{2p-1} + \frac{(p-1)n}{2p-1}$, where $p = \frac{e^{\epsilon_1}}{e^{\epsilon_1}+1}$.

Fig. 3. Refining \tilde{d} to \tilde{d}^* by MLE

distribution $f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\tilde{d}^*)^2}{2\sigma^2}}$, where the variance $\sigma^2 = n \cdot [\frac{1}{16(p-\frac{1}{2})^2} - (\frac{\tilde{d}}{n} - \frac{1}{2})^2]$ is derived in Eq. 1.² On the other hand, as \tilde{d} is obtained by adding Laplace noise to d , the probability density function of \tilde{d} follows a Laplace distribution $f_2(x) = \frac{\epsilon_2}{4} e^{-\frac{|x-\tilde{d}^*| \epsilon_2}{2}}$. The refinement, as shown in Fig. 3(b), shifts both distributions to share the same mean, i.e., the true degree, as they are both drawn from it. To estimate this mean \tilde{d}^* by MLE, we derive the joint likelihood of observing both \tilde{d} and \tilde{d}_{ABV} , and maximize it. Since they are both independently perturbed, the joint likelihood is the multiplication of individual probabilities. Formally,

$$\begin{aligned} \tilde{d}^* &= \arg \max_{\tilde{d}^*} f_1(\tilde{d}_{ABV}) \cdot f_2(\tilde{d}) \\ &= \arg \max_{\tilde{d}^*} \frac{\epsilon_2}{\sigma \cdot 4\sqrt{2\pi}} e^{-\frac{(\tilde{d}_{ABV}-\tilde{d}^*)^2 + \sigma^2 |\tilde{d}-\tilde{d}^*| \epsilon_2}{2\sigma^2}} \\ &\approx \arg \min_{\tilde{d}^*} ((\tilde{d}_{ABV} - \tilde{d}^*)^2 + \sigma^2 |\tilde{d} - \tilde{d}^*| \epsilon_2) \end{aligned}$$

By solving the above equation, we have

$$\tilde{d}^* = \text{median}(\tilde{d}_{ABV} - \frac{\sigma^2 \cdot \epsilon_2}{2}, \tilde{d}, \tilde{d}_{ABV} + \frac{\sigma^2 \cdot \epsilon_2}{2}) \quad (5)$$

4.4 Aggregation and Calibration

Upon receiving the perturbed adjacency matrix \tilde{M} and degree vector \tilde{D} ,³ the data collector can estimate the target graph metric \tilde{F} by aggregation according to Eq. 3 with a calibration function $\mathcal{R}(\cdot)$:

$$\tilde{F} = \sum_l \mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l}) \cdot g_{\psi_l}(\tilde{D})) \quad (6)$$

The calibration function aims to suppress the aggregation bias of \tilde{M} propagated by f_{ϕ_l} . On the other hand, no calibration is needed for $g_{\psi_l}(\tilde{D})$ as \tilde{D} is already an unbiased estimation of D , thanks to the Laplace Mechanism.

To derive $\mathcal{R}(\cdot)$, we regard \mathcal{R} as the mapping between $f_{\phi_l}(\tilde{M}^{k_l})$ and $f_{\phi_l}(M^{k_l})$. In other words, \mathcal{R} estimates $f_{\phi_l}(M^{k_l})$ after observing $f_{\phi_l}(\tilde{M}^{k_l})$. Formally,

$$\mathcal{R} : f_{\phi_l}(\tilde{M}^{k_l}) \rightarrow f_{\phi_l}(M^{k_l})$$

The following shows a concrete example for aggregation and calibration when estimating the number of edges in a graph. The result of this example will be used in Section 6 to estimate L_c in Eq. 2 of modularity definition.

2. Here we replace d with \tilde{d} in Eq. 1 for simplicity.

3. In the sequel, \tilde{D} denotes the refined degree \tilde{D}^* to simplify the notation.

Example 4.4. For a graph with n nodes, there are $N = \frac{1}{2}n(n-1)$ bits in its upper/lower triangular matrix, each indicating whether an edge exists or not. Let s denote the number of edges in the original graph, i.e., the number of “1”s in these N bits. These N bits are then perturbed according to *RABV* protocol by randomized response [26] with flipping probability p . To estimate s , the data collector takes the following two steps.

(1) **Aggregation.** It aggregates the number of “1”s in the perturbed N bits and uses it as an initial estimation \tilde{s} .

(2) **Calibration.** Since the mapping between s and \tilde{s} can be captured by $\tilde{s} = sp + (N-s)(1-p)$, the collector then calibrates \tilde{s} by $\mathcal{R}(\tilde{s}) = \frac{\tilde{s}}{2p-1} + \frac{p-1}{2p-1}N$, which is derived by solving the mapping function.

We can further show $\mathcal{R}(\tilde{s})$ is an unbiased estimation of s , because $\mathbb{E}[\mathcal{R}(\tilde{s})] = \frac{1}{2p-1}[sp + (N-s)(1-p) + (p-1)N] = s$.

If both $\mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l}))$ and $g_{\psi_l}(\tilde{D})$ are unbiased estimation of $f_{\phi_l}(M^{k_l})$ and $g_{\psi_l}(D)$ respectively, the following theorem guarantees \tilde{F} is an unbiased estimation of the target metric F .

Theorem 4.2. If $\mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l}))$ and $g_{\psi_l}(\tilde{D})$ are unbiased estimation of $f_{\phi_l}(M^{k_l})$ and $g_{\psi_l}(D)$ respectively, the estimated graph metric \tilde{F} is unbiased.

PROOF. According to the assumption of unbiased estimation, we have

$$\begin{aligned} \mathbb{E}[\mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l}))] &= f_{\phi_l}(M^{k_l}) \\ \mathbb{E}[g_{\psi_l}(\tilde{D})] &= g_{\psi_l}(D) \end{aligned}$$

Since the adjacency bit vector and the degree of each node are perturbed independently, we have

$$\begin{aligned} \mathbb{E}[\tilde{F}] &= \sum_l \mathbb{E}[\mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l})) \cdot g_{\psi_l}(\tilde{D})] \\ &= \sum_l \mathbb{E}[\mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l}))] \cdot \mathbb{E}[g_{\psi_l}(\tilde{D})] \\ &= \sum_l f_{\phi_l}(M^{k_l}) \cdot g_{\psi_l}(D) \\ &= F \end{aligned}$$

Therefore, \tilde{F} is unbiased. \square

4.5 Optimal Privacy Budget Allocation

The final problem in LF-GDPR is to allocate the privacy budget (step ② in Fig. 1). Formally, it divides ϵ into $\epsilon_1 = \alpha\epsilon$ and $\epsilon_2 = (1-\alpha)\epsilon$, where $\alpha \in (0, 1)$, for adjacency bit vector and node degree perturbation, respectively.

Our objective is to find the optimal α that minimizes the distance between the graph metric F and our estimation \tilde{F} . Without loss of generality, we adopt the L_2 distance [35] and set the loss function for optimization as the expectation of this distance, i.e., $\alpha = \arg \min_{\alpha \in (0,1)} \mathbb{E}[\|\tilde{F} - F\|_2^2]$.

Assuming \tilde{F} is unbiased, we have

$$\begin{aligned} \mathbb{E}[\|\tilde{F} - F\|_2^2] &= \mathbb{E}[F^2 - 2F\tilde{F} + \tilde{F}^2] \\ &= \mathbb{E}[F^2] - 2\mathbb{E}[F] \cdot \mathbb{E}[\tilde{F}] + \mathbb{E}[\tilde{F}^2] \\ &= \mathbb{E}[\tilde{F}^2] - F^2. \end{aligned}$$

Since F^2 is constant, we only need to minimize $\mathbb{E}[\tilde{F}^2]$ with respect to α :

$$\mathbb{E}[\tilde{F}^2] = \mathbb{E} \left[\left(\sum_l \mathcal{R} \left(f_{\phi_l}(\tilde{\mathcal{M}}^{k_l}) \right) \cdot g_{\psi_l}(\tilde{\mathcal{D}}) \right)^2 \right] \quad (7)$$

In the next two sections, we will demonstrate how to derive the terms in Eq. 7 with respect to α . Then we can apply numerical methods, e.g., Newton's method [36], to find α that minimizes Eq. 7. Further, the following theorem shows the accuracy guarantee of LF-GDPR.

Theorem 4.3. For a graph metric F and our estimation \tilde{F} , with at least $1 - \beta$ probability, we have

$$|F - \tilde{F}| = O(\sqrt{\mathbb{E}[\tilde{F}^2]} \cdot \log(1/\beta))$$

PROOF. For a graph metric F , and its estimated one \tilde{F} , the variance of $F - \tilde{F}$ is

$$\begin{aligned} \text{Var}[F - \tilde{F}] &= \text{Var}[\tilde{F}] = \mathbb{E}[F^2] - (\mathbb{E}[F])^2 \\ &= \mathbb{E}[\tilde{F}^2] - F^2 \leq \mathbb{E}[\tilde{F}^2] \end{aligned}$$

By Bernstein's inequality,

$$\begin{aligned} \Pr[|F - \tilde{F}| \geq \lambda] &\leq 2 \cdot \exp \left(-\frac{\lambda^2}{2\text{Var}[F - \tilde{F}] + \frac{2}{3}\lambda} \right) \\ &\leq 2 \cdot \exp \left(-\frac{\lambda^2}{2\mathbb{E}[\tilde{F}^2] + \frac{2}{3}\lambda} \right) \end{aligned}$$

By the union bound, there exists $\lambda = O(\sqrt{\mathbb{E}[\tilde{F}^2]} \cdot \log(1/\beta))$ such that $|F - \tilde{F}| < \lambda$ holds with at least $1 - \beta$ probability. \square

As will be shown in the next two sections, $\mathbb{E}[\tilde{F}^2]$ can be further expressed by ϵ , n or d for a specific graph metric.

4.6 Summary

Algorithm 1 summarizes the overall protocol of LF-GDPR. It takes three inputs — the target graph metric F , the privacy budget ϵ , and the true adjacency bit vector \mathbf{B}_i of each node i , and returns an estimation of graph metric \tilde{F} under ϵ -LDP. In Line 1, the data collector reduces F to adjacency matrix and node degree. Based on the reduction, in Line 2 the privacy budget ϵ is divided into $\alpha\epsilon$ and $(1 - \alpha)\epsilon$ by the optimal privacy budget allocation algorithm (see Section 4.5 for details), and then α is sent to each node (Line 3). On each node i , *RABV* perturbs its adjacency bit vector (Lines 5-6, see Section 4.2 for details). For each bit to perturb, it adopts RR with privacy budget $\alpha\epsilon$. Then node i further perturbs its degree d_i by adding a Laplace noise with privacy budget $(1 - \alpha)\epsilon$ (Line 7). Finally, the perturbed adjacency bit vector and node degree are sent to the data collector (Line 8). After the collector receives the perturbed adjacency matrix $\tilde{\mathcal{M}}$ and degree vector $\tilde{\mathcal{D}}$, it first completes the whole adjacency matrix by copying bits to their symmetric ones in $\tilde{\mathcal{M}}$ (Line 9), and then refines each node degree \tilde{d}_i to \tilde{d}_i^* (Line 10, see Section 4.3 for details). Finally, it applies aggregation and calibration to estimate the graph metric \tilde{F} (Line 11).

Security of Correlation. It is known that the privacy provided by differential privacy decrease significantly under correlations [37], [38]. However, correlation between

Algorithm 1 Overall protocol of LF-GDPR framework

Input: Target graph metric F
 Privacy budget ϵ
 True adjacency bit vector $\{\mathbf{B}_1, \dots, \mathbf{B}_n\}$

Output: An estimation of the graph metric \tilde{F} under ϵ -LDP

Procedure:
 //Collector side
 1: Reduce graph metric F to adjacency matrix $\mathbf{M} = \{\mathbf{B}_1, \dots, \mathbf{B}_n\}$ and degree vector \mathbf{D} derived from \mathbf{M}
 2: Calculate α for privacy budget allocation based on F and ϵ
 3: Send α to each node
 //User side
 4: **for** each node $i \in \{1, 2, \dots, n\}$ **do**
 5: $t = i \leq \lfloor \frac{n}{2} \rfloor ? \lfloor \frac{n}{2} \rfloor : \lfloor \frac{n}{2} - 1 \rfloor$
 6: **for** each $b_j \in \mathbf{B}_i$, where $i + 1 \leq j \leq (i + 1 + t) \bmod n$ **do**
 Perturb $\tilde{b}_j = \begin{cases} b_j & \text{w.p. } \frac{\epsilon\alpha\epsilon}{1+\epsilon\alpha\epsilon} \\ 1 - b_j & \text{w.p. } \frac{1}{1+\epsilon\alpha\epsilon} \end{cases}$
 7: Calculate the degree d_i from \mathbf{B}_i and then perturb it as
 $\tilde{d}_i = d_i + \text{Lap}(2/((1 - \alpha)\epsilon))$
 8: Send $\tilde{\mathbf{B}}_i$ and \tilde{d}_i to the data collector
 //Collector side
 9: Copy symmetric bits in $\tilde{\mathbf{M}} = \{\tilde{\mathbf{B}}_1, \dots, \tilde{\mathbf{B}}_n\}$
 10: Refine \tilde{d}_i to \tilde{d}_i^* of each node i according to Eq. 5
 11: Apply aggregation and calibration to estimate the graph metric \tilde{F} based on $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{D}} = \{\tilde{d}_1^*, \dots, \tilde{d}_n^*\}$
 12: **return** \tilde{F}

adjacency bit vectors and node degrees does not compromise LDP in LF-GDPR. First, there is pairwise correlation between the adjacency bit vectors of any two users, but the proposed RABV protocol is able to well address it by avoiding “double dose” of the same edge information. Second, there is correlation between the node degrees of two users who share an edge. But Theorem 4.1 proves that by setting sensitivity to 2 and adding $\text{Lap}(\frac{2}{\epsilon_2})$ noise, this correlation does not compromise ϵ_2 -LDP. Third, there is correlation between the adjacency bit vector and node degree of the same user. But since we divide the privacy budget between them, according to sequential composition, ϵ -LDP is still achieved even if they have the strongest correlation (i.e., an equivalent or causal value).

5 CLUSTERING COEFFICIENT ESTIMATION WITH LF-GDPR

In this section, we show how to use LF-GDPR to estimate the clustering coefficients of all nodes in a graph. Based on the implementation framework in Section 4, we present the details of steps ①②④. Finally, Algorithm 2 summarizes the whole process.

5.1 Implementation Details

Graph Metric Reduction (step ① in LF-GDPR). Recall that the clustering coefficient of node i , $cc_i = \frac{2t_i}{d_i(d_i - 1)}$, where t_i is the number of triangles incident to i . To count t_i , we set $k_1 = 3$ so that \mathbf{M}^3 denotes the number of 3-hop walks for all pairs of nodes. We then set projection ϕ_i to M_{ii}^3 , the i -th diagonal element of \mathbf{M}^3 that denotes the number of 3-hop walks starting and ending at node i .⁴ Note that M_{ii}^3 counts

4. The full notion of ϕ_i should be $\phi_{1,i}$. Since there is only one term in the definition of clustering coefficient, we omit the notation 1. The same applies to ψ_i .

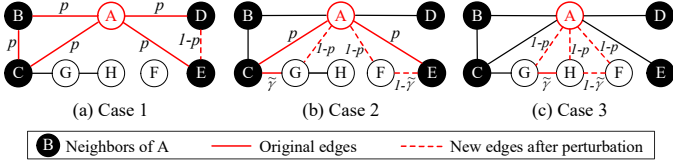


Fig. 4. Estimate number of triangles incident to node A

the triangles incident to node i twice (e.g., triangles ijk and ikj), so $M_{ii}^3 = 2t_i$, which is exactly the numerator in the above cc_i definition. As such, the aggregation function $f(\cdot)$ can be simply set to an identity function. Formally,

$$f_{\phi_i}(M^3) = f(M_{ii}^3) = M_{ii}^3 = 2t_i$$

To obtain the denominator in the above cc_i definition, we set the projection ψ_i to d_i , the i -th element of degree vector D . And the aggregation function $g(\cdot)$ is set according to the denominator in the definition of clustering coefficient:

$$g_{\psi_i}(D) = g(d_i) = \frac{1}{d_i(d_i-1)}$$

To sum up, the clustering coefficient of any node i , denoted by F_i , can be reduced to M and D as

$$F_i = f_{\phi_i}(M^3) \cdot g_{\psi_i}(D) \quad (8)$$

Aggregation and Calibration (step ④ in LF-GDPR). The data collector receives the perturbed adjacency matrix \tilde{M} and degree vector \tilde{D} . According to Eq. 6 and 8, the estimated clustering coefficient of any node i is

$$\tilde{F}_i = \mathcal{R}(f_{\phi_i}(\tilde{M}^3)) \cdot g_{\psi_i}(\tilde{D}), \quad (9)$$

where the calibration function $\mathcal{R}(\cdot)$ estimates $f_{\phi_i}(M^3)$, the number of triangles incident to node i based on the perturbed number $f_{\phi_i}(\tilde{M}^3)$. In what follows, we derive $\mathcal{R}(\cdot)$.

According to Section 4.4, to derive $\mathcal{R}(\cdot)$ we need to estimate $f_{\phi_A}(M^3)$, or equivalently $t_A = f_{\phi_A}(M^3)/2$, the number of triangles incident to node A in the original graph. Figs. 4(a)-(c) enumerate all three cases of such triangles based on whether the other two nodes of this triangle are A's neighbors in the original graph. Let d denote its degree and $p = \frac{e^{\alpha\epsilon}}{1+e^{\alpha\epsilon}}$ the perturbation probability. In each case, the edges that constitute such triangles are highlighted by red color. In particular, the red solid lines denote the original edges, and each is retained in the perturbed graph with a probability of p . The red dashed lines denote the new edges after perturbation, and each appears with a probability of $1-p$.

(1) Fig. 4(a): both nodes are neighbors of A. There are two sub-cases based on whether there exists an edge between these two nodes in the original graph. For triangles such as ABC, there is an edge between B and C in the original graph. Such triangles will be retained in the perturbed graph with probability p^3 . For triangles such as ADE, there is no edge between D and E in the original graph. Such triangles will be retained in the perturbed graph with probability $p^2(1-p)$. Summing up both sub-cases, the number of such triangles in the perturbed graph is $\tilde{t}_{A,1} = t_A \cdot p^3 + (\frac{1}{2}d(d-1) - t_A) \cdot p^2(1-p)$.

(2) Fig. 4(b): only one node is a neighbor of A, for example triangles ACG and AEF. Since d nodes are adjacent to A and $n-d-1$ nodes are not adjacent, there are $d(n-d-1)$ possible triangles. In such a triangle, the two edges incident to A will be retained in the perturbed graph with probabilities $p(1-p)$. The probability of having the third edge (e.g., CG or EF) in the perturbed graph can be approximated by the overall edge density after perturbation, i.e., $\tilde{\gamma} = \gamma p + (1-\gamma)(1-p)$, where $\gamma = \frac{\sum_{i=1}^n d_i}{n(n-1)}$ denote the edge density in the original graph. As such, the number of triangles in this case is $\tilde{t}_{A,2} = d(n-d-1) \cdot p(1-p)\tilde{\gamma}$.

(3) Fig. 4(c): neither node is a neighbor of A, for example triangles AGH and AFH. In such a triangle, the two edges incident to A will be retained in the perturbed graph with probabilities $(1-p)^2$. The probability of having the third edge (e.g., GH or FH) in the perturbed graph can also be approximated by $\tilde{\gamma}$. Since there are $\binom{n-d-1}{2} = \frac{1}{2}(n-d-1)(n-d-2)$ possible triangles, the number of triangles in this case is $\tilde{t}_{A,3} = \frac{1}{2}(n-d-1)(n-d-2) \cdot (1-p)^2\tilde{\gamma}$.

By summing up $\tilde{t}_{A,1}$, $\tilde{t}_{A,2}$, and $\tilde{t}_{A,3}$, we obtain \tilde{t}_A . Since the calibration function $\mathcal{R}(\cdot)$ maps \tilde{t}_A to t_A , i.e., $\mathcal{R}(\tilde{t}_A) = t_A$, we can solve t_A from \tilde{t}_A and derive $\mathcal{R}(\cdot)$ as⁵

$$\begin{aligned} \mathcal{R}(\tilde{t}_A) = \frac{1}{p^2(2p-1)} & \left(\tilde{t}_A - \frac{1}{2}\tilde{d}(\tilde{d}-1)p^2(1-p) \right. \\ & - \tilde{d}(n-\tilde{d}-1)p(1-p)\tilde{\gamma} \\ & \left. - \frac{1}{2}(n-\tilde{d}-1)(n-\tilde{d}-2)(1-p)^2\tilde{\gamma} \right) \end{aligned} \quad (10)$$

Privacy Budget Allocation (step ④ in LF-GDPR). According to Section 4.5, to solve α we derive and minimize $\mathbb{E}[\tilde{F}^2]$ with respect to α in Eq. 7. Theorem 5.1 below shows the closed-form solution of α .

Theorem 5.1. The optimal α for clustering coefficient estimation can be approximated by:

$$\arg \min_{\alpha \in (0,1)} \frac{e^{\alpha\epsilon} + 2}{e^{3\alpha\epsilon}(e^{\alpha\epsilon} - 1)^2} \left(1 + \frac{8(10\hat{d}^2 - 10\hat{d} + 3)}{\hat{d}^2(\hat{d}-1)^2(1-\alpha)^2\epsilon^2} \right) \quad (11)$$

where \hat{d} is a representative degree (e.g., the mean, median, or most frequent degree) of all nodes in the original graph.

PROOF. According to Eq. 7, we have

$$\begin{aligned} \mathbb{E}[\tilde{F}^2] &= \mathbb{E} \left[\left(\sum_l \mathcal{R}(f_{\phi_l}(\tilde{M}^{k_l})) \cdot g_{\psi_l}(\tilde{D}) \right)^2 \right] \\ &= \left(f_{\phi}^2(M^3) + \text{Var}[\mathcal{R}(f_{\phi}(\tilde{M}^3))] \right) \cdot \mathbb{E}[g_{\psi}^2(\tilde{D})] \end{aligned}$$

For each node i , by setting

$$f_{\phi_i}(M^3) = 2t_i \quad \text{and} \quad g_{\psi_i}(D) = \frac{1}{d_i(d_i-1)},$$

we approximate $\mathbb{E}[\tilde{F}_i^2]$ by:

$$\mathbb{E}[\tilde{F}_i^2] = 4(t_i^2 + \text{Var}[\mathcal{R}(\tilde{t}_i)]) \cdot \mathbb{E} \left[\frac{1}{\tilde{d}_i^2(\tilde{d}_i-1)^2} \right], \quad (12)$$

5. We replace d with \tilde{d} , because the former is unknown to data collector and the latter is an unbiased estimation of the former.

Algorithm 2 Collector-side clustering coefficient estimation

Input: Perturbed adjacency matrix $\tilde{M} = \{\tilde{B}_1, \dots, \tilde{B}_n\}$
 Perturbed degree vector $\tilde{D} = \{\tilde{d}_1, \dots, \tilde{d}_n\}$
 Percentage α for privacy budget allocation
Output: Estimated clustering coefficient $cc = \{cc_1, \dots, cc_n\}$
Procedure:

- 1: Calculate the edge density in perturbed graph $\tilde{\gamma} = \frac{\sum_{i=1}^n \tilde{d}_i}{n(n-1)}$
- 2: **for** each node $i \in \{1, 2, \dots, n\}$ **do**
- 3: Calculate the number of triangles \tilde{t}_i incident to node i
- 4: Calibrate \tilde{t}_i to get an unbiased one t_i according to Eq. 10, where $p = \frac{e^{\alpha\epsilon}}{1+e^{\alpha\epsilon}}$.
- 5: Estimate node i 's clustering coefficient $cc_i = \frac{2t_i}{\tilde{d}_i(\tilde{d}_i-1)}$
- 6: **return** $cc = \{cc_1, \dots, cc_n\}$

where

$$\begin{aligned} \text{Var}[\mathcal{R}(\tilde{t}_i)] &= \frac{\text{Var}[\tilde{t}_i]}{p^4(2p-1)^2} \\ &= \frac{\frac{1}{2}(n-1)(n-2)\text{Var}[\tilde{M}_{it_1}\tilde{M}_{t_1t_2}\tilde{M}_{t_2i}]}{p^4(2p-1)^2} \\ &\approx \frac{(n-d_i-1)^2(n-d_i-2)^2}{2(n-1)(n-2)} \cdot \frac{e^{\alpha\epsilon}+2}{e^{3\alpha\epsilon}(e^{\alpha\epsilon}-1)^2} \end{aligned} \quad (13)$$

Since $t_i^2 \ll O(n^2) \sim \text{Var}[\mathcal{R}(\tilde{t}_i)]$ for most cases, we omit the term of t_i^2 . As for $\mathbb{E}\left[\frac{1}{\tilde{d}_i^2(\tilde{d}_i-1)^2}\right]$, by Taylor expansion at $\mathbb{E}[\tilde{d}_i]$, we have

$$\mathbb{E}\left[\frac{1}{\tilde{d}_i^2(\tilde{d}_i-1)^2}\right] \approx \frac{1}{d_i^2(d_i-1)^2} + \frac{8(10d_i^2-10d_i+3)}{d_i^4(d_i-1)^4(1-\alpha)^2\epsilon^2} \quad (14)$$

By substituting Eq. 13 and Eq. 14 into Eq. 12, we have

$$\begin{aligned} \mathbb{E}[\tilde{F}_i^2] &= \frac{4(n-d_i-1)^2(n-d_i-2)^2}{2(n-1)(n-2)d_i^2(d_i-1)^2} \\ &\quad \cdot \frac{e^{\alpha\epsilon}+2}{e^{3\alpha\epsilon}(e^{\alpha\epsilon}-1)^2} \left(1 + \frac{8(10d_i^2-10d_i+3)}{d_i^2(d_i-1)^2(1-\alpha)^2\epsilon^2}\right) \end{aligned}$$

To minimize $\mathbb{E}[\tilde{F}_i^2]$, we omit the first item which is independent of α . To unify α for all nodes, we replace d_i with \hat{d} , a representative degree (e.g., the mean, median, or most frequent degree) of all nodes in the original graph. Therefore, we can derive α as

$$\arg \min_{\alpha \in (0,1)} \frac{e^{\alpha\epsilon}+2}{e^{3\alpha\epsilon}(e^{\alpha\epsilon}-1)^2} \left(1 + \frac{8(10\hat{d}^2-10\hat{d}+3)}{\hat{d}^2(\hat{d}-1)^2(1-\alpha)^2\epsilon^2}\right)$$

□

As for the representative degree \hat{d} , it can be estimated by a portion of privacy budget. The data collector can ask each node to consume some of its privacy budgets for a preliminary round of node degree perturbation and send back \tilde{D} to estimate \hat{d} .

5.2 Overall Algorithm

Algorithm 2 summarizes how the data collector estimates the clustering coefficients of all nodes, based on the perturbed adjacency matrix \tilde{M} and degree vector \tilde{D} . It first computes $\tilde{\gamma}$, the edge density in the perturbed graph from \tilde{D} (Line 1). Then for each node the collector calculates the

number of triangles incident to it (Line 3) and then further calibrates this number based on Eq. 10 (Line 4). Finally, its clustering coefficient is estimated based on Eq. 9 (Line 5).

Accuracy Guarantee. According to Theorem 5.1, with at least $1 - \beta$ probability, the error of clustering coefficient estimation is bounded by $O(\frac{\sqrt{\log(1/\beta)}}{d \cdot \epsilon})$.

6 COMMUNITY DETECTION WITH LF-GDPR

In this section, we show how to use LF-GDPR to estimate the modularity of any community in the graph, with only a single round of \tilde{B} and \tilde{D} collection. Based on the implementation framework in Section 4, we present the details of steps ①②④. Finally, Algorithms 3 summarizes the process of modularity estimation, which serves for further community detection.

6.1 Implementation Details

Graph Metric Reduction (step ① in LF-GDPR). Recall in Eq. 2, the modularity of a community \mathcal{C} is $q_c = \frac{L_c}{L} - \frac{K_c^2}{4L^2}$, where L_c is the number of edges in \mathcal{C} , K_c is the total degree of all nodes in \mathcal{C} , and L is the total number of edges in the whole graph. As such, we can write the graph metric $F_c = q_c$ in the form of Eq. 3 as:

$$F_c = q_c = f_{\phi_{1,c}}(\mathbf{M}) \cdot g_{\psi_{1,c}}(\mathbf{D}) - g_{\psi_{2,c}}(\mathbf{D}) \quad (15)$$

There are two terms in the above equation. In the first term, $\phi_{1,c}$ projects graph G to community \mathcal{C} , i.e., a sub-matrix \mathbf{M}_c of nodes in \mathcal{C} only, and $f_{\phi_{1,c}}(\mathbf{M}) = \frac{1}{2}\|\mathbf{M}_c\|$, half of the summation of all elements in \mathbf{M}_c . As such, $f_{\phi_{1,c}}(\mathbf{M}) = L_c$. Similarly, $g_{\psi_{1,c}}(\mathbf{D}) = \frac{1}{L} = \frac{2}{\|\mathbf{D}\|}$. The second term does not involve \mathbf{M} , so we set $f_{\phi_{2,c}}(\mathbf{M}) = -1$. To project graph G to community \mathcal{C} , we set $\psi_{2,c}$ to a sub-vector \mathbf{D}_c of nodes in \mathcal{C} only, and then $g_{\psi_{2,c}}(\mathbf{D}) = \frac{K_c^2}{4L^2} = \frac{\|\mathbf{D}_c\|^2}{\|\mathbf{D}\|^2}$.

Aggregation and Calibration (step ② in LF-GDPR). According to Eqs. 6 and 15, the data collector estimates the modularity \tilde{F} based on the perturbed adjacency matrix \tilde{M} and degree vector \tilde{D} as follows.

$$\tilde{F} = \mathcal{R}\left(f_{\phi_{1,c}}(\tilde{M})\right) \cdot g_{\psi_{1,c}}(\tilde{D}) - g_{\psi_{2,c}}(\tilde{D})$$

Note that only the first term needs calibration $\mathcal{R}(\cdot)$ as the second term does not involve \tilde{M} . To derive $\mathcal{R}(\cdot)$, we estimate $f_{\phi_{1,c}}(\mathbf{M})$ from $f_{\phi_{1,c}}(\tilde{M})$ based on the *RABV* algorithm and the fact that $f_{\phi_{1,c}}(\mathbf{M}) = \frac{1}{2}\|\mathbf{M}_c\| = L_c$. Example 4.4 shows the derivation of this estimation. By solving $f_{\phi_{1,c}}(\mathbf{M})$ in terms of $f_{\phi_{1,c}}(\tilde{M})$, we can derive $\mathcal{R}(\cdot)$ as

$$\mathcal{R}\left(f_{\phi_{1,c}}(\tilde{M})\right) = \frac{f_{\phi_{1,c}}(\tilde{M})}{2p-1} + \frac{1}{2}n_c(n_c-1)\frac{p-1}{2p-1}, \quad (16)$$

where $n_c = |\mathcal{C}|$ denotes the number of nodes in \mathcal{C} .

Privacy Budget Allocation (step ④ in LF-GDPR). Similar to clustering coefficient estimation, we derive and minimize $\mathbb{E}[\tilde{F}^2]$ with respect to α in Eq. 7. Theorem 6.1 below shows the closed-form solution of α .

Theorem 6.1. The optimal α for modularity estimation can be approximated by:

$$\arg \min_{\alpha \in (0,1)} \frac{(1-\alpha)^2\epsilon^2L^2+6n^2}{(1-\alpha)^2\epsilon^2L^4} \left(\frac{1}{16(\frac{e^{\alpha\epsilon}}{1+e^{\alpha\epsilon}}-\frac{1}{2})^2} - (\frac{2L}{n(n-1)}-\frac{1}{2})^2 \right)$$

PROOF. According to Eq. 7 and Eq. 2, we have

$$\begin{aligned}\mathbb{E}[\tilde{F}^2] &= \mathbb{E}\left[\left(\sum_l \mathcal{R}\left(f_{\phi_l}(\tilde{M}^{k_l})\right) \cdot g_{\psi_l}(\tilde{D})\right)^2\right] \\ &= \left(f_{\phi_1}^2(\mathbf{M}) + \text{Var}\left[\mathcal{R}\left(f_{\phi_1}(\tilde{M})\right)\right]\right) \cdot \mathbb{E}\left[g_{\psi_1}^2(\tilde{D})\right] \\ &\quad + \mathbb{E}\left[g_{\psi_2}^2(\tilde{D})\right] + \mathbb{E}\left[\mathcal{R}\left(f_{\phi_1}(\tilde{M})\right)\right] \mathbb{E}\left[g_{\psi_1}(\tilde{D})g_{\psi_2}(\tilde{D})\right]\end{aligned}$$

For each community \mathcal{C} , note that $\mathbb{E}[L_c] = \frac{n_c^2}{n^2}L$ and by setting

$$f_{\phi_{1,c}}(\mathbf{M}) = L_c \quad \text{and} \quad g_{\psi_{1,c}}(\mathbf{D}) = \frac{1}{L},$$

$$f_{\phi_{2,c}}(\mathbf{M}) = -1 \quad \text{and} \quad g_{\psi_{2,c}}(\mathbf{D}) = \frac{K_c^2}{4L^2},$$

we can approximate $\mathbb{E}[\tilde{F}_c^2]$ by:

$$\begin{aligned}\mathbb{E}[\tilde{F}_c^2] &= \left(\left(\frac{n_c^2 L}{n^2}\right)^2 + \text{Var}\left[\mathcal{R}\left(f_{\phi_{1,c}}(\tilde{M})\right)\right]\right) \cdot \mathbb{E}\left[g_{\psi_{1,c}}^2(\tilde{D})\right] \\ &\quad + \mathbb{E}\left[g_{\psi_{2,c}}^2(\tilde{D})\right] - \frac{2n_c^2 L}{n^2} \cdot \mathbb{E}\left[g_{\psi_{1,c}}(\tilde{D}) \cdot g_{\psi_{2,c}}(\tilde{D})\right],\end{aligned}\quad (17)$$

where

$$\begin{aligned}\text{Var}\left[\mathcal{R}\left(f_{\phi_{1,c}}(\tilde{M})\right)\right] \\ = \frac{1}{2}n_c(n_c - 1) \left(\frac{1}{16(p - \frac{1}{2})^2} - \left(\frac{2L}{n(n-1)} - \frac{1}{2}\right)^2\right)\end{aligned}\quad (18)$$

By Taylor expansion at $\mathbb{E}[\tilde{d}_i]$, we have

$$\mathbb{E}[g_{\psi_{1,c}}^2(\tilde{D})] = \frac{1}{L^2} + \frac{6n^2}{(1-\alpha)^2\epsilon^2 L^4}\quad (19)$$

$$\begin{aligned}\mathbb{E}[g_{\psi_{2,c}}^2(\tilde{D})] &= n_c^2 \left(\frac{1}{n^4} + \frac{32}{n^2(1-\alpha)^2\epsilon^2 L^2}\right. \\ &\quad \left.+ \frac{264}{(1-\alpha)^4\epsilon^4 L^4} + \frac{480n^2}{(1-\alpha)^6\epsilon^6 L^6}\right)\end{aligned}\quad (20)$$

$$\begin{aligned}\mathbb{E}[g_{\psi_{1,c}}(\tilde{D}) \cdot g_{\psi_{2,c}}(\tilde{D})] &= n_c^2 \left(\frac{1}{n^2 L} + \frac{14}{(1-\alpha)^2\epsilon^2 L^3}\right. \\ &\quad \left.+ \frac{24n^2}{(1-\alpha)^4\epsilon^4 L^5}\right)\end{aligned}\quad (21)$$

By substituting Eq. 18 to 21 into Eq. 17, we have

$$\begin{aligned}\mathbb{E}[\tilde{F}_c^2] &\approx \frac{n_c(n_c - 1)((1-\alpha)^2\epsilon^2 L^2 + 6n^2)}{2(1-\alpha)^2\epsilon^2 L^4} \left(\frac{1}{16(\frac{e^{\alpha\epsilon}}{1+e^{\alpha\epsilon}} - \frac{1}{2})^2}\right. \\ &\quad \left.- \left(\frac{2L}{n(n-1)} - \frac{1}{2}\right)^2\right) + \frac{n_c^2 - n_c^4}{n^4}\end{aligned}$$

To minimize $\mathbb{E}[\tilde{F}_c^2]$, we omit items $\frac{n_c(n_c-1)}{2}$ and $\frac{n_c^2 - n_c^4}{n^4}$ that are independent of α , and derive α as

$$\arg \min_{\alpha \in (0,1)} \frac{(1-\alpha)^2\epsilon^2 L^2 + 6n^2}{(1-\alpha)^2\epsilon^2 L^4} \left(\frac{1}{16(\frac{e^{\alpha\epsilon}}{1+e^{\alpha\epsilon}} - \frac{1}{2})^2} - \left(\frac{2L}{n(n-1)} - \frac{1}{2}\right)^2\right)$$

□

Similar to obtaining \hat{d} for clustering coefficient estimation in Section 5, the total number of edges L in graph

Algorithm 3 Collector-side modularity estimation

Input: A community \mathcal{C}
 Perturbed adjacency matrix $\tilde{\mathbf{M}} = \{\tilde{\mathbf{B}}_1, \dots, \tilde{\mathbf{B}}_n\}$
 Perturbed degree vector $\tilde{\mathbf{D}} = \{\tilde{d}_1, \dots, \tilde{d}_n\}$
 Percentage α for privacy budget allocation
 $q_c = \text{EstMod}(\cdot)$, the estimated modularity of \mathcal{C}

Output:

Procedure:

- 1: Extract a sub-matrix $\tilde{\mathbf{M}}_c$ from $\tilde{\mathbf{M}}$
- 2: Obtain \tilde{L}_c by counting and halving the number of “1”s in $\tilde{\mathbf{M}}_c$
- 3: Calibrate \tilde{L}_c to get an unbiased one L_c according to Eq. 16, where $p = \frac{e^{\alpha\epsilon}}{1+e^{\alpha\epsilon}}$.
- 4: Calculate the total number of edges in the whole graph $L = \frac{1}{2} \sum_{i=1}^n \tilde{d}_i$
- 5: Calculate the total degree of all node in \mathcal{C} : $K_c = \sum_{c \in \mathcal{C}} d_c$
- 6: Calculate the estimated modularity of \mathcal{C} : $q_c = \frac{L_c}{L} - \frac{K_c^2}{4L^2}$
- 7: **return** q_c

can be obtained by using a portion of privacy budget for a preliminary round of node degree perturbation to collect $\tilde{\mathbf{D}}$ and estimate L .

6.2 Overall Algorithm

Algorithm 3 summarizes how the data collector estimates the modularity of a given community \mathcal{C} , according to the perturbed adjacency matrix $\tilde{\mathbf{M}}$ and the degree vector $\tilde{\mathbf{D}}$. First, it obtains \tilde{L}_c , the number of edges in \mathcal{C} , by counting and halving the number of “1”s in $\tilde{\mathbf{M}}_c$, the sub-matrix of \mathcal{C} extracted from $\tilde{\mathbf{M}}$ (Lines 1-2). It then calibrates \tilde{L}_c to an unbiased estimation L_c based on Eq. 16 (Line 3). Finally, the estimated modularity is calculated according to Eq. 15 (Line 6), which is based on L_c , L (obtained from Line 4) and K_c (obtained from Line 5).

Accuracy Guarantee. According to Theorem 6.1, with at least $1 - \beta$ probability, the error of modularity estimation is bounded by $O(\frac{\sqrt{\log(1/\beta)}}{n^2 \cdot \epsilon})$.

Now that the modularity of any community can be estimated by Algorithm 3, we can adopt existing community detection methods that are based on modularity maximization [31]. In essence, they attempt to find a graph partition with the highest overall modularity of all communities. For ease of reference, Algorithm 4 presents the detailed implementation of *Louvain* method [31], a popular community detection method under LF-GDPR, where Algorithm 3 serves as the routine for modularity estimation.

As shown in Algorithm 4, there are two iterative phases in *Louvain*. In the first phase, the data collector assigns a different community to each node and calculates its modularity by invoking *EstMod*(\cdot), i.e., Algorithm 3 (Lines 1-2). Then for each node i , the data collector calculates the gain of modularity that would take place by moving i to the community of its neighbor j (Line 5). Here *EstMod*(\cdot) is invoked again to estimate the modularity of community $\{i, j\}$. Node i is then moved into the community in which this gain is positive and maximum (Line 7), and then the modularity of this community is also updated (Line 8). This process is repeated for all nodes until no individual move can improve the total modularity of the graph. The result of the first phase is a new set of communities (Line 10). In the second phase, a new graph is formed from this set of communities, and the data collector repeats the process in

Algorithm 4 Community detection under LF-GDPR with *Louvain* method

Input: Perturbed adjacency matrix $\tilde{\mathbf{M}} = \{\tilde{\mathbf{B}}_1, \dots, \tilde{\mathbf{B}}_n\}$
 Perturbed degree vector $\tilde{\mathbf{D}} = \{\tilde{d}_1, \dots, \tilde{d}_n\}$
 Privacy budget for adjacency bit vector perturbation ϵ_1

Output: A set of detected communities $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots\}$

Procedure:

- 1: Initialize n communities $\{\mathcal{C}_i | 1 \leq i \leq n\}$, each consisting of only one node
- 2: Estimate the modularity of each \mathcal{C}_i : $q_i = \text{EstMod}(\mathcal{C}_i, \tilde{\mathbf{M}}, \tilde{\mathbf{D}}, \epsilon_1)$
- 3: **for** each node $i \in \{1, 2, \dots, n\}$ **do**
- 4: **for** each node j so that $M_{ij} = 1$ **do**
- 5: Calculate gain of modularity:

$$\Delta q_{ij} = \text{EstMod}(\{i \cup j\}, \tilde{\mathbf{M}}, \tilde{\mathbf{D}}, \epsilon_1) - q_i - q_j$$
- 6: Move i to the community of j , where $j = \arg \max \{\Delta q_{ij} | \Delta q_{ij} > 0\}$
- 7: Update the modularity of \mathcal{C}_j : $q_j = \text{EstMod}(\mathcal{C}_j, \tilde{\mathbf{M}}, \tilde{\mathbf{D}}, \epsilon_1)$
- 8: Repeat Lines 3-7 until no individual move can improve the total modularity, and obtain a new set of communities \mathcal{C}^*
- 9: Build a graph from \mathcal{C}^* , and repeat Lines 3-8 to obtain \mathcal{C}
- 10: **return** \mathcal{C}

the first phase to detect the final set of communities (Line 11).

7 EXPERIMENTAL EVALUATION

In this section, we compare the performance of LF-GDPR with two alternative methods, i.e., *RABV-only* and *LDPGen* [11] in both use cases, namely, clustering coefficient estimation and modularity estimation for community detection. In LF-GDPR, the optimal α for clustering coefficient estimation and modularity estimation is derived Theorem 5.1 and 6.1, respectively. Since the derivation is independent of the ground-truth data, we use this optimal α unless stated otherwise. *RABV-only* is a baseline solution where each node spends all its privacy budget in the *RABV* protocol and then derives her node degree from the perturbed adjacency bit vector. As for *LDPGen*, since it needs the clustering coefficient to generate a synthetic graph, we choose the most favorable one for it, i.e., the ground truth value. To have a fair comparison with *RABV-only* and *LDPGen*, for LF-GDPR, we use 10% of the privacy budget to estimate the domain knowledge in both use cases, i.e., the representative degree in Theorem 5.1 and the total number of graph edges in Theorem 6.1. All experiments run in Java on a desktop computer with Intel Core i7-8700K CPU, 64G RAM running Windows 10. The code of LF-GDPR and datasets are available in GitHub at <https://github.com/Vicky-cs/LF-GDPR>.

Performance measures. For the first use case, we measure the *Mean Square Error* (MSE) of the clustering coefficients of all nodes, i.e., $\frac{1}{n} \sum_{i=1}^n (cc_i - \tilde{cc}_i)^2$. For the second use case, to evaluate the modularity estimation, we measure the *Relative Error* (RE) between the ground-truth modularity q and estimated modularity \tilde{q} of one community or all communities in a graph partition, i.e., $\frac{|q - \tilde{q}|}{q}$. To evaluate the final community detection results, we adopt the same classic metrics for cluster validation as used in [11], namely *Adjusted Random Index* (ARI) [39] and *Adjusted Mutual Information* (AMI) [40]. They measure the similarity of two clusterings, and a larger ARI or AMI value indicates more similarity between them.

Datasets. We use four public datasets [41]. The first two are used in [11], and the rest two are added to evaluate on denser and larger graphs.

- (1) *Facebook* — an undirected social network of 4,039 nodes and 88,234 edges, from a survey of participants in Facebook app.
- (2) *Enron* — an undirected email communication network of 36,692 nodes and 183,831 edges.
- (3) *AstroPh* — an undirected collaboration network of 18,772 authors and 198,110 edges indicating collaborations between authors in arXiv, who submitted papers to Astro Physical category.
- (4) *Gplus* — an undirected social network of 107,614 Google+ users and 12,238,285 edges indicating shares of social circles.⁶

7.1 Clustering Coefficient Estimation

Fig. 5 shows the clustering coefficient estimation accuracy of LF-GDPR and two alternative methods over all datasets, with privacy budget ϵ varying from 1 to 8. In all cases, LF-GDPR is the most accurate. Furthermore, it always significantly outperforms *RABV-only*, which justifies our rationale in Section 3.1 that node degree derived from perturbed adjacency bit vector is too noisy. As ϵ increases, the accuracy of LF-GDPR and *RABV-only* improves significantly while *LDPGen* does not. This is because *LDPGen* is only affected by the Laplace noise added to node degree, which is already very small when $\epsilon > 2$. In other words, *LDPGen* cannot fully exploit a large privacy budget.

To evaluate the impact of privacy budget allocation on the estimation accuracy, we compare LF-GDPR with optimal allocation (derived from Eq. 11) against LF-GDPR with four constant α , namely, 0.3, 0.5, 0.7 and 0.9 in Fig. 6. Due to the space limitation, we only show the results of *Facebook*. The optimal allocation achieves the lowest MSE in most cases. As for the constant α , we observe that a large ϵ always favors a large α , which indicates that the privacy budget needed by node degree perturbation is relatively stable, and therefore surplus budget should be mostly allocated to the adjacency bit vector. However, when privacy budget is small (e.g., $\epsilon < 2$), large α (e.g., $\alpha = 0.9$) leads to high MSE. The same observation is also made in modularity estimation, which is therefore omitted in the interest of space.

7.2 Modularity Estimation and Community Detection

In this experiment, we evaluate the modularity estimation and Louvain-based community detection of LF-GDPR against *RABV-only*, *LDPGen*, and the ground truth. To allow fair comparison, we use the same algorithms for the latter three except that the modularity is estimated from the perturbed adjacency matrix only (for *RABV-only*), or directly calculated from the synthetic graph (for *LDPGen*), or directly calculated from the original graph (for ground truth). Fig. 7 plots the RE of modularity by these three methods against ground truth in all datasets. LF-GDPR always outperforms the other two and its RE approaches 0 as ϵ increases, especially in *Facebook* and *Gplus* which have

6. The original *Gplus* dataset is a directed graph, and we convert it to an undirected graph to align with the other three datasets.

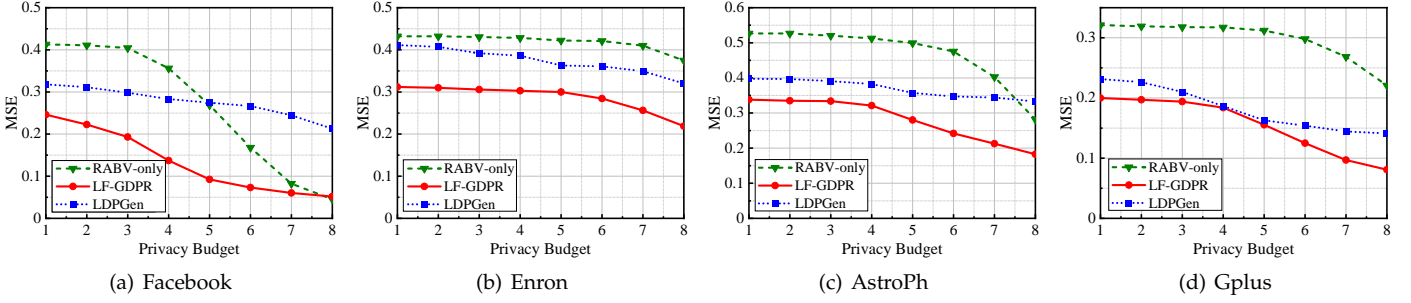
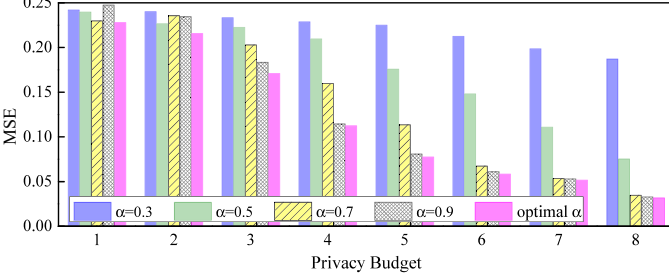


Fig. 5. Mean square error of clustering coefficient estimation

Fig. 6. Mean square error of clustering coefficient estimation, varying α

a higher mean degree than the other two datasets. *RABV-only* has the second lowest RE when ϵ is large, especially in *Facebook*, which means when the privacy budget is sufficient, adjacency bit vector alone can also estimate modularity fairly well. However, when ϵ is small, *RABV-only* has the highest RE among the three, which justifies our rationale in Section 3.1 that the estimated degree from a perturbed adjacency matrix could be too noisy to be meaningful. *LDPGen*, on the other hand, still has very high RE even when ϵ is large, which also justifies our rationale in Section 3.1 that the neighborhood information is lost in a synthetic graph.

To compare the detected communities against ground truth, we plot ARI and AMI between the estimated and ground-truth graph partitions of each method⁷ in Fig. 8. Due to space limitation, we only show the results of *Facebook* and *Enron*. LF-GDPR achieves higher ARI and AMI than *LDPGen* when $\epsilon > 1$, which means the detected communities by LF-GDPR are closer to the ground truth communities detected in the original graph. Particularly, in *Facebook* both ARI and AMI of LF-GDPR approach 1 for large ϵ (e.g., $\epsilon \geq 7$), which means that the detected communities are almost identical to the ground truth communities. We can also verify this observation from a visualization tool *Gephi* in Fig. 9, which illustrates three sets of communities detected from the original graph and from LF-GDPR ($\epsilon = 8$, $\epsilon = 1$) respectively. The sizes of top-3 communities in each set are also marked.

On the other hand, as with the RE results, *LDPGen* has steady ARI/AMI curves because it does not have the neighborhood information of the original graph. As such, it becomes significantly inferior to LF-GDPR when there

is a large privacy budget to spend. Dataset-wise, both LF-GDPR and *LDPGen* perform better in *Facebook* than in *Enron*. This is because *Enron* is more sparse and therefore has more communities — 1275 vs. 16 in *Facebook*.

In addition, we evaluate the accuracy of modularity estimation with respect to the size of a community. For datasets *Facebook* and *Enron*, we randomly select 500 small (5% of the total nodes) communities and 500 large (20% of the total nodes) communities. Then we apply both LF-GDPR and *RABV-only* to estimate the modularity of each community and measure its RE against the ground truth modularity of that community. Due to space limitation, Fig. 10 only shows the results of *Facebook* and *Enron*. LF-GDPR significantly outperforms *RABV-only* in both small and large communities, due to the excessive noise in the node degree introduced by *RABV-only*. We also observe that both methods work better for smaller communities and for the *Facebook* dataset (than the *Enron* dataset). We believe this indicates that LF-GDPR is more superior for denser graphs with more edges per node.

TABLE 2
Communication bandwidth cost (in kilobytes)

Dataset	LF-GDPR	<i>RABV-only</i>	<i>LDPGen</i>
Facebook	0.25	0.25	3.05
Enron	2.30	2.29	27.55
AstroPh	1.18	1.17	14.10
Gplus	6.73	6.73	80.73

To evaluate the communication bandwidth cost, we show the number of kilobytes (kB) between a node and the data collector for all datasets in Table 2. We observe that all three methods are proportional to the node size n , whereas *LDPGen* is also logarithmic to the number of groups g , an internal parameter of *LDPGen*. This coincides with the asymptotic complexity — $O(2n + \lceil \log g \rceil n)$ of *LDPGen* vs. $O(n)$ of LF-GDPR. As such, we expect *LDPGen* incurs even higher communication cost as the graph becomes larger due to an increasing g .

To evaluate the computation cost, we show the runtime of both metric estimation at the collector side in Table 3, with privacy budget ϵ ranging from 1 to 8. Due to the space limitation, we only show the results of *Facebook*. LF-GDPR and *RABV-only* have comparable runtime and decrease significantly with large ϵ . For small ϵ , the perturbed adjacency bit matrix is very dense, so the computation of metrics becomes time-consuming. On the other hand, *LDPGen* always needs to generate a synthetic graph with almost the same number

7. For *LDPGen**, we use the results directly from [11] because the calculation of ARI and AMI between partitions from two (similar) graphs requires an optimal node-to-cluster mapping, which is not specified in [11].

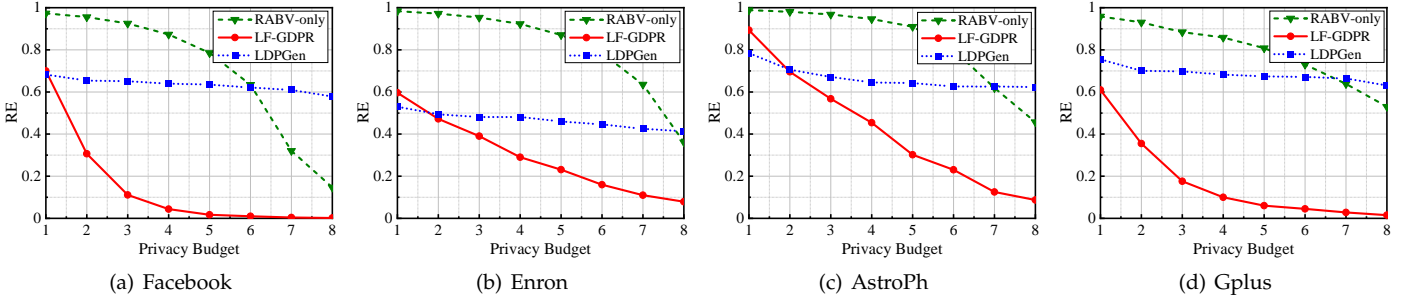


Fig. 7. Relative error of modularity of detected communities

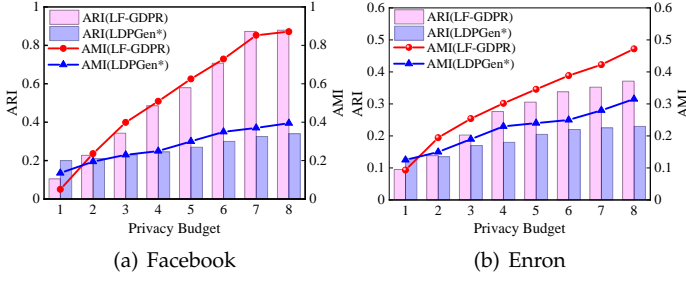


Fig. 8. Results of ARI and AMI

TABLE 3
Runtime in Data Collector Side for Facebook(in milliseconds)

Privacy Budget	Clustering Coefficient Estimation			Modularity Estimation		
	LF-GDPR	RABV-only	LDPGen	LF-GDPR	RABV-only	LDPGen
1	12686	11715	910	120	331	842
2	2273	1825	929	76	82	866
3	469	453	942	52	55	882
4	225	276	892	35	36	845
5	105	143	959	32	33	906
6	100	146	957	29	30	903
7	85	102	949	28	28	883
8	79	96	926	28	28	880

of edges as the original one, so its runtime is independent of ϵ and is outperformed by LF-GDPR and RABV-only except for small ϵ .

7.3 LF-GDPR VS. Dedicated LDP Solutions

As mentioned in the introduction, dedicated LDP solutions may provide a better utility than a general framework as LF-GDPR. In this subsection, we conduct such a comparative study on clustering coefficient estimation (CCE) and modularity estimation for community detection (CD). The main challenge is the design of local perturbation mechanisms that can provide a global view which is needed for these two graph metrics. To address this, we equip each individual user with sufficient ground truth knowledge and design two optimistic dedicated solutions, i.e., *Dedicated-CCE* and *Dedicated-CD*. In *Dedicated-CCE*, we assume each user knows the entire ground-truth adjacency matrix, based on which her clustering coefficient is calculated and then perturbed by adding Laplace noise. In *Dedicated-CD*, we assume each user knows the ground-truth graph partition, based on which her number of edges linked to her community \mathcal{C} is counted, perturbed by adding Laplace noise together with her node degree, and sent to the collector to

calculate the modularity q_c by Eq. 2. Note that these two dedicated solutions provide the same ϵ -edge LDP guarantee as LF-GDPR. But since they optimistically assume to know the ground truth, their estimation accuracy only serves as the upper bound of dedicated solutions for CCE and CD.

Figs. 11 (a) and (b) show the mean square error (MSE) of clustering coefficient estimation of *Dedicated-CCE* and LF-GDPR on *Facebook* and *Enron* datasets. In the interest of space, the results on other datasets are omitted. We observe that *Dedicated-CCE* has very large MSE when the privacy budget is small, and it gradually outperforms LF-GDPR when $\epsilon \geq 4$ (on *Facebook*) or $\epsilon \geq 3$ (on *Enron*). But its MSE is at least 64% and 16% of the MSE of LF-GDPR on two datasets when $\epsilon = 8$. Figs. 11 (c) and (d) show the relative error (RE) of modularity estimation of *Dedicated-CD* and LF-GDPR over *Facebook* and *Enron*. Similar to CCE, there is no all-winner — LF-GDPR performs better on *Facebook* when $\epsilon \geq 3$ whereas *Dedicated-CD* gains higher accuracy (but its RE is at least 20% of the RE of LF-GDPR) in other cases. To summarize, we conclude that LF-GDPR is able to obtain comparable estimation accuracy as dedicated LDP solutions for graph metric estimation.

8 RELATED WORK

There are three related fields: privacy-preserving graph release, graph analytics with differential privacy, and local differential privacy.

Privacy-Preserving Graph Release. This field studies how a data owner publishes a privacy-preserving graph. Early works focus on anonymization techniques under those privacy models derived from k -anonymity [42]. Zhou *et al.* proposed k -neighborhood anonymity to defend against neighborhood attacks [3], Liu *et al.* proposed k -degree anonymity against degree attacks [4], Zou *et al.* and Cheng *et al.* proposed k -automorphism [5] and k -isomorphism [6] respectively against structural attacks, and Xue *et al.* proposed random edge perturbation against walk-based structural identification [43]. As these approaches can be vulnerable to de-anonymization techniques [44], more rigorous privacy notions are proposed, such as L -opacity [45] and differential privacy (DP) [12]. The former ensures an adversary cannot infer whether the distance between two nodes is equal to or less than L . The latter uses a generative graph model to fit the original graph, and then produces a synthetic graph for analytics. Common graph models include dK -series [16], Stochastic Kronecker Graph (SKG) model [46], Exponential Random Graph Model (ERGM) [17], Attributed Graph

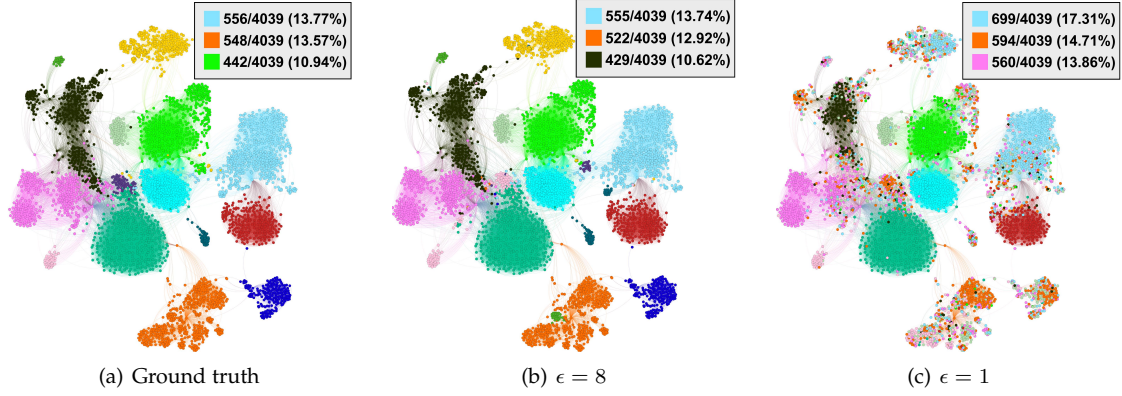


Fig. 9. Visualization of detected communities by Gephi

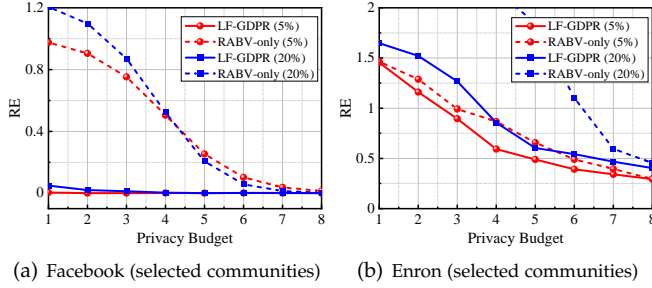


Fig. 10. Impact of community size on modularity estimation

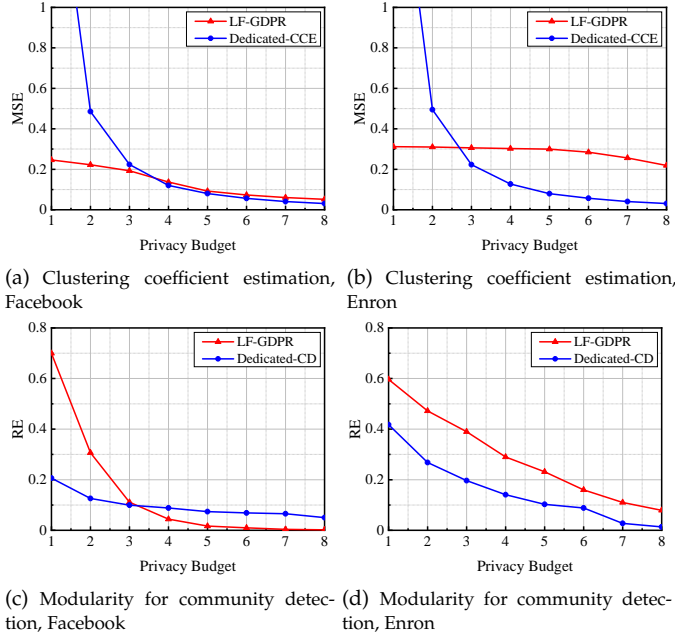


Fig. 11. Comparison with dedicated LDP solutions

Model (AGM) [18], Hierarchical Random Graph (HRG) [19], and BTER [11] (which adopts LDP).

Graph Analytics with Differential Privacy. This field studies how to estimate graph metric and statistics with differential privacy. Most of the existing work focuses on centralized differential privacy. Nissim *et al.* estimated the cost of the minimum spanning tree and the number of triangles in a graph [7]. This technique has been extended

to subgraph counting queries [20], [47] such as k -stars, k -triangles and k -cliques, and frequent subgraph mining [8], [48]. Other works estimate the distribution of node degree [14], [21] and clustering coefficient [22]. In the local setting, Sun *et al.* [47] propose to estimate subgraph counts in a decentralized graph. In our previous work [49], we briefly introduce the LF-GDPR framework that estimates generic graph metrics with local differential privacy. This work has advanced our previous work in almost all aspects. First, this work materializes all algorithms in the LF-GDPR framework. Second, it proposes a refinement strategy for degree estimation and an optimal privacy budget allocation. Third, this work shows use cases on two common graph analysis tasks, namely, clustering coefficient estimation and community detection. Last but not the least, this work comprehensively evaluates the proposed algorithms on four public datasets.

Local Differential Privacy (LDP). Due to its decentralized nature and no need of a trusted party, LDP becomes increasingly popular in privacy-preserving data collection [10], [50]. Existing works focus on estimating statistics such as frequency [32], [51], [52], mean [28], [30], heavy hitter [35], frequent itemset mining [53], k -way marginal release [54], [55], key-value data collection [29], [56] and time-series data collection [57]. Some works also focus on learning problems [30].

9 CONCLUSION

This paper presents a parameterized framework LF-GDPR for privacy-preserving graph metric estimation and analytics with local differential privacy. The building block is a user-side perturbation algorithm, and a collector-side aggregation and calibration algorithm. LF-GDPR simplifies the job of developing a practical LDP solution for a graph analysis task by providing a complete solution for all LDP steps. An optimal allocation of privacy budget between the two atomic metrics is also designed. Through theoretical and experimental analysis, we verify the privacy and data utility achieved by this framework.

As for future work, we plan to extend LF-GDPR to more specific graph types and graph analysis tasks, such as attributed graph and DAG, and influential node analysis, to demonstrate its wide applicability. We will also investigate some relaxation of DP, such as Gaussian Mechanism and

(ϵ, δ) -DP, to provide higher estimation accuracy and better utility. Graph-specific tighter bounds for the composition of DP [58] and the correlation of graph data will also be studied.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No: 62072390, U1636205, 91646203, 61941121 and 61972332), the Research Grants Council, Hong Kong SAR, China (Grant No: 15238116, 15222118, 15218919, 15203120 and C1008-16G), the Ministry of Education, Singapore (Grant No: MOE2018-T2-2-091).

REFERENCES

- [1] B. Stephanie, Facebook Scandal a 'Game Changer' in Data Privacy Regulation, *Bloomberg*, Apr 8, 2018.
- [2] Facebook, <https://developers.facebook.com/docs/graph-api/>, graph API - Facebook for Developers.
- [3] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *ICDE*. IEEE, 2008, pp. 506–515.
- [4] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *SIGMOD*. ACM, 2008, pp. 93–106.
- [5] L. Zou, L. Chen, and M. T. Özsu, "K-automorphism: A general framework for privacy preserving network publication," *PVLDB*, vol. 2, no. 1, pp. 946–957, 2009.
- [6] J. Cheng, A. W. Fu, and J. Liu, "K-isomorphism: privacy preserving network publication against structural attacks," in *SIGMOD*. ACM, 2010, pp. 459–470.
- [7] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *STOC*. ACM, 2007, pp. 75–84.
- [8] S. Xu, S. Su, L. Xiong, X. Cheng, and K. Xiao, "Differentially private frequent subgraph mining," in *ICDE*. IEEE, 2016, pp. 229–240.
- [9] C. Wei, S. Ji, C. Liu, W. Chen, and T. Wang, "Asgldp: Collecting and generating decentralized attributed graphs with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3239–3254, April 2020.
- [10] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *FOCS*. IEEE, 2013, pp. 429–438.
- [11] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren, "Generating synthetic decentralized social graphs with local differential privacy," in *CCS*. ACM, 2017, pp. 425–438.
- [12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*. Springer, 2006, pp. 265–284.
- [13] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.
- [14] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith, "Analyzing graphs with node differential privacy," in *TCC*. Springer, 2013, pp. 457–476.
- [15] J. Blocki, A. Blum, A. Datta, and O. Sheffet, "The johnson-lindenstrauss transform itself preserves differential privacy," in *FOCS*. IEEE, 2012, pp. 410–419.
- [16] A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Y. Zhao, "Sharing graphs using differentially private graph models," in *IMC*, 2011, pp. 81–98.
- [17] W. Lu and G. Miklau, "Exponential random graph estimation under differential privacy," in *KDD*. ACM, 2014, pp. 921–930.
- [18] Z. Jorgensen, T. Yu, and G. Cormode, "Publishing attributed social graphs with formal privacy guarantees," in *SIGMOD*, 2016, pp. 107–122.
- [19] Q. Xiao, R. Chen, and K. Tan, "Differentially private network data release via structural inference," in *KDD*. ACM, 2014, pp. 911–920.
- [20] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev, "Private analysis of graph structure," *PVLDB*, vol. 4, no. 11, pp. 1146–1157, 2011.
- [21] M. Hay, C. Li, G. Miklau, and D. Jensen, "Accurate estimation of the degree distribution of private networks," in *ICDM*, 2009, pp. 169–178.
- [22] Y. Wang, X. Wu, J. Zhu, and Y. Xiang, "On learning cluster coefficient of private networks," *Social network analysis and mining*, vol. 3, no. 4, pp. 925–938, 2013.
- [23] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: a structural clustering algorithm for networks," in *KDD*. ACM, 2007, pp. 824–833.
- [24] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
- [25] T. Martin, X. Zhang, and M. Newman, "Localization and centrality in networks," *Physical review E*, vol. 90, no. 5, p. 052808, 2014.
- [26] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.
- [27] C. Seshadhri, T. G. Kolda, and A. Pinar, "Community structure and scale-free collections of erdős-rényi graphs," *Physical Review E*, vol. 85, no. 5, p. 056109, 2012.
- [28] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *NIPS*, 2017, pp. 3574–3583.
- [29] Q. Ye, H. Hu, X. Meng, and H. Zheng, "PrivKV: Key-value data collection with local differential privacy," in *S&P*. IEEE, 2019, pp. 317–331.
- [30] N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu, "Collecting and analyzing multidimensional data with local differential privacy," in *ICDE*. IEEE, 2019.
- [31] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.
- [32] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *USENIX Security Symposium*, 2017, pp. 729–745.
- [33] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *ICML*. ACM, 2016, pp. 2436–2444.
- [34] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.
- [35] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *CCS*. ACM, 2016, pp. 192–203.
- [36] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.
- [37] R. Chen, B. C. Fung, S. Y. Philip, and B. C. Desai, "Correlated network data publication via differential privacy," *The VLDB Journal*, vol. 23, no. 4, pp. 653–676, 2014.
- [38] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *SIGMOD*. ACM, 2015, pp. 747–762.
- [39] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.
- [40] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *ICML*. ACM, 2009, pp. 1073–1080.
- [41] L. Jure and K. Andrej, "SNAP Datasets: Stanford large network dataset collection," <http://snap.stanford.edu/data>, 2014.
- [42] P. Samarati, "Protecting respondents identities in microdata release," *TKDE*, vol. 13, no. 6, pp. 1010–1027, 2001.
- [43] M. Xue, P. Karras, R. Chedy, P. Kalnis, and H. Pung, "Delineating social network data anonymization via random edge perturbation," in *CIKM*, 2012, pp. 475–484.
- [44] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *S&P*. IEEE, 2009, pp. 173–187.
- [45] S. Nobari, P. Karras, H. Pang, and S. Bressan, "L-opacity: Linkage-aware graph anonymization," in *EDBT*. Springer, 2014, pp. 583–594.
- [46] D. Mir and R. N. Wright, "A differentially private estimator for the stochastic kronecker graph model," in *EDBT/ICDT Workshops*. ACM, 2012, pp. 167–176.
- [47] H. Sun, X. Xiao, I. Khalil, Y. Yang, Z. Qin, H. Wang, and T. Yu, "Analyzing subgraph statistics from extended local views with decentralized differential privacy," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 703–717.
- [48] E. Shen and T. Yu, "Mining frequent graph patterns with differential privacy," in *KDD*. ACM, 2013, pp. 545–553.
- [49] Q. Ye, H. Hu, M. H. Au, X. Meng, and X. Xiao, "Towards locally differentially private generic graph metric estimation," in *ICDE*. IEEE, 2020, pp. 1922–1925.

- [50] N. Li and Q. Ye, "Mobile data collection and analysis with local differential privacy," in *MDM*. IEEE, 2019, pp. 4–7.
- [51] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *CCS*. ACM, 2014, pp. 1054–1067.
- [52] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *STOC*. ACM, 2015, pp. 127–135.
- [53] T. Wang, N. Li, and S. Jha, "Locally differentially private frequent itemset mining," in *S&P*. IEEE, 2018, pp. 127–143.
- [54] G. Cormode, T. Kulkarni, and D. Srivastava, "Marginal release under local differential privacy," in *SIGMOD*. ACM, 2018, pp. 131–146.
- [55] Z. Zhang, T. Wang, N. Li, S. He, and J. Chen, "CALM: Consistent adaptive local marginal for marginal release under local differential privacy," in *CCS*. ACM, 2018, pp. 212–229.
- [56] X. Gu, M. Li, L. Xiong, and Y. Cao, "PCKV: locally differentially private correlated key-value data collection with optimized utility," in *USENIX Security Symposium*, 2020.
- [57] Q. Ye, H. Hu, N. Li, X. Meng, H. Zheng, and H. Yan, "Beyond value perturbation: Differential privacy in the temporal setting," in *INFOCOM*. IEEE, 2021.
- [58] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *ICML*, 2015, pp. 1376–1385.

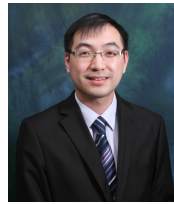


Qingqing Ye is a research assistant professor in the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. She received her PhD degree in Computer Science from Renmin University of China in 2020. She has received several prestigious awards, including China National Scholarship, Outstanding Doctoral Dissertation Award, and IEEE S&P Student Travel Award. Her research interests include data privacy and security, and adversarial machine learning.



Haibo Hu is an associate professor in the Department of Electronic and Information Engineering, Hong Kong Polytechnic University. His research interests include cybersecurity, data privacy, internet of things, and machine learning. He has published over 80 research papers in refereed journals, international conferences, and book chapters. As principal investigator, he has received over 12 million HK dollars of external research grants from Hong Kong and mainland China. He is the recipient of a number of titles

and awards, including IEEE MDM 2019 Best Paper Award, WAIM Distinguished Young Lecturer, VLDB Distinguished Reviewer, ACM-HK Best PhD Paper, Microsoft Imagine Cup, and GS1 Internet of Things Award.



Man Ho Au is an associate professor of the Department of Computer Science at the University of Hong Kong (HKU). Before joining HKU, he was an associate professor in the Department of Computing of the Hong Kong Polytechnic University. His research interests include applied cryptography, information security, blockchain technology, and related industrial applications. Dr. Au has published over 170 refereed papers in top journals and conferences, including CRYPTO, ACM CCS, ACM SIGMOD, NDSS, IEEE TIFS, TKDE. He is a recipient of the 2009 PET runner-up award for outstanding research in privacy-enhancing technologies, and best paper awards of ACISP 2016, ISPEC 2017, and ACISP 2018. He is a general chair of ASIACCS 2021, an expert member of the China delegation of ISO/IEC JTC 1/SC 27 working group 2 Cryptography and security mechanisms, and a committee member of the Hong Kong Blockchain Society R&D division.



in refereed international journals and conference proceedings including IEEE TKDE, VLDBJ, VLDB, SIGMOD, ICDE, EDBT, ACM GIS etc.

Xiaofeng Meng is a professor in School of Information, Renmin University of China. He is a CCF Fellow and the vice chair of the Special Interesting Group on Privacy of China Confidentiality Association (CCA). He has served on the program committee SIGMOD, ICDE, CIKM, MDM, DASFAA, etc., and editorial board of JCST, FCS, JoS, CRAD, etc. His research interests include web data management, cloud data management, mobile data management, and privacy protection. He has published over 200 papers



Xiaokui Xiao received his PhD degree in computer science and engineering from the Chinese University of Hong Kong in 2008. He is currently an associate professor at the School of Computing, National University of Singapore (NUS). His research interests include data privacy and algorithms for large data.

Towards Locally Differentially Private Generic Graph Metric Estimation

Qingqing Ye*, Haibo Hu†, Man Ho Au†, Xiaofeng Meng*, Xiaokui Xiao‡

*Renmin University of China; †Hong Kong Polytechnic University; ‡National University of Singapore
 yeqq@ruc.edu.cn; haibo.hu@polyu.edu.hk; csallen@comp.polyu.edu.hk; xfmeng@ruc.edu.cn; xkxiao@nus.edu.sg

Abstract—Local differential privacy (LDP) is an emerging technique for privacy-preserving data collection without a trusted collector. Despite its strong privacy guarantee, LDP cannot be easily applied to real-world graph analysis tasks such as community detection and centrality analysis due to its high implementation complexity and low data utility. In this paper, we address these two issues by presenting LF-GDPR, the first LDP-enabled graph metric estimation framework for graph analysis. It collects two atomic graph metrics — the adjacency bit vector and node degree — from each node locally. LF-GDPR simplifies the job of implementing LDP-related steps (e.g., local perturbation, aggregation and calibration) for a graph metric estimation task by providing either a complete or a parameterized algorithm for each step.

Index Terms—Local differential privacy; Graph metric; Privacy-preserving graph analysis

I. INTRODUCTION

With the prevalence of big data and machine learning, graph analytics has received great attention and nurtured numerous applications in web, social network, transportation, and knowledge base. However, recent privacy incidents, particularly the Facebook privacy scandal, pose real-life threats to any **centralized** party who needs to safeguard graph data of individuals while providing graph analysis service to third parties. In that scandal, Facebook exposed the personal profiles of 87 million users to Cambridge Analytica through Facebook API for third-party apps [11]. The main cause is that Facebook allows these apps to access the **friends list** of a user, which helps to propagate these apps easily through friends. Unfortunately, most existing privacy models assume that the trusted party cannot be compromised, which is seldom true in practice as echoed by this scandal. With General Data Protection Regulation (GDPR) enforced in EU since May 2018, there is a compelling need to find alternative privacy models without such a trusted party.

A promising model is local differential privacy (LDP) [1], [15], where each individual user **locally perturbs her share of graph metrics** (e.g., node degree and adjacency list, depending on the graph analysis task) before sending them to the data collector for analysis. As such, the data collector does not need to be trusted. A recent work *LDPPGen* [10] has also shown the potential of LDP for graph analytics. In that work, LDP is used to collect node degree for synthetic graph generation. However, such solution is usually task specific — for different tasks, such as centrality analysis and community detection,

dedicated LDP solutions must be designed from scratch. To show how complicated it is, an LDP solution usually takes four steps: (1) selecting graph metrics to collect from users for the target metric (e.g., clustering coefficient, modularity, or centrality) of this task, (2) designing a local perturbation algorithm for users to report these metrics under LDP, (3) designing a collector-side aggregation algorithm to estimate the target metric based on the perturbed data, (4) designing an optional calibration algorithm for the target metric if the estimation is biased. Obviously, working out such a solution **requires in-depth knowledge of LDP**, which hinders the embrace of LDP by more graph applications.

In this paper, we address this challenge by presenting LF-GDPR (Local Framework for Graph with Differentially Private Release), the first LDP-enabled graph metric estimation framework for general graph analysis. It simplifies the job of a graph application to design an LDP solution for a graph metric estimation task by providing complete or parameterized algorithms for steps (2)-(4) as above. As long as the target graph metric can be derived from the two atomic metrics, namely, the adjacency bit vector and node degree, the parameterized algorithms in steps (2)-(4) can be completed with ease. To summarize, our main contributions of this paper are as follows.

- This is the first LDP-enabled graph metric estimation framework for a variety of graph analysis tasks.
- We present efficient perturbation algorithms on adjacency bit vector and node degree, respectively, to address data correlation among nodes.
- We provide a complete solution for local perturbation, collector-side aggregation, and calibration.

The rest of the paper is organized as follows. Section II introduces preliminaries on local differential privacy and graph analytics. Section III presents an overview of LF-GDPR. Section IV describes the implementation details of this framework. Section V draws a conclusion with future work.

II. LOCAL DIFFERENTIAL PRIVACY ON GRAPHS

In this paper, a graph G is defined as $G = (V, E)$, where $V = \{1, 2, \dots, n\}$ is the set of nodes, and $E \subseteq V \times V$ is the set of edges. For the node i , d_i denotes its degree and $\mathbf{B}_i = \{b_1, b_2, \dots, b_n\}$ denotes its *adjacency bit vector*, where $b_j = 1$ if and only if edge $(i, j) \in E$, and otherwise $b_j = 0$. The adjacency bit vectors of all nodes constitute the *adjacency matrix* of graph G , or formally, $\mathbf{M}_{n \times n} = \{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_n\}$.

Local differential privacy (LDP) [1] is proposed to assume each individual is responsible for her own tuple in the database. In LDP, each user locally perturbs her tuple using a randomized algorithm before sending it to the untrusted data collector. Formally, a randomized algorithm \mathcal{A} satisfies ϵ -local differential privacy, if for any two input tuples t and t' and for any output t^* , $\frac{\Pr[\mathcal{A}(t)=t^*]}{\Pr[\mathcal{A}(t')=t^*]} \leq e^\epsilon$ holds. As with existing LDP works, we concern attacks where an adversary can infer with high confidence whether an edge exists or not, which compromises a user's relation anonymity in a social network. This directly leads to Definition 2.1.

Definition 2.1: (Edge local differential privacy). A randomized algorithm \mathcal{A} satisfies ϵ -edge local differential privacy (a.k.a., ϵ -edge LDP), if and only if for any two adjacency bit vectors \mathbf{B} and \mathbf{B}' that differ only in one bit, and any output $s \in \text{range}(\mathcal{A})$, $\frac{\Pr[\mathcal{A}(\mathbf{B})=s]}{\Pr[\mathcal{A}(\mathbf{B}')=s]} \leq e^\epsilon$ holds.

III. LF-GDPR: FRAMEWORK OVERVIEW

A. Design Principle

The core of privacy-preserving graph analytics often involves **estimating some target graph metric** without accessing the original graph. Under the DP/LDP privacy model, there are two solution paradigms, namely, generating a synthetic graph to calculate this metric [5], [7], [10] and designing a dedicated DP/LDP solution for such metric [4], [6], [9]. The former provides a general solution but suffers from low estimation accuracy as **the neighborhood information in the original graph is missing** from the synthetic graph. The latter can achieve higher estimation accuracy but cannot generalize such a dedicated solution to other problems — it works poorly or even no longer works if the target graph metric or graph type (e.g., undirected graph, attributed graph, and DAG) is changed [5].

LF-GDPR is our answer to both solution generality and estimation accuracy under the LDP model. It collects from each node i two atomic graph metrics that can derive a wide range of common metrics. The first is the **adjacency bit vector** \mathbf{B} , where each element j is 1 only if j is a neighbor of i . \mathbf{B} of all nodes collectively constitutes the adjacency matrix \mathbf{M} of the graph. The second metric is **node degree** d , which is frequently used in graph analytics to measure the density of connectivity [4]. Table I lists some of the most popular graph analysis tasks in the literature [3], [8], [13] and their graph metrics, all of which can be derived from \mathbf{B} , \mathbf{M} and d .

Intuitively, d can be estimated from \mathbf{B} . However, given a large graph and limited privacy budget, the estimation accuracy could be too noisy to be meaningful. To illustrate this, let us assume each bit of the adjacency bit vector \mathbf{B} is perturbed independently by the classic Randomized Response (RR) [12] algorithm with privacy budget ϵ . As stated in [12], the variance of the estimated node degree \tilde{d} is

$$\text{Var}[\tilde{d}] = n \cdot \left[\frac{1}{16(\frac{e^\epsilon}{e^\epsilon+1} - \frac{1}{2})^2} - (\frac{d}{n} - \frac{1}{2})^2 \right] \quad (1)$$

Even for a moderate social graph with extremely large privacy budget, for example, $d = 100$, $n = 1M$, and $\epsilon = 8$ (the largest

TABLE I
POPULAR GRAPH ANALYSIS TASKS AND METRICS

Graph Analysis Task	Graph Metric Concerned	Derivation from \mathbf{B} , \mathbf{M} , and d
synthetic graph generation	clustering coefficient	$cc_i = \frac{M_{ii}^3}{d_i(d_i-1)}$
community detection, graph clustering	modularity	$Q_c = \frac{\ \mathbf{M}_c\ }{\sum d} - \frac{\ \mathbf{d}_c\ ^2}{(\sum d)^2}$
node role, page rank	degree centrality	$c_i = d_i$
	eigenvector centrality	$c_i = \mathbf{B}_i \mathbf{M}^k$
connectivity analysis (clique / hub)	structural similarity	$\tau(i, j) = \frac{\ \mathbf{B}_i \cap \mathbf{B}_j\ }{\sqrt{d_i d_j}}$
node similarity search	cosine similarity	$\tau(i, j) = \frac{\mathbf{B}_i \mathbf{B}_j^T}{\sqrt{d_i d_j}}$

ϵ used in [10] is 7), $\text{Var}[\tilde{d}] \approx 435 > 4d$, which means the variance of the estimated degree is over 4 times that of the degree itself. As such, we choose to spend some privacy budget on an independently perturbed degree. This further motivates us to design an optimal privacy budget allocation between adjacency bit vector \mathbf{B} and node degree d , to minimize the distance between the target graph metric and the estimated one.

To summarize, in LF-GDPR each node sends two perturbed atomic metrics, namely, the adjacency bit vector $\tilde{\mathbf{B}}$ (perturbed from \mathbf{B}) and node degree \tilde{d} (perturbed from d), to the data collector, who then aggregates them to estimate the target graph metric.

B. LF-GDPR Overview

LF-GDPR works as shown in Fig. 1. A data collector who wishes to estimate a target graph metric F first reduces it from the adjacency matrix \mathbf{M} and degree vector \mathbf{d} of all nodes by deriving a mapping function $F = \text{Map}(\mathbf{M}, \mathbf{d})$ (step ①). Based on this reduction, LF-GDPR allocates the total privacy budget ϵ between \mathbf{M} and \mathbf{d} , denoted by ϵ_1 and ϵ_2 , respectively (step ②). Then each node locally perturbs its adjacency bit vector \mathbf{B} into $\tilde{\mathbf{B}}$ to satisfy ϵ_1 -edge LDP, and perturbs its node degree d into \tilde{d} to satisfy ϵ_2 -edge LDP (step ③). According to the composability of LDP, each node then satisfies ϵ -edge LDP. Note that this step is challenging as both \mathbf{B} and d are correlated among nodes. For \mathbf{B} , the j -th bit of node i 's adjacency bit vector is the same as the i -th bit of node j 's adjacency bit vector. For d , whether i and j has an edge affects both degrees of i and j . Sections IV-B and IV-C solve this issue and send out the perturbed \mathbf{B} and d , i.e., $\tilde{\mathbf{B}}$ and \tilde{d} . The data collector receives them from all nodes, aggregates them according to the mapping function $\text{Map}(\cdot)$ to obtain the estimated target metric \tilde{F} , and further calibrates it to suppress estimation bias and improve accuracy (step ④). The resulted \tilde{F} is then used for graph analysis. The detailed implementation of LF-GDPR for steps ①③④ will be presented in Section IV. Note that the algorithms in steps ①②④ are parameterized, which can only be determined when the target graph metric F is specified.

Example III-B. LF-GDPR against Facebook Privacy Scandal. Facebook API essentially controls how a third-party app accesses the data of each individual user. To limit the

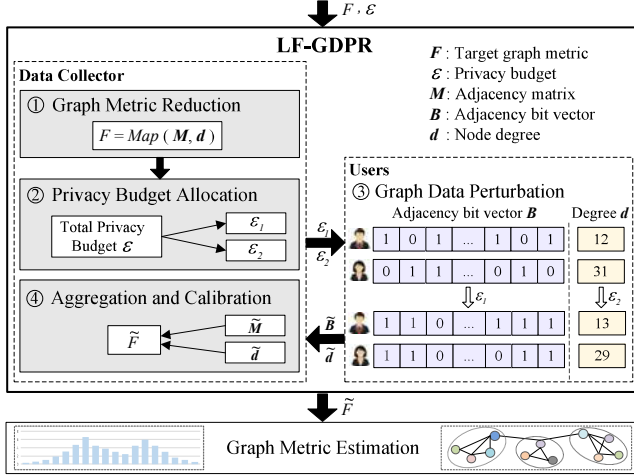


Fig. 1. An overview of LF-GDPR

access right of an average app (e.g., the one developed by Cambridge Analytica) while still supporting graph analytics, Facebook API should have a new permission rule that only allows such app to access the perturbed adjacency bit vector and degree of a user's friends list under ϵ_1 and ϵ_2 -edge LDP, respectively. In the Cambridge Analytica case, the app is a personality test, so the app developer may choose structural similarity as the target graph metric and use the estimated value for the personality test. To estimate structural similarity, the app then implements steps ①②④ of LF-GDPR. On the user side, each user u has a privacy budget ϵ_u for her friends list. If $\epsilon_u \geq \epsilon_1 + \epsilon_2$, the user can grant access to this app for perturbed adjacency bit vector and degree; otherwise, the user simply ignores this access request.

IV. LF-GDPR: IMPLEMENTATION

In this section, we present the implementation details of LF-GDPR. We first discuss graph metric reduction (step ①), followed by the perturbation protocols for adjacency bit vector and node degree, respectively (step ③). Then we elaborate on the aggregation and calibration algorithm (step ④).

A. Graph Metric Reduction

The reduction outputs a polynomial mapping function $Map(\cdot)$ from the target graph metric F to the adjacency matrix $M = \{B_1, B_2, \dots, B_n\}$ and degree vector $d = \{d_1, d_2, \dots, d_n\}$, i.e., $F = Map(M, d)$. Without loss of generality, we assume F is a polynomial of M and d . That is, F is a sum of terms F_l , each of which is a multiple of M and d of some exponents. Since F and F_l are scalars, in each term F_l , we need functions f and g to transform M and d with exponents to scalars, respectively. Formally,

$$F = \sum_l F_l = \sum_l f_{\phi_l}(M^{k_l}) \cdot g_{\psi_l}(d), \quad (2)$$

where M^{k_l} is the k_l -th power of adjacency matrix M whose cell (i, j) denotes the number of paths between node i and j of length k_l , ϕ_l projects a matrix to a cell, a row, a column

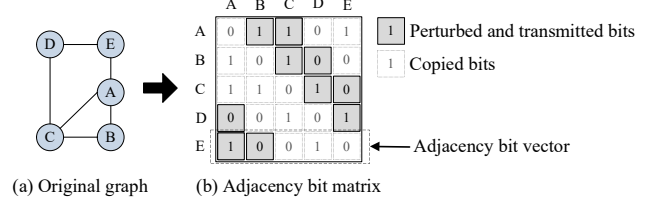


Fig. 2. Illustration of RABV protocol

or a sub-matrix, and $f_{\phi_l}(\cdot)$ denotes an aggregation function f (e.g., sum) after projection ϕ_l . Likewise, ψ_l projects a vector to a scalar or a sub-vector, and $g_{\psi_l}(\cdot)$ denotes an aggregation function g after ψ_l .

As such, the metric reduction step is to determine k_l , $f_{\phi_l}(\cdot)$, and $g_{\psi_l}(\cdot)$ for each term F_l in Eq. 2.

B. Adjacency Bit Vector Perturbation

An intuitive approach, known as *Randomized Neighbor List (RNL)* [10], perturbs each bit of the vector independently by the classic Randomized Response (RR) [12]. Formally, given an adjacency bit vector $B = \{b_1, b_2, \dots, b_n\}$, and privacy budget ϵ_1 , the perturbed vector $\tilde{B} = \{\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_n\}$ is obtained as follows:

$$\tilde{b}_i = \begin{cases} b_i & \text{w.p. } \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}} \\ 1 - b_i & \text{w.p. } \frac{1}{1+e^{\epsilon_1}} \end{cases} \quad (3)$$

RNL is proved to satisfy ϵ_1 -edge LDP for each user. However, **for undirected graphs, RNL can only achieve $2\epsilon_1$ -edge LDP for the collector**, because the data collector witnesses the same edge perturbed twice and independently. Let $\tilde{M} = \{\tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_n\}$ denote the perturbed adjacency matrix. The edge between node i and j appears in both \tilde{M}_{ij} and \tilde{M}_{ji} , each perturbed with privacy budget ϵ_1 . Then according to the theorem of composability, *RNL* becomes a $2\epsilon_1$ -edge LDP algorithm for an undirected graph, which is less private. Furthermore, *RNL* requires each user to perturb and send all n bits in the adjacency bit vector to data collector, which incurs a high computation and communication cost.

To address the problems of *RNL*, we propose a more private and efficient protocol *Randomized Adjacency Bit Vector (RABV)* to perturb edges in undirected graphs. As shown in Fig. 2(b), the adjacency matrix is composed of n rows, each corresponding to the adjacency bit vector of a node. For the first $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$ nodes, *RABV* uses RR as in Eq.3 to perturb and transmit $t = \lfloor \frac{n}{2} \rfloor$ bits (i.e., bits in grey) — from the $(i+1)$ -th bit to the $(i+1+t \bmod n)$ -th bit; for the rest nodes, *RABV* uses RR to perturb and transmit $t = \lfloor \frac{n-1}{2} \rfloor$ bits in the same way. In essence, *RABV* **perturbs one and only one bit** for each pair of symmetric bits in the adjacency matrix. The data collector can then obtain the whole matrix by copying bits in grey to their symmetric positions.

Following the same proof of *RNL*, *RABV* is guaranteed to satisfy ϵ_1 -edge LDP for the collector. Further, since each node only perturbs and transmits about half of the bits in an adjacency bit vector, *RABV* significantly reduces computation and communication cost of *RNL*.

C. Node Degree Perturbation

Releasing the degree of a node while satisfying edge ϵ -LDP is essentially a centralized DP problem because all edges incident to this node, or equivalently, all bits in its adjacency bit vector, form a database and the degree is a count function. In the literature, *Laplace Mechanism* [2] is the predominant technique to perturb numerical function values such as counts. As such, LF-GDPR adopts it to perturb the degree d_i of each node i . According to the definition of edge LDP, two adjacency bit vectors \mathbf{B} and \mathbf{B}' are two neighboring databases if they differ in only one bit. As such, the sensitivity of degree (i.e., count function) is 1, and therefore adding Laplace noise $\text{Lap}(\frac{1}{\epsilon_2})$ to the node degree can satisfy ϵ_2 -LDP. That is, $\tilde{d}_i = d_i + \text{Lap}(\frac{1}{\epsilon_2})$.

Similar to perturbing adjacency bit vector, however, in the above naive approach the data collector witnesses two node degrees d_i and d_j perturbed independently, but they share the same edge between i and j . As DP or LDP does not refrain an adversary from possessing any background knowledge, in the worst case the collector already knows all edges except for this one. As such, witnessing the two node degrees d_i and d_j is degenerated to witnessing the edge between i and j twice and independently.

Unfortunately, the remedy that works for perturbing adjacency bit vector cannot be adopted here, as direct bit copy is not feasible for degree. As such, we take an alternative approach to increase the Laplace noise. The following theorem proves that if we add Laplace noise $\text{Lap}(\frac{2}{\epsilon_2})$ to every node degree, ϵ_2 -LDP can be satisfied for the collector.

Theorem 4.1: A perturbation algorithm \mathcal{A} satisfies ϵ_2 -LDP for the collector if it adds Laplace noise $\text{Lap}(\frac{2}{\epsilon_2})$ to every node degree d_i , i.e., $\tilde{d}_i = \mathcal{A}(d_i) = d_i + \text{Lap}(\frac{2}{\epsilon_2})$.

PROOF. Please refer to our technical report [14]. \square

D. Aggregation and Calibration

Upon receiving the perturbed adjacency matrix $\tilde{\mathbf{M}}$ and degree vector $\tilde{\mathbf{d}}$,¹ the data collector can estimate the target graph metric \tilde{F} by aggregation according to Eq. 2 with a calibration function $\mathcal{R}(\cdot)$:

$$\tilde{F} = \sum_l \mathcal{R}\left(f_{\phi_l}(\tilde{\mathbf{M}}^{k_l})\right) \cdot g_{\psi_l}(\tilde{\mathbf{d}}) \quad (4)$$

The calibration function aims to suppress the aggregation bias of $\tilde{\mathbf{M}}$ propagated by f_{ϕ_l} . On the other hand, no calibration is needed for $g_{\psi_l}(\tilde{\mathbf{d}})$ as $\tilde{\mathbf{d}}$ is already an unbiased estimation of \mathbf{d} , thanks to the Laplace Mechanism.

To derive $\mathcal{R}(\cdot)$, we regard \mathcal{R} as the mapping between $f_{\phi_l}(\mathbf{M}^{k_l})$ and $f_{\phi_l}(\tilde{\mathbf{M}}^{k_l})$. In other words, \mathcal{R} estimates $f_{\phi_l}(\mathbf{M}^{k_l})$ after observing $f_{\phi_l}(\tilde{\mathbf{M}}^{k_l})$. Formally,

$$\mathcal{R} : f_{\phi_l}(\tilde{\mathbf{M}}^{k_l}) \rightarrow f_{\phi_l}(\mathbf{M}^{k_l})$$

Further, the following theorem shows the accuracy guarantee of LF-GDPR.

¹In the sequel, $\tilde{\mathbf{d}}$ denotes the refined degree $\tilde{\mathbf{d}}^*$ to simplify the notation.

Theorem 4.2: For a graph metric F and our estimation \tilde{F} , with at least $1 - \beta$ probability, we have

$$|F - \tilde{F}| = O(\sqrt{\mathbb{E}[\tilde{F}^2] \cdot \log(1/\beta)})$$

PROOF. Please refer to our technical report [14]. \square

V. CONCLUSION

This paper presents a parameterized framework LF-GDPR for privacy-preserving graph metric estimation and analytics with local differential privacy. The building block is a user-side perturbation algorithm, and a collector-side aggregation and calibration algorithm. LF-GDPR simplifies the job of developing a practical LDP solution for a graph analysis task by providing a complete solution for all LDP steps. As for future work, we plan to extend LF-GDPR to more specific graph types, such as attributed graph and DAG. We will also evaluate the performance of LF-GDPR on other graph analysis tasks such as influential node analysis to demonstrate its wide applicability.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No: 91646203, 61941121, 61572413, U1636205, 61532010, 91846204 and 61532016), the Research Grants Council, Hong Kong SAR, China (Grant No: 15238116, 15222118 and C1008-16G) (corresponding author: Xiaofeng Meng).

REFERENCES

- [1] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, pages 429–438. IEEE, 2013.
- [2] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, 2006.
- [3] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques* Morgan Kaufmann, 3 edition, 2011.
- [4] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *ICDM*, pages 169–178, 2009.
- [5] Z. Jorgensen, T. Yu, and G. Cormode. Publishing attributed social graphs with formal privacy guarantees. In *SIGMOD*, pages 107–122, 2016.
- [6] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In *TCC*, pages 457–476. Springer, 2013.
- [7] W. Lu and G. Miklau. Exponential random graph estimation under differential privacy. In *KDD*, pages 921–930. ACM, 2014.
- [8] T. Martin, X. Zhang, and M. Newman. Localization and centrality in networks. *Physical review E*, 90(5):052808, 2014.
- [9] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM, 2007.
- [10] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren. Generating synthetic decentralized social graphs with local differential privacy. In *CCS*, pages 425–438. ACM, 2017.
- [11] B. Stephanie. Facebook Scandal a ‘Game Changer’ in Data Privacy Regulation. *Bloomberg*, Apr 8, 2018.
- [12] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.
- [13] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. SCAN: a structural clustering algorithm for networks. In *KDD*, pages 824–833. ACM, 2007.
- [14] Q. Ye, H. Hu, M. H. Au, X. Meng, and X. Xiao. LF-GDPR: A framework for estimating graph metrics with local differential privacy. Technical report. <http://www.eie.polyu.edu.hk/%7Ehaibohu/papers/lfgdpr.pdf>.
- [15] Q. Ye, H. Hu, X. Meng, and H. Zheng. PrivKV: Key-value data collection with local differential privacy. In *S&P*, pages 317–331. IEEE, 2019.



A Unified Adversarial Learning Framework for Semi-supervised Multi-target Domain Adaptation

Xinle Wu^{1,2}, Lei Wang^{2(✉)}, Shuo Wang¹, Xiaofeng Meng¹, Linfeng Li^{2,3},
Haitao Huang⁴, Xiaohong Zhang⁵, and Jun Yan²

¹ School of Information, Renmin University of China, Beijing, China
{xinle.wu, shuowang, xfmeng}@ruc.edu.cn

² Yidu Cloud (Beijing) Technology Co., Ltd., Beijing, China
{lei.wang01, Linfeng.Li, jun.yan}@yiduccloud.cn

³ Institute of Information Science, Beijing Jiaotong University, Beijing, China

⁴ The second Department of Neurology, Liaoning People's Hospital, Shenyang, China

⁵ The Fourth Affiliated Hospital, China Medical University,
Taichung, Taiwan, R.O.C.

Abstract. Machine learning algorithms have been criticized as difficult to apply to new tasks or datasets without sufficient annotations. Domain adaptation is expected to tackle this problem by establishing knowledge transfer from a labeled source domain to an unlabeled or sparsely labeled target domain. Most existing domain adaptation models focus on the single-source-single-target scenario. However, the pairwise domain adaptation approaches may lead to suboptimal performance when there are multiple target domains available, because the information from other related target domains is not being utilized. In this work, we propose a unified semi-supervised multi-target domain adaptation framework to implement knowledge transfer among multiple domains (a single source domain and multiple target domains). Specifically, we aim to learn an embedded space and minimize the marginal probability distribution differences among all domains in the space. Meanwhile, we introduce Prototypical Networks to perform classification, and extend it to semi-supervised settings. On this basis, we further align the conditional probability distributions among the domains by generating pseudo-labels for the unlabeled target data and training the model with bootstrapping method. Extensive sentiment analysis experiments show that our approach significantly outperforms several state-of-the-art methods.

Keywords: Domain adaptation · Adversarial learning · Semi-supervised · Prototypical networks · Self-training · Sentiment analysis

1 Introduction

Supervised learning algorithms have achieved great success in many fields with the availability of large quantities of labeled data. However, it is costly and time-consuming to annotate such large-scale training data for new tasks or datasets.

A naive idea is directly applying the model trained on a labeled source domain to the related and sparsely labeled target domain. Unfortunately, the model usually fails to perform well in the target domain due to domain shifts [24]. Domain adaptation (DA) is proposed to address this problem by transferring knowledge from a labeled source domain to a sparsely labeled target domain.

Existing DA methods can be divided into: supervised DA (SDA) [14, 20, 26], semi-supervised DA (SSDA) [11, 21, 23, 31], and unsupervised DA (UDA) [4, 5, 9, 15]. SDA methods assume that there are some labeled data in the target domain, and perform DA algorithms only use the labeled data. Conversely, UDA methods do not need any target data labels, but they require large amounts of unlabeled target data to align the distributions between domains. Considering that it is cheap to annotate a small number of samples and a few labeled data often leads to significant performance improvements, we focus on SSDA which exploits both labeled and unlabeled data in target domains.

Typical DA methods are designed to embed the data from the source and target domains into a common embedding space, and align the marginal probability distributions between the two domains. There are two approaches to achieve this, adversarial training [4, 10, 20, 27] and directly minimizing the distance between the two distributions [15, 18, 28, 35]. Both of the methods can generate domain-invariant feature representations for input data, and the representations from the source domain are used to train a classifier, which is then generalized to the target domain. However, only aligning the marginal distributions is not sufficient to ensure the success of DA [4, 5, 12, 33], because the conditional probability distributions between the source and target domains may be different.

Most DA algorithms focus on the single-source-single-target setting. However, in many practical applications, there are multiple sparsely labeled target domains. For example, in the sentiment analysis task of product reviews, we can take the reviews of Books, DVDs, Electronics and Kitchen appliances as different domains. If we only have access to sufficient labeled data of Book reviews (source domain), and hope to transfer knowledge to the other domains, then each of the other domains can be seen as a target domain. In this case, pairwise adaptation approaches may be suboptimal, especially when there are shared features between the source and multiple target domains or the source and the target domain are associated through another target domain [9]. This is due to that these methods fail to leverage the knowledge from other relevant target domains. In addition, considering the distribution differences among multiple target domains, simply merging multiple target domains into a single one may not be the optimal solution.

To address these problems, we propose semi-supervised multi-target domain adaptation networks (MTDAN). Specifically, we use a shared encoder to extract the common features shared by all domains, and a private encoder to extract the domain-specific features of each domain. For feature representations generated by the two encoders, we train a domain discriminator to distinguish which domain they come from. To ensure that the shared representation is domain-invariant, the shared encoder is encouraged to generate the representation

cannot be correctly distinguished by the domain discriminator. Given that there are only a few labeled data in each target domain, we introduce Prototypical Networks to perform classification, which is more superior than deep classifiers in few-shot scenarios [25]. We further leverage unlabeled data to refine prototypes, and extend Prototypical Networks to semi-supervised scenarios. Moreover, we utilize the self-training algorithm to exploit unlabeled target data, and we show that it can also align the class-conditional probability distributions among multiple domains.

Contributions. Our contributions are: a) We propose a unified adversarial learning framework for semi-supervised multi-target DA. b) We show that the prototype-based classifier can achieve better performance than the deep classifier when target domains have only a few labeled data and large amounts of unlabeled data. c) We show that the self-training algorithm can effectively align the class-conditional probability distributions among multiple domains. d) Our method outperforms several state-of-the-art DA approaches on sentiment analysis dataset.

2 Related Work

Domain Adaptation. Numerous domain adaptation approaches have been proposed to solve domain shift [29]. Most of them seek to learn a shared embedded space, in which the representations of source domain and target domain cannot be distinguished [27]. Based on that, the classifier trained with labeled source data can be generalized to the target domain. There are two typical ways to learn cross-domain representations: directly minimizing the distance between two distributions [15, 17, 18] and adversarial learning [6, 7, 26, 27].

For the first method, several distance metrics have been proposed to measure the distance between source and target distributions. One common distance metric is the Maximum Mean Discrepancy (MMD) [2], which computes the norm of the difference between two domain means in the reproducing Kernel Hilbert Space (RKHS). Specifically, the DDC method [28] used both MMD and regular classification loss on the source to learn representations that are discriminative and domain invariant. The Deep Adaptation Network (DAN) [15] applied MMD to the last full connected layers to match higher order statistics of the two distributions. Most recently, [18] proposed to reduce domain shift in joint distributions of the network activation of multiple task-specific layers. Besides, Zellinger et al. proposed Center Moment Discrepancy (CMD) [32] to diminish the domain shift by aligning the central moment of each order across domains.

The other method is to optimize the source and target mappings using adversarial training. The idea is to train a domain discriminator to distinguish whether input features come from the source or target, whereas the feature encoder is trained to deceive the domain discriminator by generating representations that cannot be distinguished. [6] proposed the gradient reversal algorithm (ReverseGrad), which directly maximizes the loss of the domain discriminator by reversing its gradients. DRCN in [8] takes a similar approach in addition to learning to

reconstruct target domain images. [3] enforced these adversarial losses in a shared feature space, while learned a private feature space for each domain to avoid the contamination of shared representations.

[4, 27, 33] argued that only aligning the marginal probability distributions between the source and target is not enough to guarantee successful domain adaptation. [16] proposed to align the marginal distributions and conditional distributions between the source and target simultaneously. [20] extended the domain discriminator to predict the domain and category of the embedded representation at the same time to align the joint probability distributions of input and output, and achieved a leading effect in the supervised domain adaptive scene. [5] proposed to align the class-conditional probability distributions between the source and target.

Recently, Zhao et al. [34] introduced an adversarial framework called MDAN, which is used for multi-source-single-target domain adaptation. They utilized a multi-class domain discriminator to align the distributions between multiple source and a target domain. [9] proposed an information theoretic approach to solve unsupervised multi-target domain adaptation problem, which maximizes the mutual information between the domain labels and domain-specific features, while minimizes the mutual information between the the domain labels and the domain-invariant features. Unlike their approach, we base our method on self-training rather than entropy regularization. Moreover, we introduce prototypical networks to perform classification, which is more effective than deep classifiers in SSDA scenarios.

Semi-supervised Learning. Recently, some works treat domain adaptation as a semi-supervised learning task. [11] proposed a Domain Adaptive Semi-supervised learning framework (DAS) to jointly perform feature adaptation and semi-supervised learning. [21] applied a co-training framework for semi-supervised domain adaptation, in which the shared classifier and the private classifier boost each other to achieve better performance. [22] re-evaluated classic general-purpose bootstrapping approaches under domain shift, and proved that the classic bootstrapping algorithms make strong baselines on domain adaptation tasks.

3 Preliminaries

In this section, we introduce the notations and definitions related to single-source-multi-target DA.

Notations. We use \mathcal{D} to denote a domain, which consists of an m -dimensional feature space \mathcal{X} and a marginal probability distribution $P(\mathbf{x})$, i.e., $\mathcal{D} = \{\mathcal{X}, P(\mathbf{x})\}$, where $\mathbf{x} \in \mathcal{X}$. We use \mathcal{T} to denote a task which consists of a C -cardinality label set \mathcal{Y} and a conditional probability distribution $P(y|\mathbf{x})$, i.e., $\mathcal{T} = \{\mathcal{Y}, P(y|\mathbf{x})\}$, where $y \in \mathcal{Y}$.

Problem Formulation (Single-Source-Multi-target Domain Adaptation). Let $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ be a labeled source domain where n_s is the

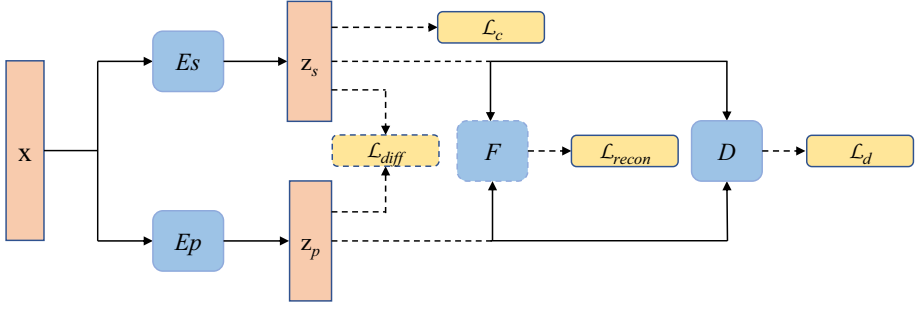


Fig. 1. The network structure of the proposed framework. The shared encoder E_s captures the common features shared among domains, while the private encoder E_p captures the domain-specific features. The shared decoder F reconstructs the input samples by using both the shared and private representations. The domain classifier D learns to distinguish which domain the input representations come from. The orthogonality constraint loss \mathcal{L}_{diff} encourages E_s and E_p to encode different aspects of the inputs. The prototype-based classifier is computed on-the-fly, and the classification loss \mathcal{L}_c is only used to optimize E_s .

number of labeled samples and let $\mathcal{D}_t = \{\mathcal{D}_{t_i}\}_{i=1}^K$ be multiple sparsely labeled target domains where $\mathcal{D}_{t_i} = \{(\mathbf{x}_l^{t_i}, y_l^{t_i})\}_{l=1}^{n_{l_i}} \cup \{\mathbf{x}_u^{t_i}\}_{u=1}^{n_{u_i}}$, K is the number of target domains, and n_{l_i} ($n_{l_i} \ll n_s$) and n_{u_i} ($n_{u_i} \gg n_{l_i}$) refer to the number of labeled and unlabeled samples of i -th target domain respectively. We assume that all domains share the same feature space \mathcal{X} and label space \mathcal{Y} , but the marginal probability distributions and the conditional probability distributions of source domain and multiple target domains are different from each other. The goal is to learn a classifier using the labeled source data and a few labeled target data, that generalizes well to the target domain.

4 Methodology

In this section, we describe each component and the corresponding loss function of the proposed framework in detail.

4.1 Proposed Approach

Our model consists of four components as shown in Fig. 1. A shared encoder E_s is trained to learn cross-domain representations, a private encoder E_p is trained to learn domain-specific representations, a shared decoder F is trained to reconstruct the input sample, and a discriminator D is trained to distinguish which domain the input sample comes from. Task classification is performed by calculating the distance from the domain-invariant representations to prototype representations of each label class.

Domain-Invariant and Domain-Specific Representations. We seek to extract domain-invariant (shared) and domain-specific (private) representations

for each input \mathbf{x} simultaneously. In our model, the shared encoder E_s and the private encoder E_p learn to generate the above two representations respectively:

$$\begin{aligned}\mathbf{z}_s &= E_s(\mathbf{x}, \boldsymbol{\theta}_s) \\ \mathbf{z}_p &= E_p(\mathbf{x}, \boldsymbol{\theta}_p)\end{aligned}\tag{1}$$

Here, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_p$ refer to the parameters of E_s and E_p respectively, \mathbf{z}_s and \mathbf{z}_p refer to the shared and private representations of the input \mathbf{x} respectively. Note that E_s and E_p can be MLP, CNN or LSTM encoders, depending on different tasks and datasets.

Reconstruction. In order to avoid information loss during the encoding, we reconstruct input samples with both shared and private representations. We use $\hat{\mathbf{x}}$ to denote the reconstruction of the input \mathbf{x} , which is generated by decoder F :

$$\hat{\mathbf{x}} = F(\mathbf{z}_s + \mathbf{z}_p, \boldsymbol{\theta}_f),\tag{2}$$

where $\boldsymbol{\theta}_f$ are the parameters of F . We use mean square error to define the reconstruction loss \mathcal{L}_{Recon} , which is applied to all domains:

$$\mathcal{L}_{Recon} = \frac{\lambda_r}{N} \sum_{i=1}^N \frac{1}{C} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2,\tag{3}$$

where C is the dimension of the input \mathbf{x} , N is the total number of samples in all domains, \mathbf{x}_i refers to the i -th sample, λ_r is the hyper-parameter controlling the weight of the loss function, and $\|\cdot\|_2^2$ is the squared L_2 -norm.

Orthogonality Constraints. To minimize the redundancy between shared and private representations, we introduce orthogonality constraints to encourage the shared and private encoders to encode different aspects of inputs. Specifically, we use \mathbf{H}_s to denote a matrix, each row of which corresponds to the shared representation of each input \mathbf{x} . Similarly, let \mathbf{H}_p be a matrix, each row of which corresponds to the private representation of each input \mathbf{x} . The corresponding loss function is:

$$\mathcal{L}_{Diff} = \lambda_{diff} \|\mathbf{H}_s^\top \mathbf{H}_p\|_F^2,\tag{4}$$

where λ_{diff} is the scale factor, $\|\cdot\|_F^2$ is the squared Frobenius norm.

Adversarial Training. The goal of adversarial training is to regularize the learning of the shared encoder E_s , so as to minimize the distance of distributions among source and multiple target domains. After that, we can apply the source classification model directly to the target representations. Therefore, we first train a domain discriminator D with the domain labels of the shared and private representations (since we know which domain each sample comes from, it is obvious that we can generate a domain label for each sample). The discriminator D is a multi-class classifier designed to distinguish which domain the

Algorithm 1. MTDAN Algorithm

Input: labeled source domain examples L_s , labeled multi-target domain examples $L_t = \{L_{t_i}\}_{i=1}^K$, unlabeled multi-target domain examples $U_t = \{U_{t_i}\}_{i=1}^K$
Hyper-parameters: coefficients for different losses: $\lambda_r, \lambda_d, \lambda_c, \lambda_{diff}$, mini-batch size b , learning rate η

```

1: initialize  $\theta_s, \theta_p, \theta_f, \theta_d$ 
2: repeat
3:   repeat
4:     Sample a mini-batch from  $\{L_s, L_t\}$ 
5:     Train  $F$  by minimizing  $\mathcal{L}_{Recon}$ 
6:     Train  $D$  by minimizing  $\mathcal{L}_D$ 
7:     Train  $E_p$  by minimizing  $\mathcal{L}_P$ 
8:     Train  $E_s$  by minimizing  $\mathcal{L}_S$ 
9:   until Convergence
10:  Apply Eq.(9) to label  $U_t$ 
11:  Select the most confident  $p$  positive and  $n$  negative predicted examples  $U_t^l$ 
    from  $U_t$ 
12:  Remove  $U_t^l$  from  $U_t$ 
13:  Add examples  $U_t^l$  and their corresponding labels to  $L_t$ 
14: until obtain best performance on the developing dataset

```

input representation comes from. Thus, D is optimized according to a standard supervised loss, defined below:

$$\mathcal{L}_D = \mathcal{L}_{D_p} + \mathcal{L}_{D_s}, \quad (5)$$

$$\mathcal{L}_{D_p} = -\frac{\lambda_d}{N} \sum_{i=1}^N \mathbf{d}_i^\top \log D(E_p(\mathbf{x}_i, \theta_p), \theta_d), \quad (6)$$

$$\mathcal{L}_{D_s} = -\frac{\lambda_d}{N} \sum_{i=1}^N \mathbf{d}_i^\top \log D(E_s(\mathbf{x}_i, \theta_s), \theta_d), \quad (7)$$

where \mathbf{d}_i is the one-hot encoding of the i -th sample's domain label, θ_d is the parameter of D , and λ_d is the scale factor.

Second, we train the shared encoder E_s to fool the discriminator D by generating cross-domain representations. We guarantee this by adding $-\mathcal{L}_{D_s}$ to the loss function of the shared encoder E_s . On the other hand, we hope the private encoder only extracts domain-specific features. Thus, we add \mathcal{L}_{D_p} to the loss function of E_p to generate representations that can be distinguished by D .

Prototypical Networks for Task Classification. The simplest way to classify the target samples is to train a deep classifier, however, it may only achieve

suboptimal performance as we can see in Table 1. The reason is that there are only a few labeled samples in each target domain, which is not enough to fine-tune a deep classifier with many parameters, so that the classifier is easy to overfit the source labeled data. Although we could generate pseudo-labeled data for target domains, the correctness of the pseudo-labels can not be guaranteed due to the poor performance of the deep classifier.

To efficiently utilize the labeled samples in target domains, we refer to the idea of prototypical networks [25]. Prototypical networks assume that there is a prototype in the latent space for each class, and the projections of samples belonging to this class cluster around the prototype. The classification is then performed by computing the distances to prototype representations of each class in the latent space. By reducing parameters of the model, the prototype-based classifier can achieve better performance than the deep classifier when labeled samples are insufficient. Note that we refine prototypes during self-training by allowing unlabeled samples with pseudo-labels to update the prototypes. Specifically, we compute the average of shared representations belonging to each class in a batch as prototypes:

$$\mathbf{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} E_s(\mathbf{x}_i, \boldsymbol{\theta}_s), \quad (8)$$

where n_k is the number of samples belonging to class k in a batch. Then we calculate a distribution by applying softmax function to distances between a shared representation with a prototype:

$$p(y = k|\mathbf{x}) = \frac{\exp(-d(\mathbf{z}_s, \mathbf{c}_k))}{\sum_{k'} \exp(-d(\mathbf{z}_s, \mathbf{c}'_k))}, \quad (9)$$

where $d(\cdot)$ is a distance measure function. We use the squared Euclidean distance in this work. The classification loss is defined as:

$$\mathcal{L}_C = -\lambda_c \log p(y = k|\mathbf{x}), \quad (10)$$

where λ_c is the scale factor.

Self-training for Conditional Distribution Adaptation. As described in [5, 19], only aligning the marginal probability distributions between source and target is not enough to guarantee successful domain adaptation. Because this only enforces alignment of the global domain statistics with no class specific transfer. Formally, we can achieve $P_s(E_s(\mathbf{x}_s)) \approx P_{t_i}(E_s(\mathbf{x}_{t_i}))$ by introducing adversarial training, but $P_s(y_s|E_s(\mathbf{x}_s)) \neq P_{t_i}(y_{t_i}|E_s(\mathbf{x}_{t_i}))$ may still hold, where $P_s(y_s|E_s(\mathbf{x}_s))$ can be regarded as the classifier trained with source data.

Here, we tackle this problem by further reducing the difference of conditional probability distributions among source domain and target domains. In practice, we replace conditional probability distributions with class-conditional probability distributions, because the posterior probability is quite involved [16].

However, it is nontrivial to adapt class-conditional distributions, as most of the target samples are unlabeled. We address this problem by producing pseudo-labels for unlabeled target samples, and train the whole model in a bootstrapping way. As we perform more learning iterations, the number of target samples with correct pseudo-labels grows and progressively enforces distributions to align class-conditionally.

To be specific, we first train our model on labeled source and target samples. Then, we use the model to generate a probability distribution over classes for each unlabeled target sample. If the probability of a sample on a certain class is higher than a predetermined threshold τ , the sample would be added to the training set with the class as its pseudo-label.

Loss Function and Model Training. We alternately optimize the four modules of our model.

For E_p , the goal of training is to minimize the following loss:

$$\mathcal{L}_P = \mathcal{L}_{Recon} + \mathcal{L}_{Diff} + \mathcal{L}_{D_p} \quad (11)$$

For E_s , the goal of training is to minimize the following loss:

$$\mathcal{L}_S = \mathcal{L}_{Recon} + \mathcal{L}_{Diff} - \mathcal{L}_{D_s} + \mathcal{L}_C \quad (12)$$

For F and D , the losses are \mathcal{L}_{Recon} and \mathcal{L}_D , respectively. The detailed training process is shown in algorithm 1.

5 Experiments

5.1 Dataset

We evaluate our proposed model on the Amazon benchmark dataset [1]. It is a sentiment classification dataset¹, which contains Amazon product reviews from four different domains: Books (B), DVD (D), Electronics (E), and Kitchen appliances (K). We remove reviews with neutral labels and encode the remaining reviews into 5000 dimensional feature vectors of unigrams and bigrams with binary labels indicating sentiment.

We pick two product as the source domain and the target domain in turn, and the other two domains as the auxiliary target domains, so that we construct 12 single-source-three-target domain adaptation tasks. For each task, the source domain contains 2,000 labeled examples, and each target domain contains 50 labeled examples and 2,000 unlabeled examples. To fine-tune the hyper-parameters, we randomly select 500 labeled examples from the target domain as the developing dataset.

¹ <https://www.cs.jhu.edu/mdredze/datasets/sentiment/>.

5.2 Compared Method

We compare MTDAN with the following baselines:

- (1) **ST**: The basic neural network classifier without any domain adaptation trained on the labeled data of the source domain and the target domain.
- (2) **CoCMD**: This is the state-of-the-art pairwise SSDA method on the Amazon benchmark dataset [21]. The shared encoder, private encoder and reconstruction decoder used in this model are the same as ours.
- (3) **MTDA-ITA**: This is the state-of-the-art single-source-multi-target UDA method on three benchmark datasets for image classification [9]. We implemented the framework and extend it to semi-supervised DA method. The shared encoder, private encoder, reconstruction decoder and domain classifier used in this model are the same as ours.
- (4) **c-MTDAN**: We combine all the target domains into a single one, and train it using MTDAN. Similarly, we also report the performance of c-CoCMD and c-MTDA-ITA.
- (5) **s-MTDAN**: We do not use any auxiliary target domains, and train MTDAN on each source-target pair.

5.3 Implementation Details

Considering that each input sample in the dataset is a tf-idf feature vector without word ordering information, we use a multilayer perceptron (MLP) with an input layer (5000 units) and one hidden layer (50 units) and sigmoid activation functions to implement both shared and private encoders. The reconstruction decoder consists of one dense hidden layer (2525 units), tanh activation functions, and relu output functions. The domain discriminator is composed of a softmax layer with n -dimensional outputs, where n is the number of the source and target domains. For MTDA-ITA, we follow the framework proposed by [9], and use the above modules to replace the original modules in the framework. Besides, the task classifier for MTDA-ITA is a fully connected layer with softmax activation functions.

The network is trained with Adam optimizer [13] and with learning rate 10^{-4} . The mini-batch size is 50. The hyper-parameters $\lambda_r, \lambda_d, \lambda_c$ and λ_{diff} are empirically set to 1.0, 0.5, 0.1 and 1.0 respectively. The threshold τ for producing pseudo-labels is set to 0.8. Following previous studies, we use classification accuracy metric to evaluate the performances of all approaches.

5.4 Results

The performances of the proposed model and other state-of-the-art methods are shown in Table 1. Key observations are summarized as follows. (1) The proposed model MTDAN achieves the best results in almost all tasks, which proves the effectiveness of our approach. (2) c-CoCMD has worse performance in all tasks compared with CoCMD, although c-CoCMD exploits labeled and unlabeled data from auxiliary target domains for training. Similar observation can

also be observed by comparing MTDA-ITA with c-MTDA-ITA and MTDAN with c-MTDAN. This demonstrates that simply combine all target domains into a single one is not an effective method to solve the multi-target DA problem. (3) Our model outperforms CoCMD by an average of nearly 2.0%, which indicates that our model can effectively leverage the labeled and unlabeled data from multiple target domains. Similarly, our model performs better than its variant, s-MTDAN, which does not leverage the data from auxiliary target domains. This also shows that it is helpful to mine knowledge from auxiliary target domains. (4) Although MTDA-ITA is also a multi-target domain adaptation method, its performance is worse than that of MTDAN. This can be due to (i) self-training is a superior method than entropy regularization to exploit unlabeled target data, (ii) the prototype-based classifier is more efficient than the deep classifier in semi-supervised scenarios, (iii) we introduce orthogonality constraints to further reduce the redundancy between shared and private representations. (5) In the K→E task, MTDAN performs slightly worse than s-MTDAN. This can be explained that domain K is closer to domain E than the other domains as shown in Fig. 2 (a), and MTDAN leads to negative transfer when using relevant target domains to help domain adaptation. (6) s-MTDAN outperforms CoCMD in 9 of the 12 tasks, note that both of them do not use the auxiliary domains. This indicates that our model is more effective than CoCMD in pairwise domain adaptation task. (7) All models achieve better performance than the basic ST model, which demonstrates that domain adaptation methods are crucial when there exist a domain gap between the source domain and the target domain.

Table 1. Average classification accuracy with 5 runs on target domain testing dataset. The best is shown in bold. c-X: combining all target domains into a single one and performing pairwise domain adaptation with model X. s-X: performing pairwise domain adaptation between the original source and target domains with model X

Method	B→D	B→E	B→K	D→B	D→E	D→K	E→B	E→D	E→K	K→B	K→D	K→E
ST	81.6	75.8	78.2	80.0	77.0	80.4	74.7	75.4	85.7	73.8	76.6	85.3
CoCMD	83.1	83.0	85.3	81.8	83.4	85.5	76.9	78.3	87.3	77.2	79.6	87.2
c-CoCMD	82.7	82.2	84.5	80.6	83.0	84.8	76.3	77.6	87.1	75.9	79.4	86.1
MTDA-ITA	83.8	83.2	83.7	81.8	83.6	85.4	76.6	78.9	87.7	77.0	78.8	86.8
c-MTDA-ITA	83.3	82.3	83.2	81.4	83.0	85.0	76.0	79.3	87.6	76.7	78.5	87.0
s-MTDAN	83.3	83.9	84.7	81.6	83.7	84.7	78.0	80.2	87.9	78.6	79.9	87.8
c-MTDAN	84.0	84.0	85.5	81.7	84.3	85.9	80.2	80.7	88.1	79.8	80.5	87.0
MTDAN	84.5	84.3	86.0	82.3	85.3	87.2	80.5	81.2	88.9	80.0	80.9	87.4

5.5 Ablation Studies

We performed ablation experiments to verify the importance of each component of our proposed model. We report the results of removing orthogonality

constraints loss (set $\lambda_{diff}=0$), self-training process, the prototype-based classifier (replaced by the deep classifier) respectively.

As we can see from Table 2, removing each of the above components causes performance degradation. To be specific, disabling self-training degrades the performance to the greatest extent, with an average decrease of 5.1%, which shows the importance of mining information from the unlabeled data of target domains. Similarly, replacing prototype-based classifiers with deep classifiers also leads to performance degradation, with an average decrease of 1.4%, which shows that the prototype-based classifiers is more effective than deep classifiers in semi-supervised scenarios. Besides, disabling the orthogonality constraints loss leads to a performance degradation of 0.7%, which indicates that encouraging the disjoint of shared and private representations can make the shared feature space more common among all domains.

We did not test the performance degradation caused by disabling reconstruction loss and multi-class adversarial training loss, because they have been proved in previous work [3, 9]. To summarize, each of the proposed components helps improve classification performance, and using all of them brings the best performance.

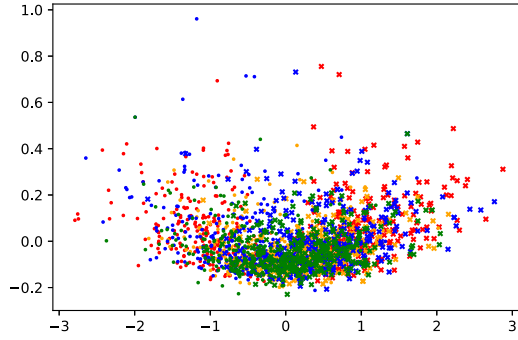
Table 2. Ablations. Performance of the proposed model when one component is removed or replaced. woDiff means without orthogonality constraints loss, woSelf means without self-training procedures, woProto means replace the prototype-based classifier with the deep classifier.

Method	B→D	B→E	B→K	D→B	D→E	D→K	E→B	E→D	E→K	K→B	K→D	K→E
MTDAN- <i>woDiff</i>	84.1	83.6	85.9	81.5	85.0	86.7	79.7	80.5	88.3	79.6	80.0	87.0
MTDAN- <i>woSelf</i>	82.8	77.6	80.0	81.6	78.8	81.5	74.8	75.6	86.7	74.5	77.3	86.7
MTDAN- <i>woProto</i>	83.3	83.8	84.1	81.8	83.8	86.9	79.2	78.9	87.9	78.0	80.3	86.4
MTDAN	84.5	84.3	86.0	82.3	85.3	87.2	80.5	81.2	88.9	80.0	80.9	87.4

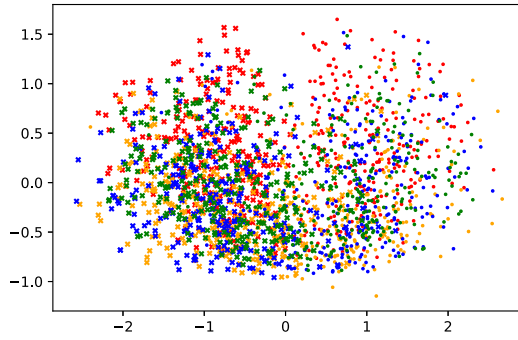
5.6 Feature Visualization

In order to understand the behavior of the proposed model intuitively, we project the shared and private encoder outputs into two-dimensional space with principle component analysis (PCA) [30] and visualize them. For comparison, we also show the visualization result of the basic ST model. Due to space constraints, we only show the visualization results of MTDAN with B as the source domain, E as the target domain, D and K as the auxiliary target domains. The results are shown in Fig. 2.

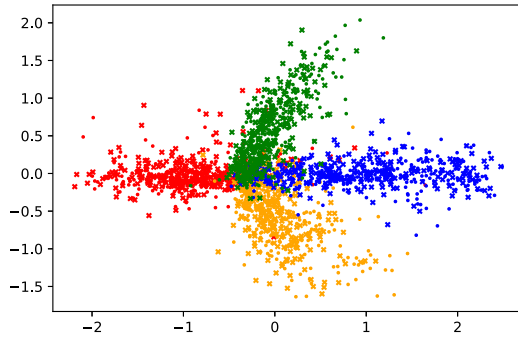
Figure 2 (a) shows the encoder output distribution of the ST model. As we can see, the distributions of domain B and domain D (called group 1) are similar and the distributions of domain E and domain D (called group 2) are similar, while the distributions of cross-group domains are relatively different. That's why the ST model gets worse classification performance when the source domain



(a) ST



(b) MTDAN-shared



(c) MTDAN-private

Fig. 2. Feature visualization for the embedding of source and target data. The red, blue, yellow and green symbols denote the samples from B, D, E and K respectively. The symbol ‘x’ is used for positive samples and ‘.’ is for negative samples. (a) the distribution of the encoder output of ST, (b) the distribution of shared representations of MTDAN, (c) the distribution of private representations of MTDAN. For ST and MTDAN, we take B as the source domain and D, E and K as the target domains.

and the target domain belong to different groups. Besides, there is no obvious boundary between positive and negative samples, which is consistent with the poor performance of the ST model.

Figure 2 (b) shows the distribution of the shared encoder output of the MTDAN model. We can see that the shared representations of the source and target domains are very close, which demonstrates that our model can effectively align the marginal distributions among the source and multiple target domains. Meanwhile, for each class of samples, the shared representations of the source and target domains are also very close, which demonstrates that our model can effectively align the class-conditional distributions among multiple domains. Comparing (a) and (b), we can find that the boundary of positive and negative samples in (b) is more obvious than that in (a), which means the shared representations of MTDAN model have superior class separability.

Figure 2 (c) shows the distribution of the private encoder output of the MTDAN model. We can see that the private representations have good domain separability, partially because the domain discriminator D encourages the private encoder E_p to generate domain-specific feature representations.

6 Conclusion

In this paper, we propose MTDAN, a unified framework for semi-supervised multi-target domain adaptation. We utilize multi-class adversarial training to align the marginal probability distributions among source domain and multiple target domains. Meanwhile, we perform self-training on target unlabeled data to align the conditional probability distributions among the domains. We further introduce Prototypical Networks to replace the deep classifiers, and extend it to semi-supervised scenarios. The experimental results on sentiment analysis dataset demonstrate that our method can effectively leverage the labeled and unlabeled data of multiple target domains to help the source model achieve generalization, and is superior to the existing methods. The proposed framework could be used for other domain adaptation tasks, and we leave this as our future work.

Acknowledgment. This work was supported by National Natural Science Foundation of China (Grant No: 91646203, 91846204, 61532010, 61941121, 61532016 and 61762082). The corresponding author is Xiaofeng Meng.

References

1. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 440–447 (2007)
2. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **22**(14), e49–e57 (2006)

3. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: *Advances in Neural Information Processing Systems*, pp. 343–351 (2016)
4. Cicek, S., Soatto, S.: Unsupervised domain adaptation via regularized conditional alignment. *arXiv preprint [arXiv:1905.10885](https://arxiv.org/abs/1905.10885)* (2019)
5. Gabourie, A.J., Rostami, M., Pope, P.E., Kolouri, S., Kim, K.: Learning a domain-invariant embedding for unsupervised domain adaptation using class-conditioned distribution alignment. In: *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 352–359. IEEE (2019)
6. Ganin, Y., Lempitsky, V.S.: Unsupervised domain adaptation by backpropagation. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015*, pp. 1180–1189 (2015)
7. Ganin, Y., et al.: Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**(1), 2030–2096 (2016)
8. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9908, pp. 597–613. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_36
9. Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., Pavlovic, V.: Unsupervised multi-target domain adaptation: An information theoretic approach. *arXiv preprint [arXiv:1810.11547](https://arxiv.org/abs/1810.11547)* (2018)
10. Guo, J., Shah, D.J., Barzilay, R.: Multi-source domain adaptation with mixture of experts. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018*, pp. 4694–4703 (2018)
11. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: Adaptive semi-supervised learning for cross-domain sentiment classification. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018*, pp. 3467–3476 (2018)
12. Hosseini-Asl, E., Zhou, Y., Xiong, C., Socher, R.: Augmented cyclic adversarial learning for low resource domain adaptation. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019* (2019)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings* (2015)
14. Koniusz, P., Tas, Y., Porikli, F.: Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4478–4487 (2017)
15. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015*, pp. 97–105 (2015)
16. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2200–2207 (2013)
17. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: *Advances in Neural Information Processing Systems*, pp. 136–144 (2016)
18. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 2208–2217 (2017). *JMLR. org*

19. Luo, Z., Zou, Y., Hoffman, J., Fei-Fei, L.F.: Label efficient learning of transferable representations across domains and tasks. In: *Advances in Neural Information Processing Systems*, pp. 165–177 (2017)
20. Motiian, S., Jones, Q., Iranmanesh, S., Doretto, G.: Few-shot adversarial domain adaptation. In: *Advances in Neural Information Processing Systems*, pp. 6670–6680 (2017)
21. Peng, M., Zhang, Q., Jiang, Y.g., Huang, X.J.: Cross-domain sentiment classification with target domain specific information. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2505–2513 (2018)
22. Ruder, S., Plank, B.: Strong baselines for neural semi-supervised learning under domain shift. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 1: Long Papers*, pp. 1044–1054 (2018)
23. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. *arXiv preprint [arXiv:1904.06487](https://arxiv.org/abs/1904.06487)* (2019)
24. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plann. Inference* **90**(2), 227–244 (2000)
25. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems*, pp. 4077–4087 (2017)
26. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4068–4076 (2015)
27. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176 (2017)
28. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: maximizing for domain invariance. *arXiv preprint [arXiv:1412.3474](https://arxiv.org/abs/1412.3474)* (2014)
29. Wang, M., Deng, W.: Deep visual domain adaptation: a survey. *Neurocomputing* **312**, 135–153 (2018)
30. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. *Chemometr. Intell. Lab. Syst.* **2**(1–3), 37–52 (1987)
31. Yao, T., Pan, Y., Ngo, C.W., Li, H., Mei, T.: Semi-supervised domain adaptation with subspace learning for visual recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2142–2150 (2015)
32. Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (CMD) for domain-invariant representation learning. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017, Conference Track Proceedings* (2017)
33. Zhao, H., des Combes, R.T., Zhang, K., Gordon, G.J.: On learning invariant representations for domain adaptation. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA*, pp. 7523–7532 (2019)
34. Zhao, H., Zhang, S., Wu, G., Moura, J.M., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: *Advances in Neural Information Processing Systems*, pp. 8559–8570 (2018)
35. Zhuang, F., Cheng, X., Luo, P., Pan, S.J., He, Q.: Supervised representation learning: Transfer learning with deep autoencoders. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015)

科研成果

1. 论文列表

● 数据智能 (Data Intelligence)

[1] Shuo Wang, Aishan Maolinyazi, Xinle Wu, and Xiaofeng Meng. Emo2Vec: Learning emotional embeddings via multi-emotion category[J]. ACM Transactions on Internet Technology (TOIT), 2020, 20(2): 1-17.

[2] Wu X, Wang L, Wang S, et al. A Unified Adversarial Learning Framework for Semi-supervised Multi-target Domain Adaptation[C], DASFAA 2020.

[3] Chen Yang, Yongjie Du, Zhihui Du, Xiaofeng Meng:Micro Analysis to Enable Energy-Efficient Database Systems. EDBT 2020: 61-72

● 数据治理 (Data Governance)

[1] Q. Ye, H. Hu, M. H. Au, X. Meng, X. Xiao. LF-GDPR: Graph Metric Estimation with Local Differential Privacy. IEEE Transactions on Knowledge and Data Engineering (TKDE).

[2] Q. Ye, H. Hu, M. H. Au, X. Meng, X. Xiao. Towards Locally Differentially Private Generic Graph Metric Estimation. Proc. of the 36th IEEE International Conference on Data Engineering (ICDE'20), Dallas, USA, Apr. 2020, pp 1922-1925.

[3] 朱敏杰, 叶青青, 孟小峰, 杨鑫. 基于权限的移动应用程序隐私风险量化[J]. 中国科学: 信息科学, accepted.

[4] 孟小峰, 刘立新. 区块链与数据治理[J]. 中国科学基金, 2020,34(1):12-17.

[5] 孟小峰, 刘立新. 基于区块链的数据透明化: 问题与挑战[J]. 计算机研究与发展, accepted.

[6] 孟小峰.破解数据垄断的几种治理模式研究[J].人民论坛,2020(27):58-61.

2. 学位论文

王硕，融合情绪心理学的情感智能计算研究（Research on Emotional Intelligence Computation Based on Psychology of Emotion），中国人民大学，博士学位论文，2020.6.6

在读时间：2015 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 22 日

答辩地点：线上答辩



叶青青，本地化差分隐私关键技术研究（Research on Local Differential Privacy），中国人民大学，博士学位论文，2020.6.6

在读时间：2015 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 22 日

答辩地点：线上答辩



杨晨，大规模科学数据的实时分析研究（Research on Real-Time Analysis for Large-Scale Scientific Data），中国人民大学，博士学位论文，2020.6.6

在读时间：2015 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 22 日

答辩地点：线上答辩



吴新乐，半监督多目标领域自适应方法研究（Research on Semi-supervised Multitarget Domain Adaptation），中国人民大学，硕士学位论文，2020.6.4

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



段志强，科学时序数据流分析的关键技术研究（Research on Key Technologies of Scientific Time Series Data Flow Analysis），中国人民大学，硕士学位论文，2020.6.4

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



杜永杰，超大空间范围查询优化研究（Research On Query Optimization of Large Spatial Range），中国人民大学，硕士学位论文，2020.6.4

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



汤庆，云环境下分布式数据管理系统的高可用调度框架（High availability scheduling framework of distributed data management system in cloud environment），中国人民大学，专业硕士学位论文，2020.6.8

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



王飞，基于 ScholarSpace 的智能指派系统（Intelligent Assignment System Based on ScholarSpace），中国人民大学，专业硕士学位论文，2020.6.5

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



吴永泰，基于知识图谱的学者推荐系统的研究与实现（Research and Implementation of Scholar Recommendation System Based on Knowledge Graph），中国人民大学，专业硕士学位论文，2020.6.15

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



杨鑫，移动应用场景下的隐私量化系统的设计与实现（Design and Implementation of Privacy Quantification System Based on Mobile Applications），中国人民大学，专业硕士学位论文，2020.3.15

在读时间：2017 年 9 月-2020 年 6 月

答辩时间：2020 年 5 月 15 日

答辩地点：线上答辩



3. 已授权专利

- 一种融合多背景知识知识图谱嵌入方法

发明名称：一种融合多背景知识知识图谱嵌入方法

申请人：孟小峰；杜治娟

专利号：ZL 2017 1 0459984.X

申请时间：2017 年 07 月 07 日

授权时间：2020 年 03 月 31 日



- 一种强适应性的知识库补全方法

发明名称：一种强适应性的知识库补全方法

申请人：孟小峰；张祎；王秋月

专利号：ZL 2017 1 0630354.8

申请时间：2017 年 07 月 28 日

授权时间：2020 年 01 月 10 日



- 一种手机 APP 隐私风险量化评估方法

发明名称：一种手机 APP 隐私风险量化评估方法

申请人：孟小峰；朱敏杰

专利号：ZL 2017 1 0623492.3

申请时间：2017 年 07 月 27 日

授权时间：2019 年 12 月 13 日



- 一种应对倾斜数据流在线连接的处理方法

发明名称：一种应对倾斜数据流在线连接的处理方法

申请人：孟小峰；王春凯

专利号：ZL 2017 1 0452086.4

申请时间：2017 年 07 月 05 日

授权时间：2019 年 11 月 15 日



活动专题

SciDI Cup: 科学数据智能发现大赛

为了寻找广袤银河中的“流浪地球”，实验室承办了首届科学数据智能发现大赛(SciDI Cup)。该竞赛是阿里云天池平台上首场研究型竞赛，吸引了全国高校和科研院所师生的广泛关注和参与。

本次竞赛的主要目标是从时域天文大数据中发现微引力透镜和恒星耀发候选体这两种短时标稀有天体光变事件，完成从光变曲线（时序数据）中发现稀有异常子序列模式的计算任务。竞赛的数据来源于中国科学院国家天文台地基广角相机阵 GWAC 所采集的真实时域天文数据，数据总量近 170 万条（其中初赛约 76 万，复赛增加约 93 万），观察时间跨度为 6 个月，能够充分支持竞赛环境。

竞赛由 ACM SIGSPATIAL 中国分会主办，中国人民大学、中国科学院国家天文台、中国科学院计算机网络信息中心承办，国家天文科学数据中心、中国科技云协办。中国人民大学的孟小峰教授、国家天文台的魏建彦研究员、中科院计算机网络信息中心的廖方宇研究员等专家和学者担任大赛的评委和指导专家。



1. 赛题背景

在广袤的银河系中，真的存在“流浪行星”吗？这些“流浪行星”本身不发光，又不绕着某一恒星转动，我们又如何发现它们呢？正因为这样，我们对于“流浪行星”知之甚少，越来越先进的现代时域天文巡天大数据，借助于先进的计算机数据智能技术，用一种称为微引力透镜的天文理论，使得更多更准确地发现这类稀有小概率天文事件成为了可能。

由于行星本身不会发光，同时质量又小，除太阳系内的行星，系外行星的直接观测非常困难。现代天文技术的发展，特别是时域天文观测技术的长足进步，越来越多的系外行星被发现。近年来，有类似木星质量的“流浪行星”被发现，刷新了人们的认知，理论预言在银河系中存在着大量的“流浪行星”，但是发现它们仍是瓶颈，一是能形成微引力透镜事件，并能被我们观测到是稀有事件，二是越小质量的“流浪行星”形成微引力透镜事件的光变时标越短。因此，从观测上就要求我们具有高时间分辨率和大天区覆盖，即能得到一批高时间分辨的时域天文观测大数据，从而提高这一稀有事件的发现效率。而计算机数据智能却是大数据中发现这一有趣天文现象的不二选择。

中国科学院国家天文台的地基光学广角相机阵（GWAC，Ground-based Wide Angle Camera）是为中法天文卫星项目（SVOM）专门建设的时域天文巡天设备，该设备能每 15 秒钟得到一个采样观测数据，并得益于其大视场覆盖能力，至今已获得具有 15 秒采样分辨率的数百万条光变样本。这为发现时标为小时量级的类地球质量的“流浪地球”目标提供了可

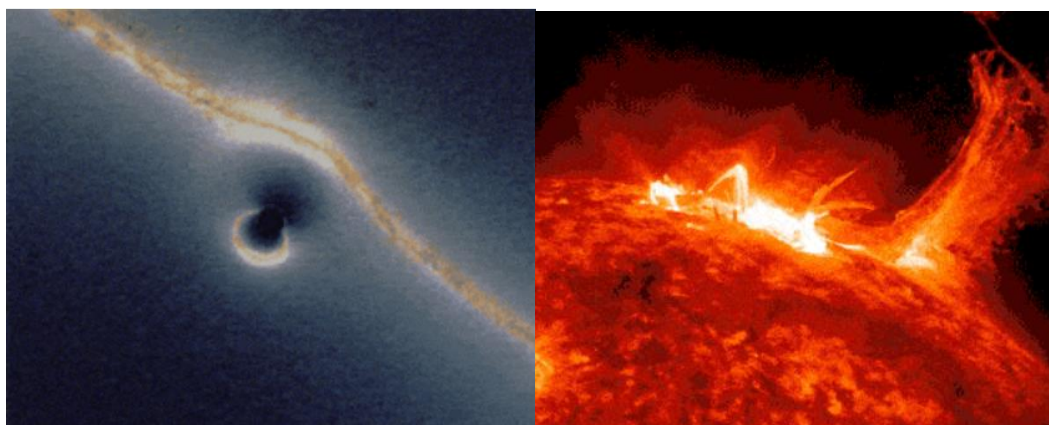
能的数据基础。

如何从这些大数据的样本中，搜索出稀有的观测目标事件，传统的人工+计算机搜索已无能为力。近年来发展并盛行起来的计算机数据智能技术正是从大数据中发现这一稀有事件的利器。

在我们 GWAC 的这些时域天文巡天大数据中包含有光度变化强烈，时标短的耀星目标，这类目标，也会通过我们的计算机数据智能得到发现。同时，还有一些未知的天文现象可能被我们的数据智能技术搜索和发现，那将对我们人类的认知打开一扇窗。期待每一位热心于天文科学探索的数据智能专家，用数据智能的方法分析我们提供的 GWAC 时域天文巡天大数据样本，获取有趣的科学发现。

2. 赛题题目

主要目标：从时域天文大数据中发现微引力透镜和恒星耀发候选体这两种短时标稀有天体光变事件，其计算任务为：从光变曲线（时序数据）中发现稀有异常子序列模式。



微引力透镜

恒星耀发

本次竞赛前期工作得到国家重点研发计划“科学大数据管理系统（2016YFB1000600）”的资助。该项目针对天文大数据、微生物大数据和高能物理大数据的管理关键技术进行研究，并致力于实现不同科学领域的大数据管理系统以加速科学发现。目前，研发的系统已成功应用于相关科学领域真实场景，并产出科学成果。

3. 数据介绍

本次竞赛的数据来源于中国科学院国家天文台的地基光学广角相机阵 GWAC，目前已获得高时间采样率的数百万条光变曲线样本，这为发现短时标稀有天体光变事件提供了数据基础。

初赛数据集共有光变曲线 76 万余条，观测时间跨度为 6 个月，光变曲线连续部分的时间采样率为 15 秒 1 个数据点。



GWAC 观测阵列

4. 奖项设置

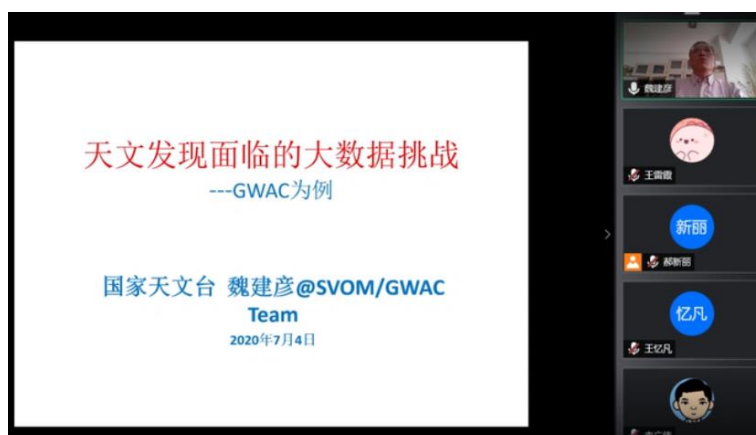
本次大赛开创了研究性大赛的先河，为鼓励激发广大师生关注和参与的热情，主办方设置了丰厚的奖项和奖励，大赛最终评审产生冠军队伍 1 支，奖金 1 万元并颁发证书、亚军队伍 2 支，奖金 0.5 万元并颁发证书、季军队伍 3 支，奖金 0.3 万元并颁发证书、优秀奖 10 支队伍，奖金 0.1 万元并颁发证书、积极参与奖 20 支队伍，纪念品并颁发证书。

奖项	奖励
冠军：1 支队伍	奖金 1 万，颁发证书
亚军：2 支队伍	奖金 0.5 万，颁发证书
季军：3 支队伍	奖金 0.3 万，颁发证书
优秀奖：10 支队伍	奖金 0.1 万，颁发证书
积极参与奖：20 支队伍	纪念品，颁发证书

5. 科学数据智能发现大赛系列讲座

2020 年 7 月 4 日，孟小峰教授主持了题为“天文发现面临的大数据挑战——以 GWAC 项目为例”的主题报告，讲座人为国家天文台的魏建彦研究员。本次讲座主要以 GWAC 项目为例，具体分析从天文大数据中获得科学发现的挑战所在，并解析本次比赛的科学意义和国际背景，帮助大家理解科学数据发现的真正意义。

本次讲座主要以 GWAC 项目（本次比赛的数据来源）为例，具体分析从天文大数据中获得科学发现的挑战所在，解析本次比赛的科学意义和国际背景，帮助大家理解科学数据智能发现大赛在天文学上的重要意义，以期选手提供更多比赛思路。此外，魏老师还为大家讲述此次大赛天文大数据背后的故事。讲座不仅获得了参赛选手的一直好评，还产生了广泛的社会影响力，给公众科普了一定的天文学知识。讲座举办期间，哔哩哔哩、腾讯会议、钉钉三个直播平台的总观看人数约为 751，其中哔哩哔哩观看峰值达到 729 人，大于比赛的报名人数。



科学数据智能发现大赛系列讲座

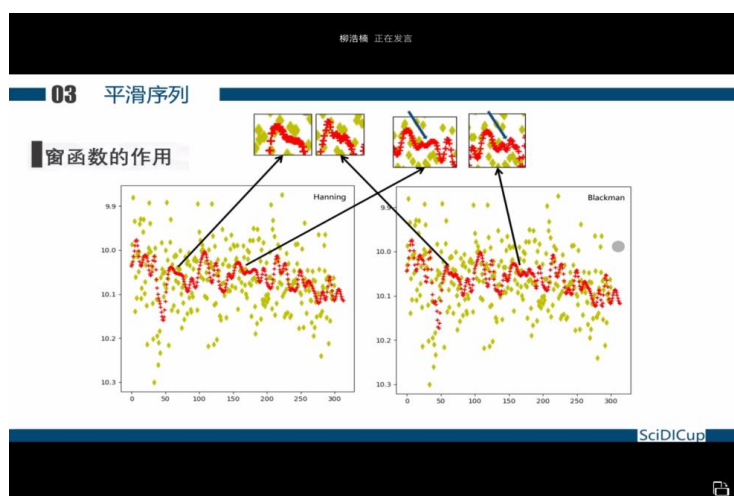
6. 科学数据智能发现大赛复赛

比赛自 2020 年 8 月 1 日正式启动以来，共有来自全国 317 支队伍 的 362 人报名参赛。参赛队伍分别来自中国人民大学、清华大学、北京大学、中科院、浙江大学、武汉大学、北京师范大学、南京大学、太原理工、陆军工程大学等 40 余所高校和科研院所。经过初赛比拼，共有 45 支队伍进入复赛。复赛的任务较初赛更具挑战性，采取机器和人工评判相结合的方式给出相应的得分及排名，最终排名靠前的 6 支队伍通过现场答辩决出冠亚季军。

7. 科学数据智能发现大赛总决赛

与初赛相比复赛的任务更具有挑战性，因为选手们要发现“主办方未知的稀有光变现象”。也因为如此，复赛我们采取机器和人工评判结合的方法给出相应的得分和排名，最终排名靠前的 6 支队伍在决赛阶段角逐冠亚季军。他们分别是“TYUTDMG”、“Thramos”、“ccjaread”、“水林淼淼”、“bnu_314”、“星河欲转千帆舞”。

2020 年 11 月 5 日，经过决赛现场答辩决出了各项奖项。决赛的总结和期间相关精彩截图如下：



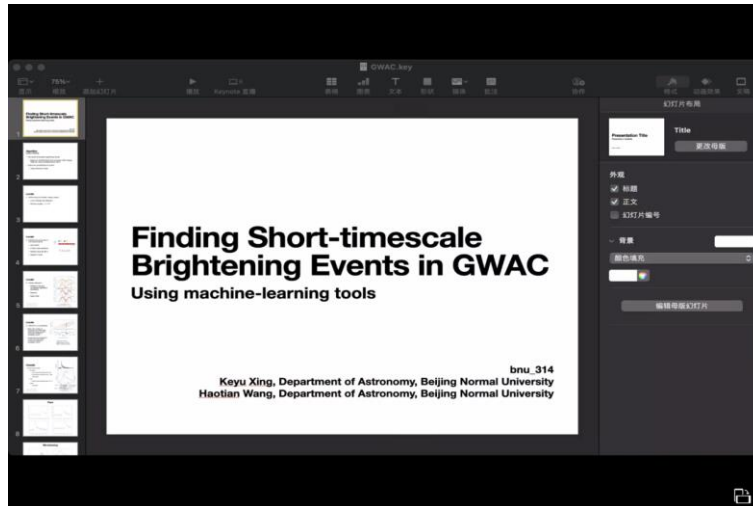
太原理工团队答辩阐述

3

- 

第三方 Python 包	
pandas	0.25.3
numpy	1.17.4+mk1
matplotlib	3.2.0rc2
scikit-learn	0.22
scipy	1.4.1
multiprocess	0.70.9
tensorflow-cpu	2.1.0rc1
Keras	2.3.1 ,tensorflow自带

复旦大学程吉安选手自我介绍



北京师范大学团队答辩阐述

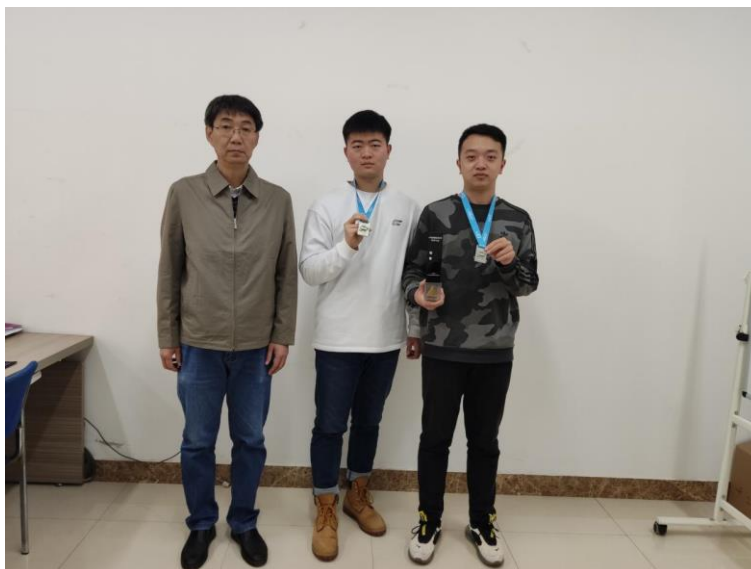
队伍名	最终排名	获奖等级
TYUTDMG	1	冠军
ccjaread	2	亚军
Thramos	3	亚军
水林淼淼	4	季军
bnu_314	5	季军
星河欲转千帆舞	6	季军

99



8. 优秀选手集锦

(1) TYUTDMG 团队



团队名称：TYUTDMG

团队成员：张旭康（太原理工大学信息与计算机学院 2017 级硕士生，导师牛保宁(图左一)），李显、柳浩楠。（太原理工大学信息与计算机学院 2019 级硕士生，分别为图左二、左三，导师牛保宁）

采用方法：对数据集采用滑动窗口的方式进行初步筛选，过滤掉部分不存在异常模式子序列的数据文件，对经过过滤的数据文件进行预处理之后，综合采用 FastDTW 算法以及基于形状的距离（SBD）进行筛选。

获奖感言：感谢各位老师以及相关工作人员，提供了如此好的平台让我们在参与比赛的同时了解到天文发现的奥妙。

(2) 星河欲转千帆舞团队



团队名称：星河欲转千帆舞

团队成员：方成龙，王撷阳，冒艳纯（南京航空航天大学，导师：许建秋）

采用方法：使用基于均值和中位数以及滑动窗口的方法来提取数据中可疑的光变事件，对可疑点的左右两边时间之和判断是否满足最小数据点的情况，通过时间之差初步判断光变事件的类型。继而通过 DTW 匹配的方法来精确判断是否为光变事件。

获奖感言：感谢主办方提供一个的这个高水平平台让我们锻炼自己。

(3) Thramos 团队



团队名称：Thramos

团队成员：翟延伟（大连理工大学）

采用方法：本次比赛通过定位、分割、特征生成、识别四步走实现光变事件的解析化分析与发现。首先，利用动态线性拟合，扩展的对时序数据进行线性化表示，找出极小值点，定位异常子序列；其次，在极小值点附近利用统计特征分割出完整的异常子序列；接着从不

同角度进行数据特征生成；最后基于数据特征实现异常子序列的分类，即光变事件的发现。

获奖感言：感谢主办方的精心组织，给了我学习交流的机会！感谢各位老师的细心指导，受益良多！

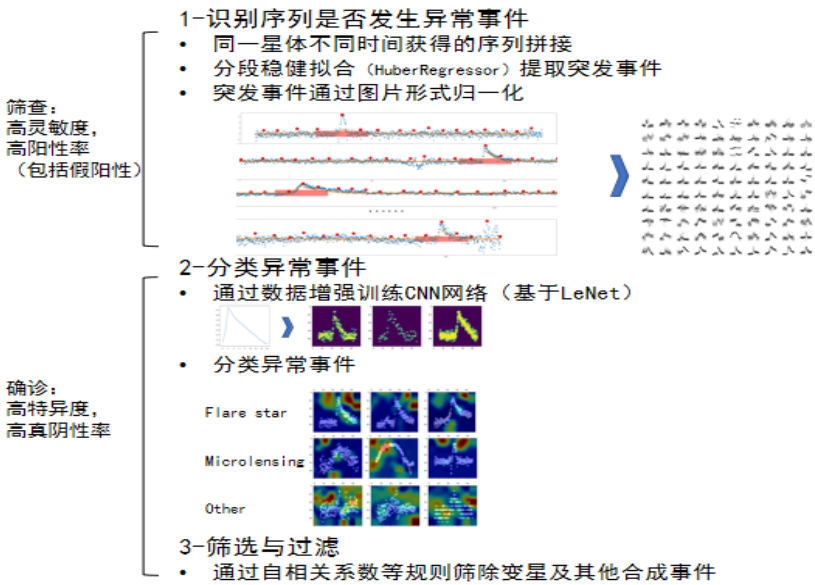
(4) ccjaread 团队



团队名称：ccjaread

团队成员：程吉安（复旦大学医学博士）

采用方法：



获奖感言：不同学科间的碰撞让我学到了很多，还希望能够参与这样的竞赛。

(5) 水林淼淼团队



团队名称：水林淼淼

团队成员：张文桂（华南理工大学模式识别与智能系统硕士）、孙瑞锦（计算机专业博士）

创新点：

- 1.自适应算法：创新的类随机森林算法，自动训练和识别。
- 2.模板优化：形状优化、长度优化、边缘优化。
- 3.训练数据分级：依据与模板相关系数及单峰、双峰。
- 4.伸缩变换：类似小波变换进行下采样。
- 5.兼顾全局与局部：采用全局相关系数和局部相关系数的调和平均数。
- 6.精准率高：通过方差约束，过零等统计特征进一步提升精准率。

获奖感言：参加比赛，受益良多。希望未来能够继续提升自己。

学术交流

一、学术活动任职

Prof. Xiaofeng Meng:

1、期刊任职

《Journal of Computer Science and Technology》编委

《Frontiers of Computer Science》编委

《计算机研究与发展》编委

《计算机科学》编委

《计算机科学与探索》编委

《信息安全研究》编委

《中国科学数据》编委

2、学术机构任职

中国计算机学会会士

ACM SIGSPATIAL 中国分会主席

中国人工智能学会社会计算与社会智能专业委员会主任

中国保密协会隐私保护专业委员会副主任

数字出版技术国家重点实验室学术委员会委员

贵州省公共大数据重点实验室学术委员会委员

中国科学院 A 类战略性先导科学专项“地球大数据科学工程”专家委员会成员

中央军委科学技术委员会联合作战实验技术国防科技专业专家组

联想研究院专家委员会委员

二、学术交流

2020.3.30-2020.4.02

博士生杨晨、硕士生杜永杰参加第 23 届扩展数据库技术国际会议(EDBT 2020)

2020 年 3 月 30 日至 4 月 2 日, 第 23 届 International Conference on Extending Database Technology(EDBT 2020)在丹麦哥本哈根举办。EDBT 是国际数据库界交流数据管理方面最新研究成果的国际学术会议。每年在欧洲召开, 会议为数据库科学家、企事业专家、数据库研发者和应用者提供学术交流的机会, 共同探讨数据管理的新思想、新技术、新工具和计算机前沿科学。是中国计算机学会(CCF)推荐国际学术会议中的 B 类会议。

本实验室博士生杨晨、硕士生杜永杰的论文“Micro Analysis to Enable Energy-Efficient Database Systems”被大会录用, 并在大会期间对该论文的研究成果进行了论文报告, 介绍了一种新的运行框架, 从而避免以往运行大型数据库系统时以性能换取能源效率的方法, 即在不损失 CPU 性能的前提下降低数据库系统的能耗是可行的。并对相关实验结果进行了展示、分析与交流。

2020.5.08-2020.5.09

孟小峰教授参加 2020 ACM SIGSPATIAL 中国空间数据智能学术会议 (SpatialDI2020)

2020 ACM SIGSPATIAL 中国空间数据智能学术会议 (SpatialDI 2020) 网络会议于 2020 年 5 月 8-9 日成功举办。会议邀请了国内外空间数据研究领域的知名华人学者以及来自百度、阿里、滴滴、华为、京东等互联网企业的代表在空间数据智能获取、管理、分析、应用等方面进行了专题研讨。各位学者和专家就高精度定位、城市众包多源感知、海量数据存储等前沿问题展开了深入的探讨。

同时, 在新冠疫情这个特殊时期, 会议还专门筹划了“空间大数据战疫”专题。特邀嘉宾就中、美、世界疫情, 针对人群移动分析、新冠肺炎传染扩散建模、防控以及可视化和决策等方面做了精彩的汇报, 显示了空间大数据在疫情分析中的重要作用。网络会议是疫情期间会议举办的主要形式, 今年中国空间智能学术年会在会上成功举办也体现了互联网作为主要的信息传输手段对科学交流支撑的重要。



2020.7.04-2020.11.05

孟小峰教授主办科学数据智能发现大赛（SciDI Cup）

为了寻找广袤银河中的“流浪地球”，我们利用现代时域天文大数据技术和天文知识理论开展了这样一场有关于“短时标稀有天体光变发现”的比赛。赛事的目的和初衷是期望来自各个不同领域的参赛选手，能够发挥各自的特长和优势，提出相关智能算法从时域天文大数据发现微引力透镜和恒星要发候选体这两种稀有的短时标光变事件。即我们可以从光变产生的时序数据中发现稀有的异常子序列。赛事所采用的所有数据来自中国科学院国家天文台地基广角相机阵 GWAC 所采集的真实时域天文数据。数据总量近 170 万条（其中初赛约 76 万，复赛增加约 93 万），观察时间跨度为 6 个月，能够充分支持竞赛环境。本次赛事由 ACM SIGSPATIAL 中国分会主办，中国人民大学、中国科学院国家天文台等单位承办。由孟小峰（中国人民大学）、魏建彦（国家天文台）、廖方宇（中科院计算机网络信息中心）等专家和学者担任大赛的评委和指导专家。



2020 年 7 月 4 日孟小峰教授主持了题为“天文发现面临的大数据挑战——以 GWAC 项目为例”的主题报告，讲座人为国家天文台魏建彦研究院。本次讲座主要以 GWAC 项目为例，具体分析从天文大数据中获得科学发现的挑战所在，并解析本次比赛的科学意义和国际背景。帮助大家理解科学数据发现的真正意义。

比赛 2020 年 8 月 1 日正式启动以来，共有来自全国 317 只队伍的 362 人参加比赛。他们分别来自中国人民大学、清华大学、北京大学、中科院、浙江大学、武汉大学、北京师范大学、南京大学、太原理工、陆军工程大学等 40 余所高校和科研院所。经过初赛比拼，共有 45 只队伍进入复赛。与初赛相比复赛的任务更具有挑战性，因为选手们要发现“主办方未知的稀有光变现象”。也因为如此，复赛我们采取机器和人工评判结合的方法给出相应的得分和排名，最终得分靠前的 6 只队伍将赛阶段通过现场答辩决出冠亚季军。

2020 年 11 月 5 日，比赛进入决赛答辩阶段。经过选手们决赛现场答辩和各位评审专家的打分，最终决出了以下奖项：

队伍名	最终排名	获奖等级
TYUTDMG	1	冠军
ccjaread	2	亚军
Thramos	3	亚军
水林淼淼	4	季军
bnu_314	5	季军
星河欲转千帆舞	6	季军

2020.8.22-2020.8.23

孟小峰教授主办 BDSC2020 第五届全国大数据与社会计算学术会议

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!

本次会议以“社会计算与社会智能”为主题,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。会议共包含 4 个大会报告,8 个专题,由来自社会学、管理学、经济学、复杂性科学、传播学、数字人文、计算机科学等多个学科专家学者,以及政府、企业等领域相关研究人员共计 44 位与大家分享了各最新成果。

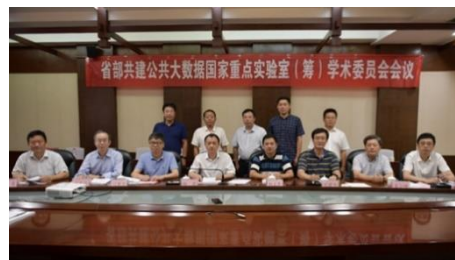
2020.9.4

孟小峰教授参加贵州大学省部共建公共大数据国家重点实验室（筹）召开第一届学术委员会第四次会议

贵州大学省部共建公共大数据国家重点实验室（筹）召开第一届学术委员会第四次会议于 2020 年 9 月 4 日在贵州省贵阳市花溪迎宾馆召开。中国人民大学孟小峰教授受邀以公共大数据国家重点实验室（筹）学术委员会委员身份参加学术委员会,出席学术委员会的还有贵州省科技厅厅长廖飞、副厅长林浩,我校党委书记李建军、副校长李军旗,北京理工大学梅宏院士、北京邮电大学杨义先教授、上海交通大学过敏意教授、中国科学院林东岱研究员、浙江大学陈刚教授、北京航空航天大学刘建伟教授、哈尔滨工业大学徐勇教授等。

贵州大学“公共大数据国家重点实验室”（State Key Laboratory of Public Big Data）（以下简称“实验室”）按照科技部对省部共建国家重点实验室的要求,针对公共大数据,特别是政府数据的开放、共享和应用中的痛点,分别从人工智能、网络安全和公共管理的角度,聚焦公共大数据融合与集成、公共大数据安全与隐私保护、块数据与区域治理三个研究方向的

基础研究、应用基础研究和地方产业服务。实验室整合全省科研力量,充分借助外部智力资源,瞄准公共大数据的“聚、通、用”需求,着力突破公共大数据应用中的关键共性技术问题,构建公共大数据“融合-安全-治理-应用”于一体交叉研发体系,组建高水平研究团队,并以贵州的国家大数据综合试验区为试验基地,推进公共大数据在治理领域的创新应用,实现大数据服务地方的功能并形成特色。



此次学术委员会第四次会议明确了大数据已上升为国家战略,数据已成为新型生产要素,实验室将在服务国家尤其是贵州省大数据产业发展中发挥核心的智力支持与科技创新助推器作用。李少波教授、梅宏教授、杨义先教授、孟小峰教授等委员对方案表示肯定,并对实验室建设提出了宝贵建议。

2020.9.19

孟小峰教授参加 2020 平行智能大会平行管理与社会计算研讨会会议

2020 年 9 月 19 日,由中国自动化学会,中国科学院自动化研究所复杂系统管理与控制国家重点实验室,国际智能科学与技术学会等主办,中国管理现代化研究会平行管理专委会,中国人工智能学会社会计算与社会智能专业委员会承办的“2020 平行管理与社会计算研讨会”在线上召开,本次会议和“2019 平行智能大会”同期举办。

此次研讨会以“平行管理助力社会计算”为主题,由中国人工智能学会社会计算与社会智能专委会主任、ACM SIGSPATIAL 中国分会主席、中国人民大学信息学院杰出学者及特聘教授 孟小峰担任论坛主席,中国人民大学教授、青岛智能产业技术研究院区块链技术研究中心主任 袁勇和洛阳师范学院信息技术学院副教授 马友忠担任共同主席。旨在构建一个自由交流的平台,为平行管理与社会计算的相结合提出更多的可能性。

来自国防科技大学、北京航空航天大学、哈尔滨工业大学、中南大学、华南师范大学、中国石油大学、北京工业大学等多所国内外知名院校、科研院所的多名专家以前沿的科学视角分享了社会计算跨学科领域的思考与洞见,并与众多的相关领域的学界精英及产业引领者共同搭建起了领域内从业者交流合作的桥梁。

2020.9.23-2020.9.25

孟小峰教授应邀参加第十七届中国信息系统及应用大会并作特邀报告

9月23日至25日，由中国计算机学会(CCF)主办、CCF信息系统专业委员会、广州大学和贵州大学共同承办的第十七届中国信息系统及应用大会(WISA 2020)在广州召开。会议本次大会采用线上线下结合的形式，来自全国各地 109 所高校、科研院所、企业的 300 余位代表参加会议，1000 余人线上同步观看大会直播。



本次大会围绕“人工智能与信息系统”主题，关注信息系统新兴应用领域，特别是人工智能与信息系统融合发展领域，聚焦关键技术难题，搭建学术、企业、政府三方参与的交流与合作平台。本次大会共收到论文投稿 165 篇，录取英文长文 42 篇、短文 16 篇，英文论文在 Springer LNCS 论文集发表，长文的录用率为 25.45%；大会设立了 14 个论文分组报告，主题包括大数据与数据挖掘、区块链与隐私安全、边缘计算与数据融合、机器学习、智能处理与决策、自然语言处理、推荐系统、智慧教育与智能决策等当前信息系统的热点领域。各个论文报告以线上线下结合的形式开展，论文作者对各自的研究工作进行了报告和交流。

其中，中国人民大学孟小峰教授应邀作了题为“中国特色的数据治理理论与实践”的大会特邀报告，针对大数据 2.0 时代在新一代信息系统构建过程中的数据治理核心问题，提出一种新的认识技术变革的方法即数据发展观，进而揭示数据治理的本质和基于数据透明的解决途径。

2020.9.25

孟小峰教授团队在人民论坛发布有关数据垄断及其治理模式的研究成果

近日,人民论坛发布了信息学院孟小峰教授有关数据垄断及其治理模式的研究成果。

作为中央主流媒体、重点党刊、思想理论传播重要平台,《人民论坛》全方位集结思想动态、深层次研判政策时局、多视角解析热点难点,互动传播名家大家和实践一线官员的精品力作与前沿思考;被读者誉为具有国际影响力的“中国第一政论期刊”,转载率、引用率、影响力、关注度名列同类期刊前茅,反响巨大。当前,《人民论坛》已成长国内领先的高端思想理论传播平台。

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。

孟小峰教授团队基于 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。孟小峰教授首先以当前数据收集者们的数据获取量为依据，分析了数据垄断的成因。然后提出了三种数据治理模式，以缓解数据垄断形势、促进数据安全与公平的共享流通。最后，孟小峰教授指出：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。

2020.10.23

孟小峰教授参加“CCF 走进高校”系列演讲公益活动

为深入学习贯彻习近平总书记重要指示和第三次中央新疆工作座谈会精神，在中国计算机学会和燕山大学的大力支持、帮助下，10 月 23 日，中国计算机学会(CCF)组织的“CCF 走进高校”系列演讲公益活动在新疆科技学院学术报告厅举行。CCF 信息系统专委孟小峰教授、王鑫教授和徐宝文教授作学术报告。



中国人民大学孟小峰教授就《大数据智能时代的人才培养》专题进行宣讲，报告围绕数据发展趋势、数据基础设施两个方面展开，以数据库背景出发，提出一种新的认识技术变革的方法即数据发展观，进而揭示数据发展的本质和对人才培养的影响，并探索新的人才培养发展路径。在报告的结尾，孟小峰教授总结了当下大数据智能时代的人才培养几个方面提出要求，培养具有数据思维的大数据人工智能人才、培养具有大数据人工智能技术的专业技术型人才、培养具有丰富的跨学科知识的人才，并对参会的同学们提出了殷切的希望。

2020.11.6

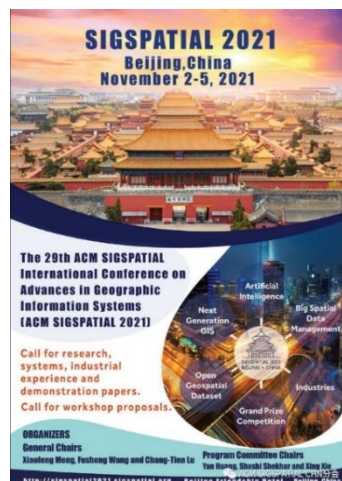
孟小峰教授参加第 28 届 ACM SIGSPATIAL 2020 国际会议

2020 年 11 月 6 日，第 28 届 ACM SIGSPATIAL 2020 国际会议以在线形式在美国西雅图落下帷幕。会议宣布下届第 29 届 ACM SIGSPATIAL 2021 国际会议将于 2021 年 11 月 2-5 日在中国北京举办。ACM SIGSPATIAL 会议被公认为空间数据管理和 GIS 领域

多学科交叉的顶级学术会议。会议在北美已成功举办了二十八届（1993-2020 年），旨在聚集空间数据和地理信息领域的研究人员、开发人员、用户和从业者，促进 GIS 全方位的跨学科讨论和研究。ACM SIGSPATIAL 2021 将是本会议历史上首次在北美以外地区举办，尤其令人期待！



ACM SIGSPATIAL 中国分会主席、中国人民大学孟小峰、美国石溪大学汪富生教授，美国弗吉尼亚理工大学吕昌田教授在星期五的“Statistics and Awards”环节上进行了 ACM SIGSPATIAL 2021 会前汇报，为正式会议的举办开展积极的宣传并做足了充分准备。ACM SIGSPATIAL 2021 由中国人民大学孟小峰、美国石溪大学汪富生教授，美国弗吉尼亚理工大学吕昌田教授担任大会主席，美国北德州大学黄艳、明尼苏达大学 Shashi Shekhar、微软亚洲研究中心谢幸担任程序委员会主席。会议主题涵盖了空间智能、空间大数据、GIS、普适计算、空间搜索、空间系统等研究方向，具有重要的学术价值与应用价值，欢迎相关领域的学者和企业积极投递论文。



此次会议还包括主会场，专题讨论，和由滴滴赞助的两万五千美金的 GIS Cup 等在内的学术及周边活动。会议亮点纷呈，既聚焦 Artificial Intelligence、Big Spatial Data Management 等主题，也开展 Industries panel、Grand Prize Competition、Open Geospatial Dataset 等与工业密切相关的活动，同时对 Next Generation GIS 等方面进行前瞻性研讨。

2020.11.6-2020.11.8

孟小峰教授应邀参加 2020 第五届中国数据安全与隐私保护大会

2020 年 11 月 6 日-8 日，由西安安盟智能科技有限公司协办的“2020 第五届中国数据安全与隐私保护大会”在西安水晶岛酒店召开，本次会议由中国保密协会隐私保护专业委员会主办，西安邮电大学承办。



近年来随着大数据技术在各领域的广泛使用，公民个人隐私的边界也频频遭遇挑战。在《中华人民共和国网络安全法》和《中华人民共和国密码法》正式实施的背景下，为了应对数据滥用、数据窃取、隐私泄露等日益增多的数据安全安全问题，本次大会特别邀请了美国罗格斯大学 Jaideep Vaidya 教授、上海交通大学来学嘉教授、中国人民大学孟小峰教授、中国科学院大学荆继武教授、新泽西理工大学唐强教授、西安电子科技大学胡予濮教授、山东大学成秀珍教授、浙江大学纪守领教授、中国地质大学朱天清教授、香港大学 Man Ho Allen Au 博士、中山大学张方国教授、杭州师范大学陈可非教授、西安电子科技大学陈晓峰教授、陕西师范大学俞勇教授等国内外知名专家学者参会，共同探讨数据隐私保护发展现状，以及所面临的关键性挑战问题和研究方向。

其中，中国人民大学孟小峰教授作了题为“大数据智能时代数据治理理论与实践”的学术报告。报告指出当前中国特色数据治理当是新一代信息系统构建的核心问题，对数据作为生产要素发挥作用亦十分关键。围绕数据治理问题，报告介绍了数据治理的理论概念、方法体系，并落地到具体实践。最后，报告表示如何通过数据智能理解数据，经由数据治理产生对数据的敬畏之心十分关键。整场报告激情澎湃，听众反响热烈。

2020.11.15

孟小峰教授应邀参加首届数字经济与人口发展研讨会

2020年11月15-16日，由工业和信息化部中国信息通信研究院、赣州市人民政府、国科技部新一代人工智能发展研究中心、国家卫生健康委中国人口与发展研究中心、赣州市大数据管理局联合举办的《首届数字经济与人口发展研讨会》在江西赣州成功召开。来自北京大学、清华大学、中国人民大学、南开大学、中山大学、国务院发展研究中心、中国科学院、中国社科院、科技部新一代人工智能发展研究中心、工信部中国信息通信研究院、工信部赛迪研究院等高校和科研院所专家学者以及科技信息领域知名企业代表应邀参加，并围绕数字经济、数字治理、智能养老、数字人口学、数字医学、数字育儿等主题发表主旨演讲和专题研讨。



其中，中国人工智能学会社会计算与社会智能专业委员会主任中国人民大学博士生导师孟小峰发表演讲数据治理的问题与挑战。孟小峰教授在报告中指出，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。

2019.12.1-2020.11.30

博士生刘俊旭赴美国埃默里大学进行为期一年的学术交流

在中国人民大学孟小峰教授的推荐下，WAMDM 实验室博士生刘俊旭成功申请国家留学基金委公派联合培养博士项目并于 2019 年 12 月 1 日至 2020 年 11 月 30 日赴美国埃默里大学 Li Xiong 教授团队开展为期一年的学术交流。

交流期间，刘俊旭围绕隐私保护的机器学习大方向，重点针对个性化隐私保护的联邦学习框架展开研究，旨在解决现有研究工作大多只能为用户提供统一隐私保护的不足，并提出 Pepsi 框架，实现了既能满足用户特定隐私需求，同时保证了模型可用性。同时，刘俊旭参与 Li Xiong 教授团队关于数据交易市场的研究课题，该研究提出一种基于模型定价的数据交易框架 Dealer，旨在分析量化数据价值，设计模型定价机制，从而在满足经济学原理的要求下使各方取得最大收益。

近年来，WAMDM 实验室在隐私保护研究领域取得了引人注目的成绩，并与很多国内外研究团队保持密切的交流与合作。此次访问进一步加深了中国人民大学 WAMDM 实验室与埃默里大学 AIMS 实验室的合作交流和真挚友谊，也为实验室未来的国际合作提供了宝贵的经验。



2020.12.14-2020.12.15

孟小峰教授担任第二届社会计算国际会议的大会共同主席

第二届社会计算国际会议（International Conference of Social Computing）于 2020 年 12 月 14 日至 2020 年 12 月 15 日以 Zoom 会议+B 站直播的形式线上召开。会议旨在促进信息科学、社会学、管理学、经济学、金融学、传播学、政治学、地理学等多学科的对话与创新。大会邀请了来自美国、德国和中国知名学者在大数据与分析、金融科技、公共卫生与社会计算等交叉领域进行主题报告。



孟小峰教授应邀担任大会共同主席并致辞。首先，孟小峰教授在开场致辞中阐述了由理论驱动的社会科学向由数据驱动的社会科学的转型相关背景和问题，简述了当前国际社会科学转型的方向、方法和路径等相关研究状况，并对未来人工智能与社会计算在社会发展与治理的应用、跨学科领域最新的突破性研究发展、新的学术思想和方法交流等作了前景展望。

附录

实验室研讨会

2020.3.13 腾讯会议	
郝新丽 (Cloud Group)	<p>报告题目：BRITS: Bidirectional Recurrent Imputation for Time Series</p> <p>报告摘要：</p> <p>Time series are widely used as signals in many classification/regression tasks. It is ubiquitous that time series contains many missing values. Given multiple correlated time series data, how to fill in missing values and to predict their class labels? Existing imputation methods often impose strong assumptions of the underlying data generating process, such as linear dynamics in the state space. In this paper, we propose BRITS, a novel method based on recurrent neural networks for missing value imputation in time series data. Our proposed method directly learns the missing values in a bidirectional recurrent dynamical system, without any specific assumption. The imputed values are treated as variables of RNN graph and can be effectively updated during the backpropagation. BRITS has three advantages: (a) it can handle multiple correlated missing values in time series; (b) it generalizes to time series with nonlinear dynamics underlying; (c) it provides a data-driven imputation procedure and applies to general settings with missing data.</p>
2020.7.3 腾讯会议	
郝新丽 (Cloud Group)	<p>报告题目：Robust and Rapid Clustering of KPIs for Large-Scale Anomaly Detection</p> <p>报告摘要：For large Internet companies, it is very important to monitor a large number of KPIs (Key Performance Indicators) and detect anomalies to ensure the service quality and reliability. However, large-scale anomaly detection on millions of KPIs is very challenging due to the large overhead of model selection, parameter tuning, model training, or labeling. In this paper we argue that KPI clustering can help: we can cluster millions of KPIs into a small number of clusters and then select and train model on a per-cluster basis. However, KPI clustering faces new challenges that are not present in classic time series clustering: KPIs are typically much longer than other time series, and noises, anomalies, phase shifts and amplitude differences often change the shape of KPIs and mislead the clustering algorithm.</p>
2020.10.13 FL1, Wing Building for Science Complex	
但唐朋 (Cloud Group)	<p>报告题目：Spatial Temporal Trajectory Similarity Join</p> <p>报告摘要：Existing works only focus on spatial dimension without the consideration of combining spatial and temporal dimensions together when processing trajectory similarity join queries, to address this problem, this paper proposes a novel two-level grid index which takes both spatial and temporal information into account when processing spatial-temporal trajectory similarity</p>

	join. A new similarity function MOGS is developed to measure the similarity in an efficient manner when our candidate trajectories have high coverage rate CR. Extensive experiments are conducted to verify the efficiency of our solution.
2020.10.20 FL1, Wing Building for Science Complex	
但唐朋 (Cloud Group)	<p>报告题目: Searching Activity Trajectories by Exemplar</p> <p>报告摘要: The rapid explosion of urban cities has modernized the residents' lives and generated a large amount of data (e.g., human mobility data, traffic data, and geographical data), especially the activity trajectory data that contains spatial and temporal as well as activity information. With these data, urban computing enables to provide better services such as location-based applications for smart cities. Recently, a novel exemplar query paradigm becomes popular that considers a user query as an example of the data of interest, which plays an important role in dealing with the information deluge. In this article, we propose a novel query, called searching activity trajectory by exemplar, where, given an exemplar trajectory τ_q, the goal is to find the top-k trajectories with the smallest distances to τ_q. We first introduce an inverted-index-based algorithm (ILA) using threshold ranking strategy. To further improve the efficiency, we propose a gridtree threshold approach (GTA) to quickly locate candidates and prune unnecessary trajectories. In addition, we extend GTA to support parallel processing. Finally, extensive experiments verify the high efficiency and scalability of the proposed algorithms.</p>
2020.11.3 FL1, Wing Building for Science Complex	
彭迎涛 (Web Group)	<p>报告题目: RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems</p> <p>报告摘要: To address the sparsity and cold start problem of collaborative filtering, researchers usually make use of side information, such as social networks or item attributes, to improve recommendation performance. This paper considers the knowledge graph as the source of side information. To address the limitations of existing embedding-based and path-based methods for knowledge-graph-aware recommendation, we propose RippleNet, an end-to-end framework that naturally incorporates the knowledge graph into recommender systems. Similar to actual ripples propagating on the water, RippleNet stimulates the propagation of user preferences over the set of knowledge entities by automatically and iteratively extending a user's potential interests along links in the knowledge graph. The multiple "ripples" activated by a user's historically clicked items are thus superposed to form the preference distribution of the user with respect to a candidate item, which could be used for predicting the final clicking probability. Through extensive experiments on real-world datasets, we demonstrate that RippleNet achieves substantial gains in a variety of scenarios, including movie, book and news recommendation, over several state-of-the-art baselines.</p>
2020.11.10 FL1, Wing Building for Science Complex	

<p>郝新丽 (Cloud Group)</p>	<p>报告题目：Opprentice: Towards Practical and Automatic Anomaly Detection Through Machine Learning</p> <p>报告摘要：Closely monitoring service performance and detecting anomalies are critical for Internet-based services. However, even though dozens of anomaly detectors have been proposed over the years, deploying them to a given service remains a great challenge, requiring manually and iteratively tuning detector parameters and thresholds. This paper tackles this challenge through a novel approach based on supervised machine learning. With our proposed system, Opprentice (Operators' apprentice), operators' only manual work is to periodically label the anomalies in the performance data with a convenient tool. Multiple existing detectors are applied to the performance data in parallel to extract anomaly features. Then the features and the labels are used to train a random forest classifier to automatically select the appropriate detector-parameter combinations and the thresholds. For three different service KPIs in a top global search engine, Opprentice can automatically satisfy or approximate a reasonable accuracy preference ($\text{recall} \geq 0.66$ and $\text{precision} \geq 0.66$). More importantly, Opprentice allows operators to label data in only tens of minutes, while operators traditionally have to spend more than ten days selecting and tuning detectors, which may still turn out not to work in the end.</p>
<p>马超红 (Cloud Group)</p>	<p>报告题目：Learning Multi-dimensional Indexes</p> <p>报告摘要：Scanning and filtering over multi-dimensional tables are key operations in modern analytical database engines. To optimize the performance of these operations, databases often create clustered indexes over a single dimension or multidimensional indexes such as R-Trees, or use complex sort orders (e.g., Z-ordering). However, these schemes are often hard to tune and their performance is inconsistent across different datasets and queries. This paper introduce Flood, a multi-dimensional in-memory read-optimized index that automatically adapts itself to a particular dataset and workload by jointly optimizing the index structure and data storage layout. Flood a new multi-dimensional primary index that is jointly optimized using both the underlying data and query workloads.</p>
<p>但唐朋 (Cloud Group)</p>	<p>报告题目：Neural circuit policies enabling auditable autonomy</p> <p>报告摘要：A central goal of artificial intelligence in high-stakes decision-making applications is to design a single algorithm that simultaneously expresses generalizability by learning coherent representations of their world and interpretable explanations of its dynamics. Here, we combine brain-inspired neural computation principles and scalable deep learning architectures to design compact neural controllers for task-specific compartments of a full-stack autonomous vehicle control system. We discover that a single algorithm with 19 control neurons, connecting 32 encapsulated input features to outputs by 253 synapses, learns to map high-dimensional inputs into steering commands. This system shows superior generalizability, interpretability and robustness compared with orders-of-magnitude larger black-box learning systems. The obtained neural agents enable high-fidelity autonomy for task-specific parts of a complex autonomous system.</p>

2020.11.17 FL1, Wing Building for Science Complex	
郝新丽 (Cloud Group)	<p>报告题目：Active Model Selection for Positive Unlabeled Time Series Classification</p> <p>报告摘要：Positive unlabeled time series classification (PUTSC) refers to classifying time series with a set P of positive labeled examples and a set U of unlabeled ones. Model selection for PUTSC is a largely untouched topic. In this paper, we look into PUTSC model selection, which as far as we know is the first systematic study in this topic. Focusing on the widely adopted self-training one-nearest-neighbor (ST-1NN) paradigm, we propose a model selection framework based on active learning (AL). We present the novel concepts of self-training label propagation, pseudo label calibration principles and ultimately influence to fully exploit the mechanism of ST-1NN. Based on them, we develop an effective model performance evaluation strategy and three AL sampling strategies. Experiments on over 120 datasets and a case study in arrhythmia detection show that our methods can yield top performance in interactive environments, and can achieve near optimal results by querying very limited numbers of labels from the AL oracle.</p>
马超红 (Cloud Group)	<p>报告题目：DBOS: A Database-oriented operating system</p> <p>报告摘要：Current operating systems are complex systems that were designed before today's computing environments. This makes it difficult for them to meet the scalability, heterogeneity, availability, and security challenges in current cloud and parallel computing environments. To address these problems, this paper propose a radically new OS design based on data-centric architecture: all operating system state should be represented uniformly as database tables, and operations on this state should be made via queries from otherwise stateless tasks. This design makes it easy to scale and evolve the OS without whole-system refactoring, inspect and debug system state, upgrade components without downtime, manage decisions using machine learning, and implement sophisticated security features. Everything is table, every request is query.</p>
2020.11.24 FL1, Wing Building for Science Complex	
彭迎涛 (Web Group)	<p>报告题目：Knowledge Graph Convolutional Networks for Recommender Systems</p> <p>报告摘要：To alleviate sparsity and cold start problem of collaborative filtering based recommender systems, researchers and engineers usually collect attributes of users and items, and design delicate algorithms to exploit these additional information. In general, the attributes are not isolated but connected with each other, which forms a knowledge graph (KG). In this paper, we propose Knowledge Graph Convolutional Networks (KGCN), an end-to-end framework that captures inter-item relatedness effectively by mining their associated attributes on the KG. To automatically discover both high-order structure information and semantic information of the KG, we sample from the neighbors for each entity in the KG as their receptive field, then combine neighborhood information with bias when</p>

	calculating the representation of a given entity. The receptive field can be extended to multiple hops away to model high-order proximity information and capture users' potential long-distance interests. Moreover, we implement the proposed KGCN in a minibatch fashion, which enables our model to operate on large datasets and KGs. We apply the proposed model to three datasets about movie, book, and music recommendation, and experiment results demonstrate that our approach outperforms strong recommender baselines
2020.12.1 FL1, Wing Building for Science Complex	
彭迎涛 (Web Group)	<p>报告题目: KGAT: Knowledge Graph Attention Network for Recommendation</p> <p>报告摘要: To provide more accurate, diverse, and explainable recommendation, it is compulsory to go beyond modeling user-item interactions and take side information into account. Traditional methods like factorization machine (FM) cast it as a supervised learning problem, which assumes each interaction as an independent instance with side information encoded. Due to the overlook of the relations among instances or items (e.g., the director of a movie is also an actor of another movie), these methods are insufficient to distill the collaborative signal from the collective behaviors of users. In this work, we investigate the utility of knowledge graph (KG), which breaks down the independent interaction assumption by linking items with their attributes. We argue that in such a hybrid structure of KG and user-item graph, high-order relations — which connect two items with one or multiple linked attributes — are an essential factor for successful recommendation. We propose a new method named Knowledge Graph Attention Network (KGAT) which explicitly models the high-order connectivities in KG in an end-to-end fashion. It recursively propagates the embeddings from a node's neighbors (which can be users, items, or attributes) to refine the node's embedding, and employs an attention mechanism to discriminate the importance of the neighbors. Our KGAT is conceptually advantageous to existing KG-based recommendation methods, which either exploit high-order relations by extracting paths or implicitly modeling them with regularization. Empirical results on three public benchmarks show that KGAT significantly outperforms state-of-the-art methods like Neural FM and RippleNet. Further studies verify the efficacy of embedding propagation for high-order relation modeling and the interpretability benefits brought by the attention mechanism.</p>
2020.12.08 FL1, Wing Building for Science Complex	
但唐朋 (Cloud Group)	<p>报告题目: Sample Factory: Egocentric 3D Control from Pixels at 100000 FPS with Asynchronous Reinforcement Learning</p> <p>报告摘要: Increasing the scale of reinforcement learning experiments has allowed researchers to achieve unprecedented results in both training sophisticated agents for video games, and in sim-to-real transfer for robotics. Typically such experiments rely on large distributed systems and require expensive hardware setups, limiting wider access to this exciting area of research. In this work we aim to solve this problem by optimizing the efficiency and resource utilization of</p>

	<p>reinforcement learning algorithms instead of relying on distributed computation. We present the "Sample Factory", a high-throughput training system optimized for a single-machine setting. Our architecture combines a highly efficient, asynchronous, GPU-based sampler with off-policy correction techniques, allowing us to achieve throughput higher than 10^5 environment frames/second on non-trivial control problems in 3D without sacrificing sample efficiency. We extend Sample Factory to support self-play and population-based training and apply these techniques to train highly capable agents for a multiplayer first-person shooter game.</p>
2021.01.05 FL1, Wing Building for Science Complex	
彭迎涛 (Web Group)	<p>报 告 题 目 : IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models</p> <p>报告摘要:</p> <p>This paper provides a unified account of two schools of thinking in information retrieval modelling: the generative retrieval focusing on predicting relevant documents given a query, and the discriminative retrieval focusing on predicting relevancy given a query-document pair. We propose a game theoretical minimax game to iteratively optimise both models. On one hand, the discriminative model, aiming to mine signals from labelled and unlabelled data, provides guidance to train the generative model towards fitting the underlying relevance distribution over documents given the query. On the other hand, the generative model, acting as an attacker to the current discriminative model, generates difficult examples for the discriminative model in an adversarial way by minimising its discrimination objective. With the competition between these two models, we show that the unified framework takes advantage of both schools of thinking: (i) the generative model learns to fit the relevance distribution over documents via the signals from the discriminative model, and (ii) the discriminative model is able to exploit the unlabelled data selected by the generative model to achieve a better estimation for document ranking.</p>

实验室成员

Faculty Members



Meng Xiaofeng

孟小峰

博士 教授 博导

WAMDM 实验室负责人

Ph.D. Candidates



Liu Lixin

刘立新



Liu Junxu

刘俊旭



Ai Shan

艾山



Ma ChaoHong

马超红



Wang Leixia

王雷霞



Hao XinLi

郝新丽



Peng Yingtao

彭迎涛



Dan Tangpeng

但唐朋

M.Sc. Students



Tang Zili

唐子立



Fan Zhuoya

范卓娅

实验室新生感言

但唐朋 2020 级博士研究生

“不积跬步，无以至千里；不积小流，无以成江海”。继承和积累是科研学习的诀窍，坚持与信心是向前进步的法则。进入 WAMDM 这一位列国际顶尖实验室的同时也应该对自己有新的要求，新的目标和新的期待。当激动，紧张，期待都融汇一体时，剩下的只有久久的感动。感动着，来到新的一个出发点，感动着，认识那些生命里注将相遇相识的人。道阻且长，行则必至。新的挑丰富多彩惹人眼花缭乱，无时无刻在锻炼自己的综合能力。而我相信，我所经历的一切都是在都在为自身和社会的精彩添砖加瓦。

诚然，未来的道路上肯定会遇到各种各样的困难，但我相信在孟老师和实验室师兄师姐的帮助下自己一定能够不怕困难坚持初心，砥砺前行，不言放弃，保持进步。



彭迎涛 2020 级博士研究生

时光荏苒，悄然而逝，今天是 2021 年的第一天，博士入学已有近三个月的时间。人大在我心目中，不仅是学术的殿堂，更是心仪之选。正所谓知晓愈多，印象愈深，志向愈坚，情感愈浓，来到 WAMDM 实验室后，在这里聆听孟老师的谆谆教诲，在这里结识出类拔萃的师兄师姐，在这最美的年华相聚在最好的地方，体验和谐温馨的生活环境，享受丰富多彩的学术资源。

我知道，选择这里就是选择了与优秀的人在为伍，选择了攻读博士学位，就是选择了与笃定的自己为伴。因此，作为一名博士新生，我不仅要多和朋辈们多多交流，在侃侃而谈中碰撞出思想的火花，更要学会与自己独处，在孜孜不倦中探寻科研的奥秘。还记得在一篇文章中看到：“读博这条路艰辛、漫长、疲惫，充满了沮丧和失败，当梦想照进现实，你是否真的有勇气坚持到底？站在博士学习生涯的起点上扪心自问，也许很难给出一个肯定的答案”。但即使前路漫漫，并非坦途，我也应该始终保持乐观的精神、健康的体魄、求知的热情和专注的态度，大胆前行，不言放弃，正谓之“是非经过不知难，成如容易却艰辛”。

希望在博士阶段，自己能够仰望星空、脚踏实地、无愧于心，顺利通过大小论文两个难关，完成对自我思想的重塑，升华对数据挖掘领域的认识，具备最严谨有效解决问题的能力。同样希望博士毕业之后，无论在学术之田深耕不辍，还是在业界机构大展身手，自己都能够带着博士期间收获的宝贵财富，在未来的道路上愈走愈远。



范卓娅 2020 级硕士研究生

从 2019 年 2 月 10 日决定加入实验室到现在已经快两年了，时间真是过得飞快。在这两年中，我不仅学到了知识技能，也学到了很多做人的道理，非常感恩师兄师姐的朝夕相伴和孟老师的关怀指导。

2020 年经历了太多事情，我发现原来稀松平常、不以为意的生活点滴都那么可贵。转眼间，这一年也像被按了快进键一样变成了过去。希望自己能更加珍惜短暂的研究生生活，以优秀的师兄师姐为榜样，在有限的时间里做出更多有价值的事情。



毕业生寄语

叶青青 2020 届博士毕业 香港理工大学，研究助理教授

时间飞逝，不知不觉中距离毕业已过半载。庆幸五年博士生涯的训练，让自己能够继续从事所热爱的科研工作。

至今仍清楚地记得五年前第一次来到 WAMDM 实验室的情景，那是第一次参加实验室例会，充分感受到属于这个集体的满满的能量。自那以后，每周一次的大组会成为了惯例，或讨论，或报告，或聆听，不仅锻炼了演讲和汇报的能力，而且拓宽了认知和研究的角度。在实验室的几年时间里，以自律和努力为习惯，以学习和研究为目标，终于成长为一名合格的 PhD。回首五年读博生涯，似乎弹指一挥间，却也是一段格外美好的经历。平淡的日子枯燥乏味，奋斗的日子则格外充实。读博就是这样一个奋斗的过程，不乏初入课题时的迷茫，发现 idea 时的欣喜，推倒重来的勇气，锲而不舍的坚持，论文录用时的激动和豁然开朗的心境。如此循环往复，锻炼科研能力，完成读博使命，收获人生阅历。心中万分感激孟老师所给予的机会和栽培，感谢 WAMDM 实验室的支持，感念实验室同学们一路相伴前行，最终收获一段珍贵的人生经历。此外，在香港理工大学的交流经历，结识胡老师这样一位良师益友实乃人生之幸，学习能力的培养，科研能力的提升，工作生活的技巧，胡老师亦教会了我许多，心里由衷感激。



五年博士生涯，大都在实验室和宿舍的两点一线间度过，虽单调，却为后续的发展打下了坚实的基础；中途的闲暇时光虽然短暂，却未曾错过沿路的风景。博一入学的第一个月去了哈尔滨参加大创年会，第一次论文报告，甚至不知道需要准备 slides。博二上学期，首届隐私保护会议的筹备和举办工作，虽然心力交瘁，却真真切切锻炼了工作能力；博二下学期第一次来到香港，全身心专注科研工作，茶余饭后的一点时间，喜欢漫步在维港边上感受海风徐徐，在天星小轮上眺望无垠的大海，在太平山顶欣赏云雾缭绕，在观景台上俯瞰香港的美丽夜景。博三回到了实验室，这一年里最大的收获莫过于 S&P 的论文录用，体验到科研转化为成果的欣喜。博四再次来到香港，研究方面完成了图数据和流数据两个工作，投稿过程可谓一路坎坷，庆幸最终去到 ICDE, TKDE 和 INFOCOM，也算是个好归宿。大半年在香港的时间里，每周少有的闲暇时光，或流连于山水风光，或徜徉在丛林小径，像一只快乐的小小鸟，黄金海岸海风习习，船湾淡水湖细雨蒙蒙，金山郊野公园可爱的小猴子漫山遍野，邮轮码头公园里的一席交流，最高峰大帽山上的一番感悟，溜西洲高尔夫球场的一次尝试。当然，还有美国 S&P 之行，不亦乐乎。博五回到人大，在校时间其实只有短短一个学期，新冠疫情提前结束了在校生活。

2020 是个很特殊的年份，疫情对我们的生活产生了巨大的影响，无论是对老师和学生，还是对家庭和个人。但对于我，2020 保留了一份“仁慈”，毕业、工作和生活都很顺利。来到香港理工大学任教，工作内容无外乎教学和科研。人生中第一次以老师的身份站在“讲台”上是在今年的教师节，依然记得上课前心里格外紧张，上完课心里也格外激动。科研上，除了课题研究，还需要申请研究经费，这是与博士阶段的研究很大的不同。繁忙之中半年的时光悄然而逝，庆幸从中学习了很多，体验了很多，在学习和体验的过程中不断成长。

最后衷心祝愿孟老师身体安康，祝愿 WAMDM 实验室越来越好！

王硕 2020 届博士毕业 河北大学

转眼间毕业已经半年了，上半年虽然经历了新冠疫情的干扰没能返校，但是还是会经常想起人大校园里的银杏、樱花和白玉兰，有时梦里还会坐在实验室的工位上忙碌着，五年的求学之路如同电影一般浮现在眼前。

2015 年的金秋季节，我第一次和实验室的老师和同学们坐在会议室中开会，当时对 NLP 这个英文缩写还十分陌生，毫无概念；接下来的两年，通过艰苦学习和探索，自己终于掌握了自然语言处理的相关技术和知识融合的一般方法；孟老师也没有看低我这名门外汉，带我出国参加了很多会议，这在我读博之前是从来没有想到过的历练和经验；实验室的研究工作在国内科研领域也十分活跃，定期召开的研讨会和项目工作会议使我见识了科研团队间的交流与合作，参加的各类全国学术会议让我积累了很多专业知识和学术上的朋友；在实验室的日常事务工作中，孟老师也给了我锻炼自己的机会，使我可以从容面对很多自己从前并不熟悉的事务，增强了自己的时间管理和规划能力。

我在实验室紧张而有节奏的学习与生活中不断成长，直到成为别人口中的师兄，5 年的成长和历练使我变得更加自信、自律和自强，实验室不仅让我在学术上有了进步，也让我在生活与工作中有了更大的动力。感谢 WAMDM 实验室，愿你一如既往地坚持自己的道路，阔步向前，也希望 WAMDM 实验室的同学们更加努力，为实验室谱写新的篇章。

毕业有感而作：

语言今日成吾念，
寄迹学海渡华年，
业在君心苦亦得，
毕生何处似此贤。



杨晨 2020 届博士毕业 中国人民银行清算总中心

五年的博士学习生涯，在毕业之后回首看来依然是五味杂陈，其终究在我的生命中留下浓墨重彩的一笔。孟子有云“故天将降大任于是人也，必先苦其心志，劳其筋骨，饿其体肤，空乏其身，行拂乱其所为，所以动心忍性，曾益其所不能。”从修行的角度而言，想要“益其所不能”，前面的“苦其心志”是必不可少的，这也是读博期间所必须经历的，也感谢孟老师和实验室给予了这样一个环境，让我初探了“从心所欲而不逾矩”的创新思维能力，获得了人生中宝贵的一段经历。

在实验室工作期间，首先深深地感谢我的导师孟小峰教授。首先，孟老师引领我走入了大数据分析的研究领域，并结合实际的项目，能够边研究边实践，这是一段宝贵的经验，让我在走入工作岗位不惧任何挑战，既可以探究创新理论问题，又可以实际参与具体研发工作。此外，孟老师对新知识不断追求的激情、严谨的治学态度、一丝不苟的做事方式和认真负责的做事态度，无时无刻不在教育和感染着我，这些都将使我受益终生。在此，谨向孟老师表示最诚挚的谢意！

写下这段感言的时候，也深深地怀念实验室的兄弟姐妹们，大家在一起学习工作中能够守望相助，建立了很深的情谊。在大家集中开发的日子里，虽然很辛苦，但反而那段日子留下了深刻的记忆，历经弥新。和实验室同学在一起的日子总是充满了欢声笑语，能够与你



们一起度过博士生涯是我的荣幸。

在实验室学习期间,虽然过程是曲折的,但是毕业后能明显感觉到自身的能力的成长。因此,在这里也希望还在实验室耕读的师弟师妹们,守住本心,坚持下去,在孟老师的引领下,实验室必定以后会越来越越好,大家的前途必定可期。

吴新乐 2020 届硕士毕业 奥尔堡大学读博

毕业半年,回想起在 WAMDM 实验室度过的时光,心中依旧充满不舍。非常感谢孟老师三年来对我的教导,教给我很多知识和做人的道理,孟老师始终是我科研之路上的 lighthouse,指引我激励我。WAMDM 实验室像一个温馨融洽的大家庭,我在其中度过了充实难忘的三年,感谢在此期间各位师兄姐妹们的帮助,我会一直想念你们的。非常荣幸自己曾经是 WAMDM 实验室的一份子,也希望实验室发展得越来越好。



杜永杰 2020 届硕士毕业 中国农业银行

时光飞逝,转眼三年。过去的三年里,有和师兄师姐一起在实验室学习的快乐,也有和大家一起赴怀柔集中开发的艰辛,还有登山看星辰大海的激动。在实验室里,总有人在你困难的时候帮你克服,总有人在你成功的时候与你分享喜悦,总有人在你懈怠的时候督促你前行。在研究生学习生涯中,不仅收获了宝贵的知识财富,而且很庆幸能跟着孟老师学到了很多书本之外的东西,还很开心能结识一群充满活力的同学。但遗憾的是大家只能云答辩和云毕业,没能在毕业季相聚实验室,没能和孟老师挥手道别。希望以后大家常聚实验室,也希望这个大家庭越来越好。



段志强 2020 届硕士毕业 中国农业银行

很感谢这三年的学习和奋斗,让我能够接触到这么多优秀的人和能磨炼我的事情,我在这三年当中学习到了更专业的计算机知识,学到了如何进行科学研究。我在这三年中经历的每一件事情都会深刻的影响到我以后的生活,接触到的每一个人都会让我更加具有奋进的精神。这些都不断的促使我一步步的上升。首先要感谢我的导师孟小峰教授,我幸运的在研究生期间能遇到孟老师这样一位对待学术严谨认真、对待学生严格关怀的导师。感谢孟老师对我一直以来的指导和提点,孟老师的对学术的态度是我终身学习和追求的榜样,也是日后职业生涯中始终引导我的灯塔。感谢 WAMDM 的全体成员。尤其是杨晨师兄和杜永杰,没有杨晨师兄的帮助我无法在严苛的学术道路上一直前进,他一直在我学术和生活的道路上带着我前进,帮助我克服一个又一个的难题。感谢永杰一直以来的陪伴,我们同入实验室,在这三年的学习和生活当中互帮互助,共同前进。最后祝愿 WAMDM 实验室越来越好。

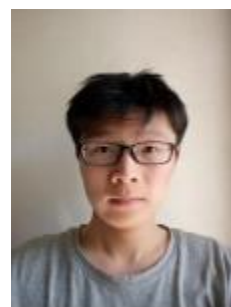


吴永泰 2020 届硕士毕业 生命奇点（北京）科技有限公司

时光飞逝，三年的学习时光一转眼就过去了。感谢孟小峰教授，作为领路人，带领我们进入学术的殿堂。孟老师给我最深刻的印象就是渊博的知识、对待学术的严谨态度、诲人不倦的教育理念、对待行业大趋势的敏锐洞察力和风趣的说话方式，让我在实验室三年的时间受益匪浅，感谢孟老师的言传身教和三年的培养。

同时，我也要感谢这三年陪伴我走过来的师兄师姐和小伙伴。三年来，我们朝夕相处，一起学习，相互鼓励，共同进步，让三年的生活和学习时光充满了欢声和笑语。

三年的时间，回忆种种，满满的都是收获。希望未来的自己更成熟、努力，希望实验室在孟老师的带领下继续创造更辉煌的未来。



王飞 2020 届硕士毕业 京东

两年的时间，时光飞逝，WAMDM 实验室哺育了我，使我从初入硕士懵懂与初探科学的迷茫到找到自己道路，坚定自己的信念前进。当真正走向社会，总是珍惜学生时光，每当看到校园总是回忆起当时的一一点一滴，仿佛就在昨日。尤其是人大的生活，三点一线的节奏，纵向的知识探索与横向的综合提升，感恩机遇让我来到人大，来到 WAMDM 实验室。“惠施多方，其书五车”，孟老师的给我们的不仅仅是计算机领域的专业知识，更是哲学明义的人生道理，这一切促使着我养成了良好的学习研究习惯与严谨务实的工作态度。北京是一个神奇的城市，有视野蒙蒙的雾霾弥漫也有艳阳高照的晴空万里。孟老师曾说过“每个人心中都要有自己的灯塔”。从当年的懵懂与研究生阶段的寻觅，到现今的明晰，心中有灯塔，纵使雾霾弥漫，也能拨开云雾见光明。



杨鑫 2020 届硕士毕业 上海蔚来汽车有限公司

从踏进人大校园开始，就注定我会有一段奇妙的经历。美国政治家罗斯福说过：“对明天的认识的唯一限度，取决于我们今天的怀疑”。我的硕士生活也是从怀疑开始，怀疑我的选择，怀疑未来是否坦途，怀疑我是否能行……

攻读硕士期间，孟老师将我引领入一个全新的学术领域，他严谨认真，创新活跃的科研精神让我受益匪浅。至今对孟老师的一句话印象深刻：“我的重点不是教给你们具体的技术，尤其是计算机学科，那个很快就过时了，我想教给你们的是种思维方式和学习能力，这会让您们经受住任何考验”。

转眼间，毕业半年。在新的工作岗位，依然践行实验室严谨、创新的精神，一丝不苟，脚踏实地，努力在自己的工作岗位干出成绩。



汤庆 2020 届硕士毕业 商汤科技开发有限公司

我很荣幸能够来到 WAMDM 实验室学习深造,能和这样一群优秀、友爱又有趣的人度过研究生生涯的三年时光。感谢孟老师的教诲,也感谢师兄师姐师弟师妹们的帮助。毕业了,我们有着过去的不舍,也有着对未来的憧憬,就像孟老师在会议室里边挂的字一样,"日日是好日",愿我们都能好好珍惜每一天,享受每一天。



2020 年毕业生去向

姓名	学历	时间	毕业去向
叶青青	博士	2020 年 7 月	香港理工大学, 研究助理教授
王硕	博士	2020 年 7 月	河北大学
杨晨	博士	2020 年 7 月	中国人民银行清算总中心
吴新乐	硕士	2020 年 7 月	奥尔堡大学读博
杜永杰	硕士	2020 年 7 月	中国农业银行
段志强	硕士	2020 年 7 月	中国农业银行
吴永泰	硕士	2020 年 7 月	生命奇点(北京)科技有限公司
王飞	硕士	2020 年 7 月	京东
杨鑫	硕士	2020 年 7 月	上海蔚来汽车有限公司
汤庆	硕士	2020 年 7 月	商汤科技开发有限公司

历年年报回顾

WAMDM Report 2019



周有光老先生在 2009 年 5 月 8 日（时年一百零四岁）写了一篇文章《全球化时代的世界观》，其中写道“全球化时代的世界观，跟过去不同，主要是：过去从国家看世界，现在从世界看国家。过去的 worldview 没有看到整个世界，现在的世界看到了整个世界。在全球化时代，由于看到了整个世界，一切事物都要重新评估（transvaluation）。”读来十分受益！

中美之间的冲突其实是世界观的冲突，美国仍抱有从国家看世界的 worldview，代表的是过去；中国已经在从世界看国家，代表的是未来。科技创新也要具备全球化时代的世界观，能够看到整个世界即多学科交叉融合；而产业创新的全球化时代的世界观要求从单一产品创新过渡到产业链创新，在产业链创新中才能掌握源头，从而彻底解决卡脖子的问题。

在过去的 2019 年中，实验室在技术研究与系统开发上多点开花，总结撰写了一系列 AI、DB、SC（Social Computing）学术前沿和交叉研究的报告，包括机器学习化数据库、机器学习隐私保护问题、机器学习可解释问题、以及人工智能下的社会计算科学等，并在大数据实时分析、数据融合、隐私保护三个方面取得了阶段性成果。

WAMDM Report 2018

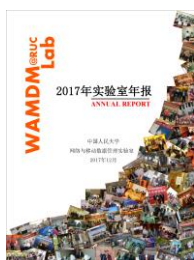


丰子恺的散文《剪网》中提到“我仿佛看见这世间有一个极大而极复杂的网。大小大小的一切事物，都被牢结在这网中，所以我想把握某一种事物的时候，总要牵动无数的线，带出无数的别的事物来，使得本物不能孤独地明晰地显现在我的眼前，因之永远不能看见世界的真相，大娘舅在大世界里。只将其与“钱”相结的一根线剪断，已能得到满足而归来。所以我想找一把快剪刀，把这个网尽行剪破，然后来认识这世界的真相。”

万物互联的时代已经来临，当下的技术还停留在互联网上，多是以“织网”为主，万物互联时代“织网”终将不成问题，人们需要的是“剪网”。“大娘家白相了大世界”，只将与“钱”相关的线剪断，便能感到快乐，满足。在大数据时代、人工智能时代，有了的网，我们生活在一个“万物互联”的时代，未来的发展如何，可能我们更多时候必须要剪掉“网中的线”，才能够感到快乐。谁能找到这把快剪刀，谁就能制胜未来。

在过去的 2018 年中，实验室在技术研究与系统开发上多点开花，在大数据实时分析、数据融合、隐私保护三个方面取得一些阶段性成果。

WAMDM Report2017



故宫博物院院长单霁翔在去年《朗读者》节目中朗读到：“在中国传统文化中，最高的状态是意会的境界。大，意味着多。多，意味着无穷无尽，无穷无尽就是空。既无穷莫测，故实则虚之。实则虚之，是中国人的文化密码，投射到每个人的心中。”想来感触颇多。对当下大数据和人工智能热潮，更多呈现了“虚则实之”（虚张声势），甚至是“虚而虚之”（空城计）的状况。

在过去的 2017 年中，实验室在技术研究与系统开发上多点开花，在大数据实时分析、数据融合、隐私保护三个方面取得一些阶段性成果，力图达到“实则虚之”的境界。

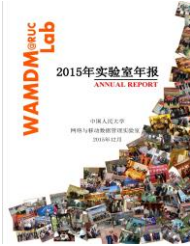
WAMDM Report 2016



2006-2016 一晃十年！当 2006 年即将过去的时候，我和我的学生们讲，我们是否应该总结些什么，总结过去，展望未来，或许对我们自己，对他人，对社会都是一种责任，一种鼓舞，一种鞭策。由此有了实验室的第一份年报。之后成为实验室的一种惯例，每年一册，算是对过去一年的一个交代，也是对未来一年的一个期盼。

过去十年 IT 技术突飞猛进，本实验室“网络与移动数据管理”（Web and Mobile Data management, WAMDAM）仍秉承萨师煊、王珊教授所一贯倡导的学术研究与系统开发并重的传统，以创新数据管理系统的研究为目标，立足云计算和大数据技术背景，将研究定位在数据融合与知识融合、大数据实时分析与交互分析、大数据隐私管理等方面，在十三五开局之年把握机遇，迎接挑战！

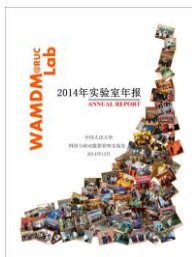
WAMDM Report 2015



陈寅恪先生说：“一时代之学术，必有其新材料与新问题。取用此材料，以研求问题，则为此时学术之新潮流。治学之士，得预于此潮流者，谓之预流（借用佛教初果之名）。其未得预者，谓之未入流。”对今天的信息技术而言，“新材料”即为大数据，而“新问题”则是产生于“新材料”之上的新的应用需求。

对数据库领域而言，真正的“预流”是 Jim Gray 和 Michael Stonebraker 等大师们。十三年前面对“数据库领域还能再活跃 30 年吗”这一问题，Jim Gray 给出的回答是：“不可能。在数据库领域里，我们已经非常狭隘。”但其转面的回答是：“SIGMOD 这个词中的 MOD 表示‘数据管理’。对我来说，数据管理包含很多工作，如收集数据、存储数据、组织数据、分析数据、表示数据，特别是数据的表示部分。现在人们已经拥有太多的数据，而我对许多人说我们仅仅希望拥有更多的时间。所以，整个数据收集、数据分析和数据简单化的工作，就是能准确地给予人们所要的数据，而不是把所有的数据都提供给他们。这个问题不会消失，而是会变得越来越重要。”（见《数据库大数据访谈录》）。其实十三年前大师们已看到了“新材料”，而且指出了“新问题”。面对大数据浪潮，各种提法众多，我坚持用“大数据管理”概括抽象这个领域的研究，也是源于此。Jim Gray 早于 1998 年因事务处理方面的成就获得图灵奖。Michael Stonebraker 也于 2015 年因系统创新的成果如愿获得这迟到的殊荣，实乃众望所归，实至名归，可喜可贺。

WAMDM Report 2014



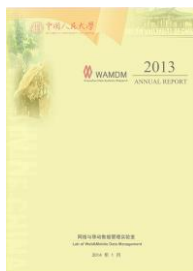
“浅浅的知识比无知更使人栗六不安，深深的知识使人安定，我们无非是落在这一片深深浅浅之中”。木心的语言总是那么平静而深邃。

过去的一年大数据话题仍是热热闹闹，深深浅浅。凭借坚守数据管理的理念，思考着大数据管理的机遇与挑战，试图能给出一个比较“深深的知识”。年底刚刚完成《计算机研究与发展》的一个专题：大数据管理。在此专题的前言中总结了自己近期的一些思考。

已故的图灵奖得主 Jim Gray 在其《事务处理》一书中提到：6000 年以前，苏美尔人（Sumerians）就使用了数据记录的方法，已知最早的数据是写在土块上，上面记录着皇家税收、土地、谷物、牲畜、奴隶和黄金等情况。随着社会的进步和生产力的提高，类似土块的处理系统演变了数千年，经历了殷墟甲骨文、古埃及纸莎草纸、羊皮纸等。19 世纪后期，打孔卡片出现，用于 1890 年美国人口普查，用卡片取代土块，使得系统可以每秒查找或更新一个“土块”（卡片）。可见，用数据记录社会由来已久，而数据的多少和系统的能力是与当时的社会结构的复杂程度和生产力水平密切相关的。

随着人类进入 21 世纪，尤其是互联网和移动互联网技术的发展，使得人与人之间的联系日益密切，社会结构日趋复杂，生产力水平得到极大提升，人类创造性活力得到充分释放，与之相应的数据规模和处理系统发生了巨大改变，从而催涌了当下众人热议的大数据局面。

WAMDM Report 2013



“自从阴错阳差地掉进这片海里，不知不觉我也已经游了 8 年，并且于公元 1998 年惊奇发现，好像是真的已经游到了‘胜利彼岸’。眼下，我衣衫褴褛、筋疲力尽，孤零零坐在岸边，而心里最想做的就是召唤故乡那些智慧勇敢的人们，一起来到这片肥沃而又辽阔的新大陆，跑马圈地，共建家园。”十五年前的这段话，描写当下的心境也颇为合适。大数据的浪潮过去一年一浪高过一浪，自己凭着一点点了解到各处做了若干场大数据的报告，想起来犹如那位衣衫褴褛者，但无需召唤，蜂拥而至的人们一起涌向了这片新大陆，但只见跑马圈地，未见共建家园。

这一年仍在不停歇地思考大数据的根本问题。年初发表在计算机研究与发展上的综述“大数据管理：概念、技术与挑战”获得了同行的广泛关注，下载次数一致高居中国知网的首位，由此可见大数据的热度非同一般。文中通过对数据源产生的演化分析，揭示了数据管理需求和任务的不断变化，促使数据管理系统不断推陈出新。回顾数据管理技术的发展，一脉相承的是追求在系统中提供尽可能贴近用户的数据抽象，数据抽象越到位，用户使用越方便。数据管理系统便是要实现这种抽象机制：面对企业数据的管理，DBMS 提供了物理模式到逻辑模式（关系模型）的抽象；面对互联网数据的管理，数据集成（数据空间）提供了局部模式到全局模式的抽象；如今面对大数据管理，到底要实现什么样的数据抽象，苦苦思索，仍不得要领，但隐隐感觉这是正确的思考方向。

WAMDM Report 2012



“这是一个最美好的时代，也是最糟糕的时代；这是智慧的年代，也是愚昧的年代；这是信仰的时期，也是怀疑的时期；这是光明的季节，也是黑暗的季节；这是希望的春天，也是失望的冬天；我们前途无量，同时又感到希望渺茫；我们一起奔向天堂，我们全都走向另一个方向……”这是狄更斯在其《双城记》中开篇之语，很耐人寻味。比照着，我们可以有：这是大数据（BD）的时代，也是小数据（DB）的时代；这是创新的时期，也是怀疑（钱学森之问）的时期；这是实干（兴邦）的季节，也是空谈（误国）的季节；这是中国的春天，也是世界的冬天；我们前途无量，同时又感到希望渺茫……

大数据确实是当下最热的词汇，各种概念、判断、论调纷争。今年暑期去了一趟河南安阳的殷墟遗址，对大数据的内涵颇有感悟。河南安阳殷墟遗址的最大发现就是青铜器司母戊鼎和甲骨文。尤其是甲骨文，目前已出土了十四万片，在当年的发掘中发现了一个“甲骨文大坑”，其中散落了一万七千余片的甲骨文残片，这些残片数量众多，其上所刻的文字内容繁杂。由于到目前为止还无法完全了解每个文字所代表的准确含义，所以整个甲骨文的解读仍处于一个相对初级的阶段。倘若我们能够发现一种方法，可以有效的对“甲骨文大坑”残片上的文字进行解读，并从中归纳出不同残片上文字之间的关联，那么就极有可能在此基础上整理出甲骨文的完整体系，从而最大程度的还原出当时的社会面貌，即体现出其价值所在。

WAMDM Report 2011



去年我们将自己未来十年的研究概括为创新数据管理研究 2.0，涉及云计算、闪存存储、隐私保护、移动互联网等关键词，试图探索为下一代计算技术和应用所需的数据管理技术。过去的一年研究使我们更坚定了这一定位。基于闪存、PCM 等新型存储技术的数据库系统研究有可能产生基础性的创新，隐私保护是未来众多技术发展中不可逾越的障碍，移动互联网的普及同样有若干关键问题需要解决。倒是目前最热的云计算、物联网目前还未找到实质的感觉，看来“云里雾（物）里和海里”的探索还是有些飘忽不定，需要扎到应用中去积累。至少能解决一些现实问题，理论创新不敢想。

过去一年基于人大的学科背景，在中国人民大学重大基础研究计划的支持下，着重开展了社会计算的研究，并召开相关多学科交叉的研讨。社会计算是沟通社会科学和计算机科学的桥梁。社会计算要支持社会科学研究，从信息获取、分析、建模、实验、决策和平台等层面突破目前交叉学科借鉴的困境，为社会科学提供新的研究框架与工具，也为信息技术提供新思路。当然我们也看到社会计算在社会科学家中扩散所面临的困难，其中既有技术知识、研究经费等物质条件的因素，也有知晓度低、与社会科学传统方法和“技术怀疑主义”的不兼容的原因。不过我们还是觉得在以人文社会科学为主体的人民大学开展这一研究工作还是非常有意义的。因为真正交叉学科的研究是未来产生创新的机会所在。

WAMDM Report 2010



新世纪以来，数据库界普遍面临的一个问题是，在传统的数据库技术成熟之后，数据库研究应向何处去？凭借自己对当时技术趋势的判断，我将研究目标定位在解决数据库技术与 Web 计算和移动计算交叉结合所产生的挑战性问题，即结构多样的 Web 数据管理，半结构化 XML 数据的管理，以及移动环境下的数据管理问题，并创立了“网络与移动数据管理实验室（Web and Mobile Data Management）”，致力于这方面的研究，取得了一些国内外所共知的研究成果。我把这一阶段的研究概括为创新数据管理研究 1.0。

数据库系统发展经历了三十年，大致呈现出了“分久必合、合久必分”规律。六七十年代广泛的应用需求的出现促成了各类数据库系统的产生。八九十年代大型网络分布计算环境的普及使得政府、企业的应用需求趋同，导致几大数据库系统的“大一统”局面出现。当下互联网特别是云计算的出现，使得应用需求再趋多样化，人们更期盼与自己的需求功能相宜的数据库系统，而不是面面俱到的“大拼盘”系统，多样化时代重新到来。最近日渐火爆的“NoSQL”运动正是迈向这一目标的尝试。我们在本年度报告里试图把这些我们观察到的、看明白或没看明白的一些问题总结成短文，与大家交流，抛砖引玉。

WAMDM Report 2009



在过去的十年间，随着互联网的迅速发展，整个 Web 的数据量已经超过了 200,000TB，并仍在快速地增长，这使其成为人们获取有用信息的最重要的途径之一。另一方面，随着 3G 时代的到来，大量的手机、移动设备需要频繁访问互联网，以从互联网上获取丰富的信息，这是一个必然的趋势。而 3G 所带来的高带宽，使得未来手机将不再是一个简单的通话工具，人们从互联网上获取信息将越来越依赖于手机和以及其它移动设备。如何解决面向移动用户的 Web 数据集成问题，成为实验室今后关注的一个新的研究领域，目前研究界还缺乏有关的研究成果，我们认为这是一个创新的机遇。

云计算是当今信息产业最受关注的一种计算模式，在这种模式下，企业和个人可以根据自己的需要购买存储设备和计算能力，而非花费巨资购买大规模高性能计算机。作为云计算的一项关键技术，云数据存储和云数据管理为业界带来巨大的潜在价值。随着信息产业的发展，企业 and 公司产生的数据量快速增长，通常数据规模可以达到 TB 甚至 PB 级别。如何管理和分析海量数据是目前很多领域所面临的问题，例如在医疗、通信和互联网领域。对此实验室的提出的新的研究课题是云计算环境下数据库技术，实现一种具有高可用性、高容错性、可扩展性和高性能的云数据库系统。

WAMDM Report 2008



图灵奖获得者 Jim Gray 曾在 1998 年的获奖演说中，对未来数据量急剧增长的规律做过这样的预言：未来每 18 个月产生的数据量等于有史以来的数据量之和！最近，我们又看到美国《未来学家》杂志根据世界未来学会年度预测，对未来世界发展前景进行了展望，其中认为未来的数据将以佑字节（Yottabyte，即十亿 GB）为单位进行存储。这种发展趋势将引伸出网络环境下数据管理新的科学问题：即以佑字节为单位的数据管理！这不同于通常所说的海量数据管理，它将面临完全不同的应用需求和完全不同的存储。

当一个产业的根本需求和底层架构发生如此重大变化的时候，与挑战同时到来的是巨大的机遇，能否抓住这次机遇，在这片崭新的天地写下属于我们自己的一笔，这正是我们实验室近期研究的巨大动力。在国家自然科学基金重点项目和 863 计划探索项目的支持下，近期我们开展了闪存数据库技术和个人数据空间管理的研究工作。一年即将过去，在继过去两年有关实验室科研情况的年度报告的基础上，再次整理了 2008 年的年度报告，内容涉及技术展望，系统研发，论文精选和学术交流等。

WAMDM Report 2007



正像去年在序中所说的，“总结过去，展望未来，或许对我们自己，对他人，对社会都是一种责任，一种鼓舞，一种鞭策。”所以在 2007 年结束之际，我们又编辑了这样一份报告，是对大家的感谢，也是完成对自己的承诺。

本年度报告的结构仍延续去年的风格，报告第一部分汇集了我们实验室的技术综述，展示我们对数据管理技术发展；本年度报告的第二部分汇集了我们的系统工作；报告的第三部分是论文汇集，本年我们在 VLDB2007 发表一篇长文，在 DASFAA2007 发表 3 篇文章，论文质量有所提高；实验室一贯重视国际学术的交流，实验室几乎每一位在读学生都有机会出国合作研究或参加国际会议，经常有国外学者来实验室交流访问。

WAMDM Report 2006



当 2006 年即将过去的时候，我和我的学生们讲，我们是否应该总结些什么，总结过去，展望未来，或许对我们自己，对他人，对社会都是一种责任，一种鼓舞，一种鞭策。经过近一个月的努力，我们终于有了手头的这部集子，算是对过去一年的一个交代，也是对未来一年的一个期盼。

过去五六年间，我们的研究工作始终围绕数据库技术与网络计算与移动计算环境的结合。因此实验室的名字为“网络与移动数据管理”（Web and Mobile Data management, WAMDM）。实验室的研究风格秉承萨师煊、王珊教授所一贯倡导的学术研究与系统开发并重的传统，以创新数据管理系统的研究为目标。

内部资料，妥善保存

网络与移动数据管理实验室
地址：中国人民大学理工配楼一层
网址：<http://idke.ruc.edu.cn>
电话：010-62512719
传真：010-62512719
编辑：范卓娅，但唐朋，彭迎涛