中国科学: 信息科学 2020 年 第 50 卷 第 1 期: 25–66
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

大数据安全技术研究进展

陈性元[1,2*], 高元照[1,2], 唐慧林[1], 杜学绘[1]

1. 信息工程大学三院, 郑州 450001
2. 密码科学技术国家重点实验室, 北京 100094
* 通信作者. E-mail: chxy302@vip.sina.com

收稿日期: 2019–04–18; 接受日期: 2019–06–27; 网络出版日期: 2020–01–08

国家重点研发计划 (批准号: 2018YFB0803603) 和科技创新特区 (批准号: 18-H863-01-ZT-005-017-01) 资助项目

摘要 大数据是经济发展的新动能, 社会发展的新引擎, 塑造国家竞争力的战略制高点, 对人民生活具有重大影响. 然而随着社会对数据价值认知的提升和大数据平台建设的蓬勃发展, 大数据安全问题日益成为阻碍大数据应用推广的瓶颈. 同时, 由于大数据技术、框架仍在不断演变当中, 研究人员对大数据安全内涵的核心认知和关键特征理解还存在差异, 尚未形成相对统一的大数据安全框架. 当前亟需对大数据安全技术发展现状进行梳理, 为大数据安全重点问题的研究和突破提供参考. 本文结合典型大数据系统技术框架, 围绕大数据安全需求, 构建了大数据安全技术框架. 在此框架下, 从大数据安全共享与可信服务、大数据平台安全和大数据安全监管 3 个方面系统梳理了大数据安全关键技术的研究现状, 囊括了大数据业务流程和大数据系统技术框架所涉及的主要安全机制. 最后总结了大数据安全技术有待解决的核心问题和发展趋势.

关键词 大数据安全, 安全技术框架, 数据安全共享, 平台安全, 安全监管

1 引言

自 “大数据” 概念由全球数据科学权威维克托 · 迈尔 - 舍恩伯格 (Viktor Mayer-Schönberger) 在《大数据时代》一书中明确提出以来, 大数据引起了学术界、产业界和各国政府的广泛关注. *Nature* 和 *Science* 等期刊发表专刊探讨大数据带来的机遇与挑战. 达沃斯世界经济论坛发布报告《Big Data, Big Impact》宣称数据已经成为一种新的经济资产类别. 美英等世界上主要国家相继发布大数据战略[1)], 我国国务院也于 2015 年发布《促进大数据发展行动纲要》, 以国家为主导的大数据快速发展时代已经到来. 当前, 大数据应用正不断向电子商务、智慧城市、国防建设、科学研究等众多领域推广.

1) 世界主要国家的大数据战略和行动. 2015. http://www.cac.gov.cn/2015-07/03/c_1115812491.htm.

引用格式: 陈性元, 高元照, 唐慧林, 等. 大数据安全技术研究进展. 中国科学: 信息科学, 2020, 50: 25–66, doi: 10.1360/N112019-00077
Chen X Y, Gao Y Z, Tang H L, et al. Research progress on big data security technology (in Chinese). Sci Sin Inform, 2020, 50: 25–66, doi: 10.1360/N112019-00077

大数据在给生产、生活方式带来变革的同时, 其安全问题也日益凸显. 2017 年美国征信机构 Equifax 数据泄露, 导致几乎全美一半人口的个人敏感信息掌握在黑客手中, 同年美国国家安全局 (National Security Agency, NSA) 数据泄露, 美军超过 100 GB 的绝密数据暴露在亚马逊云上 [1]. 2018 年, 剑桥分析公司非法收集 Facebook 用户信息, 并基于分析结果干预美国大选[2)]. 上述安全事件表明, 加快大数据安全技术研究, 已成为保障信息化建设和数字经济稳步向前推进的迫切要求.

然而, 不同于传统数据, 大数据具有体量大 (volume)、多样性 (variety) 和速度快 (velocity) 等特性 [2], 这为安全技术在大数据环境下的应用带来了极大挑战. 同时, 为实现大数据的有效处理还引入了分布式的计算与存储框架. 这些新型框架也带来了新的安全威胁. 当前, 大数据安全研究仍处于初期, 研究人员对大数据安全的核心认知和关键特征理解还存在差异, 理论成果同实际应用要求之间还存在差距, 亟待对大数据安全技术的发展现状进行系统梳理, 为大数据安全重点问题的研究和突破提供参考.

本文结合大数据业务流程和大数据系统技术框架, 分析了大数据安全关键技术的研究进展, 在此基础上总结大数据安全中有待解决的挑战性问题和解决思路, 探讨大数据安全的发展趋势. 第 2 节在总结大数据安全技术主流分类视角和主要内容的基础上, 提出一种大数据安全技术框架; 在该框架下, 第 3 节从数据价值链的角度梳理了大数据安全共享与可信服务中的安全技术; 第 4 节则从 IT 价值链的视角梳理了大数据平台安全技术; 第 5 节综合数据、服务与平台 3 个方面梳理了大数据安全监管技术; 第 6 节总结全文并对大数据安全技术研究的未来趋势进行展望.

2 大数据安全技术框架

建立有效的大数据安全技术框架, 能够为大数据系统的安全技术研究与部署提供指导. 本节首先对当前有代表性的大数据安全技术分类视角和包含内容进行介绍, 在此基础上, 提出一种大数据安全技术框架.

2.1 大数据安全技术的分类视角

对大数据安全技术进行合理分类与组织, 是构建大数据安全技术框架的基础. 当前, 代表性的分类视角主要有基于大数据生命周期的分类和基于大数据系统技术框架的分类.

基于大数据生命周期的分类方式按照生命周期的不同阶段对相关安全技术进行分类 [3~6]. 由于所关注安全需求的不同, 不同研究人员对大数据生命周期的划分和每个阶段涉及技术的理解差异较大. 例如文献 [4] 将大数据生命周期分为数据生成、存储和处理 3 个阶段. 数据生成阶段主要采取访问控制、数据伪造等手段防止数据的非授权采集和隐私泄露; 数据存储安全包括数据加密和完整性验证; 处理安全包括隐私保护数据发布和隐私保护数据挖掘. 文献 [5] 将大数据生命周期分为数据发布、数据存储、数据分析和数据使用 4 个阶段. 其中, 数据发布与分析阶段分别对应于文献 [4] 的数据生成和处理阶段, 而在所包含的安全技术上, 发布和分析两个阶段的安全技术与文献 [4] 处理阶段的安全技术基本一致. 在数据使用阶段, 主要涉及多种访问控制技术.

基于大数据生命周期的分类方式与大数据业务处理流程耦合比较紧密, 对建立大数据安全技术框架有一定的指导意义, 但考虑到在大数据的生命周期中, 一些安全技术可能出现在多个阶段 (例如: 数据在存储、分析、发布等过程中的安全, 都会涉及到访问控制), 建立大数据安全技术框架并不能完全

2) Facebook 数据泄露. 2018. http://opinion.china.com.cn/event_5141_1.html.

按照生命周期的方式. 并且这种方式主要关注数据的安全与隐私, 缺乏对大数据平台安全与安全监管的考虑, 而这也是大数据安全的重要组成部分.

基于大数据系统技术框架的分类方式对大数据系统面临的安全挑战进行归纳, 按照保护对象以及所要解决的安全问题对安全技术进行分类. 该分类方式中, 有代表性的是国际云安全联盟 (Cloud Security Alliance, CSA) [7] 和美国国家标准与技术研究院 (National Institute of Standards and Technology, NIST) [8] 提出的分类. CSA 将大数据安全技术划分为基础设施安全、数据隐私、数据管理以及完整性和被动安全 (reactive security) 4 个方面. NIST 将大数据安全分为设备与应用注册、用户身份与访问管理、数据治理、基础设施管理, 以及风险与追责 5 个方面. CSA 和 NIST 所提分类基本涵盖了大数据与大数据平台所涉及的主要安全技术, 但没有结合大数据系统技术框架对安全技术进行进一步组织或明确阐述不同安全技术之间的关系, 难以很好地指导安全技术在大数据系统中的部署. 阿里巴巴、奇虎 360、IBM 等众多互联网企业也从大数据系统技术框架的角度对安全技术进行了归纳 [9], 但所提安全技术框架主要结合自身业务特点, 通用性不强.

此外, 文献 [10~12] 等介绍大数据安全的综述性文献对大数据安全所需关注的挑战性问题进行了阐述, 但没有提出明确的安全技术分类, 这些文献在学术研究上有很好的借鉴意义, 但在大数据安全技术框架的构建上指导性较弱.

综上所述, 基于大数据生命周期和基于大数据系统技术框架的分类方式各有优点, 但也存在一定局限性. 2.2 小节将综合运用这两种安全技术的组织方式, 并结合 NIST 提出的大数据参考架构, 提出一种大数据安全技术框架.

2.2 技术框架

2015 年, NIST 提出一种大数据参考架构, 将大数据系统参与者划分为数据提供者、数据消费者、大数据应用提供者和大数据框架提供者 4 种角色 [8]. 其中, 应用提供者执行数据的采集、预处理、分析、可视化和访问, 框架提供者提供数据的处理、存储框架和基础设施.

NIST 大数据参考架构在国际国内都有较大的影响力, 国际标准化组织/国际电工委员会下的大数据工作组、我国信息技术标准化技术委员会在建立大数据参考架构时都参考了 NIST 所提架构 [13, 14]. 因此, 本文基于 NIST 架构, 结合大数据业务流程和大数据系统技术框架组成特点, 兼顾数据安全与平台安全、安全防御与安全管理, 提出符合大数据业务特点的安全技术框架, 如图 1 所示.

该框架将大数据安全技术划分为大数据安全共享与可信服务、大数据平台安全和大数据安全监管 3 部分, 在对大数据安全技术进行比较系统的学术归纳的同时, 更重要的是考虑了大数据安全技术框架与大数据系统框架间的耦合. 该框架将大数据安全技术与大数据系统的主要组成部分进行对应, 使得大数据安全需求更加明确, 安全技术的服务对象也更加明确, 有利于指导大数据安全技术的研究与部署.

大数据共享与服务主要面向用户, 为用户提供数据共享与分析的接口. 大数据安全共享与可信服务是大数据安全的根本, 主要解决数据共享与服务过程中的价值激励、安全信任与隐私保护等问题, 具体包括基于数据迁移的数据安全共享、基于计算迁移的多中心协同可信服务和数据服务隐私保护与脱敏. 其中, 数据迁移和计算迁移代表了数据共享的两种方式, 计算迁移采用移动计算而非移动数据的方式实现数据的间接共享.

大数据平台为大数据共享与服务提供数据存储与处理的基础支撑, 负责数据存储、处理和访问等的实际执行. 大数据平台安全是大数据安全的基础, 主要包括大数据处理安全、大数据存储安全、基础设施安全和大数据访问控制. 其中, 大数据存储安全是平台安全的重中之重, 并以密码技术作为核

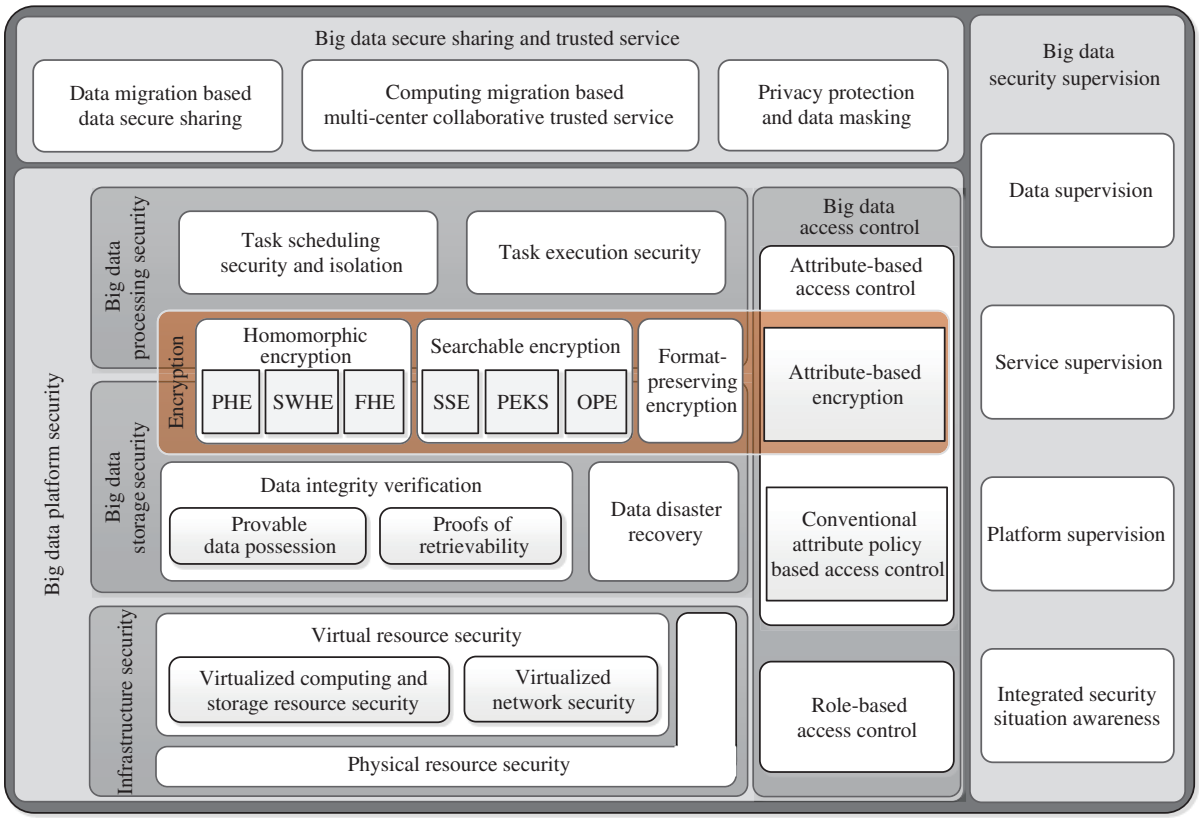

图 1　(网络版彩图) 大数据安全技术框架

Figure 1　(Color online) Big data security technology framework

心. 同态加密、可搜索加密、保留格式加密和属性加密等密码机制不仅能够提供数据机密性保护, 还能为密态数据的统计、分析、搜索和访问控制等提供支持.

大数据安全监管是大数据安全的保证, 主要解决数据自身、大数据服务和大数据平台安全的监控与评估等问题, 具体包括数据监管、平台监管、服务监管和综合安全态势感知.

本文后续章节将依据提出的大数据安全技术框架对各安全技术的研究现状和存在的问题进行详细介绍.

3　大数据安全共享与可信服务

本节首先介绍大数据安全共享与可信服务面临的利益藩篱和隐私泄露等问题, 然后介绍针对利益藩篱等问题的两种数据共享方式: 基于数据迁移的数据安全共享和基于计算迁移的多中心协同可信服务, 最后介绍针对隐私泄露问题的数据服务隐私保护与脱敏.

3.1　面临的问题与解决方案

数据的高度共享与充分利用是实现大数据价值、提升大数据效能的核心目标. 然而由于利益藩篱和数据安全担忧, 在数据共享过程中需要考虑价值激励、责任认定、安全信任等问题, 这些问题的难以解决导致了大数据 “不愿、不敢、不能” 共享的难题. 此外, 鉴于大数据的广泛集成和深度挖掘, 隐私

泄露风险的显著增加也成为数据拥有者共享数据时的一大担忧.

针对利益藩篱和安全信任等问题导致的数据共享难题, 通过技术手段而不是仅凭承诺或管理来解决价值激励、责任认定、安全信任等问题, 对促进数据共享十分必要. 当前主要有两种实现方式. 一种方式通过 "数据迁移", 即数据交易、数据流转或数据交换等, 实现数据的直接共享; 另一种方式无须移动数据, 而是采用 "计算迁移" 的方式实现数据的间接共享. 在这两种方式中, 区块链以其可追溯、不可篡改、去中心化的信任建立等特性, 能在数据安全共享与可信服务中发挥重要作用.

针对隐私泄露问题, 当前研究主要采用隐私保护数据发布、隐私保护数据挖掘和数据脱敏等技术, 确保数据共享过程中与商业秘密和个人隐私等有关的数据不被泄露, 提升大数据服务的可信性.

3.2 基于数据迁移的数据安全共享

建立大数据交易市场进行数据交易是基于 "数据迁移" 实现数据共享的有效方式 [15], 但当前数据市场无法有效保护数据拥有者权益和隐私, 并非完全可信, 并且存在单点故障等问题, 难以满足数据安全共享的需求 [16, 17]. 区块链是一种分布式总账系统 [18], 整合时间戳、非对称加密、共识机制、智能合约等多种技术, 具有可追溯、不可篡改、去中心化的信任建立、多方共同维护等特性. 基于区块链实现大数据交易, 能够从技术上解决激励与价值认可、安全与责任认定、分析与全维共享等问题, 实现安全共享与可信服务.

当前研究人员已经就基于区块链实现去中心化的大数据交易开始了初步探索. Chen 和 Xue [16] 提出一种基于区块链的大数据交易生态系统, 无需任何可信第三方, 不建立数据市场, 生态系统的所有参与方通过基于区块链组成的数据交易网络连接, 并且由部分选定的参与者来制定共识协议、数据质量评估方法和数据交易规则等. Missier 等 [19] 针对物联网数据提出一种基于区块链的数据交易市场, 但该市场并不由任何人拥有, 而是采用智能合约等技术, 以公开、透明、可信的方式制定市场治理和数据交易规则, 但该文并没有明确说明智能合约或市场治理规则由谁制定、如何生成的问题. Nasonov 等 [20] 则提出一种由服务提供商支持的基于区块链的数据市场, 除了数据交换, 还提供数据分析服务. 该文主要关注数据交易的完整性, 基于区块链采用分布式的方式设计了完整性管理器, 以有效维护数据交易完整性, 并且能解决传统数据市场的单点失效问题. Molinajimenez 等 [21] 在肯定区块链有效性的基础上, 指出由于区块链在可扩展性、性能等方面的局限性和应用需求的多样性, 在某些情况下不使用区块链, 而由可信第三方以集中化的方式执行交易是更好的选择. 为充分发挥区块链与可信第三方两者的优势, 提出一种二者结合的数据交易方式, 交易中的部分操作由区块链支持, 部分操作则采用集中式的方式执行.

此外, 在某些行业或领域中, 数据拥有方希望通过数据共享促进该行业的发展, 但由于受到数据管理政策等严格限制, 数据难以有效共享. 区块链能够提供方便、安全、快捷的数据共享途径以解决上述问题. 例如: 针对医疗数据监管严格、数据共享审查周期长等问题, Azaria 等 [22]、Castaldo 等 [23] 基于区块链提出医疗数据的跨机构、跨境共享方案.

综上所述, 区块链是解决价值激励、安全信任等问题的有效方法, 当前基于区块链的数据交易和共享研究刚刚起步, 在是否建立数据市场或引入可信第三方, 如何改进当前数据市场等问题上提出不同的思路. 一方面, 研究人员所提框架、协议等需要进一步完善和测试, 针对多样化大数据的数据质量与价值评估方法 [15]、数据拥有者对数据流通权限的控制与数据确权 [24] 等问题有待进一步研究. 另一方面, 区块链在共识机制和智能合约等自身关键技术中存在安全漏洞 [25], 在大数据服务中引入区块链后产生的安全问题有待进一步探索.

3.3 基于计算迁移的多中心协同可信服务

若数据拥有者不愿采用 “数据迁移” 的方式进行数据的直接共享, 或者由于欧盟《通用数据保护条例》等法律法规的约束导致无法实现数据的直接交换与聚合, 采取 “计算迁移” 的方式实现数据的间接共享, 或者称为信息、知识的共享是实现数据价值的另一种有效途径, 其核心思想是不移动数据, 将原本在数据需求方完成的计算任务迁移到数据所在地完成.

Dong 等 [26] 采用计算迁移的思想, 基于区块链提出一种数据共享模型. 数据提供方在区块链上发布数据信息, 需求方根据自身需求编写计算合约, 数据分析则在数据提供方控制的数据空间内进行, 系统中所有的事务通过区块链进行记录以确保不可篡改和不可伪造. 该模型还采用安全多方计算和差分隐私确保计算隐私和输出隐私, 其中利用多方计算可以适用于多个数据提供方协同为需求方提供计算的场景. 该系统目前仅支持选择、连接和排序 3 种基本运算, 需要根据不同的应用场景及数据类型, 研究高效的计算挖掘算法. Nasonov 等 [20] 提出一种基于区块链的数据市场平台, 除了直接的数据共享, 还提供基于计算迁移的数据分析服务. 该平台的灵活性在于数据拥有者可自行选择将数据存储在私有基础设施或大数据交易平台中, 相应的数据计算任务依据数据所在地进行迁移. 香港科技大学 Yang 教授 [27] 针对不同数据拥有者无法直接共享数据给人工智能发展带来的挑战, 提出一种数据不出本地即可实现数据建模的新的机器学习方法 —— 联邦迁移学习. 联邦学习是指协作各方首先利用自己拥有的数据建立模型, 而后通过加密机制下的参数交换, 建立一个虚拟的共有模型, 主要适用于协作方数据特征相同而样本不同或者样本相同而特征不同的情况. 迁移学习是针对两个存在某种关联或共同点但数据样本与特征均不相同的领域, 利用两个领域间的关联, 将在一个领域已经建立的模型迁移到另一个领域 [28]. 联邦迁移学习的目标是在数据不出本地的情况下, 所建立的模型效果要和把数据真正聚合在一起一样. Yang 教授还提出可以利用区块链共识机制对各参与方的贡献进行评估, 以实现价值激励.

“计算迁移” 是在无法实现数据的直接共享时发挥数据价值的一种有效方法, 当前虽然研究成果较少, 但值得关注并深入研究. 计算结果的真实可信是需要考虑的首要问题, 一方面, 在多中心协同的场景下, 对各中心分析结果的聚合与直接将原始数据聚合分析的结果要尽量保持一致; 另一方面需要采用技术手段防止不诚实的数据提供者或第三方交易平台提供错误结果. 由于将计算任务迁移到数据所在地, 保护被迁移计算模型或算法的安全性以及数据需求方的隐私或敏感信息不被泄露也是一个需要考虑的问题, 例如 Yang 教授在联邦学习中提出参数的交换过程需要加密, 并且要保证不能被反推参与者的特征信息 [27].

3.4 数据服务隐私保护与脱敏

隐私保护问题在大数据时代备受关注 [2]. 采用隐私保护与数据脱敏技术, 是促进数据安全流通与共享、确保大数据服务可信的重要手段. 当前数据服务隐私保护与脱敏主要包括 3 个方面: 隐私保护数据发布、隐私保护数据挖掘和数据脱敏.

3.4.1 隐私保护数据发布

隐私保护数据发布 (privacy preserving data publishing, PPDP) 的核心是在数据发布前对其进行处理, 防止敏感信息泄露, 同时确保数据能够用于分析挖掘 (即可用性). 当前的 PPDP 技术主要包括数据匿名化发布与基于差分隐私的数据发布.

(1) 数据匿名化发布. 依据保护对象的差异, 匿名化技术主要包括结构化数据匿名化、图数据匿名化和位置数据匿名化.

针对结构化数据, 典型的匿名方案是 Sweeney 提出的 k- 匿名 [29]. 该方案将数据集中记录的准标识符值进行泛化、压缩处理, 使得所有记录被划分到若干个等价类, 每个等价类至少包含 k 个具有相同准标识符的记录, 从而实现标识信息的隐藏. 此后研究人员又提出 l-diversity 匿名、t-closeness 匿名等优化方案 [30]. 上述方案主要面向静态数据, 针对大数据发布的动态性, Byun 等 [31] 首次提出支持新增数据重发布的匿名技术, Xiao 和 Tao [32] 提出支持数据插入与删除的 m-invariance 匿名技术, Bu 等 [33] 进而提出支持数据插入、删除与修改的匿名技术. 这些匿名技术能够抵御攻击者联合历史数据的分析与推理.

图数据匿名化技术主要用于社交网络, 不仅要隐藏用户的标识与属性信息, 还要隐藏用户间的关系. 图数据匿名方法主要有基于结构变换的方法和基于超级节点的方法 [34]. 最典型的结构变换匿名方案是子图 k- 匿名, 主要包括 k- 度匿名 [35] 和 k- 同构子图匿名 [36], 分别针对基于节点度数与基于子图信息的身份识别攻击. 上述两种纯结构匿名的方法不能完全保护节点 – 属性的关联, 因此 Yuan 等 [37] 通过在原图中添加噪声节点, 提出一种 "k- 度 -l- 多样性" 匿名模型. 基于超级节点的匿名方案利用超级节点对图结构进行分割与聚类 [38,39]. 其中 Fu 等 [38] 针对现有方法对属性分布与社交结构之间的关联扰动不足的问题, 对节点的属性连接与社交连接进行分割, 以提升节点的匿名性. 此外, 一般图匿名方法都是针对无权图, Skarkala 等 [40] 基于超点和超边的匿名机制提出带权图的数据隐私保护方法, 该方法主要预防用户身份的泄露, 但缺乏对属性信息的考虑.

位置大数据包括位置数据与轨迹数据. 早期的匿名化技术研究主要针对位置数据. Gruteser 和 Grunwald [41] 最先把 k- 匿名的概念引入位置隐私保护领域, 提出了位置 k- 匿名, 确保用户查询的位置是至少包含 k 个不同用户的隐形区域. 此后研究人员又提出了 l-diversity, t-closeness 和 m-invariance 等改进方案 [42]. 然而仅针对位置的匿名技术并不能有效解决轨迹的泄露问题. 轨迹是某个对象的位置数据按时间排序的序列 [43]. 针对任意一条轨迹, 如果在任意时刻, 至少有 $k-1$ 条轨迹在采样位置上与该轨迹处于同一区域, 才称这些轨迹满足 k- 匿名. 轨迹数据 (也适用于位置数据) 的匿名化发布主要应用于两种场景: 一是针对位置服务 (location based service, LBS) 提供商对用户数据的收集分析, 二是用户使用 LBS 时主动查询 [43]. 前者主要针对静态数据, 一种方法是对整条轨迹进行匿名, 寻找在时空上相近的 k 条轨迹形成 k- 匿名集, 另一种方法则是对轨迹的准标识符进行匿名 [44]. 而当用户使用 LBS 服务时, 由于面向动态数据, 匿名技术需要在轨迹开始时就确定 k- 匿名集, 当前主要有基于轨迹划分 [45] 和基于历史轨迹 [46] 的方法 [43].

综上所述, 当前的数据匿名化技术主要是在最初的 k- 匿名方案 [29] 的基础上进行改进. 一方面 k- 匿名方案的隐私保护效果易受到数据分布的影响, 经匿名处理的数据其可用性可能严重下降; 另一方面 k- 匿名方案通常假设攻击者拥有的背景知识, 然而大数据场景下攻击者能够从多种渠道获得未知的背景知识 [47]. 针对不同的应用场景, 结合具体的数据可用性与服务质量需求, 对现有匿名方案的缺陷进行改进, 以实现隐私保护效果与数据/服务质量之间的平衡, 是未来数据匿名化研究需要重点解决的问题.

(2) 基于差分隐私的数据发布. 差分隐私 (differential privacy) [48] 是 Dwork 最先提出的一种建立在坚实的数学基础上的隐私保护方法, 无需考虑攻击者可能具有的任何背景知识. 差分隐私最初被应用于数据库领域, 由于其良好的特性, 目前已广泛应用于社交网络 [49]、位置数据 [47] 等领域中数据发布的隐私保护. 根据数据隐私化处理实施者的不同, 差分隐私可分为中心化差分隐私 (centralized differential privacy, CDP) 和本地化差分隐私 (local differential privacy, LDP), 还有一些研究将本地化差分隐私称为分布式差分隐私 (distributed differential privacy) [50].

中心化差分隐私应用于先将数据收集到数据中心, 而后由数据中心进行隐私化处理的场景. CDP

的代表性方案由 Dwork [48] 提出, 其基本原理是对于两个相差一条记录的数据集 D_1 和 D_2, 在处理函数 F 对它们进行处理后, 采用一个随机函数 M 对 F 的结果添加噪声, 使得 D_1 和 D_2 几乎以相同的概率输出同一结果, 达到任何一个记录是否在数据集中对最终的输出结果几乎没有影响的效果, 从而保护隐私信息不被泄露. 当前 CDP 主要包含 Laplace 机制和指数机制两种噪声机制, 前者适用于数值型数据, 后者适用于非数值型数据 [51]. 基于 CDP 的数据发布根据应用环境不同可分为交互式和非交互式数据发布 [52]. 在交互式环境下, 数据中心在用户发起查询后对数据添加噪声反馈给用户, 研究主要集中在发布机制和基于直方图的发布方法上. 在非交互式环境下, 数据中心针对所有可能的查询一次性发布所有数据. 研究主要集中在批查询、列联表发布、基于分组的发布方法以及净化数据集发布方法上.

本地化差分隐私是针对第三方数据管理者的非可信性提出的, 由用户在本地进行满足差分隐私的数据扰动后, 再将数据发送给收集者. LDP 扰动机制主要包括随机响应、信息压缩和扭曲, 其中随机响应是 LDP 中的主流扰动机制, 其基本原理是针对某敏感问题, 以一定的概率回答自身真实情况, 通过这种不确定性保护个体隐私信息, 而后对收集的数据通过校正函数还原真实的统计情况 [53]. 基于 LDP 的数据发布也可分为交互式和非交互式数据发布 [54]. 交互式框架适用于后续输出结果对之前输出结果有依赖关系的情况, 非交互式框架则适用于前后输出结果无依赖关系的场景. 当前 LDP 主要支持两种数据发布形式: 针对离散型数据的频数统计与针对连续型数据的均值统计 [55], 远不能满足大数据场景下复杂的查询需求.

CDP 与 LDP 在大数据环境下面临一些共同的挑战性问题. 首先, 大数据复杂异构, 数据集的记录间还可能存在关联, 尤其针对图数据, 当前 CDP 和 LDP 方法都不能在确保数据较高可用性的前提下实现隐私保护处理 [54]. 其次, 高维数据的发布由于扰动误差和计算复杂性的增加成为差分隐私方法的一大瓶颈. 当前的解决方法主要是对高维数据进行降维. 例如 Xu 等 [56] 通过随机投影提出了针对 CDP 的数据降维方法. Ren 等 [57] 通过属性划分提出了 LDP 下的数据降维方法, 但没有有效解决高维数据发布的巨大通信代价问题. 此外, CDP 在实际中面临动态的数据发布 (如增量更新), 文献 [58, 59] 提出用于动态数据发布的差分隐私技术, 但面对不断增长的数据量, 存在隐私预算耗尽、累积噪声大等问题.

3.4.2 隐私保护数据挖掘

隐私保护数据挖掘 (privacy preserving data mining, PPDM) [60] 旨在挖掘有价值模式或规律的同时避免敏感数据泄露. 当前 PPDM 技术主要分为两类: 基于数据失真的技术和基于数据加密的技术.

(1) 基于数据失真的 PPDM 技术. 数据失真技术是对原始数据进行扰动, 使攻击者不能发现原始数据, 同时确保失真的数据仍能用于数据挖掘 [61]. 传统的数据失真技术包括随机扰动 (random perturbation)、阻塞 (blocking)、交换 (swapping)、凝聚 (condensation) 等, 但这些方法的有效性缺乏严格证明 [52]. 差分隐私保护技术是一种严格可证明的隐私保护模型 [48]. 当前差分隐私已用于频繁项集挖掘 [62]、分类 (主要包括决策树 [63]、SVM [64] 和回归 [65])、聚类 [66] 等隐私保护挖掘研究中.

面向高维海量数据, 同时确保算法的隐私保护程度和数据的高可用性是大数据挖掘中的差分隐私保护必须解决的问题 [67], 当前研究主要针对特定的挖掘算法进行改进. Li 等 [68] 利用一个频繁项集的所有子集也是频繁项集的性质, 提出了频繁项集挖掘的降维方法, 但存在计算结果发生偏差的问题. Lin 等 [69] 引入动态噪声阈值的概念来解释噪声和数据集大小之间的关系, 使其提出的差分隐私方案更适合大数据分析. 此外, 在将满足差分隐私的数据挖掘算法应用于大数据处理框架的研究上, Roy 等 [70] 提出了将差分隐私和分布式信息流控制集成于 MapReduce 的方法, 并对 k-means 和贝叶斯

(Bayes) 分类算法进行了实验, 但总体上当前将满足差分隐私的挖掘算法应用于分布式计算环境的研究较少.

(2) 基于加密的 PPDM 技术. 基于加密的 PPDM 技术主要依赖于同态加密 (homomorphic encryption) 和安全多方计算 (secure multiparty computation, SMC).

同态加密的特点是对密文的计算结果进行解密, 即可得到对应明文的计算结果 [71]. 将挖掘算法用于同态密文的挖掘能实现原始数据的隐私保护. 当前针对同态密文的挖掘算法包括分类 [72]、聚类 [73]、频繁项集挖掘和关联规则挖掘 [74] 等. 其中文献 [74] 方案允许多个数据所有者通过数据共享共同挖掘关联规则而不牺牲数据隐私, 但 Wang 等 [75] 证明了该方案的加密密钥可以通过连分数 (continued fraction) 算法和欧几里得算法恢复出来. 此外, Gilad-Bachrach 等 [76] 还提出了针对同态密文的神经网络学习方法. 当前, 针对同态密文的数据挖掘研究已逐渐成为一个热点问题, 并主要致力于数据处理能力、效率与安全性的提升.

安全多方计算是指两个或多个参与者共享自己的数据进行联合秘密计算, 该计算方式确保各个参与者只能得到既定的输出结果, 参与者的任何私有信息不会被泄露 [77,78]. SMC 能够用于分布式大数据环境下隐私保护的数据挖掘, 包括隐私保护的分类 [79]、聚类 [80]、集合交集计算 [81] 等. SMC 是实现隐私保护数据挖掘的有效手段, 但 SMC 协议的安全执行也面临外部攻击者或不诚实的内部参与者的威胁, 通过与同态加密相结合, 实现密文的多方计算, 能进一步提升 SMC 协议的安全性 [82]. 全同态加密 (fully homomorphic encryption, FHE) 支持对密文任意的加法和乘法操作, Asharov 等 [83]、López-Alt 等 [84] 分别利用门限 FHE 方案和多密钥 FHE 方案构造了 SMC 协议, 但实用性上受到 FHE 的效率制约. 因而 Peter 等 [85]、Damgard 等 [86] 分别提出了基于单同态 (partial homomorphic encryption, PHE) 和类同态 (somewhat homomorphic encryption, SWHE) 的 SMC 协议, 以提升其在某些特定计算任务中的实用性.

综上所述, 差分隐私、同态加密和多方计算等是隐私保护数据挖掘采取的主要手段, 当前研究虽取得了一系列成果, 但大数据环境下, 数据规模急剧增长, 数据类型和数据分析任务多样化、复杂化, 支持高维数据、流数据、图数据等的安全高效 PPDM 算法以及 PPDM 算法在大数据处理框架下的实用化需要进一步研究.

3.4.3 数据脱敏

数据脱敏 (data masking) 是对数据中包含的敏感信息进行标定和处理, 以达到数据变形的效果, 使得恶意攻击者无法从已脱敏数据中获得敏感信息. 数据脱敏与隐私保护并不完全一致. 隐私是一种个人信息, 强调数据与个体的关联, 而敏感数据并不局限于个人、集体或其他特定对象, 而是强调数据的敏感性, 并且数据脱敏与隐私保护在应用场景和采用的技术等方面存在一定差异. 例如数据脱敏经常需要对非结构化文档中的敏感字词进行脱敏处理, 但这并不属于差分隐私等隐私保护方法的研究范畴.

(1) 敏感信息标定. 在面向大的数据量和访问量时, 数据脱敏需要采用自动化的方式进行. 因此敏感信息标定要在实际的脱敏处理前识别数据对象中的敏感信息并对其位置进行记录 [87]. 自动化敏感信息识别需要生成不同领域的敏感知识库, 仅依靠人工定义的方式会造成遗漏, 采用机器学习方法从已经标定了敏感信息的数据对象中分析可能的敏感信息并添加到知识库中是一种有效的弥补方法 [88].

进行敏感信息识别首先要对具体的数据对象格式 (如 PDF 文档、JPG 图像等) 进行解析, 识别数据对象内容, 而后根据知识库对敏感信息进行匹配. 文档内容通常采用自然语言描述, 单纯的字符串匹配不能完全识别文档中的敏感信息, 采用自然语言处理技术和基于语义特征的文档分类算法, 能够

提升文档内容识别能力, 并通过将文档划分到一个或多个领域, 实现特定领域的敏感信息识别而非全局匹配, 进而提升识别效率 [87]. 图像中的敏感信息包括敏感图像与文字, 通用的敏感信息识别采用形状、纹理、颜色等特征进行, 并且可以采用机器学习算法自动学习敏感特征. 针对文字的结构特点, 还有基于连通域、基于边缘的敏感文字识别方法 [89]. 此外, 在实际应用中敏感知识库是动态更新的, 建立一种低开销的标定更新机制也十分必要.

(2) 数据脱敏. 按照应用场景的不同, 数据脱敏处理的实现可以分为静态数据脱敏 (static data masking, SDM) 和动态数据脱敏 (dynamic data masking, DDM) [88]. SDM 主要用于开发、测试等非生产环境下对数据的脱敏, 防止开发人员对数据的滥用; DDM 则用于生产环境, 当低权限用户访问包含敏感信息的数据时, 对数据进行实时脱敏. SDM 与 DDM 的区别在于是否在使用敏感数据时才脱敏, 因此 DDM 更适用于大数据服务场景中用户访问数据时的敏感数据共享与保护, 同时对脱敏技术的时效性要求更高.

按照脱敏数据能否恢复为原始数据, 脱敏技术分为可恢复类和不可恢复类 [88]. 针对文本类数据, 可恢复类技术主要采用加密机制实现, 典型代表是保留格式加密 (format preserving encryption, FPE) [90], 其特点是密文与明文具有相同的格式, 能有效应用于对数据存储格式有严格要求的存储系统. 不可恢复类技术主要包括采用虚构的数据或特殊字符替换敏感信息、针对数值型数据的混洗 (shuffle) 和均值化等. 针对图像数据的脱敏通常要求具有不可恢复性, 当前虽不乏针对图像进行模糊化处理、添加马赛克的技术, 但在防止恶意人员恢复出原始图像等方面的可靠性有待进一步验证.

当前, 专门针对大数据服务的数据脱敏已经引起 IBM, Informatic 等公司的关注, 并逐渐成为补充其安全技术体系的重要方向 [91]. 在大数据服务中, 面向多样化数据和海量用户请求, 单一的脱敏方法显然无法满足需求, 建立支持结构化和非结构化数据、支持 SDM 和 DDM 的综合性脱敏系统, 以自动化的方式高效、可靠地去除数据中的敏感信息, 同时控制对大数据正常业务的影响, 是一个十分复杂但亟待解决的问题 [92].

4 大数据平台安全

本节主要介绍大数据平台安全的 4 个方面: 大数据处理安全、大数据存储安全、基础设施安全和大数据访问控制.

4.1 大数据处理安全

大数据处理框架定义了大数据计算和处理的方式, 为大数据应用提供必要的基础软件支持. 为满足从批量大规模数据处理到近实时 (near real time) 数据处理的广泛需求, 大数据平台通常需要集成多种处理框架. NIST 从用户视角把大数据处理框架分为批处理、流处理和交互式处理 3 类 [93]. 典型的大数据处理框架包括 MapReduce, Storm 和 Spark 等, 这些框架在得到广泛应用的同时, 由于其最初设计缺乏安全方面的考虑, 面临非授权访问、信息泄露等诸多安全威胁 [94]. 因此, 如何在充分发挥大数据处理平台核心功能和效能的同时, 保证处理任务调度与执行的安全、实现处理结果可信是大数据处理安全面临的主要问题.

4.1.1 任务调度安全与隔离

任务调度安全是指任务调度除考虑吞吐量、资源利用率等性能因素外, 还需要考虑任务执行过程中的数据安全需求以保障大数据处理安全. 任务隔离则是考虑到共享环境下不同租户的任务在相同计

算节点中执行时, 为防止恶意用户非法访问其他用户的数据或任务信息, 需要对多租户任务进行隔离.

在任务调度安全的研究上, Zhang 等 [95] 设计了混合云架构下防止敏感信息泄露的 MapReduce 任务调度机制. 首先对文件中包含的敏感数据进行标定或者依据文件的访问权限将整个文件标定为敏感数据, 在任务执行时, 将去除敏感信息的数据外包给公有云, 敏感数据则总是在私有云中处理. 该方法仅考虑了原始数据集的敏感度, 并且所有 Reduce 任务都在私有云中进行, 没有充分利用公有云的计算资源. Zhang 等 [96] 进一步提出一种混合云下的 Tagged-MapReduce, 为数据集中的每个 key/value 对设置敏感度标签, 能够处理 Map 或 Reduce 任务中间输出结果的敏感度, 并且能够将不包含敏感数据的 Reduce 任务调度给公有云以提升效率. 相比于文献 [95], Tagged-MapReduce 能够支持更加丰富的安全策略和迭代型任务等复杂的 MapReduce 计算任务. Oktay 等 [97] 认为 Tagged-MapReduce 在处理 Map 任务的输出结果时仍不够高效, 通过改进公有云和私有云上 Reduce 任务的执行方式以进一步提升计算性能. 此外, Shen 等 [98] 针对多租户任务的安全隔离, 提出一种基于动态域划分的安全调度策略, 将待调度节点在逻辑上划分为与不同租户作业关联的冲突域、可信域或调度域, 通过调度策略确保存在冲突的多租户任务不会被同时分配在相同计算节点上, 防止存在竞争关系的用户间发生信息泄露或篡改.

针对多租户任务隔离, 除了采用安全调度的实现方式, 也可以基于硬件实现. 现有研究主要运用了 Intel SGX (software guard extensions) 技术. SGX [99] 是 Intel 指令集架构的新扩展, 能够提供程序粒度的可信隔离环境. SGX 允许将合法程序的某些安全操作封装在一个被称为 Enclave 的容器中, 容器内代码和数据的机密性和完整性能得到有效保护. Schuster 等 [100] 提出基于 SGX 的可验证保密云计算框架 VC3. 在该框架下, 用户采用 C++ 语言自行编写 Map 和 Reduce 函数, 并在将它们加密后再上传到云平台. 密钥交换协议、Map 和 Reduce 函数解密以及数据处理等敏感操作在 SGX 提供的内存隔离区完成, 以保证 MapReduce 计算安全. Pires 等 [101] 认为 VC3 方案采用复杂的 C++ 语言容易产生内存非法访问等潜在问题, 考虑到轻量级编程语言 Lua 的代码库较小, 出错概率较低, 并且更易于维护, 因而采用 Lua 语言编写 MapReduce 程序, 提出一种基于 SGX 的轻量 MapReduce 框架. 虽然 SGX 能提供可信的隔离空间, 但仅依赖于 SGX 并不能保证整个任务处理过程的安全 (例如 MapReduce 框架中 Map 节点和 Reduce 节点的通信安全 [102]), 同时 SGX 自身也面临侧信道攻击等安全威胁 [103], 加强 SGX 等安全硬件与其他安全机制在大数据处理框架下的有机结合有待深入研究.

4.1.2 任务执行安全

当前任务执行安全的研究重点围绕两方面内容. 一是有效控制任务在执行过程中能够访问的数据, 防止敏感信息泄露或非法访问其他用户数据; 二是基于可验证计算 (verifiable computation, VC) 等理论技术解决远程、非可信环境下的任务执行结果可信问题.

MapReduce, Spark 等主流大数据处理框架目前仅支持比较简单的访问控制机制 (如访问控制列表 ACL) [104,105], 难以满足海量异构数据的多样化访问控制需求. 针对大数据处理框架访问控制机制的改善, Ulusoy 等 [106] 提出 MapReduce 框架下 key/value 对级别的细粒度访问控制, 防止将整个数据集的访问权限授予潜在的恶意任务导致数据集中的部分敏感记录泄露. Preuveneers 和 Joosen [107] 提出了针对 Spark 流处理扩展 Streaming 的 ABAC 方案, 提供了对流式隐含数据定义访问控制策略的能力. 但该方案需要将 RDD 操作链作为一个整体判别其是否满足用户对数据的访问权限, 在实现上十分困难, 并且该方案只针对 Spark 的流数据处理场景. Ning 等 [105] 进一步提出了一种支持不同数据源的细粒度访问控制方法 GuardSpark. GuardSpark 基于 Spark 的声明式编程接口和 Catalyst 可扩展优化器, 实现了细粒度的访问对象识别、标识和访问控制. 通过统一的声明式编程接口与基于规则

的树状图变换, 实现了访问控制与数据源和用户应用代码的解耦合, 以有效地对各种数据源施加集中化控制. 此外, Hadoop 生态中现在也出现了提供访问控制功能的补丁式中间件, 如 Apache Accumulo, Apache Sentry 等, 但它们并不支持 MapReduce 编程模型中复杂的访问控制, 也不适用于流数据处理场景 [105].

可验证计算 VC 是大数据场景下验证分布式计算结果可信性的重要措施. 依据采用的证明系统分类, VC 协议主要包括基于交互式证明系统的 VC 协议和基于论证系统的有预处理的 VC 协议 [108]. 近年来虽涌现出很多新的 VC 方案, 但当前研究成果还很难应用于 MapReduce 等分布式计算框架的远程计算结果验证, 一方面是由于采用复杂性理论和密码学理论构造的 VC 协议性能开销过大, 另一方面是由于大部分 VC 协议不支持远程输入, 即验证者必须处理所有的输入和输出, 无法表示 MapReduce 程序. 当前仅 Pantry 协议 [109] 能够支持远程输入, 具备了 MapReduce 远程计算的验证能力, 但存在不适用于内存密集型程序等缺陷 [108].

当前分布式并行计算结果的可信性验证仍主要采用冗余计算、可信硬件等相对传统的方法, 验证方式包括内部环境验证和外部计算结果验证两种 [110]. 内部验证方面, Wei 等 [111] 提出使用多个副本对 Map 阶段的计算结果进行验证. 在此基础上, Wang 和 Wei [112] 针对多 Map 副本的共谋问题, 在计算模型中引入了 Verifier 角色, 对通过多副本验证的计算结果进行抽样复算. Xiao 等 [113] 则针对计算结果欺骗问题设计了一种可记录的 MapReduce 平台, 利用可信审计节点组记录并验证 MapReduce 各阶段产生的结果. 外部验证方面, Huang 等 [114] 提出一种水印植入和随机抽样相结合的方法, 基于事先插入的水印验证计算结果的完整性. Ruan 和 Martin [115] 利用 TCG (trusted computing group) 可信计算基础设施使用远程证明机制来保证计算结果的完整性. Wang 等 [116] 则提出了跨云的 MapReduce 构架, 利用私有可信云对公有非可信云下的计算结果进行验证.

综上所述, 当前大数据处理安全的研究主要面向 MapReduce, 对于 Spark, Storm 等新兴数据处理框架的安全研究较少, 处理安全机制的研究滞后于不断演化的数据处理框架. 同时, 大数据处理对计算性能的要求使得可验证计算等理论在大规模分布式计算框架下的实用性有待进一步检验和提升.

4.2 大数据存储安全

大数据存储安全是大数据平台安全的重中之重, 主要目标是确保存储数据的机密性、完整性和可用性, 具体实现机制包括数据加密、数据完整性证明和数据容灾备份等.

4.2.1 数据加密

数据加密是确保大数据存储安全的核心技术. 当前研究的热点包括同态加密、可搜索加密、属性加密和保留格式加密等. 这些加密机制不仅能够提供数据机密性保护, 还能为密态数据的统计、分析、搜索和访问控制等提供支持. 其中属性加密主要用于基于属性的访问控制, 将在 4.4 小节中介绍, 本小节重点介绍同态加密、可搜索加密和保留格式加密的研究现状.

(1) 同态加密. 同态加密是一种能够有效解决密文计算的加密机制, 对经过同态加密的密文进行计算, 而后将输出解密, 所得结果与对相应明文进行相同计算得出的结果是一样的. 同态加密的思想最早是在 1978 年由 Rivest 等 [71] 提出, 经历了单同态加密 PHE、类同态加密 SWHE 的发展, 直至 2009 年才由 Gentry 在全同态加密 FHE 上取得突破 [117]. PHE 只支持加法或乘法一种同态运算, SWHE 则只能支持有限次加和乘的同态运算, FHE 可实现任意次加和乘的同态运算, 是同态加密研究的主要方向.

密文噪声的控制是实现全同态的关键问题. 按照 FHE 采用的噪声管理技术, 当前的 FHE 方案可

分为无限层 FHE 和层次型 FHE [118].

无限层 FHE 以 Gentry 基于理想格的方案为代表, 其关键是采用基于同态解密的 bootstrapping 技术, 然而同态解密的计算开销、密钥和密文尺寸过大, 难以实用, 并且 Coron [119] 指出该方案在量子计算环境下已不再安全. 为改进 bootstrapping 技术以提升效率, Ducas 和 Micciancio [120] 提出了一种针对单比特运算的 bootstrapping 方法, 能在 0.5 s 以内完成 PC 机上的密文刷新过程, Chillotti 等 [121] 以 external product 的形式来表示 Ducas 方案的密文, 进一步将时间缩短到 0.1 s 以内. IBM 最新的研究成果提出 [122], 通过减少 bootstrapping 所采用的线性变换中自同构的数量与单个自同构的开销, 将 HElib 库中算法的效率最高提升 75 倍, 但该库当前仍仅面向同态加密的研究人员, 无法满足大数据的时效性要求. 针对基于格的方案描述较为复杂的问题, van Dijk 等 [123] 构造了一个基于整数的更加简洁的 FHE 方案, 仅采用基本的模运算, 其安全性依赖于近似最大公约数问题. 该方案遵循 Gentry 的构造蓝图, 能实现任意深度的同态操作 [124]. 此后, Coron, Cheon 等研究人员在效率、安全性以及能处理的数据类型等方面进行了改进 [125], 但由于基于整数的 FHE 方案仍需要 bootstrapping 技术以实现任意深度的同态操作, 与实际的大数据应用仍有一定距离.

层次型 FHE 寻求无限深度同态操作与噪声管理之间的平衡. 基于 "错误学习" (learning with error, LWE) 的 FHE 方案和基于 NTRU (number theory research unit) 的 FHE 方案在层次型方案中最具代表性, 其中基于 LWE 的方案被认为能够抵抗量子攻击 [119]. 典型的 LWE 方案包括 BV [126], BGV [127] 和 GSW 方案 [128] 等. NTRU 方案的典型代表是 López-Alt 等 [84] 提出的 LTV 方案. BV, BGV 和 LTV 方案都采用模交换技术控制密文噪声膨胀; GSW 方案则采用矩阵近似特征向量构造 FHE, 避免了复杂的模交换技术. 在避免使用模交换进行噪声控制的研究上, Brakerski [129] 还提出利用张量乘积 (tensor product) 技术构造一个标量不变 (scale-invariant) 的 FHE 方案, 即模 q 与初始化噪声 B 的比例保持不变, 从而无需再利用模交换技术控制噪声增长. Fan 和 Vercauteren [130]、Bos 等 [131] 分别基于 BGV 和 LTV 方案提出了标量不变的 FHE 方案 FV 和 YASHE. Lepoint 和 Naehrig [132] 又进一步将轻量分组密码 SIMON 与 FV, YASHE 结合, 以解决大数据应用中庞大的密文扩展问题. 此外, 研究人员在不断改进层次型 FHE 降噪技术的同时, 也在积极扩展同态加密所能支持的数据类型, 以提升实用性. 例如 Cheon 等 [133] 针对当前 FHE 方案主要支持离散空间上 (如有限域) 的精确计算的问题, 基于 Ring-LWE 问题, 提出了一种能够对实数和复数进行近似计算的层次型 FHE 方案, 之后又采用 Gentry 的 bootstrapping 技术, 提出了无限层 FHE 方案 [134].

同态加密由于能实现密文处理, 与其他数据安全或隐私保护技术的结合能进一步促进大数据的安全, 成为当前的研究热点. 当前研究主要包括同态加密在隐私保护数据挖掘、隐私信息检索 (private information retrieval, PIR)、可搜索加密 (searchable encryption, SE) 等领域中的应用. 其中, PIR 系统用于保护查询者的隐私 [135], SE 则用于保护被查询数据的安全. 同态加密在隐私保护数据挖掘中的应用已在 3.4 小节中介绍, 此处不再赘述. 在 PIR 方面, Brakerski 和 Vaikuntanathan [126] 基于 LWE 假设提出了首个支持 FHE 的 PIR 系统. Sunar 等 [136] 则利用 NTRU 方案设计了 PIR 系统. 在 SE 方面, 麻省理工学院 (Massachusetts Institute of Technology) 的研究人员基于 PHE 设计了商用化的密文数据库 CryptDB [137]. Cheon 等 [138] 则基于 FHE 提出了能够同时支持密文搜索与计算的密文处理框架. 此外, 同态加密在基于身份的加密、基于属性的加密等其他密码学原语中也有着丰富的应用 [139].

综上所述, 自 Gentry 在 FHE 上取得突破以来, FHE 在效率、安全性和可理解性等方面取得了很大进步, 但其在大数据平台中的实用化仍面临诸多挑战性问题 [124, 125, 139].

(a) Bootstrapping 技术带来的效率问题严重制约 FHE 在大数据场景下的应用. 虽然层次型 FHE 针对某些特定应用在效率和安全性上已具备一定实用性, 但要实现无限深度的 FHE 以适用于通用应

用场景, 仍要依赖于 bootstrapping 技术. Yagisawa [140] 和 Liu [141] 提出了 “无噪声” 的 FHE 方案, 即不采用 bootstrapping 也能实现无限深度的同态操作, 但 Wang [142] 指出这两种方案并不安全. 因此, 优化 FHE 方案的噪声管理技术、研究安全的无噪声 FHE 方案仍是当前需要研究的重点内容.

(b) FHE 方案的安全性影响其在大数据平台中的应用. FHE 由于自身的延展性 (malleability), 无法达到适应性选择密文攻击下安全性, 当前大部分的 FHE 方案只能证明选择明文攻击安全, 能证明非适应性选择密文攻击安全的 FHE 方案目前还没有实现.

(c) 在大数据场景下, 为应对复杂多样的数据分析任务, 并促进 FHE 在其他数据安全技术中的应用, 能够支持实数和复数运算的 FHE 方案、支持多比特加密的 FHE 方案以及支持不限用户数量的多密钥 FHE 方案等都有待进一步研究.

(2) 可搜索加密. 可搜索加密 (searchable encryption, SE) 主要解决在密文上进行关键词搜索的问题, 按照构造方法可分为对称可搜索加密机制 (symmetric searchable encryption, SSE) 和公钥可搜索加密机制 (public key encryption with keyword search, PEKS). 与 PEKS 相比, SSE 算法的计算开销较小, 但由于数据加解密和关键词陷门生成需要相同的密钥, 更适合于个人数据存储的场景. 而 PEKS 算法中公私钥分离, 更适合于数据共享, 在大数据环境下应用更广泛 [143].

SSE 算法主要是基于伪随机函数构造的. Song 等 [144] 在 2000 年首次提出了基于顺序描述的 SSE 构建方法, 但由于在搜索关键词时要进行全文扫描, 效率较低. 为此, Goh [145] 和 Curtmola 等 [146] 提出了基于索引的 SSE 构建方法. Goh [145] 采用 “文件 – 关键词” 的方式构建索引, 文件更新时效率较高, 但查询效率低, Curtmola 等 [146] 采用 “关键词 – 文件” 的方式构建索引, 查询效率高, 但在文件更新时由于要重建索引, 效率较低. 针对密文的动态更新, van Liesdonk 等 [147] 提出一种新型 SSE 算法, 但为实现快速的索引重建, 产生的服务器与客户端交互开销较大. Kamara 等 [148] 基于 Curtmola 方案, 并结合 “文件 – 关键词” 的索引构建思想, 首次提出了兼顾搜索效率、动态性和安全性的 SSE 算法. 上述介绍的 SSE 算法只支持简单的关键词匹配, 而在大数据场景下, 通常面临复杂的查询请求, 需要对算法的搜索功能进行扩展. Golle 等 [149] 提出了支持多关键词搜索的 SSE 算法, 能够对以逻辑连接词连接的多个关键词进行直接搜索. Cao 等 [150] 提出支持多关键词排序的算法, 以进一步优化查询结果. Li 等 [151] 提出了支持模糊关键词搜索的算法, 能够容忍用户输入在文字或格式上的细微错误或差别, 搜索能力更强. 在安全性优化方面, Chai 和 Gong [152] 提出可验证的 SSE 算法, 以确保搜索结果的正确性与完整性.

PEKS 算法主要是基于双线性对构造的. Boneh 等 [153] 在 2004 年首次提出了 PEKS 算法, 但该算法存在严重的安全问题. Abdalla 等 [154] 提出了 PEKS 方案完美一致性、统计一致性和计算一致性的定义, 指出 Boneh 方案 [153] 仅满足计算一致性, 进而提出一个满足统计一致性的 PEKS 算法. Abdalla 等还提出了通过基于身份的匿名加密方案构造 PEKS 的一般方法. 针对 Boneh 方案 [153] 易受关键词猜测攻击的问题, Xu 等 [155] 通过先在服务器进行模糊关键词搜索, 而后在本地执行精确搜索的方法来抵御该攻击. Chen 等 [156] 提出一种双服务器的 PEKS 方案, 通过两个独立的服务器来执行匹配搜索过程以抵御来自恶意服务器的关键词猜测攻击. 针对 Boneh 方案 [153] 需要安全通道的问题, Baek 等 [157] 设计了在随机预言模型下可证安全的无需安全通道的 dPEKS 方案. 此外, Zheng 等 [158] 提出一种可验证的基于属性的 PEKS 算法, 允许用户根据访问控制策略搜索数据拥有者外包的加密数据并验证搜索结果的正确性. 在提升 PEKS 算法的效率方面, Bellare 等 [159] 提出了采用确定性加密 (即针对同一公钥与明文, 加密得出的密文相同) 实现高效的 PEKS 方案. 然而大规模系统中, 攻击人员可能会得到用户的额外信息, 威胁确定性加密算法的安全性. 因此 Regev [160] 基于格上 LWE 假设设计了一种具有额外输入的确定性加密方案. 在搜索能力的扩展方面, 与 SSE 算法类似, 主要解决多关键

词搜索 [161] 和模糊关键词搜索问题 [155].

上述可搜索加密算法能够实现关键词的匹配搜索, 但不支持区间搜索. 针对该问题, Agrawal 等 [162] 提出保序加密 (order-preserving encryption, OPE), 一种密文能够保持明文顺序的加密方式, 能实现密文的区间搜索. Agrawal 根据明文空间数据的数量 P, 在用户提供的目标分布中随机抽取 P 个数据, 并对它们进行排序, 得到密钥表格 T. 明文空间中第 i 个数据 p_i 的加密密文 $c_i = T[i]$. 由于 T 中数据是经过排序的, 因此密文能够保持明文的顺序. 但该方案没有给出严格的安全性证明, 需要用户根据使用的数据集生成密钥, 并且对数据集更新的支持度也比较差. Boldyreva 等 [163] 首次对 OPE 进行了严格安全定义, 并提出理想安全性的概念, 即除了明文顺序不泄露其他任何明文信息, 但 Boldyreva 方案并没有达到理想安全性. Popa 等 [164] 采用可变密文 (mutable ciphertexts) 技术, 即少量明文的密文会随着时间推移发生变化, 提出了第一个能满足理想安全性的 OPE 方案. 但该方案存储密文的查找树会泄露明文的出现次数, 无法防御统计分析攻击. 针对该问题, Kerschbaum [165] 通过随机化密文提出了一种隐藏明文频率的改进方案. 此外, Boneh 等 [166] 认为 OPE 方案无法实现最佳语义安全, 进而提出了顺序可见加密 (order revealing encryption, ORE), 通过一个专用函数对密文进行比较从而得出对应明文的大小关系, 但密文并不一定保序, 能够提供最佳语义安全性. 根据密文是否有陷门, 当前 ORE 方案可以分为无陷门 ORE [166] 和陷门 ORE [167]. 其中, 陷门 ORE 确保只有拥有密钥的人才可以通过密文比较判定对应明文的大小.

SE 算法自提出以来, 在安全性和效率上已取得了很大进步, 但新的攻击方法也层出不穷 [168], 并且大数据服务面向的搜索请求规模更大且趋于复杂化, 因此构造更加安全高效可验证且具备多样化搜索能力 (如模糊搜索、多关键词搜索、搜索结果排序等) 的 SE 方案以适应大数据应用场景仍是当前需要解决的主要问题 [143,169]. 此外, 在 OPE 方面, 当前的 OPE 方案难以兼顾理想安全性与检索效率, 并且不支持多维场景的高效查询, 对这些问题的研究对于 OPE 的实用化十分重要 [170].

(3) 保留格式加密. 保留格式加密 (format-preserving encryption, FPE), 部分文献也称之为保持数据类型加密 (data-type preserving encryption) [171], 是一种能够保证密文与明文具有相同格式的加密方式, 其设计初衷是确保在数据库中进行数据加密后, 无需改变数据库结构或更改任何应用程序. 随着研究的发展, FPE 也可应用于数据脱敏、网络数据安全传输 [172] 等领域.

FPE 的早期研究主要致力于探索简单问题域上保留格式加密的实现方法, 最早可以追溯到 1981 年的 FIPS74 标准 [173]. FIPS74 提出基于 DES 算法实现 FPE 的设计, 但密文取值范围仅限于数字和英文字母, 且算法的安全强度较低. 2002 年, Black 和 Rogaway [90] 首次从密码学角度研究了 FPE 问题, 并提出了 3 种 FPE 构造方法: Prefix, Cyclewalking 和 Generalized-Feistel. 其中, Feistel 网络是当前 FFSEM, FF1 和 FF3 等典型 FPE 模型设计对称密码部分所采用的主要方法 [172]. 随着 FPE 研究的深入, 研究人员认识到 FPE 问题的复杂性除了表现在保留数据格式上外, 还表现在待解决消息空间的复杂性上, 这种复杂性使得研究人员很难发现广泛适用于不同消息空间的通用解决办法. 当前普遍采取的思路是通过把问题域转换到等价的具有较低复杂度的整数域上, 以解决复杂问题域上的 FPE 问题 [174,175].

当前, FPE 在大数据领域的应用研究相对较少, 一方面是由于 FPE 的效率问题. Cui 等 [176] 提出基于开源大数据处理平台、综合多种 FPE 算法的数据脱敏框架, 在实验室环境下进行了初步验证, 实验表明该框架下 FPE 算法的执行效率仍是亟待解决的主要问题. 另一方面主要是由于 Feistel 网络的安全性. 大多数 FPE 算法的构造基于 Feistel 网络, Feistel 网络的安全性直接影响算法的安全性. 2015 年 Biryukov 等 [177] 提出针对 5 轮 Feistel 网络的一般性选择明文/密文攻击, 能够有效恢复 Feistel 函数的完整描述, 并采用基于循环 (cycle) 的 Yoyo 分析 [178] 推广至对 6 轮、7 轮 Feistel 网络的攻击.

2016 年 CCS 会议上, Bellare 等 [179] 提出了小域 (small domain) 上针对 FPE 的已知明文攻击, 但该方法每次只能恢复一个密文信息, 需要多个调整因子 (tweak), 并且要求解密目标与已知明文之间具有一定关联. 2017 年 CRYPTO 会议上, Durak 和 Vaudenay [180] 提出针对 4 轮 Feistel 网络的一般性已知明文攻击, 并结合 FF3 坏域分离 (bad domain separation) 的设计缺陷, 提出了小域上针对 FF3 的攻击, 该方法仅需两个调整因子, 但修复 FF3 模式的上述缺陷使该攻击失效也并不困难. 2018 年 CRYPTO 会议上, Hoang 等 [181] 进一步改进了对基于 Feistel 网络的 FPE 算法的已知明文攻击, 改善了多目标场景下的摊销复杂度, 能同时解密多个密文, 并且该攻击无需假设已知明文同解密目标之间的关联性. Hoang 等还发现了 FF3 算法在处理奇数长度域时存在的漏洞, 并基于此提升了攻击速度. 此外, 该文还提出了针对两种非 Feistel FPE 算法的攻击方法: 针对 Cisco 公司提出的 FNR 算法 [182] 的明文恢复攻击和针对基于 DTP 结构 (Brightwell 和 Smith 在文献 [171] 中提出) 的 FPE 算法的唯密文攻击. 其中后者已部署在美国数据安全公司 Protegrity 的商业应用中. 上述攻击表明, 当前的 FPE 算法并未达到期望的安全级别. 由于 Durak 和 Vaudenay 提出的攻击方式, NIST 暂时不鼓励 FF3 的使用甚至可能取消对 FF3 的批准[3)].

FPE 在传统关系型数据库上的应用研究较多, 落地的产品也主要针对特定类型的数据. 然而由于大数据结构、类型的多样性, 如何有效综合运用多种 FPE 算法以适应大数据应用场景有待进一步研究 [183], 这需要综合考虑安全、效率等多个方面. 首先, 由于近几年陆续有学者提出针对 NIST SP 800-38G 中算法 [184] 的攻击方法, 设计可行的高安全 FPE 成为一个挑战性问题 [181]. 其次, 大数据环境下, 如何设计高效的编解码算法, 以将 FPE 应用于复杂问题域是一个具有实际意义的问题. 此外, FPE 是一种对称密码, 安全的密钥分配和管理也成为影响其应用的因素.

(4) 密文去重. 数据去重 (data deduplication) [185] 是一种用于消除冗余数据的技术, 能够用于解决大数据存储过程中数据的重复存储问题, 节约存储空间. 但为保证数据的安全性, 数据通常加密存储, 如何解决密文去重问题又成为了新的挑战. 传统的加密技术一般具有语义安全性, 即相同的明文经加密后生成不同的密文, 导致服务器无法检测到数据的重复性从而无法实现数据去重. 为了实现对密文数据的去重, 研究人员提出了支持密文去重的加密原语, 主要包括收敛加密 (convergent encryption) [186] 和消息锁加密 (message-locked encryption) [187], 但这类方法通常只通过数据的指纹信息验证用户要上传的数据是否存储在服务器中, 而并不验证用户是否真的拥有该数据, 因此攻击人员能够通过蛮力攻击等手段获得数据的指纹信息进而从服务器上获取数据. 针对该问题, González-Manzano 和 Orfila [188] 结合所有权证明 (proof of ownership) 机制提出一种基于收敛加密的安全去重方法, 若用户要上传的数据已存在, 服务器则向用户发送挑战, 只有用户确实拥有该数据才能生成正确的应答并获取数据在服务器上的指针, 从而防止上述攻击, 但该方案不具有语义安全性. 此外, 大部分支持密文去重的加密算法在实现其安全目标的同时都会产生较大的计算与通信开销. 当前, 设计具有较低开销、抵抗蛮力攻击并能够达到语义安全性的实用化密文去重方法仍是一个难题 [185, 189].

4.2.2 数据完整性证明

数据拥有者将数据上传至大数据平台, 对数据的实际控制能力被大大削弱, 需要在不取回完整数据的情况下, 高效可靠地验证数据的完整性是否被破坏. 按照是否能够恢复原始数据划分, 当前的数据完整性验证机制可分为数据持有性证明机制 (provable data possession, PDP) [190] 和数据可恢复证明机制 (proofs of retrievability, POR) [191]. PDP 机制主要通过数据块签名验证数据的完整性. POR 机制除能验证数据的完整性外, 还通过纠错码技术提供对损坏数据的恢复, 实现难度更大. 完整性验证机

3) Recent Cryptanalysis of FF3. 2017. https://csrc.nist.gov/News/2017/Recent-Cryptanalysis-of-FF3.

制通常需要满足没有验证次数限制、无需待验证数据的本地副本、无状态验证等条件. 在大数据环境下, 还应当支持动态操作、共享数据验证、公开验证等.

(1) 支持动态操作的完整性验证机制. 大数据平台中, 数据更新频繁, 只针对静态数据的验证机制无法满足实际的应用需求, 因此研究人员提出支持动态操作的验证机制. 如果验证机制能够支持任意的数据插入、删除和修改操作, 则称为支持全动态操作 [192].

针对 PDP 机制, Erway 等 [192]、Ateniese 等 [190] 引入动态数据结构, 分别基于跳表和 Merkle Hash 树提出支持全动态操作的 PDP 机制, 但这两种方法在更新与验证过程中通信开销较大. Shen 等 [193] 提出了一种由双链信息表和位置数组组成的新型动态结构, 能更高效地支持数据动态性. 针对大数据存储中的数据去重问题, Wu 等 [194] 提出一种支持数据去重和动态操作的 PDP 方案.

针对 POR 机制, Wang 等 [195] 基于 Merkle Hash 树提出了支持全动态操作的 POR 机制. Ren 等 [196] 基于 range-based 2-3 树提出一种更高效的支持全动态操作的 POR 方案. 总体上当前支持全动态操作的 POR 机制相对较少.

(2) 支持共享数据的完整性验证机制. 针对大数据环境下共享数据的完整性验证问题, 当前研究主要关注 PDP 机制. Tate 等 [197] 基于可信硬件提出了一种支持共享动态数据的 PDP 方案, 但没有考虑已撤销用户对数据的篡改问题. 针对该问题, Wang 等 [198] 基于群密钥共享的思想提出共享动态数据的 PDP 方案, 但防止已撤销用户篡改数据的管理开销过大. Wang 等 [199] 提出一种基于代理重签名的 PDP 方案, 但该方案没有考虑服务提供商与撤销用户的共谋问题. Jiang 等 [200] 进而基于向量承诺 (vector commitment) 和验证者本地撤销组签名提出了能够抵抗服务提供商和撤销用户之间共谋的 PDP 方案. 但 Wang[201] 认为按照 Jiang 等的方案, 被撤销用户虽然无法与服务提供商合谋获取合法用户的私钥, 但仍能欺骗公开验证方, 篡改数据而不被发现.

(3) 支持公开验证的完整性验证机制. 完整性验证机制有私有验证和公开验证两种方式. 其中, 公开验证引入第三方验证者代替数据拥有者完成验证任务, 帮助用户摆脱繁重的验证负担, 成为近年来研究的热点. 一个好的公开验证方案应当在满足前文所述完整性验证需求的基础上支持用户隐私保护和批量验证.

在针对 PDP 的公开验证方案上, Ateniese 等 [190]、Zhuo 等 [202] 提出基于 RSA 困难问题的支持公开验证的 PDP 机制, 但基于 RSA 问题设计的 PDP 机制通信开销和存储开销较大. Wang 等 [203] 结合基于公钥的同态认证和随机掩码, 提出具有隐私保护功能、支持批量验证的 PDP 公开验证方案, 并且基于 Merkle Hash 树以支持数据动态操作. Shen 等 [193] 采用基于 BLS 签名的同态可验证认证器 (homomorphic verifiable authenticator), 提出一种隐私保护的可公开验证的 PDP 机制, 并提出一种由双链接信息表和位置数组组成的新型动态结构, 能更高效地支持数据的动态操作和批量验证. 考虑到用户为提高数据可用性, 可能选择多个服务提供商来存储数据的情况, Zhu 等 [204] 针对多个服务商协作存储用户数据的场景, 提出支持公开验证和隐私保护的协作 PDP 机制. 该场景下, 用户将数据分块后上传到多个服务提供商, 由可信第三方进行完整性验证. 但该方法只支持静态数据, 并且通信开销较大. Yang 等 [205] 基于双线性对的特性设计了一种更高效的隐私保护的公开 PDP 验证机制, 能够对多个用户存储在多个服务提供商服务器上的数据进行批量验证, 并且支持数据的动态操作. 此外, 针对基于证书的公钥密码体制来实现完整性验证的方案证书管理开销大的问题, Wang [206] 和 Yu 等 [207] 提出基于身份的 PDP 公开验证机制, 并能实现零知识证明. 针对基于身份的方案存在密钥托管的问题, He 等 [208] 又进一步提出无证书的 PDP 公开验证方案.

针对 POR 的公开验证研究相对较少, Wang 等 [195, 209] 基于 BLS 短签名和 Merkle Hash 树设计了支持动态操作和公开验证的 POR 方案, 其中文献 [209] 支持批量验证, 但两者都不具备隐私保护

功能. Zhu 等 [210] 提出了支持零知识证明的公开 POR 方案, 防止被验证数据的内容泄露给验证者, 并通过建立数据分块的索引表以支持动态操作, 但不支持批量验证. Liu 等 [211] 提出一种基于再生码 (regenerating code) 的 POR 方案, 支持公开验证和批量验证, 并通过伪随机函数随机化编码系数实现数据隐私保护, 但不支持数据的动态操作.

综上所述, 随着数据完整性验证机制研究的深入, 能够高效支持数据动态操作、隐私保护、公开验证和批量验证的完整性验证机制仍是当前研究的主要方向. 尤其 POR 机制能实现被损坏数据的恢复, 相对 PDP 具有更高的实用价值, 但当前能够同时满足上述要求的 POR 机制还十分缺乏. 此外, 随着数据共享和多数据中心协作的发展, 支持共享数据验证和跨数据中心验证的方案也是值得深入研究的问题.

4.2.3 数据容灾备份

容灾备份是保证大数据平台高可用性的有效措施, 按照对系统的保护程度, 容灾系统分为数据容灾系统和应用容灾系统. 其中, 应用容灾是对整个应用系统的备份, 实现复杂性极大, 当前研究主要针对数据容灾. 数据容灾的基本思路是 "数据冗余 + 异地分布" [212], 并通常以数据恢复点目标 (recovery point objective, RPO) 和恢复时间目标 (recovery time objective, RTO) 作为容灾系统的评价指标. 其中 RPO 指业务系统所能容忍的最大数据丢失量, RTO 指灾难发生后所能容忍的最长恢复时间 [213].

在数据冗余方面, 当前分布式文件系统 (distributed file system, DFS) (如 HDFS, Amazon S3 等) 通常采用 3 副本的方式来实现数据冗余, 这种方式简单易行, 但存储开销较大. 纠删码 (erasure code) [214] 是一种空间开销更小的容错编码技术, 通常可用一个三元组 (n, k, k') 表示, 其中 $n > k \geqslant k'$. 纠删码将包含 k 个数据块的原始数据 O 编码为 n 个编码块, 通过 n 个编码块中的任意 k' 个即能恢复 O. 但纠删码在 DFS 中应用时, 由于一个文件可能被分割成多个数据块, 这些数据块分布在多个节点上, 对一个文件的容灾与恢复需要多个节点的协作. 在确保容灾能力的前提下, 纠删码的使用面临如何降低数据读取量与传输量, 以提升数据恢复效率的挑战, 在编码实现、数据恢复和数据更新等方面有待进一步研究 [215]. 此外, Xu 等 [216] 针对虚拟机映像数据, 提出一种基于遗传算法的自适应备份策略, 通过合理组合不同的备份策略以最小化备份的空间开销和备份过程中去冗余的时间开销.

在数据异地分布的研究上, Chang [217] 认为采用单一技术或将数据备份到单一目的站点的容灾方案不适用于大数据平台, 因而提出一种采用多种容灾技术 (包括基于 TCP/IP 的技术、快照技术等) 并将数据备份到多个站点的方案, 能够降低数据的恢复时间并达到接近 100% 的数据恢复率. Wood 等 [218] 提出基于云计算和流水线同步复制技术的容灾系统 PipeCloud, 利用云计算丰富的存储资源对自身数据进行异地备份, 以降低容灾成本并缩短灾难恢复时间, 但该方法只能将数据备份到单一目的站点. Zhong 等 [219] 和 Gu 等 [220] 进一步提出基于多云平台的富云容灾模式以提升容灾可靠性. 但 Zhong 等的方法主要针对私有云数据进行容灾, Gu 等也没有分析该类方法在面向大规模的大数据平台时的可行性.

综上所述, 大数据平台具有动态可扩展的特性, 在规模和复杂性上持续增长. 当前大多数容灾策略与实施方法仍是针对较小规模的数据容灾 [221], 针对大数据平台的体系化容灾方案亟待进一步研究 [8].

4.3 大数据基础设施安全

大数据基础设施为大数据平台组件的运行提供所需的计算、存储和网络等资源, 包括物理资源和虚拟化资源. 由于信息系统软件的复杂性导致漏洞及脆弱性不可避免, 大数据基础设施安全面临着极

大挑战. 其主要安全需求是应对资源共享与虚拟化带来的安全威胁, 包括虚拟机和虚拟机监控器的安全、虚拟化网络 SDN 和 NFV 的安全等.

4.3.1 虚拟机和虚拟机监控器安全

基于云计算基础设施构建大数据系统是大数据发展的一大趋势 [8]. 云计算的虚拟化技术在有效提升资源利用率、实现资源共享的同时, 也导致了安全边界弱化、攻击面增大等问题. 当前针对云虚拟化安全的研究主要包括虚拟机 (virtual machine, VM) 安全和 VM 监控器 (VM monitor, VMM, 也称作 Hypervisor) 安全.

(1) VM 安全. VM 安全防护机制包括 VM 操作系统完整性保护、VM 隔离、监控, 以及安全迁移等. VM 操作系统完整性保护主要采用可信计算技术确保操作系统可信启动 [222] 和基于 Hypervisor 进行恶意代码注入攻击防御 [223]. 关于 VM 隔离, 主要关注运行在同一物理机上的 VM 的隔离问题, 当前研究包括针对共享内存的隔离粒度与隔离性能的提升 [224] 和针对层出不穷的 Cache 侧信道攻击方式的防御 [225]. VM 监控主要采用 VM 自省 (VM introspection, VMI) 的方法, 通过 Hypervisor 或其他 VM 来监控和管理目标 VM 的运行状态 [226], 能够对 VM 操作系统完整性保护等发挥重要作用, 但 VMI 技术的实用性如可移植性和性能等有待进一步提升 [227]. VM 动态迁移是确保云服务高可用性的重要手段. VM 迁移中, 一方面需要维护 VM 迁移性能, 另一方面需要防止迁移过程中的数据泄露隐患, 因此确保迁移过程安全可信的同时降低性能开销是当前 VM 安全迁移研究的主要方向 [228].

(2) Hypervisor 安全. Hypervisor 是 VM 的管理核心, 许多 VM 保护措施都是通过 Hypervisor 实施, 确保 Hypervisor 的安全对于虚拟化安全至关重要. 针对 Hypervisor 自身的安全性, 当前研究主要关注 Hypervisor 攻击面的减小和完整性保护 [224].

在减小攻击面方面, Li 等 [229] 针对 XEN 提出通过将 Dom0 从虚拟化层的可信计算基中去除从而减小 Hypervisor 被攻击的可能性. Azab 等 [230] 提出将 Hypervisor 的部分功能从 CPU root 模式的内核层移到用户层, 但该方法只是将 Hypervisor 所受攻击的影响限制在一个域内, 而并未根除. Szefer 等 [231] 则提出去掉 Hypervisor, 利用硬件的虚拟化特性, VM 可直接访问硬件, 但若 VM 需要的硬件资源不支持硬件虚拟化, 该方法无法实行.

在 Hypervisor 完整性保护方面, 完整性度量与验证是当前确保 Hypervisor 不被篡改的主要手段. HyperCheck [232] 和 HyperSentry [230] 是两个 Hypervisor 完整性度量框架, 前者的度量操作要由 Hypervisor 自身触发, HyperSentry 则可隐秘触发度量操作, 防止攻击者隐藏其攻击行为. 这两种框架只能对 Hypervisor 的静态数据或代码进行保护, HyperSafe [233] 通过不可被绕过的内存锁定机制和受限的指针索引, 能够提供全生命周期内 Hypervisor 控制流的完整性保护, 但 HypeSafe 机制的灵活性较差.

综上所述, 云虚拟化安全目前在国内外研究广泛, 并取得了很多成果, 但由于虚拟化软件栈自身的复杂性以及由此产生的脆弱性和攻击方式的多样性, “共享与隔离”、“安全与性能” 之间的矛盾目前尚未得到很好的解决.

4.3.2 虚拟化网络安全

为应对大数据环境下网络架构的可扩展性需求, 以软件定义网络 (software defined network, SDN) 和网络功能虚拟化 (network function virtualization, NFV) 为代表的新型网络虚拟化技术近来发展迅速 [93]. SDN 和 NFV 通过在一个基础网络架构中实现多种异构的虚拟网络, 能够极大地提高基础网络设备的资源利用率, 但也引入了新的安全问题.

(1) SDN 安全. SDN 将网络设备的控制面和数据面分离, 集中控制和开放可编程的特性在提升网络性能的同时也带来新的安全威胁 [234]. SDN 整体架构主要包括应用层、控制层、数据层和南北向接口 (南向接口负责控制层与数据层通信, 北向接口负责控制层与应用层通信). 当前针对 SDN 的安全研究主要以控制层为中心涉及上述 5 个部分和 SDN 整体架构的安全优化.

针对 SDN 北向接口的开放性可能引发漏洞暴露和接口滥用的问题, 需要对应用层设置有效的认证和授权机制 [235]. 通过北向接口的标准化以减少安全漏洞也是目前亟待解决的问题 [236]. 控制器是 SDN 攻击的焦点, 针对控制器的单点失效问题, 研究人员提出采用分布式控制器的解决方法 [237]. DoS 攻击是造成控制器失效的一种常见攻击, 根据交换机流表规则无法匹配时的两种处理方式, 分别有在控制层和数据层进行防御的方法 [235], 并以轻量、高效、准确作为防御方法的主要目标 [238]. 针对控制层与数据层的安全通信问题, 部分 SDN 方案支持 TLS 协议, 但不能完全确保通信安全 [239]. 此外, 在 SDN 整体架构的安全优化上, Varadharajan 等 [240] 提出一种策略驱动的安全体系架构, 用于保护跨多个 SDN 域的端到端服务, 并且专门开发了一种策略语言用于设计 SDN 安全策略. Hu 等 [241] 则提出了一种确保 SDN 网络策略能够安全接收并正确执行的安全框架.

(2) NFV 安全. NFV 通过虚拟化技术采用软件实现网络功能, 摆脱专用硬件的束缚, 具有低成本、弹性灵活的优势 [234]. NFV 架构包含 3 个关键要素: 网络功能虚拟化基础设施 (network function virtualization infrastructure, NFVI)、虚拟网络功能 (virtual network function, VNF) 和管理编排 (management and orchestration, MANO). 当 NFV 公开部署时, VNF 通常外包给第三方虚拟化平台, 如云平台. 共享、非可信和虚拟化的环境为 NFV 的管理编排和安全运行带来很大威胁 [242].

虚拟环境下 NFVI 和 VNF 的安全管理和编排是确保 NFV 安全的一大挑战 [243]. 研究人员提出结合 SDN 来增强 NFV 控制力 [244] 和引入安全编排器以增强 NFV 安全管理 [245] 的方法. 当前的 MANO 主要是集中式的, 研究更加智能、支持配置动态更新的分布式 MANO 是未来需要解决的重要问题 [246]. 针对外包 VNF 的安全运行, 应用于 VM 的安全隔离与监控等保护方法在 NFV 中同样适用 [247]. 还有一些研究人员提出采用 Intel SGX 技术 [248] 或可信计算技术 [247] 为 VNF 提供可信的运行环境. 此外, Melis 等 [249] 提出对外包的 VNF 进行加密以防止非可信环境下的数据泄露.

综上所述, SDN 和 NFV 等新型网络架构不但能提升网络的灵活性和可扩展性, 其在网络安全领域的应用也成为当前研究的重要方向 [250]. 然而, 它们也带来了新的安全威胁. 当前针对 SDN 和 NFV 安全挑战的分析日渐深入, 但在 DDoS 等网络攻击的防御、SDN 和 NFV 自身安全机制的增强以及接口的标准化等方面仍需进一步研究 [246, 251]. 尤其在大数据环境下, 大规模的网络流量和复杂的网络事务, 对 SDN 和 NFV 的性能、可扩展性、智能化和安全性等提出了更高要求 [252].

4.4 大数据访问控制

访问控制是确保大数据分析、数据流转服务等过程中多元异构海量数据安全的重要机制. 在大数据场景下, 数据、应用和用户规模激增, 用户的访问请求复杂多变, 跨数据中心、跨安全域的数据共享越来越频繁, 访问控制面临海量数据的细粒度访问控制和跨域访问控制的挑战. 基于属性的访问控制模型 ABAC 和基于角色的访问控制模型 RBAC 是大数据环境下主要应用的访问控制模型, 针对上述挑战, 当前研究主要围绕基于属性加密的访问控制和角色挖掘展开.

4.4.1 基于属性加密的访问控制

ABAC 包括基于常规属性策略的访问控制和基于属性加密 (attribute-based encryption, ABE) 的访问控制. 在基于常规属性策略的访问控制中, 用户权限和数据都由服务提供商进行分配和管理, 但

服务提供商并不完全可信, 并且用户对数据的每次访问都需要与服务器进行交互, 当面向海量用户时, 将严重影响系统效率. 而 ABE 机制将密文、用户密钥与属性关联, 数据拥有者可根据需要灵活地制定加密策略, 产生的密文只有属性满足加密策略的用户才可以解密, 能够实现针对密文的细粒度非交互式访问控制 [253]. 相比于基于常规属性策略的访问控制, 基于 ABE 的访问控制能更好地应用于大数据访问控制, 当前研究主要涉及 ABE 访问结构设计、属性撤销、可追踪 ABE、多机构 ABE 等.

典型的 ABE 方案包括密钥策略 ABE (key-policy attribute-based encryption, KP-ABE) [254] 和密文策略 ABE (ciphertext-policy attribute-based encryption, CP-ABE) [255]. 在 KP-ABE 中, 数据拥有者对被访问数据的控制能力较弱; 而在 CP-ABE 中, 数据拥有者可以利用属性制定访问控制策略, 因此 CP-ABE 更适合于大数据环境下的访问控制 [256], 但同时也面临访问结构设计的难题. 基于 ABE 的访问控制通过访问结构表示访问控制策略, 策略的灵活性会导致访问结构的复杂化. 而在 CP-ABE 中, 访问结构复杂度的增大又将导致系统公钥设计复杂度的增加, 增大了系统的计算代价和通讯代价, 因此 CP-ABE 中的访问结构设计成为一大难题. 当前访问结构设计主要采用基于门限、树结构和线性秘密共享方案的方法, 但都无法实现访问结构不受限制且基于标准复杂假设可证安全的 CP-ABE 方案 [253, 257]. 此外, 由于 CP-ABE 中访问结构与密文相关联, 数据解密时存在访问控制策略泄露的风险, 因此 Kapadia 等 [258] 提出了策略隐藏的 CP-ABE 方案, 但现有方案无法在保证访问结构表达能力的同时确保算法效率 [259].

针对用户密钥泄露或用户权限变更等情形, 属性撤销是基于 ABE 的访问控制必须要考虑的问题, 但属性撤销将引起密钥的撤销和相关数据的重新加密, 产生很大的计算开销. 大数据环境下大规模的用户和属性则进一步加剧了该问题. 因此高效的属性撤销是 ABE 中一个重要的研究方向. 按照撤销执行者的不同, 当前主要有直接撤销和间接撤销两种方式 [257]. 直接撤销由数据拥有者在加密数据时规定撤销用户列表, 但当前的最新研究仍主要针对用户级的撤销 (即撤销用户的全部权限而不是只撤销部分属性) [260, 261], 缺乏有效的属性级撤销方案, 并且当用户及属性量很大时, 用户权限列表的管理负担非常大, 难以在大数据场景下应用. 间接撤销由属性授权机构 (attribute authority, AA) 执行, 能够实现属性级撤销. 早期方案主要由属性授权机构采用周期性的方式撤销 [262], 但授权机构在密钥更新时的工作量随着用户量的增长急剧增长, 并且撤销具有滞后性. 为减轻授权机构工作量并实现实时属性撤销, 研究人员引入第三方机构执行属性撤销, 而针对第三方机构并非完全可信的问题, 当前研究主要结合代理重加密和延迟重加密 (lazy re-encryption) 实施属性撤销. 其中仅采用代理重加密的方法会给代理服务器带来较大开销, 配合使用延迟重加密即在用户下一次访问数据时再进行重加密是降低计算和通讯开销的有效方法 [263, 264].

可追踪 ABE 主要是针对 ABE 中的密钥滥用问题提出的. 在 ABE 中, 用户私钥只与其属性相关, 而不包含任何用户特有的信息, 由于拥有不同属性的用户通过非法合谋获得新的属性、恶意授权机构非法分配密钥甚至恶意用户私自泄露密钥等原因都可能导致密钥滥用 [257]. 研究具备可追踪性的 ABE 机制, 实现密钥滥用的追责, 是推动基于 ABE 的访问控制在大数据环境下广泛应用所必须解决的问题. 可追踪 ABE 的主要思路是在用户的解密密钥中嵌入用户特有的标识信息以提供可追踪性, 实现方式包括白盒追踪和黑盒追踪 [265]. 白盒追踪将事先设计好的密钥作为追踪算法的输入, 进而追踪到设计该密钥的用户. 而黑盒追踪仅需要给定解密设备, 无需知道其中封装的解密密钥甚至解密算法, 通过向该解密设备提供只有具有可疑标识的用户才能解密的密文, 实现对构造该解密设备的用户的追踪 [266]. 可追踪 ABE 已引起研究人员的重视, 当前研究主要致力于在确保可追踪性的基础上实现对丰富的访问策略、无限集合 (large universe) 属性和用户撤销的支持 [267, 268].

大数据场景下, 单授权机构 (AA) 的 ABE 方案无法满足大规模分布式应用对不同机构协作的需求, 巨大的管理负担使 AA 成为系统的性能瓶颈, 并且 AA 还易于受到集中攻击, 因此研究人员提出了多 AA 协同合作的 ABE 机制 [253]. 根据多个 AA 之间是否采用中央授权机构 (central authority, CA) 来确保解密的正确性, 当前研究可分为采用 CA 的多机构 ABE [269] 和无 CA 的多机构 ABE 两类 [270], 在安全性方面需要考虑 AA 被腐化以及用户合谋恢复 AA 密钥的问题. Chase [269] 最先提出采用 CA 的多机构 ABE, 通过利用多个独立 AA 分别管理用户一部分属性的方式解决单个 AA 被腐化的问题, 并采用用户全局唯一标识 (globally unique identifier, GID) 防止用户间的合谋, 但该方法需要保证 CA 完全可信. 为避免采用 CA 带来的安全脆弱性, 研究人员提出无 CA 的多机构 ABE, 但早期方案只能支持有限集合 (small universe) 的属性 [257]. 2015 年 Li 等 [271]、Rouselakis 和 Waters [272] 使用素数阶双线性群分别构造了支持无限集合属性和单调访问结构的无 CA 多机构 KP-ABE 和 CP-ABE 方案, Zhang 等 [268] 进而提出支持白盒追踪的 CP-ABE 方案, 但这些方案都没有考虑用户撤销的问题.

综上所述, ABE 的效率和安全性是影响其实用性的主要因素. 在效率方面, 除了访问结构, 密文和密钥长度、加解密算法、支持的属性集合也是影响计算效率和通讯效率的重要因素 [253]. 在安全性方面, 仅以得到可证明适应性安全方案为目标是不够的, 还需要考虑用户可追踪、可撤销等多样化安全需求. 此外, 针对大数据场景下分布式应用的协作需求, 多机构 ABE 也是需要考虑的重要问题. 综合考虑上述因素, 当前基于 ABE 的访问控制的研究成果距离大数据环境下的实际应用还有一定距离. 具备丰富表达能力, 支持更短密文、密钥和无限集合属性, 可追踪、可撤销、多机构的安全高效 ABE 方案需要进一步研究 [273, 274].

4.4.2 角色挖掘

由于大数据系统的复杂性, 角色设计复杂化是大数据环境下应用 RBAC 模型面临的主要难题. 当前研究主要采用角色挖掘的方法实现角色的自动化生成. 角色挖掘 [275] 用于解决如何产生角色, 并建立 “用户 – 角色”、“角色 – 权限” 映射的问题, 在已有 “用户 – 权限” 关系的基础上, 利用数据挖掘 [276]、机器学习 [277] 等手段实现自动化角色定义和权限管理, 以缓解 RBAC 模型在大数据应用中存在的过度授权或授权不足现象 [278]. 根据候选角色集合生成结果的不同, 角色挖掘主要包括仅生成候选角色的方法和能够同时生成角色集合与完整角色状态的方法 [279].

在仅生成候选角色的方法中, Kuhlmann 等 [276] 采用聚类和关联规则挖掘的方法, 首先利用聚类算法对用户分类, 每类用户对应一种角色, 而后利用关联规则算法挖掘角色对应的权限. 该方法预先定义生成的集群数量, 缺乏角色生成后的有效选择方法, 并且没有建立角色的层次结构. Vaidya 等 [280] 指出采用聚类方法实现角色挖掘, 一种权限只能被授予一种角色, 但实际应用中一项权限可能被赋予多个角色, 因而提出基于枚举的角色挖掘算法, 识别系统中所有可能的角色集合, 最后按照包含的用户数量对枚举出的集合进行排序和选择. Zhang 等 [281] 则采用频繁模式挖掘的方法, 在当前 “用户 – 权限” 的分配关系中识别经常被一起分配的权限集作为候选角色. 上述仅生成候选角色集合的方法实现相对简单, 但缺乏对生成角色优劣的有效度量.

角色状态不仅包含角色集合, 还包括 “用户 – 角色”、“角色 – 权限”、“用户 – 权限” 的分配关系以及角色间的层次关系 [279]. 能够生成角色状态的方法通常将角色状态与原有 RBAC 配置的一致性、系统复杂性、角色的语义等作为生成角色优劣的度量指标. 在角色状态与原始 RBAC 配置的一致性上, 很多角色挖掘方法都能够满足 [282]. 在系统复杂性上, Vaidya 等 [283] 将生成角色的数量作为衡量指标. Zhang 等 [284] 采用图论方法, 将角色挖掘问题转化为图的优化问题, 以 “用户 – 角色”、“角色 – 权限” 的分配关系作为图中的边, 将角色与边的数量之和作为系统复杂度的衡量指标. Molloy

等 [282] 则提出更一般的度量方法, 将角色状态中所有关系的总和作为复杂度衡量指标, 并将不同种类的关系赋予不同权重. 在角色的语义意义上, Molloy 等 [282] 提出在已知信息只有 “用户 – 权限” 分配关系时, 引入形式化概念格 (formal concept lattices), 采用形式化概念分析挖掘有语义意义的角色; 如果还能获得用户的属性信息, 则通过创建能被用户属性表达式解释的角色, 确保角色的现实意义. Frank 等 [285] 结合已有的 “用户 – 权限” 分配关系以及与角色挖掘相关的业务信息, 采用概率模型推断最可能的权限分配关系, 该方法确保产生的角色集合在业务上是可解释的. 此外, Jafarian 等 [286] 提出一种通用的角色挖掘方法, 将角色挖掘问题转化为约束满足问题, 能够支持自定义角色度量指标, 进而求解该度量下最优的角色挖掘结果.

综上所述, 角色挖掘是解决 RBAC 模型在大数据场景中面临挑战的有效方法. 当前, 如何充分利用大数据环境下丰富的数据资源和强大的计算能力, 解决带噪声数据的挖掘、具有语义意义角色的挖掘等问题还需进一步研究 [278, 287].

5 大数据安全监管

大数据安全监管主要包括数据监管、服务监管和平台监管. 其中数据监管在方法、技术上和传统安全监管有着显著不同, 基于数据世系的数据安全监管将成为未来的重要发展方向. 服务与平台安全监管则面临海量异构的安全监管数据带来的诸多挑战, 引入大数据技术实现安全态势准确掌控和威胁快速发现成为当前研究的热点问题. 因此, 本节主要介绍基于数据世系的数据安全监管和基于大数据技术的服务与平台安全监管.

5.1 基于数据世系的数据安全监管

大数据应用的核心是数据价值的挖掘, 如果数据在生成、处理过程中的安全性无法保证, 将会直接影响到大数据服务的可信性. 因此, 在大数据场景下, 实施数据监管, 解决识别和排查不可信数据源、跟踪并诊断数据安全威胁等问题十分必要 [288].

数据世系 (data provenance) 描述了数据从初始产生到演进的过程, 包含了给定数据对象的源数据, 以及该数据对象的演进所经历的处理过程 [289]. 数据世系能够应用于数据质量分析与评估、程序调试、科学分析过程的可重复性保证、数据访问权限衍生等, 同时也是解决数据监管问题的有效途径 [290]. 传统数据世系方法主要面向关系型数据库或工作流系统 [291]. 近年来, 随着大数据的发展和数据价值利用需求的快速增长, 大数据世系逐渐得到研究人员的关注 [292]. 为充分发挥世系的数据监管作用, 首先需要解决大数据场景下的世系模型构建, 世系数据的采集、存储、融合和查询, 以及世系数据自身的安全保护等问题. 然后基于世系数据的分析为数据监管提供支持.

世系模型以形式化的方式定义世系中包含的要素、要素间的关系, 以及施加在这两者上的规则以有效表达数据的演进过程, 为世系数据的采集、组织和分析等过程提供指导, 并为世系数据的充分共享和利用提供支撑 [293]. 自世系提出以来, 研究人员对世系模型不断丰富 [294], 当前的代表性工作是 Moreau 等 [293] 提出的开放世系模型 (open provenance model, OPM). 该模型定义了 3 类节点 (artifact, process 和 agent) 和节点间的 5 种关系 (used, wasGeneratedBy, wasControlledBy, wasTriggeredBy 和 wasDerivedFrom) 以描绘数据的演进过程. 然而, OPM 等模型没有考虑数据类型和数据处理模式的多样化等大数据特点, 对这些模型在大数据技术框架下的扩展是一个亟待解决的问题 [295].

在世系数据的采集、存储、融合和查询中, 世系采集是研究人员的主要关注. 在大数据场景下, 世系数据本身也是一种大数据, 若不加选择地采集所有能够得到的世系数据, 将为系统带来巨大的时间

和空间开销. Gehani 等 [296] 认为根据目标或需求确定世系采集的覆盖范围、抽象级别和时间粒度等是控制系统开销的有效方法. 例如 Fu 等 [297] 在云数据世系系统 Progger [298] 的基础上, 针对 Hadoop 设计了满足数据泄露取证需求的世系系统 HProgger, 通过减少所需记录的数据类别, 提升世系系统性能.

在世系采集的实现方面, 依据所面向对象的不同, 当前研究主要涉及在大数据存储、处理组件中的世系采集. 在大数据存储组件方面, Kulkarni [299] 提出了在非关系型键值存储系统中采集世系数据的两种方式: 直接改造数据库本身以支持世系采集和建立第三方世系采集系统集成到数据库中. Alkhaldi 等 [300] 采用第 2 种方式设计了针对 NoSQL 的元数据系统 WASEF 进而支持数据世系. Chacko 等 [301] 则通过跟踪 NoSQL 自身记录的日志捕获世系数据. 在大数据处理组件方面, 当前研究主要针对 MapReduce 框架. Park 等 [302] 基于包装器 (wrapper) 采用避免对 MapReduce 框架进行改造的方式, 设计了细粒度世系数据捕获模型 RAMP, 该模型对世系数据的捕获和处理都是在 MapReduce 任务执行过程中, 性能开销较大. 针对 RAMP 性能问题, Akoush 等 [303] 通过改造 Hadoop 将世系追踪集成为 MapReduce 框架的内在特征, 并通过将世系图的构建推迟到世系查询阶段以提升效率.

由于世系数据所包含信息的敏感性以及世系在数据监管等应用中发挥的作用, 保护世系数据的安全性是世系研究的重要问题之一 [304]. 研究人员在世系数据安全的形式化定义、世系数据的加密、签名、访问控制和隐私保护等诸多方面进行了研究 [305]. 其中, 由于世系数据的特殊性, 针对世系数据的访问控制与传统访问控制方法有较大差异, 是当前研究的重点. 世系数据表达了源数据到目的数据的演变关系. 这种数据演变关系以及参与其中的实体构成一个有向无环图 (directed acyclic graph, DAG). 世系数据的保护不但要保护数据项本身 (DAG 中的点), 还要保护数据项之间的演变关系 (DAG 中的边或子图). Braun 和 Shinnar [306] 针对 DAG 的访问控制提出了边模型和节点模型两个相互独立的访问控制模型, 这种方式简化模型的构建, 便于理解和实施. Cadenheda 等 [307] 则提出了支持世系子图的世系数据访问控制策略语言. Danger 等 [308] 在 Cadenheda 工作的基础上采用世系过滤技术解决了 Cadenheda 工作无法选择性地公开世系子图的问题. 上述研究主要处于理论层面, 考虑到大数据环境下世系数据规模的增大和世系图结构的复杂化, 世系图访问控制策略的正确性、策略冲突检测等问题需要进一步研究. 此外, 利用区块链的不可篡改、多方共同维护等特性, 实现可信的世系数据收集、管理和验证成为世系安全领域近年来一个新的研究方向 [309, 310]. 这类方法中, 世系数据的安全性在很大程度上依赖于区块链智能合约、共识机制等的安全性.

在基于世系的数据安全监管方面, Muniswamy-Reddy 等 [311] 通过具体案例说明从操作系统和文件系统获取的世系数据能够用于监测对数据进行的操作是否按预期执行, 为取证追责提供证据. Suen 等 [312] 提出一种分布式环境下以数据为中心的事件记录机制, 通过将在文件和数据块级别捕获的数据事件链接起来形成数据事件序列, 描述数据在整个生命周期中的世系记录, 进而基于世系数据检测内部威胁、数据泄露等安全事件. Alabi 等 [313] 则提出基于世系数据可以检测 Hadoop 集群中的数据泄露, 但没有提出预测性的检测方法. Bates 等 [314] 针对数据库, 提出利用世系数据阻止基于 SQL 注入的数据泄露攻击的方法, 该方法通过查询被请求数据的祖先是否在白名单 (或黑名单) 中, 以判定是否允许数据的传输. 此外, 除了世系自身的监管功能, Appelbaum [315] 提出日志等其他审计证据的世系数据能为这些审计证据的可靠性提供支撑.

综上所述, 当前大数据世系的研究虽然取得了一些成果, 但仍有很多问题有待进一步研究, 具体体现在 [290, 316, 317]: (1) 大数据的异构特性导致对世系数据进行统一建模和结构化描述更加困难; 不同世系系统产生的世系数据间的交互融合面临多种世系格式的兼容性问题; (2) 大数据的海量特性使得世系数据的采集、传输和存储加重了大数据平台在计算、通信和存储方面的负担, 一些研究人员提出采

用世系压缩的方法降低存储和传输开销, 但数据的压缩和解压也带来新的计算开销 [288]; (3) 大数据平台的封装和透明化, 增加了世系数据采集的实现难度, 尤其对流数据的世系采集、存储等研究仍处于起步阶段, 最新研究成果较少; (4) 世系数据的安全已引起研究人员关注, 但所提方法在大数据场景下的实用性有待进一步验证和提升; (5) 目前基于世系的数据安全监管研究大多只分析了世系数据用于数据监管的可行性并通过案例进行说明, 但缺乏基于世系检测数据泄露等安全威胁的自动化方法.

5.2 基于大数据技术的服务与平台安全监管

大数据服务与平台安全监管通过融合处理大数据服务和平台的运行状态数据、日志数据和网络数据流等多源安全数据, 及时发现服务与平台面临的安全威胁, 为系统监管和态势掌控提供支撑. 大数据环境下, 平台与服务的监管和分析使得传统威胁分析方法的数据采集、融合和分析模式遭遇瓶颈, 难以快速发现安全威胁并精准掌控安全态势. 与此同时, 各种外部和内部攻击手段不断升级, 对安全监管系统的威胁发现和及时响应能力提出了更高要求. 针对上述问题和挑战, 引入大数据技术成为安全数据分析和威胁发现的重要方向 [318]. 当前研究主要针对大数据服务与平台面临的外部网络攻击和内部威胁等典型安全威胁展开, 网络攻击检测又包括基于大数据技术的网络入侵检测系统 NIDS 构建和针对特定网络攻击的检测.

(1) 基于大数据技术的 NIDS 系统构建. 以 Hadoop 生态圈为代表的大数据平台, 能够为海量数据的存储、分析提供有力支撑 [318,319]. 当前基于 Hadoop 的 NIDS 构建已取得一定成果. Jeong 等 [320] 针对网络流量的爆炸性增长给 NIDS 系统数据处理带来的挑战, 提出在 Hadoop 框架下建立 NIDS 系统的构想. Cheon 和 Choe [321] 基于 Snort 和 Hadoop 搭建了 8 工作节点的分布式 NIDS 框架, 以提升 Snort 警报信息的处理效率. Baker 等 [322] 基于 Hadoop 构建了网络安全监控平台 PacketPig, 能够在捕获所有数据包的情况下开展深度数据包检测和深度网络分析. Rathore 等 [323] 针对高速网, 在 Hadoop 之上使用 Spark 执行高速网络流的实时分析以实现实时的 NIDS 系统. Marchal 等 [324] 针对多源网络数据构建了分布式数据关联分析系统, 在 4 种离线数据处理场景下, 对比了 MapReduce, Hive, Pig, Spark 和 Shark 5 种数据处理框架的性能差异, 其中 Spark 和 Shark 的表现最好.

(2) 基于大数据技术的网络攻击检测. APT 攻击是针对大数据系统的最具代表性的网络攻击之一. 由于 APT 检测的一大挑战就是要对大量多源异构的数据进行长期分析, 大数据分析技术特别适合于 APT 攻击的检测 [318]. AT&T 研究人员 [325] 基于 MapReduce 建立了大规模分布式计算框架, 采用基于签名、异常或策略的多种检测算法, 高效地处理具有长时间跨度的多种数据 (包括系统日志、NIDS 和防火墙监控数据等) 以进行 APT 检测. Bhatt 等 [326] 也利用 Hadoop 对从 NIDS 等多个数据源收集的数据进行存储和关联分析, 并基于 Intrusion Kill Chain 模型 (描述了攻击者规划和实施攻击的 7 个阶段), 对 APT 的多阶段攻击进行建模和识别. Sharma 等 [327] 提出一种包含 4 个并行分类器的分布式 APT 检测框架, 这些分类器对相同的网络数据独立分析, 并经过事件关联形成 APT 攻击的 4 个检测结果, 然后通过投票机制确定系统是否受到了 APT 攻击. 此外, 为应对 DDoS 洪泛攻击的大规模网络流量, 采用 MapReduce, Spark 等分布式计算框架提升数据处理效率成为当前 DDoS 攻击检测的重要手段 [328,329]. Francois 等 [330] 还提出了基于 Hadoop 进行大规模网络流量记录分析进而检测僵尸网络的方法.

(3) 基于大数据技术的内部威胁检测. 内部威胁是大数据安全面临的最严重威胁之一 [331], 其造成的危害远高于外部安全事件, 例如棱镜门事件. 同时, 由于大数据平台的数据资源和访问接口异常复杂, 内部人员的恶意行为能够隐藏在大量的正常数据或数据访问中, 常规的安全策略难以防范. 因此, 内部威胁检测成为大数据安全面临的极具挑战的问题.

当前的内部威胁检测主要从主观要素和客观要素两个方面研究. 主观要素方面主要利用心理学和社会学技术监控并分析内部人员的心理状态和人格特征等, 并对可能产生内部威胁的人员进行预测和防范 [332]. 客观要素方面通过采集命令执行、网络访问、文件操作和鼠标键盘使用等数据中的内部人员行为痕迹, 运用数据挖掘算法分析人员行为特征, 检测并预防内部威胁.

大数据环境下, 为精准刻画人员特征, 提升内部威胁检测的正确率, 需要扩大数据采集的深度和广度, 并且对主客观要素进行联合分析 [332], 在 Greitzer 等 [333] 最新的研究成果中, 甚至将分析对象的心理、行为等个人因素进一步扩展到组织因素. 高效的海量数据处理、内部威胁特征分析和精确的内部威胁检测还需要大数据技术的支持. 例如, IBM 基于 Hadoop 平台开发出一款名为 IBM 大数据安全智能的安全工具, 可以扫描公司内部数十年以来的电子邮件、社交网络等网络流量, 并利用 Hadoop 进行模式分析, 检测心怀不满的员工, 预防数据泄露[4)]. Böse 等 [334] 采用 Spark Streaming 对海量异构流数据进行可扩展实时分析, 通过检测事件流中的异常模式实现近实时的内部威胁检测.

(4) 大数据技术在其他威胁发现研究中的应用. 除了上述应用, 研究人员还将大数据技术应用到系统漏洞挖掘、恶意软件检测等领域. 代表性成果包括: 赛门铁克 (Symantec) 研究人员建立大数据分析平台 WINE, 通过分析全球 1100 万个主机上二进制文件的下载情况, 识别了 18 个 0day 漏洞 [335]. Win 等 [336] 提出一种云环境下基于大数据技术的恶意软件和 rootkit 攻击实时检测框架, 利用 HDFS 存储从客户 VM 中采集的海量数据, 利用基于图的事件关联分析和基于 MapReduce 解析器的潜在攻击路径识别方法提取攻击特征, 进而利用基于逻辑回归和置信传播 (belief propagation) 的机器学习算法实现攻击检测.

综上所述, 基于大数据技术实现平台及服务的安全监管, 能够分析的数据深度和广度更大, 时间跨度更长, 而且能够检测未知的安全攻击或威胁, 将成为大数据安全监管的一大趋势. 然而基于大数据技术对安全数据进行分析也面临诸多挑战. Ullah 和 Babar [337] 从系统架构视角总结了大数据安全分析系统的研究现状, 提出互操作性、可修改性、适应性、通用性、隐蔽性和隐私保护这些重要属性在当前研究中缺乏明确的体系架构支持; 不同架构策略之间的权衡和依赖性需要更深入的探索; 学术界和工业界普遍缺乏有效协作以支持分析系统的建立; 不同大数据处理框架如 Hadoop, Spark 和 Storm 在大数据监管与态势分析中的性能需要进行更多的比较分析. 此外, 监管数据采集的全面性与可信性、监管数据的可视化能力等也是影响基于大数据技术的安全监管系统效能的重要因素, 相关研究有待进一步加强 [318, 338].

6 结论

大数据作为一种颠覆性创新技术, 是信息化发展的新阶段, 影响重塑社会形态、改变人们的生活工作方式, 其蕴含的巨大价值使其极易成为攻击的重点目标. 但由于大数据海量、异构和高速等特点, 传统的安全技术难以满足大数据安全保障的高效、实时、动态、跨域等需求, 大数据的安全保护面临巨大挑战. 同时, 为实现大数据的高效处理, 大数据平台引入了新的数据存储与处理框架, 带来了新的安全威胁, 而这些新型框架所采取的安全机制却十分薄弱, 使大数据面临的安全问题更加严重. 近年来, 大数据安全事件频繁发生, 涉及公民隐私、军事机密甚至政治政权, 针对大数据的安全技术研究已刻不容缓.

本文在总结现有大数据安全技术分类与框架的基础上, 提出了一种符合大数据业务流程特点和大数据系统技术框架组成特点的大数据安全技术框架, 并基于所提框架从大数据安全共享与可信服务、

4) IBM 开发出新型安全工具: 运用大数据识别安全威胁和不满员工. 2013. https://36kr.com/p/201176.html.

大数据平台安全和大数据安全监管 3 个方面总结了大数据安全技术的研究进展, 囊括了大数据业务流程和大数据系统技术框架所涉及的主要安全机制. 研究发现, 随着大数据应用的推广, 当前大数据安全技术的研究还远远落后于大数据应用的安全需求. 大数据安全面临的重大挑战主要体现在以下 3 个方面.

(1) 大数据安全防护水平与大数据担负的时代使命不相适应. 大数据平台是数据的集中存放地, 平台安全是确保大数据安全的基础, 既要防外, 更要防内, 其核心解决方案必须依靠密码. 其中, 同态密码内外兼防, 为存储安全提供根本性解决方案; 属性密码为共享环境访问控制提供了有效的新方法, 但效率成为制约它们实用化的最大挑战. 如何在不影响大数据应用效能的前提下, 充分发挥密码的作用, 实现密码的实用化成为学术研究和商业化亟待攻克的难题. 突破同态密码、属性密码等密码体制的基础理论和关键技术在效率等方面的瓶颈是当前研究的重点.

(2) 大数据安全共享程度与大数据应发挥的重要作用不相适应. 为实现大数据价值、提升大数据效能, 必须解决由利益藩篱、安全信任等导致的大数据 “不愿、不敢、不能” 共享的问题. 区块链技术是解决激励与价值认可、安全与责任认定、分析与全维共享等问题的关键, 为大数据安全共享与可信服务提供了新的途径. 当前, 如何利用区块链技术解决大数据安全共享和可信服务问题的研究刚刚起步. 基于区块链的数据共享体系架构构建, 多样化数据、信息与知识的质量与价值评估, 智能合约、密码安全实现和运行安全等问题有待进一步研究.

(3) 大数据安全监管能力与大数据所处的重要地位不相适应. 面对日益增长的海量异构数据和大数据平台组件, 如何对数据、服务和平台进行有效监管, 以破解数据使用情况不清、系统安全态势不明的难题成为迫切需要解决的问题. 数据世系技术为解决数据监管难题提供了理论、技术和工程等经验, 基于数据世系的数据安全监管是未来的重要发展方向, 但在大数据环境下, 也面临着世系理论模型和世系采集、存储、融合、分析、安全等技术挑战. 大数据技术在解决大数据环境下威胁快速发现和安全态势准确掌控的问题上具有天然优势, 其在安全监管中的应用研究亟待加强.

参考文献

1 Computer Emergency Rediness Team. 2017 security report–data breach. Qihoo 360 Technology Co. Ltd., 2018 [360 网络安全响应中心. 2017 年度安全报告 —— 数据泄密. 奇虎 360 科技有限公司, 2018]

2 Li X L, Gong H G. A survey on big data systems. Sci Sin Inform, 2015, 45: 1–44 [李学龙, 龚海刚. 大数据系统综述. 中国科学: 信息科学, 2015, 45: 1–44]

3 Alshboul Y, Wang Y. Big data lifecycle: threats and security model. In: Proceedings of the 21st Americas Conference on Information Systems, Fajardo, 2015. 3623–3629

4 Mehmood A, Natgunanathan I, Xiang Y, et al. Protection of big data privacy. IEEE Access, 2016, 4: 1821–1834

5 Fang B X, Jia Y, Li A P, et al. Privacy preservation in big data: a survey. Big Data Res, 2016, 2: 1–18 [方滨兴, 贾焰, 李爱平, 等. 大数据隐私保护技术综述. 大数据, 2016, 2: 1–18]

6 National Information Security Standardization Technical Committee. Information Security Technology – Big Data Security Management Guide (Draft for Comments). 2017 [全国信息安全标准化技术委员会. 信息安全技术大数据安全管理指南 (征求意见稿). 2017

7 Anant B, Yu C, Adam F, et al. Expanded top ten big data security and privacy challenges. Cloud Security Alliance Big Data Working Group, 2013. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

8 Chang W L, Roy A, Underwood M, et al. NIST big data interoperability framework: volume 4, security and privacy. NIST Special Publication 1500-4, 2015

9 Wang J M, Chen X S, Liu X G, et al. Big data security standardization white paper (2017). National Information Security Standardization Technical Committee SWG-BDS, 2017 [王建民, 陈兴蜀, 刘贤刚, 等. 大数据安全标准化白皮书 (2017). 全国信息安全标准化技术委员会大数据安全标准特别工作组, 2017]

10 Tankard C. Big data security. Netw Secur, 2012, 2012: 5–8
11 Matturdi B, Zhou X, Li S, et al. Big data security and privacy: a review. China Commun, 2014, 11: 135–145
12 Bertino E, Ferrari E. Big data security and privacy. In: A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Berlin: Springer, 2018. 425–439
13 Walshe R, Boyd D. Big Data Reference Architecture - Part 3: Reference Architecture (2nd Working Draft). ISO/IEC JTC1 WG9, 2016
14 Mei H, Gao L, Dai H, et al. Information Technology - Big Data - Technical Reference Model. National Information Technoloty Standardization Technical Committee, 2017 [梅宏, 高林, 代红, 等. 信息技术大数据技术参考模型. 全国信息技术标准化技术委员会, 2017]
15 Liang F, Yu W, An D, et al. A survey on big data market: pricing, trading and protection. IEEE Access, 2018, 6: 15132–15154
16 Chen J C, Xue Y Z. Bootstrapping a blockchain based ecosystem for big data exchange. In: Proceedings of the 2017 IEEE International Congress on Big Data, Hawaii, 2017. 460–463
17 Liang J, Han W L, Guo Z Q, et al. DESC: enabling secure data exchange based on smart contracts. Sci China Inf Sci, 2018, 61: 049102
18 Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. 2008. https://static.coinpaprika.com/storage/cdn/whitepapers/215.pdf
19 Missier P, Bajoudah S, Capossele A, et al. Mind my value: a decentralized infrastructure for fair and trusted IoT data trading. In: Proceedings of the 7th International Conference on the Internet of Things, Linz, 2017. 15
20 Nasonov D, Visheratin A A, Boukhanovsky A. Blockchain-based transaction integrity in distributed big data marketplace. In: Proceedings of the International Conference on Computational Science, Wuxi, 2018. 569–577
21 Molinajimenez C, Solaiman E, Sfyrakis I, et al. On and off-blockchain enforcement of smart contracts. In: Euro-Par 2018: Parallel Processing Workshops. Berlin: Springer, 2018. 342–354
22 Azaria A, Ekblaw A, Vieira T, et al. MedRec: using blockchain for medical data access and permission management. In: Proceedings of the 2nd International Conference on Open and Big Data, Vienna, 2016. 25–30
23 Castaldo L, Cinque V. Blockchain-based logging for the cross-border exchange of ehealth data in Europe. In: Proceedings of the International ISCIS Security Workshop, London, 2018. 46–56
24 Yan S, Qing S D, Wei K. Application of blockchain in data circulation. Big Data Res, 2018, 4: 3–12 [闫树, 卿苏德, 魏凯. 区块链在数据流通中的应用. 大数据, 2018, 4: 3–12]
25 Lin I-C, Liao T-C. A survey of blockchain security issues and challenges. Int J Netw Secur, 2017, 19: 653–659
26 Dong X Q, Guo B, Shen Y, et al. An efficient and secure decentralizing data sharing model. Chin J Comput, 2018, 41: 1021–1036 [董祥千, 郭兵, 沈艳, 等. 一种高效安全的去中心化数据共享模型. 计算机学报, 2018, 41: 1021–1036]
27 Yang Q. The challenge of GDPR to AI and the countermeasures based on federated transfer learning. CAAI Trans Intell Tech, 2018, 8: 1–8 [杨强. GDPR 对 AI 的挑战和基于联邦迁移学习的对策. 中国人工智能学会通讯, 2018, 8: 1–8]
28 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359
29 Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst, 2002, 10: 557–570
30 Feng D G, Zhang M, Li H. Big data security and privacy protection. Chin J Comput, 2014, 37: 246–258 [冯登国, 张敏, 李昊. 大数据安全与隐私保护. 计算机学报, 2014, 37: 246–258]
31 Byun J W, Sohn Y, Bertino E, et al. Secure Anonymization for Incremental Datasets. Berlin: Springer, 2006
32 Xiao X K, Tao Y F. M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Beijing, 2007. 689–700
33 Bu Y, Fu A W C, Wong R C W, et al. Privacy preserving serial data publishing by role composition. Proc VLDB Endow, 2008, 1: 845–856
34 Fu Y Y, Fu H, Xie X. Social network anonymization and privacy protection. Commun CCF, 2014, 10: 51–58 [付艳艳, 付浩, 谢幸, 等. 社交网络匿名与隐私保护. 中国计算机学会通讯, 2014, 10: 51–58]
35 Liu P, Li X X. An improved privacy preserving algorithm for publishing social network data. In: Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, Zhangjiajie, 2013. 888–895
36 Zou L, Chen L, Özsu M T. k-automorphism: a general framework for privacy preserving network publication. Proc

VLDB Endow, 2009, 2: 946–957
37 Yuan M X, Chen L, Philip S Y, et al. Protecting sensitive labels in social network data anonymization. IEEE Trans Knowl Data Eng, 2013, 25: 633–647
38 Fu Y Y, Zhang M, Feng D G, et al. Attribute privacy preservation in social networks based on node anatomy. J Sotfw, 2014, 25: 768–780 [付艳艳, 张敏, 冯登国, 等. 基于节点分割的社交网络属性隐私保护. 软件学报, 2014, 25: 768–780]
39 Tassa T, Cohen D J. Anonymization of centralized and distributed social networks by sequential clustering. IEEE Trans Knowl Data Eng, 2013, 25: 311–324
40 Skarkala M E, Maragoudakis M, Gritzalis S, et al. Privacy preservation by k-anonymization of weighted social networks. In: Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012. 423–428
41 Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, San Francisco, 2003. 31–42
42 Dewri R, Ray I, Ray I, et al. Query m-invariance: preventing query disclosures in continuous location-based services. In: Proceedings of the 11th International Conference on Mobile Data Management, Kansas City, 2010. 95–104
43 Huo Z, Meng X F. A survey of trajectory privacy-preserving techniques. Chin J Comput, 2011, 34: 1820–1830 [霍峥, 孟小峰. 轨迹隐私保护技术研究. 计算机学报, 2011, 34: 1820–1830]
44 Poulis G, Skiadopoulos S, Loukides G, et al. Distance-based km-anonymization of trajectory data. In: Proceedings of the 14th International Conference on Mobile Data Management, Milan, 2013. 57–62
45 Gidofalvi G, Huang X, Pedersen T B. Privacy-preserving data mining on moving object trajectories. In: Proceedings of the 8th International Conference on Mobile Data Management, Mannheim, 2007. 60–68
46 Xu T, Cai Y. Exploring historical location data for anonymity preservation in location-based services. In: Proceedings of the 27th Conference on Computer Communications, Phoenix, 2008. 547–555
47 Huo Z, Meng X F. A trajectory data publication method under differential privacy. Chin J Comput, 2018, 41: 400–412 [霍峥, 孟小峰. 一种满足差分隐私的轨迹数据发布方法. 计算机学报, 2018, 41: 400–412]
48 Dwork C. Differential privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Venice, 2006. 1–12
49 Sala A, Zhao X, Wilson C, et al. Sharing graphs using differentially private graph models. In: Proceedings of ACM SIGCOMM Conference on Internet Measurement Conference, 2011. 81–98
50 Wagner I, Eckhoff D. Technical privacy metrics. ACM Comput Surv, 2018, 51: 1–38
51 Friedman A, Schuster A. Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, 2010. 493–502
52 Xiong P, Zhu T Q, Wang X F. A survey on differential privacy and applications. Chin J Comput, 2014, 37: 101–122 [熊平, 朱天清, 王晓峰. 差分隐私保护及其应用. 计算机学报, 2014, 37: 101–122]
53 Warner S L. Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc, 1965, 60: 63–69
54 Ye Q Q, Meng X F, Zhu M J, et al. Survey on local differential privacy. J Softw, 2018, 29: 1981–2005 [叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述. 软件学报, 2018, 29: 1981–2005]
55 Qin Z, Yang Y, Yu T, et al. Heavy hitter estimation over set-valued data with local differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, 2016. 192–203
56 Xu C G, Ren J, Zhang Y X, et al. DPPro: differentially private high-dimensional data release via random projection. IEEE Trans Inform Forensic Secur, 2017, 12: 3081–3093
57 Ren X, Yu C M, Yu W, et al. LoPub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Trans Inform Forensic Secur, 2018, 13: 2151–2166
58 Fan L Y, Xiong L. An adaptive approach to real-time aggregate monitoring with differential privacy. IEEE Trans Knowl Data Eng, 2014, 26: 2094–2106
59 Chan T H, Shi E, Song D. Private and continual release of statistics. In: Proceedings of International Colloquium Conference on Automata, Languages and Programming, 2010. 405–417

60 Agrawal R, Srikant R. Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, 2000. 439–450

61 Zhou S G, Li F, Tao Y F, et al. Privacy preservation in database applications: a survey. Chin J Comput, 2009, 32: 847–861 [周水庚, 李丰, 陶宇飞, 等. 面向数据库应用的隐私保护研究综述. 计算机学报, 2009, 32: 847–861]

62 Cheng X, Su S, Xu S Z, et al. DP-apriori: a differentially private frequent itemset mining algorithm based on transaction splitting. Comput Secur, 2015, 50: 74–90

63 McSherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence, 2009. 19–30

64 Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. J Mach Learn Res, 2011, 12: 1069–1109

65 Zhang J, Zhang Z J, Xiao X K, et al. Functional mechanism. Proc VLDB Endow, 2012, 5: 1364–1375

66 Dwork C. A firm foundation for private data analysis. Commun ACM, 2011, 54: 86

67 Kang H Y, Ma Y L. Survey on application of data mining via differential privacy. J Shandong Univ (Nat Sci), 2017, 52: 16–23 [康海燕, 马跃雷. 差分隐私保护在数据挖掘中应用综述. 山东大学学报: 理学版, 2017, 52: 16–23]

68 Li N H, Qardaji W, Su D, et al. PrivBasis: frequent itemset mining with differential privacy. Proc VLDB Endow, 2012, 5: 1340–1351

69 Lin C, Song Z H, Song H B, et al. Differential privacy preserving in big data analytics for connected health. J Med Syst, 2016, 40: 97

70 Roy I, Setty S T, Kilzer A, et al. Airavat: security and privacy for MapReduce. In: Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation, San Jose, 2010. 297–312

71 Rivest R L, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms. Foundations Secure Comput, 1978, 4: 169–180

72 Graepel T, Lauter K, Naehrig M. ML confidential: machine learning on encrypted data. In: Proceedings of the 15th International Conference on Information Security and Cryptology, Seoul, 2012. 1–21

73 Almutairi N, Coenen F, Dures K. K-means clustering using homomorphic encryption and an updatable distance matrix: secure third party data clustering with limited data owner interaction. In: Proceedings of the 19th International Conference on Big Data Analytics and Knowledge Discovery, Lyon, 2017. 274–285

74 Li L C, Lu R X, Choo K K R, et al. Privacy-preserving-outsourced association rule mining on vertically partitioned databases. IEEE Trans Inform Forensic Secur, 2016, 11: 1847–1861

75 Wang B C, Zhan Y, Zhang Z L. Cryptanalysis of a symmetric fully homomorphic encryption scheme. IEEE Trans Inform Forensic Secur, 2018, 13: 1460–1467

76 Gilad-Bachrach R, Dowlin N, Laine K, et al. Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In: Proceedings of the 33rd International Conference on Machine Learning, New York, 2016. 201–210

77 Ben-Or M, Goldwasser S, Wigderson A. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In: Proceedings of the 20th Annual ACM Symposium on Theory of Computing, Chicago, 1988. 1–10

78 Zhou S F, Dou J W, Guo Y M, et al. Secure multiparty vector computation. Chin J Comput, 2017, 40: 1134–1150 [周素芳, 窦家维, 郭奕旻, 等. 安全多方向量计算. 计算机学报, 2017, 40: 1134–1150]

79 Catak F Ö. Secure multi-party computation based privacy preserving extreme learning machine algorithm over vertically distributed data. In: Proceedings of the 22nd International Conference on Neural Information Processing, Istanbul, 2015. 337–345

80 Inan A, Kaya S V, Saygin Y, et al. Privacy preserving clustering on horizontally partitioned data. Data Knowl Eng, 2007, 63: 646–666

81 Kamara S, Mohassel P, Raykova M, et al. Scaling private set intersection to billion-element sets. In: Proceedings of the 18th International Conference on Financial Cryptography and Data Security, Barbados, 2014. 195–215

82 Jiang H, Xu Q L. Secure multiparty computation in cloud computing. J Comput Res Develop, 2016, 53: 2152–2162 [蒋瀚, 徐秋亮. 基于云计算服务的安全多方计算. 计算机研究与发展, 2016, 53: 2152–2162]

83 Asharov G, Jain A, López-Alt A, et al. Multiparty computation with low communication, computation and interaction via threshold FHE. In: Proceedings of the 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, 2012. 483–501

84 López-Alt A, Tromer E, Vaikuntanathan V. On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In: Proceedings of the 44th Annual ACM Symposium on Theory of Computing, New York, 2012. 1219–1234

85 Peter A, Tews E, Katzenbeisser S. Efficiently outsourcing multiparty computation under multiple keys. IEEE Trans Inform Forensic Secur, 2013, 8: 2046–2058

86 Damgard I, Pastro V, Smart N, et al. Multiparty computation from somewhat homomorphic encryption. In: Advances in Cryptology-CRYPTO 2012. Berlin: Springer, 2012. 643–662

87 Liu M H, Zhang N, Zhang Y X, et al. Research on sensitive data protection technology on cloud computing. Telecommun Sci, 2014, 30: 2–8 [刘明辉, 张尼, 张云勇, 等. 云环境下的敏感数据保护技术研究. 电信科学, 2014, 30: 2–8]

88 Chen T Y, Chen J F. Intelligent data masking system for big data productive environment. Commun Tech, 2016, 49: 915–922 [陈天莹, 陈剑锋. 大数据环境下的智能数据脱敏系统. 通信技术, 2016, 49: 915–922]

89 Jin J, Ping X J, Zhang T, et al. Survey of text localization techniques in images. Appl Res Comput, 2007, 24: 8–11 [晋瑾, 平西建, 张涛, 等. 图像中的文本定位技术研究综述. 计算机应用研究, 2007, 24: 8–11]

90 Black J, Rogaway P. Ciphers with arbitrary finite domains. In: Proceedings of the Cryptographers' Track at the RSA Conference, San Jose, 2002. 114–130

91 Joseph F, Brian L. Magic Quadrant for Data Masking Technology. G00247005, 2013

92 Wang J M, Liu X G, Jin T, et al. Big Data Security Standardization White Paper (2018). National Information Security Standardization Technical Committee SWG-BDS, 2018 [王建民, 刘贤刚, 金涛, 等. 大数据安全标准化白皮书 (2018 版). 全国信息安全标准化技术委员会大数据安全标准特别工作组, 2018]

93 Chang W L. NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. NIST Special Publication 1500-6, 2015

94 Das D, O'Malley O, Radia S, et al. Adding Security to Apache Hadoop. Hortonworks Technical Report 1, 2011

95 Zhang K, Zhou X Y, Chen Y, et al. Sedic: privacy-aware data intensive computing on hybrid clouds. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, Chicago, 2011. 515–526

96 Zhang C, Chang E C, Yap R H C. Tagged-MapReduce: a general framework for secure computing with mixed-sensitivity data on hybrid clouds. In: Proceedings of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Chicago, 2014. 31–40

97 Oktay K Y, Mehrotra S, Khadilkar V, et al. SEMROD: secure and efficient MapReduce over HybriD clouds. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, 2015. 153–166

98 Shen Q N, Qing S H, Wu Z H, et al. Securely redundant scheduling policy for MapReduce based on dynamic domains partition. J Commun, 2014, 35: 34–46 [沈晴霓, 卿斯汉, 吴中海, 等. 基于动态域划分的 MapReduce 安全冗余调度策略. 通信学报, 2014, 35: 34–46]

99 Mckeen F, Alexandrovich I, Berenzon A, et al. Innovative instructions and software model for isolated execution. In: Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy, Tel-Aviv, 2013. 10

100 Schuster F, Costa M, Fournet C, et al. VC3: trustworthy data analytics in the cloud using SGX. In: Proceedings of the 36th IEEE Symposium on Security and Privacy, San Jose, 2015. 38–54

101 Pires R, Gavril D, Felber P, et al. A lightweight MapReduce framework for secure processing with SGX. In: Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Madrid, 2017. 1100–1107

102 Ohrimenko O, Costa M, Fournet C, et al. Observing and preventing leakage in MapReduce. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, 2015. 1570–1581

103 Schwarz M, Weiser S, Gruss D, et al. Malware guard extension: using SGX to conceal cache attacks. In: Proceedings of the 2017 International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Bonn, 2017. 3–24

104 Sharma P P, Navdeti C P. Securing big data hadoop: a review of security issues, threats and solution. Int J Comput Sci Inf Tech, 2014, 5: 2126–2131

105 Ning F X, Wen Y, Shi G. GuardSpark: access control enforcement in spark. J Cyber Secur, 2017, 2: 70–81 [宁方潇, 文雨, 史岗. GuardSpark: Spark 访问控制增强机制. 信息安全学报, 2017, 2: 70–81]
106 Ulusoy H, Colombo P, Ferrari E, et al. GuardMR: fine-grained security policy enforcement for MapReduce systems. In: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, Singapore, 2015. 285–296
107 Preuveneers D, Joosen W. SparkXS: efficient access control for intelligent and large-scale streaming data applications. In: Proceedings of the 11th International Conference on Intelligent Environments, Prague, 2015. 96–103
108 Wang J H, Liu C Y, Wang G F, et al. Review of trusted cloud computing based on proof-based verifiable computation. Chin J Comput, 2016, 39: 286–304 [王佳慧, 刘川意, 王国峰, 等. 基于可验证计算的可信云计算研究. 计算机学报, 2016, 39: 286–304]
109 Braun B, Feldman A J, Ren Z, et al. Verifying computations with state. In: Proceedings of the 24th ACM Symposium on Operating Systems Principles, Farmington, 2013. 341–357
110 Ding Y, Wang H M, Shi P C, et al. Trusted cloud service. Chin J Comput, 2015, 38: 133–149 [丁滟, 王怀民, 史佩昌, 等. 可信云服务. 计算机学报, 2015, 38: 133–149]
111 Wei W, Du J, Yu T, et al. Securemr: a service integrity assurance framework for mapreduce. In: Proceedings of the Annual Computer Security Applications Conference, Honolulu, 2009. 73–82
112 Wang Y Z, Wei J P. Viaf: verification-based integrity assurance framework for mapreduce. In: Proceedings of the 2011 IEEE International Conference on Cloud Computing, Washington, 2011. 300–307
113 Xiao Z F, Xiao Y. Accountable MapReduce in cloud computing. In: Proceedings of Computer Communications Workshops, 2011. 1082–1087
114 Huang C, Zhu S C, Wu D H. Towards trusted services: result verification schemes for mapreduce. In: Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Ottawa, 2012. 41–48
115 Ruan A, Martin A. Tmr: towards a trusted mapreduce infrastructure. In: Proceedings of the IEEE 8th World Congress on Services, Honolulu, 2012. 141–148
116 Wang Y Z, Wei J P, Srivatsa M. Result integrity check for mapreduce computation on hybrid clouds. In: Proceedings of the IEEE 6th International Conference on Cloud Computing, Santa Clara, 2013. 847–854
117 Gentry C. Fully homomorphic encryption using ideal lattices. In: Proceedings of the 41st ACM Symposium on Theory of Computing, Washington, 2009. 169–178
118 China Association for Science and Technoloty. 2014–2015 Report on Advances in Cryptology. Beijing: China Science and Technology Press [中国科学技术协会. 2014–2015 密码学学科发展报告. 北京: 中国科学技术出版社, 2016]
119 Coron J-S. Survey of Existing SHE Schemes and Cryptanalytic Techniques. ICT-644209, 2015
120 Ducas L, Micciancio D. FHEW: bootstrapping homomorphic encryption in less than a second. In: Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, 2015. 617–640
121 Chillotti I, Gama N, Georgieva M, et al. Faster fully homomorphic encryption: bootstrapping in less than 0.1 seconds. In: Proceedings of International Conference on the Theory and Application of Cryptology and Information Security, Hanoi, 2016. 3–33
122 Halevi S, Shoup V. Faster homomorphic linear transformations in HElib. In: Advances in Cryptology-CRYPTO 2018. Berlin: Springer, 2018
123 van Dijk M, Gentry C, Halevi S, et al. Fully homomorphic encryption over the integers. In: Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques, Monaco and Nice, 2010. 24–43
124 Martins P, Sousa L, Mariano A. A survey on fully homomorphic encryption. ACM Comput Surv, 2018, 50: 83
125 Acar A, Aksu H, Uluagac A S, et al. A survey on homomorphic encryption schemes. ACM Comput Surv, 2018, 51: 79
126 Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (standard) LWE. In: Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science, Washington, 2011. 97–106
127 Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) fully homomorphic encryption without bootstrapping. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, 2012. 309–325
128 Gentry C, Sahai A, Waters B. Homomorphic encryption from learning with errors: conceptually-simpler, asymptotically-faster, attribute-based. In: Proceedings of the 33rd Annual Cryptology Conference, Santa Barbara,

2013. 75–92
129 Brakerski Z. Fully homomorphic encryption without modulus switching from classical GapSVP. In: Proceedings of Advances in Cryptology-Crypto 2012, Santa Barbara, 2012. 868–886
130 Fan J, Vercauteren F. Somewhat practical fully homomorphic encryption. IACR Cryptol ePrint Arch, 2012, 2012: 144
131 Bos J W, Lauter K, Loftus J, et al. Improved security for a ring-based fully homomorphic encryption scheme. In: Proceedings of the 14th IMA International Conference on Cryptography and Coding, Oxford, 2013. 45–64
132 Lepoint T, Naehrig M. A comparison of the homomorphic encryption schemes FV and YASHE. In: Proceedings of International Conference on Cryptology in Africa, Marrakesh, 2014. 318–335
133 Cheon J H, Kim A, Kim M, et al. Homomorphic encryption for arithmetic of approximate numbers. In: Proceedings of International Conference on the Theory and Application of Cryptology and Information Security, Hong Kong, 2017. 409–437
134 Cheon J H, Han K, Kim A, et al. Bootstrapping for approximate homomorphic encryption. In: Proceedings of the 37th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tel Aviv, 2018. 360–384
135 Chor B, Goldreich O, Kushilevitz E, et al. Private information retrieval. In: Proceedings of the 36th Annual Symposium on Foundations of Computer Science, Milwaukee, 1995. 41–50
136 Doröz Y, Sunar B, Hammouri G. Bandwidth efficient PIR from NTRU. In: Proceedings of International Conference on Financial Cryptography and Data Security, Barbados, 2014. 195–207
137 Popa R A, Redfield C, Zeldovich N, et al. CryptDB: protecting confidentiality with encrypted query processing. In: Proceedings of the 23rd ACM Symposium on Operating Systems Principles, Cascais, 2011. 85–100
138 Cheon J H, Kim M, Kim M. Search-and-compute on encrypted data. In: Proceedings of International Conference on Financial Cryptography and Data Security, San Juan, 2015. 142–159
139 Li Z P, Ma C G, Zhou H S. Overview on Fully Homomorphic Encryption. J Cryptologic Res, 2017, 4: 561–578 [李增鹏, 马春光, 周红生. 全同态加密研究. 密码学报, 2017, 4: 561–578]
140 Yagisawa M. Fully homomorphic encryption without bootstrapping. IACR Cryptol ePrint Arch, 2015, 2015: 474
141 Liu D X. Practical Fully Homomorphic Encryption without Noise Reduction. IACR Cryptol ePrint Arch, 2015, 2015: 468
142 Wang Y G. Notes on two fully homomorphic encryption schemes without bootstrapping. IACR Cryptol ePrint Arch, 2015, 2015: 519
143 Qin Z G, Xu J, Nie X Y, et al. A survey of public-key encryption with keyword search. J Cyber Secur, 2017, 2: 1–12 [秦志光, 徐骏, 聂旭云, 等. 公钥可搜索加密体制综述. 信息安全学报, 2017, 2: 1–12]
144 Song D X, Wagner D, Perrig A. Practical techniques for searches on encrypted data. In: Proceedings of 2000 IEEE Symposium on Security and Privacy, Berkeley, 2000. 44–55
145 Goh E-J. Secure indexes. IACR Cryptol ePrint Arch, 2003, 2003: 216
146 Curtmola R, Garay J, Kamara S, et al. Searchable symmetric encryption: improved definitions and efficient constructions. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, Alexandria, 2006. 79–88
147 van Liesdonk P, Sedghi S, Doumen J, et al. Computationally efficient searchable symmetric encryption. In: Proceedings of Workshop on Secure Data Management, Seattle, 2010. 87–100
148 Kamara S, Papamanthou C, Roeder T. Dynamic searchable symmetric encryption. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, Raleigh, 2012. 965–976
149 Golle P, Staddon J, Waters B. Secure conjunctive keyword search over encrypted data. In: Proceedings of International Conference on Applied Cryptography and Network Security, Yellow Mountains, 2004. 31–45
150 Cao N, Wang C, Li M, et al. Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Trans Parallel Distrib Syst, 2014, 25: 222–233
151 Li J, Wang Q, Wang C, et al. Fuzzy keyword search over encrypted data in cloud computing. In: Proceedings of the 29th Conference on Computer Communications, San Diego, 2010. 1–5
152 Chai Q, Gong G. Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. In: Proceedings of the 2012 IEEE International Conference on Communications, Ottawa, 2012. 917–922

153 Boneh D, Di Crescenzo G, Ostrovsky R, et al. Public key encryption with keyword search. In: Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, 2004. 506–522
154 Abdalla M, Bellare M, Catalano D, et al. Searchable encryption revisited: consistency properties, relation to anonymous IBE, and extensions. In: Proceedings of Annual International Cryptology Conference, Santa Barbara, 2005. 205–222
155 Xu P, Jin H, Wu Q H, et al. Public-key encryption with fuzzy keyword search: a provably secure scheme under keyword guessing attack. IEEE Trans Comput, 2013, 62: 2266–2277
156 Chen R M, Mu Y, Yang G M, et al. Dual-server public-key encryption with keyword search for secure cloud storage. IEEE Trans Inform Forensic Secur, 2016, 11: 789–798
157 Baek J, Safavi-Naini R, Susilo W. Public key encryption with keyword search revisited. In: Proceedings of International Conference on Computational Science and Its Applications, Perugia, 2008. 1249–1259
158 Zheng Q J, Xu S H, Ateniese G. VABKS: verifiable attribute-based keyword search over outsourced encrypted data. In: Proceedings of the 33rd Annual IEEE International Conference on Computer Communications, Toronto, 2014. 522–530
159 Bellare M, Boldyreva A, O'Neill A. Deterministic and efficiently searchable encryption. In: Proceedings of the Annual International Cryptology Conference, Santa Barbara, 2007. 535–552
160 Regev O. On lattices, learning with errors, random linear codes, and cryptography. J ACM, 2009, 56: 1–40
161 Boneh D, Waters B. Conjunctive, subset, and range queries on encrypted data. In: Proceedings of the 4th Theory of Cryptography Conference, Amsterdam, 2007. 535–554
162 Agrawal R, Kiernan J, Srikant R, et al. Order preserving encryption for numeric data. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, Paris, 2004. 563–574
163 Boldyreva A, Chenette N, Lee Y, et al. Order-preserving symmetric encryption. In: Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cologne, 2009. 224–241
164 Popa R A, Li F H, Zeldovich N. An ideal-security protocol for order-preserving encoding. In: Proceedings of the 2013 IEEE Symposium on Security and Privacy, San Francisco, 2013. 463–477
165 Kerschbaum F. Frequency-hiding order-preserving encryption. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, 2015. 656–667
166 Boneh D, Lewi K, Raykova M, et al. Semantically secure order-revealing encryption: multi-input functional encryption without obfuscation. In: Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, 2015. 563–594
167 Lewi K, Wu D J. Order-revealing encryption: new constructions, applications, and lower bounds. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, 2016. 1167–1178
168 Ning J T, Xu J, Liang K T, et al. Passive attacks against searchable encryption. IEEE Trans Inform Forensic Secur, 2019, 14: 789–802
169 Fu Z, Wu X, Guan C, et al. Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement. IEEE Trans Inform Forensic Secur, 2016, 11: 2706–2716
170 Guo J J, Miao M X, Wang J F. Research and progress of order preserving encryption. J Cryptol Res, 2018, 5: 182–195 [郭晶晶, 苗美霞, 王剑锋. 保序加密技术研究与进展. 密码学报, 2018, 5: 182–195]
171 Brightwell M, Smith H. Using datatype-preserving encryption to enhance data warehouse security. In: Proceedings of the 20th National Information Systems Security Conference, Baltimore, 1997. 141–149
172 Liu Z L, Jia C F, Li J W. Research on the format-preserving encryption modes. J Commun, 2011, 32: 184–190 [刘哲理, 贾春福, 李经纬. 保留格式加密模型研究. 通信学报, 2011, 32: 184–190]
173 Bellare M, Ristenpart T, Rogaway P, et al. Format-preserving encryption. In: Proceedings of the International Workshop on Selected Areas in Cryptography, Alberta, 2009. 295–312
174 Liu Z L, Jia C F, Li J W. Research on the format-preserving encryption techniques. J Softw, 2012, 23: 152–170 [刘哲理, 贾春福, 李经纬. 保留格式加密技术研究. 软件学报, 2012, 23: 152–170]
175 Liu Z L, Jia C F, Li J W, et al. Format-preserving encryption for datetime. In: Proceedings of the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen, 2010. 201–205
176 Cui B J, Zhang B H, Wang K Y. A data masking scheme for sensitive big data based on format-preserving encryption. In: Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering & Embedded

and Ubiquitous Computing, Guangzhou, 2017. 518–524
177 Biryukov A, Leurent G, Perrin L. Cryptanalysis of Feistel networks with secret round functions. In: Proceedings of the International Conference on Selected Areas in Cryptography, New Brunswick, 2015. 102–121
178 Biham E, Biryukov A, Dunkelman O, et al. Initial observations on skipjack: cryptanalysis of skipjack-3XOR. In: Proceedings of the International Workshop on Selected Areas in Cryptography, Kingston, 1998. 362–375
179 Bellare M, Hoang V T, Tessaro S. Message-recovery attacks on Feistel-based format preserving encryption. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, 2016. 444–455
180 Durak F B, Vaudenay S. Breaking the FF3 format-preserving encryption standard over small domains. In: Proceedings of the Annual International Cryptology Conference, Santa Barbara, 2017. 679–707
181 Hoang V T, Tessaro S, Trieu N. The curse of small domains: new attacks on format-preserving encryption. In: Proceedings of the Annual International Cryptology Conference, Santa Barbara, 2018. 221–251
182 Naor M, Reingold O. On the construction of pseudorandom permutations: Luby-Rackoff revisited. J Cryptol, 1999, 12: 29–66
183 Moniruzzaman A B M, Hossain S A. NoSQL database: new era of databases for big data analytics-classification, characteristics and comparison. 2013. ArXiv: 1307.0191
184 Dworkin M. Recommendation for block cipher modes of operation: methods for formatpreserving encryption. NIST Special Publication 800-38G, 2016
185 Shin Y, Koo D, Hur J. A survey of secure data deduplication schemes for cloud storage systems. ACM Comput Surv, 2017, 49: 74
186 Douceur J R, Adya A, Bolosky W J, et al. Reclaiming space from duplicate files in a serverless distributed file system. In: Proceedings of the 22nd International Conference on Distributed Computing Systems, Vienna, 2002. 617–624
187 Bellare M, Keelveedhi S, Ristenpart T. Message-locked encryption and secure deduplication. In: Proceedings of the 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, 2013. 296–312
188 González-Manzano L, Orfila A. An efficient confidentiality-preserving proof of ownership for deduplication. J Netw Comput Appl, 2015, 50: 49–59
189 Xiong J B, Zhang Y Y, Li F H, et al. Research progress on secure data deduplication in cloud. J Commun, 2016, 37: 169–180 [熊金波, 张媛媛, 李凤华, 等. 云环境中数据安全去重研究进展. 通信学报, 2016, 37: 169–180]
190 Ateniese G, Burns R, Curtmola R, et al. Provable data possession at untrusted stores. In: Proceedings of the 14th ACM Conference on Computer and Communications Security, Alexandria, 2007. 598–609
191 Juels A, Kaliski Jr B S. PORs: proofs of retrievability for large files. In: Proceedings of the 14th ACM Conference on Computer and Communications Security, Alexandria, 2007. 584–597
192 Erway C, Küpcü A, Papamanthou C, et al. Dynamic provable data possession. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, 2009. 213–222
193 Shen J, Shen J, Chen X F, et al. An efficient public auditing protocol with novel dynamic structure for cloud data. IEEE Trans Inform Forensic Secur, 2017, 12: 2402–2415
194 Wu Y, Jiang Z L, Wang X, et al. Dynamic data operations with deduplication in privacy-preserving public auditing for secure cloud storage. In: Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, 2017. 562–567
195 Wang Q, Wang C, Li J, et al. Enabling public verifiability and data dynamics for storage security in cloud computing. In: Proceedings of the 14th European Symposium on Research in Computer Security, Saint-Malo, 2009. 355–370
196 Ren Z W, Wang L N, Wang Q, et al. Dynamic proofs of retrievability for coded cloud storage systems. IEEE Trans Serv Comput, 2018, 11: 685–698
197 Tate S R, Vishwanathan R, Everhart L. Multi-user dynamic proofs of data possession using trusted hardware. In: Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy, San Antonio, 2013. 353–364
198 Wang B, Chow S S, Li M, et al. Storing shared data on the cloud via security-mediator. In: Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems Philadelphia, 2013. 124–133

199 Wang B Y, Li B C, Li H. Panda: public auditing for shared data with efficient user revocation in the cloud. IEEE Trans Serv Comput, 2015, 8: 92–106

200 Jiang T, Chen X F, Ma J F. Public integrity auditing for shared dynamic cloud data with group user revocation. IEEE Trans Comput, 2016, 65: 2363–2373

201 Wang Z H. Research on several security mechanisms for cloud storage service. Dissertation for Ph.D. Degree. Beijing: Beijing Jiaotong University, 2016 [王中华. 云存储服务的若干安全机制研究. 博士学位论文. 北京: 北京交通大学, 2016]

202 Zhuo H, Sheng Z, Yu N H. A privacy-preserving remote data integrity checking protocol with data dynamics and public verifiability. IEEE Trans Knowl Data Eng, 2011, 23: 1432–1437

203 Wang C, Chow S S, Wang Q, et al. Privacy-preserving public auditing for secure cloud storage. IEEE Trans Comput, 2013, 62: 362–375

204 Zhu Y, Hu H X, Ahn G J, et al. Cooperative provable data possession for integrity verification in multicloud storage. IEEE Trans Parallel Distrib Syst, 2012, 23: 2231–2244

205 Yang K, Jia X H. An efficient and secure dynamic auditing protocol for data storage in cloud computing. IEEE Trans Parallel Distrib Syst, 2013, 24: 1717–1726

206 Wang H Q. Identity-based distributed provable data possession in multicloud storage. IEEE Trans Serv Comput, 2015, 8: 328–340

207 Yu Y, Au M H, Ateniese G, et al. Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. IEEE Trans Inform Forensic Secur, 2017, 12: 767–778

208 He D B, Kumar N, Wang H Q, et al. Privacy-preserving certificateless provable data possession scheme for big data storage on cloud. Appl Math Comput, 2017, 314: 31–43

209 Wang Y Z, Wei J P. VIAF: verification-based integrity assurance framework for MapReduce. In: Proceedings of IEEE International Conference on Cloud Computing, 2011. 300–307

210 Zhu Y, Wang H X, Hu Z X, et al. Zero-knowledge proofs of retrievability. Sci China Inf Sci, 2011, 54: 1608–1617

211 Liu J, Huang K, Rong H, et al. Privacy-preserving public auditing for regenerating-code-based cloud storage. IEEE Trans Inform Forensic Secur, 2015, 10: 1513–1528

212 Xiang F, Liu C Y, Fang B X, et al. Novel "rich cloud" based data disaster recovery strategy. J Commun, 2013, 34: 92–101 [项菲, 刘川意, 方滨兴, 等. 新的基于云计算环境的数据容灾策略. 通信学报, 2013, 34: 92–101]

213 Wood T, Cecchet E, Ramakrishnan K K, et al. Disaster recovery as a cloud service: economic benefits & deployment challenges. HotCloud, 2010, 10: 8–15

214 Weatherspoon H, Kubiatowicz J D. Erasure coding vs. replication: a quantitative comparison. In: Proceedings of International Workshop on Peer-to-Peer Systems, Cambridge, 2002. 328–337

215 Wang Y J, Xu F L, Pei X Q. Research on erasure code-based fault-tolerant technology for distributed storage. Chin J Comput, 2017, 40: 236–255 [王意洁, 许方亮, 裴晓强. 分布式存储中的纠删码容错技术研究. 计算机学报, 2017, 40: 236–255]

216 Xu J W, Zhang W B, Wang T, et al. A genetic algorithm based ataptive strategy for image backup of virtual machines. Chin J Comput, 2016, 39: 351–363 [徐继伟, 张文博, 王焘, 等. 一种基于遗传算法的虚拟机镜像自适应备份策略. 计算机学报, 2016, 39: 351–363]

217 Chang V. Towards a big data system disaster recovery in a private cloud. Ad Hoc Netw, 2015, 35: 65–82

218 Wood T, Lagar-Cavilla H A, Ramakrishnan K, et al. PipeCloud: using causality to overcome speed-of-light delays in cloud-based disaster recovery. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, Cascais, 2011. 17

219 Zhong R M, Liu C Y, Wang C L, et al. Cost-aware data reliability provision algorithm for the cloud providers. J Softw, 2014, 25: 1874–1886 [钟睿明, 刘川意, 王春露, 等. 一种成本相关的云提供商数据可靠性保证算法. 软件学报, 2014, 25: 1874–1886]

220 Gu Y, Wang D S, Liu C Y. DR-cloud: multi-cloud based disaster recovery service. Tsinghua Sci Technol, 2014, 19: 13–23

221 Colman-Meixner C, Develder C, Tornatore M, et al. A survey on resiliency techniques in cloud computing infrastructures and applications. IEEE Commun Surv Tut, 2016, 18: 2244–2281

222 Perez R, Sailer R, van Doorn L. vTPM: virtualizing the trusted platform module. In: Proceedings of the 15th USENIX Security Symposium, Vancouver, 2006. 305–320

223 Hua J, Sakurai K. Barrier: a lightweight hypervisor for protecting kernel integrity via memory isolation. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, Trento, 2012. 1470–1477

224 Zhu M, Tu B B, Meng D. The security research of virtualization software stack. Chin J Comput, 2017, 40: 481–504 [朱民, 涂碧波, 孟丹. 虚拟化软件栈安全研究. 计算机学报, 2017, 40: 481–504]

225 Ainapure B S, Shah D, Rao A A. Understanding perception of cache-based side-channel attack on cloud environment. In: Proceedings of Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, 2018. 9–21

226 Garfinkel T, Rosenblum M. A virtual machine introspection based architecture for intrusion detection. In: Proceedings of the 2003 Network and Distributed System Security Symposium, San Diego, 2003. 191–206

227 Hebbal Y, Laniepce S, Menaud J-M. Virtual machine introspection: techniques and applications. In: Proceedings of the 10th International Conference on Availability, Reliability and Security, Toulouse, 2015. 676–685

228 Noshy M, Ibrahim A, Ali H A. Optimization of live virtual machine migration in cloud computing: A survey and future directions. J Netw Comput Appl, 2018, 110: 1–10

229 Li C, Raghunathan A, Jha N K. Secure virtual machine execution under an untrusted management OS. In: Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing, Miami, 2010. 172–179

230 Azab A M, Ning P, Wang Z, et al. HyperSentry: enabling stealthy in-context measurement of hypervisor integrity. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, Chicago, 2010. 38–49

231 Szefer J, Keller E, Lee R B, et al. Eliminating the hypervisor attack surface for a more secure cloud. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, Chicago, 2011. 401–412

232 Wang J, Stavrou A, Ghosh A. Hypercheck: a hardware-assisted integrity monitor. In: Proceedings of the International Workshop on Recent Advances in Intrusion Detection, Ottawa, 2010. 158–177

233 Wang Z, Jiang X X. Hypersafe: a lightweight approach to provide lifetime hypervisor control-flow integrity. In: Proceedings of 2010 IEEE Symposium on Security and Privacy, Berkeley, 2010. 380–395

234 Mijumbi R, Serrat J, Gorricho J L, et al. Network function virtualization: state-of-the-art and research challenges. IEEE Commun Surv Tut, 2016, 18: 236–262

235 Sezer S, Scott-Hayward S, Chouhan P K, et al. Are we ready for SDN? Implementation challenges for software-defined networks. IEEE Commun Mag, 2013, 51: 36–43

236 Yu Y, Wang Z L, Bi J, et al. Survey on the languages in the northbound interface of software defined networking. J Softw, 2016, 27: 993–1008 [于洋, 王之梁, 毕军, 等. 软件定义网络中北向接口语言综述. 软件学报, 2016, 27: 993–1008]

237 Zaalouk A, Khondoker R, Marx R, et al. Orchsec: an orchestrator-based architecture for enhancing network-security using network monitoring and SDN control functions. In: Proceedings of Network Operations and Management Symposium (NOMS), New York, 2014. 1–9

238 Wang T, Chen H C. SGuard: a lightweight SDN safe-guard architecture for DoS attacks. China Commun, 2017, 14: 113–125

239 Kreutz D, Ramos F, Verissimo P. Towards secure and dependable software-defined networks. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking, 2013. 55–60

240 Varadharajan V, Karmakar K, Tupakula U, et al. A policy-based security architecture for software-defined networks. IEEE Trans Inform Forensic Secur, 2019, 14: 897–912

241 Hu Z Y, Wang M W, Yan X Q, et al. A comprehensive security architecture for SDN. In: Proceedings of the 18th International Conference on Intelligence in Next Generation Networks, Paris, 2015. 30–37

242 Han B, Gopalakrishnan V, Ji L, et al. Network function virtualization: challenges and opportunities for innovations. IEEE Commun Mag, 2015, 53: 90–97

243 Yang W, Fung C. A survey on security in network functions virtualization. In: Proceedings of NetSoft Conference and Workshops (NetSoft), New York, 2016. 15–19

244 Gember-Jacobson A, Viswanathan R, Prakash C, et al. OpenNF: enabling innovation in network function control. In: Proceedings of ACM SIGCOMM Computer Communication Review, 2014. 163–174

245 Jaeger B. Security orchestrator: introducing a security orchestrator in the context of the etsi nfv reference architecture. In: Proceedings of Trustcom/BigDataSE/ISPA, New York, 2015. 1255–1260

246 Pattaranantakul M, He R, Song Q P, et al. NFV security survey: from use case driven threat analysis to state-of-the-art countermeasures. IEEE Commun Surv Tut, 2018, 20: 3330–3368

247 Firoozjaei M D, Jeong J P, Ko H, et al. Security challenges with network functions virtualization. Future Gener Comput Syst, 2017, 67: 315–324
248 Wang J, Hao S R, Li Y, et al. Challenges towards protecting VNF with SGX. In: Proceedings of the 2018 ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization, Tempe, 2018. 39–42
249 Melis L, Asghar H J, de Cristofaro E, et al. Private processing of outsourced network functions: feasibility and constructions. In: Proceedings of the 2016 ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization, 2016. 39–44
250 Bonfim M S, Dias K L, Fernandes S F L. Integrated NFV/SDN architectures. ACM Comput Surv, 2019, 51: 1–39
251 Wang M M, Liu J W, Chen J, et al. Software defined networking: security model, threats and mechanism. J Softw, 2016, 27: 969–992 [王蒙蒙, 刘建伟, 陈杰, 等. 软件定义网络: 安全模型、机制及研究进展综述. 软件学报, 2016, 27: 969–992]
252 Rawat D B, Reddy S R. Software defined networking architecture, security and energy efficiency: a survey. IEEE Commun Surv Tut, 2017, 19: 325–346
253 Feng D G, Chen C. Research on attribute-based cryptography. J Cryptol Res, 2014, 1: 1–12 [冯登国, 陈成. 属性密码学研究. 密码学报, 2014, 1: 1–12]
254 Goyal V, Pandey O, Sahai A, et al. Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, 2006. 89–98
255 Bethencourt J, Sahai A, Waters B. Ciphertext-policy attribute-based encryption. In: Proceedings of IEEE Symposium on Security and Privacy, 2007. 321–334
256 Fugkeaw S, Sato H. Scalable and secure access control policy update for outsourced big data. Future Gener Comput Syst, 2018, 79: 364–373
257 Fang L, Yin L H, Guo Y C, et al. A survey of technologies in attribute-based access control scheme. Chin J Comput, 2017, 40: 1680–1698 [房梁, 殷丽华, 郭云川, 等. 基于属性的访问控制关键技术研究综述. 计算机学报, 2017, 40: 1680–1698]
258 Kapadia A, Tsang P P, Smith S W. Attribute-based publishing with hidden credentials and hidden policies. In: Proceedings of the 14th Annual Network & Distributed System Security Symposium, San Diego, 2007. 179–192
259 Cui H, Deng R H, Lai J, et al. An efficient and expressive ciphertext-policy attribute-based encryption scheme with partially hidden access structures, revisited. Comput Netw, 2018, 133: 157–165
260 Wang H, Zheng Z H, Wu L, et al. New directly revocable attribute-based encryption scheme and its application in cloud storage environment. Cluster Comput, 2017, 20: 2385–2392
261 Liu J K, Yuen T H, Zhang P, et al. Time-based direct revocable ciphertext-policy attribute-based encryption with short revocation list. In: Proceedings of the 16th International Conference on Applied Cryptography and Network Security, Leuven, 2018. 516–534
262 Pirretti M, Traynor P, McDaniel P, et al. Secure attribute-based systems. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, Alexandria, 2006. 99–112
263 Sun W H, Yu S C, Lou W J, et al. Protecting your right: verifiable attribute-based keyword search with fine-grained owner-enforced search authorization in the cloud. IEEE Trans Parallel Distrib Syst, 2016, 27: 1187–1198
264 Sookhak M, Yu F R, Khan M K, et al. Attribute-based data access control in mobile cloud computing: taxonomy and open issues. Future Gener Comput Syst, 2017, 72: 273–287
265 Ning J T, Dong X L, Cao Z F, et al. White-box traceable ciphertext-policy attribute-based encryption supporting flexible attributes. IEEE Trans Inform Forensic Secur, 2015, 10: 1274–1288
266 Liu Z, Cao Z F, Wong D S. Blackbox traceable CP-ABE: how to catch people leaking their keys by selling decryption devices on eBay. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, Berlin, 2013. 475–486
267 Liu Z H, Duan S, Zhou P, et al. Traceable-then-revocable ciphertext-policy attribute-based encryption scheme. Future Gener Comput Syst, 2019, 93: 903–913
268 Zhang K, Li H, Ma J F, et al. Efficient large-universe multi-authority ciphertext-policy attribute-based encryption with white-box traceability. Sci China Inf Sci, 2018, 61: 032102

269 Chase M. Multi-authority attribute based encryption. In: Proceedings of the 4th Theory of Cryptography Conference, Amsterdam, 2007. 515–534

270 Chase M, Chow S S. Improving privacy and security in multi-authority attribute-based encryption. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, 2009. 121–130

271 Li Q, Ma J F, Li R, et al. Large universe decentralized key-policy attribute-based encryption. Secur Comm Netw, 2015, 8: 501–509

272 Rouselakis Y, Waters B. Efficient statically-secure large-universe multi-authority attribute-based encryption. In: Proceedings of the 19th International Conference on Financial Cryptography and Data Security, San Juan, 2015. 315–332

273 Can Z F, Dong X L, Zhou J, et al. Research advances on big data security and privacy preserving. J Comput Res Develop, 2016, 53: 2137–2151 [曹珍富, 董晓蕾, 周俊, 等. 大数据安全与隐私保护研究进展. 计算机研究与发展, 2016, 53: 2137–2151]

274 Kumar P P, Kumar P S, Alphonse P J A. Attribute based encryption in cloud computing: a survey, gap analysis, and future directions. J Netw Comput Appl, 2018, 108: 37–52

275 Kuhlmann M, Shohat D, Schimpf G. Role mining - revealing business roles for security administration using data mining technology. In: Proceedings of the 8th ACM Symposium on Access Control MODELS and Technologies, 2003. 179–186

276 Kuhlmann M, Shohat D, Schimpf G. Role mining-revealing business roles for security administration using data mining technology. In: Proceedings of the 8th ACM Symposium on Access Control Models and Technologies, Como, 2003. 179–186

277 Molloy I, Park Y, Chari S. Generative models for access control policies: applications to role mining over logs with attribution. In: Proceedings of the 17th ACM Symposium on Access Control Models and Technologies, Newark, 2012. 45–56

278 Li H, Zhang M, Feng D G, et al. Research on access control of big data. Chin J Comput, 2017, 40: 72–91 [李昊, 张敏, 冯登国, 等. 大数据访问控制研究. 计算机学报, 2017, 40: 72–91]

279 Molloy I, Li N, Li T, et al. Evaluating role mining algorithms. In: Proceedings of the 14th ACM Symposium on Access Control Models and Technologies, Stresa, 2009. 95–104

280 Vaidya J, Atluri V, Warner J. RoleMiner: mining roles using subset enumeration. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, Alexandria, 2006. 144–153

281 Zhang D N, Ramamohanarao K, Ebringer T, et al. Permission set mining: discovering practical and useful roles. In: Proceedings of the 24th Annual Computer Security Applications Conference, Anaheim, 2008. 247–256

282 Molloy I, Hong C, Li T C, et al. Mining roles with semantic meanings. In: Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, Estes Park, 2008. 21–30

283 Vaidya J, Atluri V, Guo Q. The role mining problem: finding a minimal descriptive set of roles. In: Proceedings of the 12th ACM Symposium on Access Control Models and Technologies, Sophia Antipolis, 2007. 175–184

284 Zhang D, Ramamohanarao K, Ebringer T. Role engineering using graph optimisation. In: Proceedings of the 12th ACM Symposium on Access Control Models and Technologies, Sophia Antipolis, 2007. 139–144

285 Frank M, Streich A P, Basin D A, et al. A probabilistic approach to hybrid role mining. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, 2009. 101–111

286 Jafarian J H, Takabi H, Touati H, et al. Towards a general framework for optimal role mining: a constraint satisfaction approach. In: Proceedings of the 20th ACM Symposium on Access Control Models and Technologies, Vienna, 2015. 211–220

287 Mitra B, Sural S, Vaidya J, et al. A survey of role mining. ACM Comput Surv, 2016, 48: 1–37

288 Phua T W, Ko R K L. Data provenance for big data security and accountability. In: Encyclopedia of Big Data Technologies. Berlin: Springer, 2018. 1–6

289 Buneman P, Khanna S, Tan W C. Why and where: a characterization of data provenance. In: Proceedings of 2001 International Conference on Database Theory, London, 2001. 216–330

290 Glavic B. Big data provenance: challenges and implications for benchmarking. In: Proceedings of the 2012 Workshop on Big Data Benchmarks, Pune, 2012. 72–80

291 Cheney J, Chong S, Foster N, et al. Provenance: a future history. In: Proceedings of the 24th ACM SIGPLAN

Conference Companion on Object Oriented Programming Systems Languages and Applications, Orlando, 2009. 957–964
292 Labrinidis A, Jagadish H V. Challenges and opportunities with big data. Proc VLDB Endowment, 2012, 5: 2032–2033
293 Moreau L, Clifford B, Freire J, et al. The open provenance model core specification (v1.1). Future Gener Comput Syst, 2011, 27: 743–756
294 Sahoo S S, Barga R S, Goldstein J, et al. Provenance algebra and materialized view-based provenance management. In: Proceedings of the 2nd International Provenance and Annotation Workshop, Salt Lake City, 2008. 531–540
295 Wang J, Crawl D, Purawat S, et al. Big data provenance: challenges, state of the art and opportunities. In: Proceedings of 2015 IEEE International Conference on Big Data, Santa Clara, 2015. 2509–2516
296 Gehani A, Kazmi H, Irshad H. Scaling spade to "big provenance". In: Proceedings of the 8th USENIX Conference on Theory and Practice of Provenance, Washington, 2016. 26–33
297 Fu X, Gao Y, Luo B, et al. Security threats to Hadoop: data leakage attacks and investigation. IEEE Netw, 2017, 31: 67–71
298 Ko R K, Will M A. Progger: an efficient, tamper-evident kernel-space logger for cloud data provenance tracking. In: Proceedings of the IEEE 7th International Conference on Cloud Computing, Anchorage, 2014. 881–889
299 Kulkarni D. A provenance model for key-value systems. In: Proceedings of the 5th Workshop on the Theory and Practice of Provenance, Lombard, 2013. 1–4
300 Alkhaldi A, Gupta I, Raghavan V, et al. Leveraging metadata in no SQL storage systems. In: Proceedings of the 8th IEEE International Conference on Cloud Computing, New York, 2015. 57–64
301 Chacko A M, Fairooz M, Kumar S M. Provenance-aware NoSQL databases. In: Proceedings of the International Symposium on Security in Computing and Communication, Jaipur, 2016. 152–160
302 Park H, Ikeda R, Widom J. Ramp: a system for capturing and tracing provenance in mapreduce workflows. Proc VLDB Endowment, 2011, 4: 1351–1354
303 Akoush S, Sohan R, Hopper A. HadoopProv: towards provenance as a first class citizen in MapReduce. In: Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance, Lombard, 2013. 1–4
304 Zafar F, Khan A, Suhail S, et al. Trustworthy data: a survey, taxonomy and future trends of secure provenance schemes. J Netw Comput Appl, 2017, 94: 50–68
305 Cheney J. A formal framework for provenance security. In: Proceedings of the 24th IEEE Computer Security Foundations Symposium, Cernay-la-Ville, 2011. 281–293
306 Braun U, Shinnar A. A Security Model for Provenance. Harvard Computer Science Group Technical Report TR-04-06. 2006
307 Cadenhead T, Khadilkar V, Kantarcioglu M, et al. A language for provenance access control. In: Proceedings of the 1st ACM Conference on Data and Application Security and Privacy, San Antonio, 2011. 133–144
308 Danger R, Curcin V, Missier P, et al. Access control and view generation for provenance graphs. Future Gener Comput Syst, 2015, 49: 8–27
309 Liang X, Shetty S, Tosh D, et al. ProvChain: a blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In: Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Madrid, 2017. 468–477
310 Ramachandran A, Kantarcioglu M. SmartProvenance: a distributed, blockchain based data provenance system. In: Proceedings of the 8th ACM Conference on Data and Application Security and Privacy, Tempe, 2018. 35–42
311 Muniswamy-Reddy K K, Holland D A, Braun U, et al. Provenance-aware storage systems. In: Proceedings of 2006 USENIX Annual Technical Conference, Boston, 2006. 43–56
312 Suen C H, Ko R K L, Yu S T, et al. S2Logger: end-to-end data tracking mechanism for cloud data provenance. In: Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, 2013. 594–602
313 Alabi O, Beckman J, Dark M, et al. Toward a data spillage prevention process in hadoop using data provenance. In: Proceedings of the 2nd Workshop on Changing Landscapes in HPC Security, Portland, 2015. 9–13
314 Bates A, Butler K, Dobra A, et al. Retrofitting applications with provenance-based security monitoring. 2016. ArXiv: 1609.00266
315 Appelbaum D. Securing big data provenance for auditors: the big data provenance black box as reliable evidence. J

Emerg Tech Account, 2016, 13: 17–36
316 Ghoshal D, Plale B. Provenance from log files: a BigData problem. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, 2013. 290–297
317 Cuzzocrea A. Provenance research issues and challenges in the big data era. In: Proceedings of the IEEE 39th Annual Computer Software and Applications Conference, Taichung, 2015. 684–686
318 Cardenas A A, Manadhata P K, Rajan S P. Big data analytics for security. IEEE Secur Privacy, 2013, 11: 74–76
319 Zuech R, Khoshgoftaar T M, Wald R. Intrusion detection and big heterogeneous data: a survey. J Big Data, 2015, 2: 3
320 Jeong H-D J, Hyun W, Lim J, et al. Anomaly teletraffic intrusion detection systems on hadoop-based platforms: a survey of some problems and solutions. In: Proceedings of the 15th International Conference on Network-Based Information Systems, Melbourne, 2012. 766–770
321 Cheon J, Choe T-Y. Distributed processing of snort alert log using hadoop. Int J Eng Tech, 2013, 5: 2685–2690
322 Baker M, Turnbull D, Kaszuba G. Finding needles in haystacks (the size of countries). In: Proceedings of Black Hat Europe 2012, Amsterdam, 2012. 1–13
323 Rathore M M, Paul A, Ahmad A, et al. Hadoop based real-time intrusion detection for high-speed networks. In: Proceedings of the 2016 IEEE Global Communications Conference, Washington, 2016. 1–6
324 Marchal S, Jiang X, State R, et al. A big data architecture for large scale security monitoring. In: Proceedings of the 2014 IEEE International Congress on Big Data, Anchorage, 2014. 56–63
325 Giura P, Wang W. Using large scale distributed computing to unveil advanced persistent threats. Sci J, 2012, 1: 93–105
326 Bhatt P, Yano E T, Gustavsson P. Towards a framework to detect multi-stage advanced persistent threats attacks. In: Proceedings of the 8th IEEE International Symposium on Service Oriented System Engineering, Oxford, 2014. 390–395
327 Sharma P K, Moon S Y, Moon D, et al. DFA-AD: a distributed framework architecture for the detection of advanced persistent threats. Cluster Comput, 2017, 20: 597–609
328 Hameed S, Ali U. Efficacy of live ddos detection with hadoop. In: Proceedings of 2016 IEEE/IFIP Operations and Management Symposium, Istanbul, 2016. 488–494
329 Terzi D S, Terzi R, Sagiroglu S. Big data analytics for network anomaly detection from netflow data. In: Proceedings of International Conference on Computer Science and Engineering, Bangkok, 2017. 592–597
330 Francois J, Wang S, Bronzi W, et al. Botcloud: detecting botnets using mapreduce. In: Proceedings of the 2011 IEEE International Workshop on Information Forensics and Security, Iguacu Falls, 2011. 1–6
331 Jon-Michael B, Scot F, Dave S, et al. The treacherous 12: cloud computing top threats in 2016. Cloud Security Alliance, 2016. https://downloads.cloudsecurityalliance.org/assets/research/top-threats/Treacherous-12_Cloud-Computing_Top-Threats.pdf
332 Yang G, Ma J G, Yu A M, et al. Survey of insider threat detection. J Cyber Secur, 2016, 1: 21–36 [杨光, 马建刚, 于爱民, 等. 内部威胁检测研究. 信息安全学报, 2016, 1: 21–36]
333 Greitzer F, Purl J, Leong Y M, et al. SOFIT: sociotechnical and organizational factors for insider threat. In: Proceedings of 2018 IEEE Security and Privacy Workshops, San Francisco, 2018. 197–206
334 Böse B, Avasarala B, Tirthapura S, et al. Detecting insider threats using RADISH: a system for real-time anomaly detection in heterogeneous data streams. IEEE Syst J, 2017, 11: 471–482
335 Bilge L, Dumitras T. Before we knew it: an empirical study of zero-day attacks in the real world. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, Raleigh, 2012. 833–844
336 Win T Y, Tianfield H, Mair Q. Big data based security analytics for protecting virtualized infrastructures in cloud computing. IEEE Trans Big Data, 2018, 4: 11–25
337 Ullah F, Babar M A. Architectural tactics for big data cybersecurity analytics systems: a review. J Syst Softw, 2019, 151: 81–118
338 Alguliyev R, Imamverdiyev Y. Big data: big promises for information security. In: Proceedings of the IEEE 8th International Conference on Application of Information and Communication Technologies, Kazakhstan, 2014. 1–4

Research progress on big data security technology

Xingyuan CHEN[1,2*], Yuanzhao GAO[1,2], Huilin TANG[1] & Xuehui DU[1]

1. *Third Academy, Information Engineering University, Zhengzhou* 450001*, China*;
2. *State Key Laboratory of Cryptology, Beijing* 100094*, China*
* Corresponding author. E-mail: chxy302@vip.sina.com

Abstract As a new and energetic realm of economic development, an innovative engine of social development, and a strategic tool for shaping national competitiveness, big data significantly effects people's lives. However, improved social awareness of data value and vigorous development of big data platforms mean that big data security is increasingly hindering the promotion of big data applications. Meanwhile, as big data technology and framework continue to evolve, researchers still have different understandings of the core ideas and key features of big data security, and a unified big data security framework has yet to be established. Currently, determining the state-of-the-art of big data security technology is urgently needed to provide reference for research aimed at solving key big data security issues. Following a typical big data system technology framework, this review builds a novel big data security technology framework around big data security requirements. With this framework, state-of-the-art key big data security technologies are systematically summarized from three aspects: big data secure sharing and trusted services, big data platform security, and big data security supervision, which includes the main security mechanisms involved in big data business processes and system technology frameworks. Finally, big data security technology's core issues and development trends are summarized.

Keywords big data security, security technology framework, data secure sharing, platform security, security supervision

Xingyuan CHEN was born in 1963. He received his Ph.D. degree in communication and information systems from the Information Engineering University, Zhengzhou, in 2003. He is currently a professor at the Information Engineering University. His research interests include network and information security.

Yuanzhao GAO was born in 1992. He is a Ph.D. candidate at the Information Engineering University. His research interests include big data security.

Huilin TANG was born in 1981. He is a Ph.D. candidate at the Information Engineering University. His research interests include threat intelligence awareness.

Xuehui DU was born in 1968. She received her Ph.D. degree in cryptography from the Information Engineering University, Zhengzhou, in 2001. Currently, she is a professor at the Information Engineering University. Her research interests include information system multi-level security and cloud computing security.