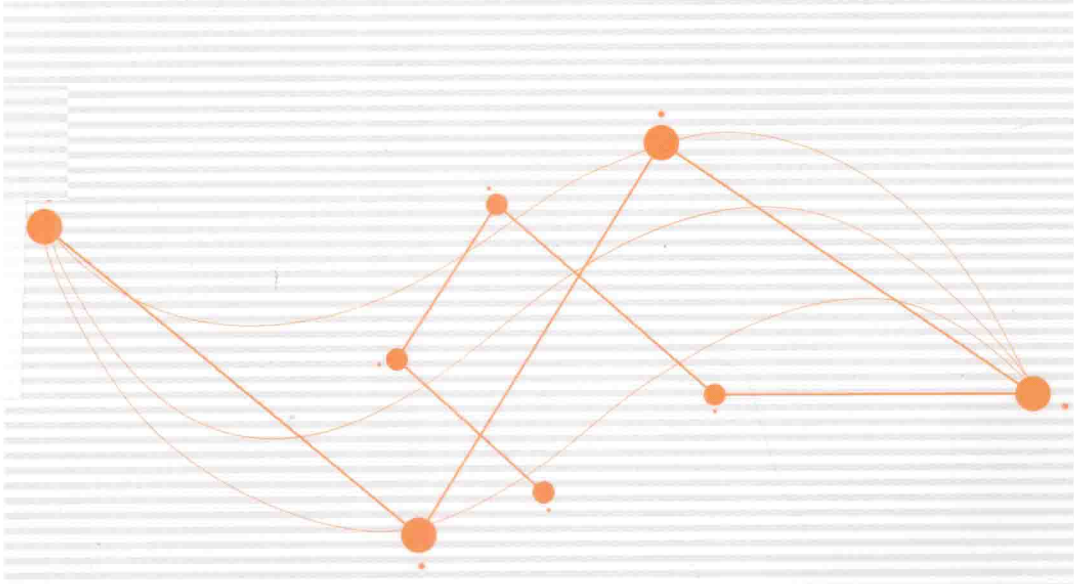
nnovative Research

中国联通研究院创新研究系列丛书

大数据安全技术与应用

张尼　张云勇　胡坤　刘明辉　宫雪　陶冶　等　编著



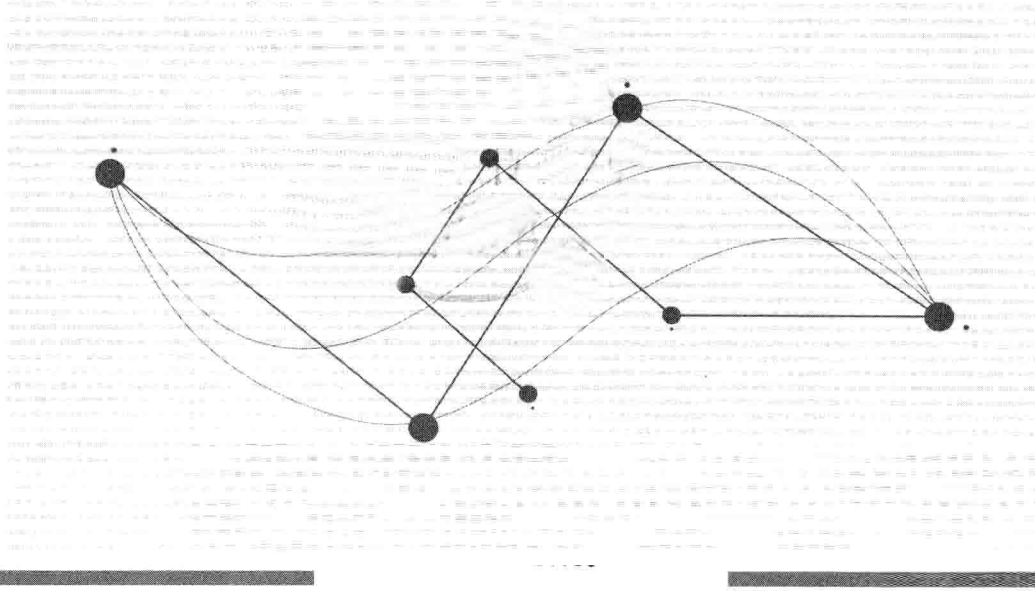
人民邮电出版社
POSTS & TELECOM PRESS

nnovative Research

中国联通研究院创新研究系列丛书

大数据安全技术与应用

张尼　张云勇　胡坤　刘明辉　宫雪　陶冶　等　编著



人民邮电出版社

北京

图书在版编目（CIP）数据

大数据安全技术与应用 / 张尼等编著. -- 北京 : 人民邮电出版社, 2014.5
（中国联通研究院创新研究系列丛书）
ISBN 978-7-115-34329-1

Ⅰ. ①大… Ⅱ. ①张… Ⅲ. ①数据处理－安全技术 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2013)第319220号

内 容 提 要

本书以大数据发展历史、特征、发展趋势为切入点，分析各领域面临的大数据安全威胁和需求，归纳总结大数据安全的科学内涵和技术研究方向。在此基础上，引出大数据安全的关键技术和应用实践。随后对大数据安全的产业动态、法律法规、标准研究进行系统梳理，预测大数据安全的发展趋势。

本书融通俗性、完整性、实用性、丰富性于一体，有助于广大读者理解大数据安全的基本内容、核心技术、使用机制等。本书可作为高等院校信息安全专业本科生和研究生的参考教材，也可作为IT工程技术人员、大数据应用研究人员、信息安全从业人员的参考书。

◆ 编　著　张　尼　张云勇　胡　坤　刘明辉　宫　雪　陶　冶　等
责任编辑　邢建春
责任印制　杨林杰

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京铭成印刷有限公司印刷

◆ 开本：700×1000　1/16
印张：14　　2014 年 5 月第 1 版
字数：274 千字　　2014 年 5 月北京第 1 次印刷

定价：59.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

丛书编委会名单

本书编写组

主编： 张　尼　张云勇　胡　坤　刘明辉　宫　雪

编著： 陶　冶　刘　镝　陈　豪　李　正　申珉宇

王笑帝　房秉毅　魏进武　郭志斌　程　莹

徐　雷　李素粉　汪　芳　王淑玲　周　巍

张　立　冯伟斌　张　园　李　卫　李　丹

张基恒　汤雅妃

序

信息化时代瞬息万变，当移动互联网、云计算、物联网、社交网络等新技术发展方兴未艾之时，大数据技术又开始悄然兴起。当前，企业的数据每隔 1～2 年就增加一倍，呈现出数据量大（Volume）、产生速度快（Velocity）、数据来源复杂（Variety）、潜在价值高（Value）等特性。随着大数据的发展和渗透，人们在生产与实践中逐渐认识到，通过数据的开放、整合和分析，能够发现新知识、创造新价值，从而为社会带来科技进步、经济增长和创新发展。

据 IDC 研究显示，到 2015 年大数据市场规模将达到 169 亿美元，大数据正在催生新的经济增长点，并从商业行为上升到国家发展战略。作为大数据的策源地和创新引领者，美国政府一直在积极推动大数据的发展。2012 年，奥巴马政府发布了“大数据研究和发展倡议”，联邦政府的多个部门（国家科学基金委、国家卫生研究院、能源部、国防部、国防部高级研究计划局、地质勘探局等）都在部署研究大数据的相关项目。与此同时，我国的大数据产业生态环境也在加速形成与完善，许多互联网公司、电信企业都拥有大数据，有迫切且明确的应用需求，具备了大数据的运营条件。综观国内外大数据的发展趋势可知，大数据战略在本质上应该是“需求牵引、效益驱动、技术支撑”三位一体的，在战略中，企业、政府和研究机构的定位应各有侧重，以企业为主、政府为辅，这是与大数据的应

用牵引本质相符的。

大数据安全涉及两个方面：大数据自身的安全和大数据助力于信息安全。其中，大数据自身的安全包括 4 个层面。一是设备可靠。处理大规模数据涉及的设备众多，设备可靠性成为大数据安全的基础问题。二是系统安全。一方面，大数据平台庞大的计算环境存在系统复杂，运行不稳定的风险；另一方面，大数据分析过程中产生的知识和价值容易引发黑客攻击，因此，大数据系统需要完善安全机制。三是数据可信。大数据挖掘通常需要依赖云计算平台的存储和计算能力，因此，可能会出现数据被云服务商破坏和窃取的情况，而大数据来源的繁杂性，也使得对数据的合规性和真实性检查成为必要，典型的例子是网络恶评，这类数据的真实性、客观性有待商榷。四是隐私保护。这是大数据大量、多源特征引发的新问题。过去人们发布数据时只是简单地隐藏部分敏感信息，但大数据技术出现后，一些较为隐秘的信息都有可能被挖掘出来，因此，亟需更为先进、强大的技术手段，能够在不侵犯用户隐私的前提下对大数据进行有效地分析、开放和共享。

简言之，大数据出现之后，数据处理规模更大，对系统可靠性的要求更高；要处理的数据更为集中，使得安全风险更大；对外部系统的依赖更明显，使得自主可控的能力更弱；用户的数据收集工作更加完备，大数据的处理能力导致用户隐私暴露的风险更大。可见，大数据使得信息安全风险显著增加，成为信息安全的新挑战。

与此同时，大数据的商业价值和市场需求也在推动大数据处理理论与技术快速发展。信息安全领域通常面临着数据量大、产生速度快、数据类型各异、计算复杂、挖掘难度大等问题，一旦大数据处理理论与技术出现了质的突破，则意味着信息安全的处理能力也会有质的飞跃。例如，对海量数据的分析将会有助于更好地刻画网络异常行为，从而找出网络中的风险点。这就是信息安全在大数据时

代所面临的机遇，大数据与信息安全的整合能够贯穿于产业链的各环节，有效推动信息安全技术的进步，加速信息安全技术与其他学科技术的融合与发展。

综上所述，伴随着大数据技术的快速发展，大数据安全将成为当下研究和关注的热点。

本书的主要作者张尼曾经是我的博士研究生，他及所在的安全团队在中国联通研究院长期从事信息安全研发工作。近年来，他们又代表联通公司在国内、外标准化组织、国家项目的示范应用中积极开展了大数据系统安全、用户隐私保护等应用实践。为了迎接大数据发展契机，中国联通研究院撰写了《大数据安全技术与应用》一书，这是对大数据安全技术的背景、内涵、技术研究方向、产业现状及实践的一次深入分析和高度概括，为国内的相关从业人员提供了一份全面的参考资料。

中国工程院院士 方滨兴

序 言

由硬件摩尔定律、云计算、数据挖掘作为技术推动，移动计算、社交网络作为业务推动，催生了大数据技术的产生，并建立起了迅猛发展的生态体系。大数据已经逐渐成为当今时代最关键的生产要素和产品形态，也是当前社会从工业经济向知识经济转变的重要特征。大数据将代表着信息技术未来发展的战略走向，以大数据为代表的数据密集型科学将成为新一代技术变革的基石。

大数据带来的不仅仅是数据体量的变化，更是思维的变革——样本数据或局部数据向全体数据的变革，结果数据向过程数据的变革，静态存储数据向动态流处理数据的变革。在大量数据产生、收集、存储和分析的过程中，既会涉及一些传统安全问题，也涉及一些新的安全问题，并且这两类问题会随着数据规模、处理过程、安全要求等因素而被放大。同时，大数据的 4V+1C 特性，也使得大数据在技术、管理和法律等方面面临新的安全威胁与挑战。此外，大数据也引发了信息安全技术的革命，通过对结构化和非结构化数据的分析与应用，能够从海量数据中发现潜在的风险，预测未来的趋势与变化。因此，随着大数据技术的研究与应用，大数据安全问题成为学术界、产业界争相关注的焦点。

本书从各个行业的大数据安全需求说起，提出大数据安全体系架构，阐述大数据安全的关键技术、标准现状、行业应用实践等，并结合信息产业发展现状对

大数据安全发展趋势进行展望。全书是对大数据安全的一次深度总结和分析。

如今，大数据已经发展成为全社会资源，各个行业既是大数据的创造者，也是大数据的消费者。面对未来，各个行业、各个领域应积极盘活大数据资产，发挥大数据的社会和经济价值，为拉动信息消费、形成以大数据服务产业为核心的高黏性信息服务产业生态，做出应有的贡献。

前 言

随着信息技术的飞速发展，全球数据信息量呈现指数式爆炸增长之势，数据体量从 PB 级跃升至 ZB 级。根据国际数据公司（IDC）发布的 2012 年研究报告显示，2011 年全球创建和复制的数据总量为 1.8 ZB，并且以每两年翻一番的速度快速增长；预计到 2020 年，全球产生的数据总量将超过 40 ZB，将是地球上所有海滩上沙粒数量的 57 倍。

由于数据体量的激增、结构类型的复杂、单数据的低密度价值以及处理速度的提升等新特性的出现，促使人们对大数据进行研究与实践。对大数据进行全面、深入、实时的分析和应用，能够使企业更加精准地洞察客户需求，提升企业自身智能化水平和行业信息化服务能力，并对外提供数据挖掘和分析的新业务及服务。因此，已成为现代企业应对新挑战、开创新模式、拓展新市场的重要工作内容。现阶段，大数据已逐渐渗透到各个行业和业务职能领域，成为重要的生产因素，为企业增强竞争优势、拓展蓝海业务带来新的机会。可以这样讲，大数据正以前所未有的速度，颠覆人们探索世界的方法，驱动产业间的融合与分立。

但是，在大量数据产生、收集、存储和分析的过程中，会面临数据保密、用户隐私、商业合作等一系列问题。这既涉及一些传统安全问题，例如：物理安全、设备安全、网络安全、数据库安全、系统安全等，也涉及一些新的安全问题，例

如：因数据散乱在众多系统中，信息来源十分庞杂而带来的数据采集安全；因数据种类和业务类型众多而带来的数据整合与存储安全；因外部数据需求和用户隐私保护而带来的数据审计和安全发布问题等。而现有的数据安全机制并不能满足大数据的安全需求，数据的分布式、精细化处理进一步加大了数据泄露的风险，企业存储的大数据将成为黑客攻击的显著目标，并成为高级可持续攻击的载体。安全问题正成为制约大数据技术发展的瓶颈。因此，对大数据进行全面而完整的安全防护变得尤为重要。

此外，大数据技术也为信息安全注入新的活力。通过大数据技术可以有效识别病毒、木马，甚至 APT 攻击，为移动互联网安全、云计算安全提供新型解决方案。针对安全数据的挖掘与分析可以表征企业整体运行情况，及时发现安全隐患，同时也能够表现安全运行趋势，预测业务走向，反应系统闲忙，再现网络使用情况等潜在信息。基于流量数据的大数据分析可以感知网络态势，反映网络流量突变、快速定位异常、显示未来趋势。而通过对视频监控数据的大数据分析能够实现视频图像模糊查询、快速检索、精准定位等功能。

因此，对于各行各业的用户而言，无论是构建大数据安全保障体系，还是开发大数据安全应用产品，都需要了解、掌握大数据安全的科学内涵和关键技术。

本书以此为背景，从大数据的基本知识说起，分析各个行业的大数据安全需求，提出大数据安全体系架构，阐述大数据安全的关键技术、标准现状、行业应用实践等。最后，结合信息产业发展现状对大数据安全发展趋势进行展望。本书内容可以为大数据平台部署提供有益参考，对于完善大数据技术体系，保障大数据技术推广应用具有重要的意义。

本书结构如下：第 1 章和第 2 章分析了大数据的基本理论、发展现状与云计算的关系；第 3 章介绍大数据梳理大数据安全的产业动态、法律法规、标准研究

情况；第 4 章和第 5 章介绍了大数据在不同层面面临的安全威胁与挑战，分析了各领域存在的大数据安全需求，归纳总结大数据安全的科学内涵和技术研究方向；第 6～9 章从大数据安全保障和大数据应用安全两个方面阐述大数据安全的技术原理、实现细节、产业实践；第 10 章对大数据安全的后续发展进行展望，并给出合理建议。

本书适合 IT 工程技术人员、大数据应用研究人员、信息安全从业人员阅读，也可作为高等院校信息安全专业本科生和研究生的参考教材。

本书具有如下 3 大特色。

1）通俗性

本书介绍了大数据安全的基本知识，涵盖了从具体技术原理到实际应用案例的相关知识。读者只需具备基本的 IT 知识即可。每章的标题就是对该章内容的高度概括，在接下来的内容中对其进行的解释已尽可能做到了准确、详实。

2）完整性

本书从大数据安全的科学内涵、技术解析、关键要素、实现细节到具体应用案例都进行了周详的论述。

3）实用性

本书紧密结合大数据应用实际，从安全需求、科学分析到技术支持、应用实践等各方面进行分析和论述。

本书由丛书编委会负责策划和统稿。第 1 章由张尼、胡坤编写，第 2、3、6 章由张云勇、张尼编写，第 7、10 章由刘明辉、陶冶编写，第 4、5、8、9 章由胡坤、宫雪编写。

参加研究和写作的成员还有：刘镝、陈豪、李正、申珉宇、王笑帝、孙兆欣、匡斌、房秉毅、魏进武、郭志斌、程莹、徐雷、李素粉、汪芳、王淑玲、周巍、

张立、冯伟斌、张园、李卫、马书惠、李丹、张基恒、汤雅妃、贾宝军。

本书能够出版，需要感谢中国联通研究院陈赤航副院长、孙海滨副院长、黄文良副院长、信息室孙兆欣主任、匡斌副主任、范云杰编辑，中国联通集团技术部顾闵霞经理、裴小燕经理、彭久生经理、黄文利、徐克航、林敏，陈博，中国联通集团产品创新部许海翔经理、于鹏经理、李楠经理、顾芳、马丽，中国联通集团电子商务与信息化事业部娄瑜经理、王志山经理、刘海舟经理、刘险峰，北京邮电大学姚海鹏、黄韬、郭达副教授、四川大学计算机学院彭舰院长的帮助。

本书凝聚了笔者长期的安全实践经验以及研究思考的成果。在本书的编写过程中，广泛收集了国内外相关材料，参考了一些最新论著，并引用了部分材料，在此向其著作人表示感谢。人民邮电出版社的邢建春编辑为此书倾注了大量的心血，在此致以诚挚的谢意。

本书内容是作者的大胆探索和思考，仅代表个人观点，与任何机构的立场无关。我们希望通过大家的共同努力，理清大数据安全的科学内涵、关键技术、应用机制及其未来发展趋势，为大数据业务的安全应用与创新发展贡献一份力量。由于作者水平有限，加之时间仓促，书中难免有错误、不当之处，恳请广大专家、学者不吝批评指正。

作者

2014 年 2 月于北京

目录 Contents