数据安全与隐私保护技术研究

Research on Data Security and Privacy Protection Technology

刘明辉,张　玮,陈　洯,王　然(中国信息通信研究院,北京 100191)
Liu Minghui,Zhang Wei,Chen Tian,Wang Ran(China Academy of Information and Communication Technology,Beijing 100191,China)

摘　要:
数字经济时代,新业务、新架构、新技术对数据安全和隐私保护提出了新的挑战。研究了目前主流数据安全与隐私保护技术的发展现状与面临的问题,数据安全监控和防泄露技术相对成熟,而数据的共享安全、非结构化数据库的安全防护以及数据泄露溯源技术亟待改进。在个人隐私保护方面,技术的发展明显无法满足当前迫切的隐私保护需求,大数据应用场景下的个人信息隐私保护问题需要构建法律、技术、经济等多重手段相结合的保障体系。最后给出了研究建议。

关键词:
数据安全;隐私保护;脱敏;数据溯源;去标识化
doi:10.12045/j.issn.1007-3043.2019.04.006
中图分类号:TN915.08
文献标识码:A
文章编号:1007-3043(2019)04-0025-05

Abstract:
In the digital economy era,new services,new architectures and new technologies bring new challenges to data security and privacy protection. It studies the current development status and problems of mainstream data security and privacy protection technologies. Data security monitoring and data leak prevention technology are relatively mature,but data sharing security,unstructured database security protection and data leak traceability technology need to be improved urgently. In terms of personal privacy protection,the development of technology obviously can not meet the urgent needs of privacy protection. In the scene of big data application,a guarantee system combining legal,technical and economical means needs to be built for the privacy protection of personal information. Finally,the research suggestions are given.

Keywords:
Data security;Privacy protection;Desensitization;Data traceability;De-identification

引用格式:刘明辉,张玮,陈洯,等. 数据安全与隐私保护技术研究[J]. 邮电设计技术,2019(4):25-29.

0 引言

数字经济时代来临,数据价值急剧攀升,促使数据安全与国家安全、经济运行安全、社会公共安全、个人合法权益之间的关联日趋紧密。同时,数据面临的安全威胁日益严重,重大数据安全事件频发。数据泄露和隐私问题已经成为制约数字经济发展的关键因素,建立数据安全保障体系成为数字经济健康、稳定发展的重要环节。本文从分析数据安全面临的威胁和挑战入手,研究数据安全需求和解决数据安全问题的关键技术,分析现有安全技术存在的问题,提出数据安全技术的研究方向和研究建议。

收稿日期:2019-02-19

1 数据安全与隐私保护面临的威胁与挑战

大数据的体量大、种类多等特点,使得大数据环境下的数据安全出现了有别于传统数据安全的新威胁[1]。同时,数据采集、处理、分析的方式、工具和能力对传统个人隐私保护框架和技术能力亦带来了严峻挑战。

1.1 数据泄漏问题日趋严重

大数据因其蕴藏的巨大价值和集中化的存储管理模式,成为网络攻击的重点目标,针对大数据的勒

索攻击和数据泄露问题日趋严重，重大数据安全事件频发。金雅拓（Gemalto）《2018年数据泄露水平指数（Breach Level Index）》调查报告显示：仅2018年上半年，全球就发生了944起数据泄露事件，共计导致45亿条数据泄露[2]；从2013年到2018年全球数据泄露的数目呈现逐年上涨的趋势。

1.2 数据采集环节成为影响决策分析的新风险点

在数据采集环节，大数据体量大、种类多、来源复杂的特点为数据的真实性和完整性校验带来困难，目前，尚无严格的数据真实性、可信度鉴别和监测手段，无法识别并剔除虚假甚至恶意的数据。若黑客利用网络攻击向数据采集端注入脏数据，会破坏数据真实性，故意将数据分析的结果引向预设的方向，进而实现操纵分析结果的攻击目的。

1.3 数据处理过程中的机密性保障问题逐渐显现

数字经济时代，越来越多的企业或组织需要参与产业链协同，以数据流动与合作为基础进行生产活动。企业或组织在开展数据合作和共享的应用场景中，数据将突破组织和系统的边界进行流转，产生跨系统的访问或多方数据汇聚进行联合运算。保证个人信息、商业机密或独有数据资源在合作过程中的机密性，是企业或组织参与数据共享合作的前提，也是数据有序流动必须要解决的问题。

1.4 数据流动路径复杂化导致追踪溯源变得异常困难

大数据应用体系庞杂，频繁的数据共享和交换促使数据流动路径变得交错复杂，数据从产生到销毁不再是单向、单路径的简单流动模式，也不再仅限于组织内部流转，而会从一个数据控制者流向另一个控制者。在此过程中，实现异构网络环境下跨越数据控制者或安全域的全路径数据追踪溯源变得更加困难，特别是数据溯源中数据标记的可信性、数据标记与数据内容之间捆绑的安全性等问题更加突出。2018年3月的“剑桥分析”事件中，Facebook即是因为对第三方使用数据缺乏有效的管理和追责机制，最终导致8 700万名用户资料被滥用，造成股价暴跌、信誉度下降等严重后果。

1.5 传统隐私保护技术因大数据超强的分析能力面临失效的可能

在大数据环境下，企业对多来源多类型数据集进行关联分析和深度挖掘，可以复原匿名化数据，进而能够识别特定个人或获取其有价值的个人信息[3]。在传统的隐私保护中，数据控制者针对单个数据集孤立地选择隐私保护技术和参数来保护个人数据，特别是利用去标识、掩码等技术的做法，无法应对上述大数据场景下多源数据分析挖掘引发的隐私泄露问题。

1.6 传统隐私保护技术难以适应大数据的非关系型数据库

在大数据技术环境下，数据呈现动态变化、半结构化和非结构化数据居多的特性，对于占数据总量80%以上的非结构化数据，通常采用非关系型数据库（NoSQL）存储技术完成对大数据的抓取、管理和处理。而非关系型数据库目前尚无严格的访问控制机制及相对完善的隐私保护工具，现有的隐私保护技术，如脱敏、匿名化技术[4]等，多适用于关系型数据库。

2 数据安全技术

2.1 数据安全技术发展现状

目前，数据安全的防护一般是从数据生命周期防护的视角出发，设置分级分类的动态防护策略，降低已知风险的同时考虑减少对业务数据流动的干扰与伤害。

2.1.1 建立覆盖全生命周期的防御体系是数据安全防护的首要模式

随着大数据云计算技术的普及，业务系统数据融合共享成为趋势，数据流动性已经成为数据的基本特征，也引发了数据安全体系建设的变化[5]。从数据应用系统的角度看，数据从采集、传输到应用、共享直至销毁的全生命周期的各个环节均面临安全风险，数据安全防御体系必须贯穿生命周期的始终。在数据采集传输的安全防护环节，主要通过采集白名单、数据源操作权限管理、事前敏感字段标注、安全级别设置、静态脱敏、传输加密等技术来实现采集数据源、采集流程以及传输通道的安全防护；在数据存储处理环节，主要通过透明加密、数据完整性检验提高数据存储安全性，通过细粒度的权限管控、动态脱敏等技术保障数据应用的安全；在数据共享环节，针对数据流接口的方式，采用接口操作权限管理、流量管控、接口认证等方式保障接口的安全，针对文件共享的方式，通过数字水印等方式，实现共享数据泄露后的追踪（见图1）。

2.1.2 敏感数据识别和标注技术作为数据安全监控的必要技术条件逐步实现自动化

在敏感数据的监控方案中，基础部分就是从海量的数据中挑选出敏感数据，完成对敏感数据的识别，

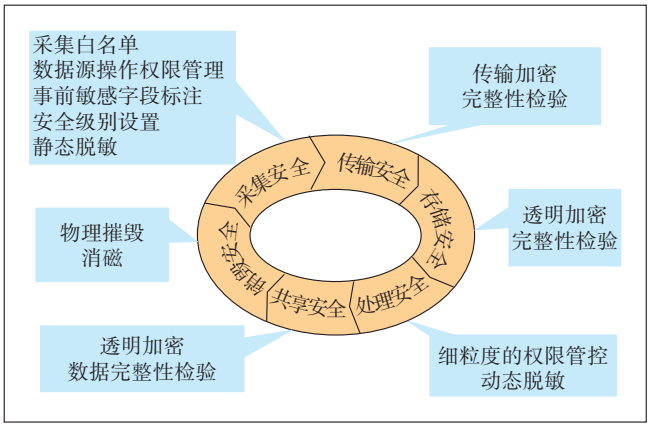

图1　数据全生命周期安全防护架构图

进而建立系统的总体数据视图,并采取分类分级的安全防护策略保护数据安全。传统的数据识别方法采用关键字、字典和正则表达式匹配等方式,通常结合模式匹配算法,该方法简单实用,但人工参与相对较多,自动化程度较低,随着人工智能识别技术的引入,通过机器学习可以实现大量文档的聚类分析,自动生成分类规则库,内容自动化识别程度正逐步提高。

2.1.3 数据防泄露技术发展相对成熟并向智能化演进

DLP是指通过一定的技术手段,防止用户的指定数据或信息资产以违反安全策略规定的形式流出企业的一类数据安全防护手段。针对数据泄露的主要途径,DLP采用的主要技术如下:针对使用泄露和存储泄露,通常采用身份认证管理、进程监控、日志分析和安全审计等技术手段,观察和记录操作员对计算机、文件、软件和数据的操作情况,发现、识别、监控计算机中的敏感数据的使用和流动,对敏感数据的违规使用进行警告、阻断等。针对传输泄露,通常采取敏感数据动态识别、动态加密、访问阻断和数据库防火墙等技术,监控服务器、终端以及网络中动态传输的敏感数据,发现和阻止敏感数据通过聊天工具、网盘、微博、FTP、论坛等方式泄露出去。目前的DLP,普遍引入了自然语言处理、机器学习、聚类分类等新技术,将数据管理的颗粒度进行了细化,对敏感数据和安全风险进行智能识别。"智能安全"将会成为DLP技术发展的趋势,大数据分析技术、机器学习算法的发展与演进将推动数据泄露防护的智能化发展,DLP将实现用户行为分析与数据内容的智能识别,实现数据的智能化分层、分级保护,并提供终端、网络、云端协同一体的敏感数据动态集中管控体系。

2.1.4 结构化数据库安全防护技术基本成熟,非结构化数据库安全防护亟需加强

结构化的数据安全技术主要是指数据库安全防护技术,可以分为事前评估加固、事中安全管控和事后分析追责3类,其中评估主要采用数据库漏洞扫描技术,安全管控主要采用数据库防火墙、数据加密、脱敏技术,事后分析追责主要采用数据库审计技术。目前数据库安全防护技术发展逐步成熟。而在针对云环境和大数据环境的安全方面,针对非结构化数据库的防护方案已经由一些技术领先的厂商提出,但技术成熟度较低。

2.1.5 密文计算技术因多源数据计算机密性需求成为研究热点

随着多源数据计算场景的增多,在保证数据机密性的基础上实现数据的流通和合作应用一直是困扰产业界的难题,同态加密和安全多方计算等密文计算方法为解决这个难题提供了一种有效的解决思路[6]。

同态加密提供了一种对加密数据进行处理的功能,对经过同态加密的数据处理得到一个输出,将这一输出进行解密,其结果与同一方法处理未加密的原始数据得到的输出结果一致。也就是说,其他人可以对加密数据进行处理,但是处理过程不会泄露任何原始内容。同时,拥有密钥的用户对处理过的数据进行解密后,得到的正好是处理后的结果。因为这一良好特性,同态加密特别适合在大数据环境中应用,既能满足数据应用的需求,又能保护用户隐私不被泄露,是一种理想的解决方案。2009年,Gentry提出了第1个全同态加密体制使得该方面的研究取得突破性进展,随后许多密码学家在全同态加密方案研究的基础上做出了有意义的工作,促进了全同态加密向实用化的发展[7-8],但是目前同态加密算法的计算开销过高,尚未应用到实际生产中。

安全多方计算(SMPC——Secure Multi-Party Computation)解决了一组互不信任的参与方之间保护隐私的协同计算问题,SMPC要确保输入的独立性,计算的正确性,同时不泄露各输入值给参与计算的其他成员。安全多方计算的这一特点,对于大数据环境下的数据机密性保护有独特的优势。通用的安全多方计算协议虽然可以解决一般性的安全多方计算问题,但是计算效率很低,尽管近年来研究者努力进行实用化技术的研究,并取得一些成果,但是离真正的产业化应用还有一段距离。

2.1.6 数字水印和数据溯源技术尚不满足实际需求

以上的数据识别、密文计算、安全监控和防护是“事前”和“事中”的安全保障技术,随着数据泄露事件的频繁发生,“事后”追踪和溯源技术变得越来越重要。安全事件发生后泄露源头的追查和责任的判定是及时发现问题、查缺补漏的关键,同时对安全管理制度的执行也会起到一定的威慑作用。目前常用的追踪溯源技术包括数字水印和数据溯源技术。

数字水印技术是为了保持对分发后的数据流向追踪[9],在数据泄露行为发生后,对造成数据泄露的源头可进行回溯。对于结构化数据,在分发数据中掺杂不影响运算结果的数据,采用增加伪行、增加伪列等方法,拿到泄密数据的样本,可追溯数据泄露源。对于非结构化数据,数字水印可以应用于数字图像、音频、视频、打印、文本、条码等数据信息中,在数据外发的环节加上隐蔽标识水印,可以追踪数据扩散路径。但目前的数字水印方案大多还是针对静态数据集,满足数据量巨大、更新速度极快的水印方案尚不成熟。

数据溯源技术又称为数据血缘、起源、谱系(Lineage,Provenance,Pedigree),是指数据产生的链路,数据溯源可以记载对数据处理的整个历史,包括数据的起源和处理这些数据的所有后继过程(数据产生并随着时间推移而演变的整个过程)。通过数据溯源,可以获得数据在数据流中的演化过程[10]。当数据发生异常时,通过数据溯源分析能追踪到异常发生的原因,把风险控制在适当的水平。目前数据溯源分析模型和技术方案多是针对组织内部的数据流动与溯源,无法应用于数据跨组织流动的溯源场景[11]。

2.2 数据安全技术问题分析

目前,数据安全监控和防泄露技术相对成熟,数据的共享安全、非结构化数据库的安全防护以及数据泄露溯源技术亟待改进。目前,数据泄露问题在技术上可以得到较完备的解决,敏感数据自动化识别为防泄露提供了基础技术;人工智能、机器学习等技术的引入,使得数据防泄露向智能化方向演进;数据库防护技术的发展也为数据防泄露提供了有力的技术保障。密文计算技术、数据泄露追踪技术的发展仍无法满足实际的应用需求,难以解决数据处理过程的机密性保障问题和数据流动路径追踪溯源问题。具体而言,密文计算技术的研究仍处在理论阶段,运算效率远未达到实际应用的要求;数字水印技术无法满足大数据环境下大量、快速更新的应用需求;数据溯源技术无法应用于数据跨组织流动的溯源场景。

3 个人隐私保护技术

3.1 个人隐私保护技术发展现状

数据安全技术提供了数据机密性、完整性和可用性的防护基础,隐私保护是在此基础上,保证个人隐私信息不发生泄露或不被外界知悉。目前应用最广泛的是数据脱敏技术,学术界也提出了同态加密、安全多方计算等可用于隐私保护的密码算法。

3.1.1 数据脱敏技术发展成熟,是目前应用最广泛的隐私保护技术

数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形,实现对个人数据的隐私保护,是应用最广泛的隐私保护技术。目前的脱敏技术主要分为3种:第1种加密方法,是指标准的加密算法,加密后完全失去业务属性,属于低层次脱敏,算法开销大,适用于机密性要求高、不需要保持业务属性的场景;第2种基于数据失真的技术,最常用的是随机干扰、乱序等,是不可逆算法,通过这种算法可以生成“看起来很真实的假数据”,适用于群体信息统计或(和)需要保持业务属性的场景;第3种可逆的置换算法,兼具可逆和保证业务属性的特征,可以通过位置变换、表映射、算法映射等方式实现,表映射方法应用起来相对简单,也能解决业务属性保留的问题,但是随着数据量的增大,相应的映射表同量增大,应用局限性高。算法映射方法不需要做映射表,通过自行设计的算法来实现数据的变换,这类算法都是基于密码学的基本概念自行设计的,通常的做法是在公开算法的基础上做一定的变换,适用于需要保持业务属性或(和)需要可逆的场景。数据应用系统在选择脱敏算法时,可用性和隐私保护的平衡是关键,既要考虑系统开销,满足业务系统的需求,又要兼顾最小可用原则,最大限度地保护用户隐私。

3.1.2 去标识化将成为解决隐私保护问题的有效途径

去标识化(de-identification)是指通过对个人信息的技术处理,在不借助额外信息的情况下,无法识别个人信息主体的过程。去标识化是隐私保护数据发布的主要工具之一,通过去除数据集中隐私属性和数据主体之间的关联关系,并且具有足够的防止重识别能力后,数据集的某些属性就可以共享发布,供外部业务系统进行处理分析。常用的去标识化技术包括统计技术、密码技术、抑制技术、假名化技术、泛化技术、随机化技术和数据合成技术,常用的去标识化模

型包括K-匿名模型和差分隐私模型。

国际标准化组织（ISO）和国际电工委员会（IEC）2018年发布的ISO/IEC 20889规定了去标识化有关的术语、技术以及应用原则。NIST于2016年发布NIST SP 800-188，为政府机构提供数据去标识化技术指导，包括建立和改进去标识化程序、去标识化的技术步骤、去标识化工具的需求以及评价去标识化工具方法等内容。国内方面，TC260正在编制的《个人信息去标识化指南》，借鉴国内外个人信息去标识化的最新研究理论，研究个人信息去标识化的目标、原则、技术、模型、过程和组织措施，提出符合我国信息化发展需要的个人信息去标识化指南。

金融行业在推进个人信息去标识化方面走在前列。国际芯片卡标准化组织（EMVCo）2014年发布支付令牌化技术框架，提出了在支付场景中使用一个不同的号码串替换银行卡主账号的过程规范。中国银联于2016年发布《中国银联支付标记化技术指引》，给出了使用支付令牌代替银行卡号进行交易验证的框架、技术要求和应用场景。

学术界在去标识化算法的改进方面取得了一些进展。学术界的研究方向主要集中在对算法的抗攻击、计算效率提高、降低用户设备开销和减少对数据特性的影响等方面。

3.2 个人隐私保护技术问题分析

在个人隐私保护方面，技术的发展明显无法满足当前迫切的隐私保护需求，大数据应用场景下的个人信息隐私保护问题需要构建法律、技术、经济等多重手段相结合的保障体系。目前，应用广泛的数据脱敏技术受到多源数据汇聚的严重挑战可能面临失效，去标识化等前沿技术目前正在不断研究和推进中，但普遍存在运算效率过低、开销过大等问题，还需要在算法的优化方面进行持续改进，以满足大数据环境下的隐私保护需求。同时，大数据应用与个人信息隐私保护之间的突出矛盾不单是技术问题，尤其是在缺乏技术保障的当下，更需要通过加快立法、加强执法，规范大数据应用场景下的个人信息收集、使用行为，尽快构建政府管理、企业履责、社会监督、网民自律等多主体共同参与的个人信息保护制度体系。

4 结论

数据在流动中发挥价值，大数据环境下，数据应用生态环境日益复杂，数据生命周期各环节都面临新的安全保障需求，数据的采集和溯源成为突出的安全风险点，跨组织数据合作的广泛开展触发了多源汇聚计算的机密性保障需求。应加强数据采集、运算、溯源等关键环节的保障能力建设，强化数据安全监测、预警、控制和应急处置能力，以数据安全关键环节和关键技术的研究为突破点，完善大数据安全技术体系，促进整个大数据产业的健康发展。

在大数据应用场景下，数据利用和隐私保护是天然矛盾的两端，同态加密、去标识化等技术可以实现这两者良好的平衡，是解决大数据应用过程中隐私保护问题的理想技术，隐私保护核心技术方面的进展必然会极大推动大数据应用的发展。目前，隐私保护技术的核心问题是效率，应加强降低计算开销和存储开销的研究，提升大数据环境下隐私保护技术水平。

参考文献：

[1] 刘艺，邓青，彭雨苏. 大数据时代数据主权与隐私保护面临的安全挑战[J]. 管理现代，2019(01)：104-107.
[2] GEMALTO. Breach Level Index Report H1-2018-Gemalto[EB/OL]. [2019-01-22]. https://bbs.pinggu.org/a-2623705.html.
[3] 冯登国，张敏，李昊. 大数据安全与隐私保护[J]. 计算机学报，2014(1)：246-258.
[4] 方滨兴，贾焰，李爱平，等. 大数据隐私保护技术综述[J]. 大数据，2016(1)：1-18.
[5] 王 康，王晓慧. 国内数据安全研究热点与前沿分析[J]. 新世纪图书馆，2018(9)：88-91.
[6] 蒋瀚，徐秋亮. 实用安全多方计算协议关键技术研究进展[J]. 计算机研究与发展，2015(10)：2247-2257.
[7] 曹珍富，董晓蕾，周俊，等. 大数据安全与隐私保护研究进展[J]. 计算机研究与发展，2016(10)：2137-2151.
[8] 刘明洁，王安. 全同态加密研究动态及其应用概述[J]. 计算机研究与发展，2014(12)：2593-2603.
[9] 朱倩，李雪燕. 数字水印技术在大数据安全保护中的应用[J]. 软件导刊，2016(1)：153-155.
[10] 戴超凡，王涛，张鹏程. 数据溯源技术发展研究综述[J]. 计算机应用研究，2010，27(9)：3216-3221.
[11] 明华，张勇，符小辉. 数据溯源技术综述[J]. 小型微型计算机系统，2012(9)：1917-1923.

作者简介：
刘明辉，毕业于北京邮电大学，高级工程师，博士，主要研究方向为数据安全与隐私保护技术、大数据平台安全、物联网安全与智能卡；张玮，毕业于北京邮电大学，工程师，硕士，主要研究方向为数据安全风险评估、数据安全技术与隐私保护；陈涤，毕业于北京邮电大学，高级工程师，硕士，主要研究方向为数据安全与隐私保护技术、数据安全与隐私法律与政策；王然，毕业于北京科技大学，工程师，硕士，主要研究方向为数据安全、大数据安全技术咨询服务以及数据安全相关标准。