大数据安全与隐私保护

冯登国 张 敏 李 昊

(中国科学院软件研究所 可信计算与信息保障实验室 北京 100190)

摘 要 大数据(Big Data)是当前学术界和产业界的研究热点,正影响着人们日常生活方式、工作习惯及思考模式.但目前大数据在收集、存储和使用过程中面临着诸多安全风险,大数据所导致的隐私泄露为用户带来严重困扰,虚假数据将导致错误或无效的大数据分析结果.该文分析了实现大数据安全与隐私保护所面临的技术挑战,整理了若干关键技术及其最新进展.分析指出大数据在引入安全问题的同时,也是解决信息安全问题的有效手段.它为信息安全领域的发展带来了新的契机.

关键词 大数据;大数据安全;隐私保护;信息安全

中图法分类号 TP309 **DOI号** 10.3724/SP.J.1016.2014.00246

Big Data Security and Privacy Protection

FENG Deng-Guo ZHANG Min LI Hao

(*Trusted Computing and Information Assurance Laboratory*, *Institute of Software*, *Chinese Academy of Sciences*, *Beijing* 100190)

Abstract Nowadays big data has become a hot topic in both the academic and the industrial research. It is regarded as a revolution that will transform how we live, work and think. However, there are many security risks in the field of data security and privacy protection when collecting, storing and utilizing big data. Privacy issues related with big data analysis spell trouble for individuals. And deceptive or fake information within big data may lead to incorrect analysis results. This paper summarizes and analyzes the security challenges brought by big data, and then describes the key technologies which can be exploited to deal with these challenges. Finally, this paper argues that big data brings not only challenges, but also technical revolution in the field of information security.

Keywords big data; big data security; privacy protection; information security

1 引 言

当今,社会信息化和网络化的发展导致数据爆炸式增长.据统计,平均每秒有200万用户在使用谷歌搜索,Facebook用户每天共享的东西超过40亿,Twitter每天处理的推特数量超过3.4亿.同时,科学计算、医疗卫生、金融、零售业等各行业也有大量数据在不断产生.2012年全球信息总量已经达到2.7ZB,而到2015年这一数值预计会达到8ZB.

这一现象引发了人们的广泛关注.在学术界,图灵奖获得者Jim Gray提出了科学研究的第四范式,即以大数据为基础的数据密集型科学研究;2008年《*Nature*》推出了大数据专刊对其展开探讨;2011年《*Science*》也推出类似的数据处理专刊.IT产业界行动更为积极,持续关注数据再利用,挖掘大数据的潜

收稿日期:2013-07-22;最终修改稿收到日期:2013-11-26.本课题得到国家自然科学基金(91118006,61232005,61100237)、国家"八六三"高技术研究发展计划项目基金(2011AA0123824001)资助.冯登国,男,1965年生,博士,研究员,主要研究领域为信息安全与密码学、可信计算与信息保障.E-mail:fengdg@263.net.张 敏,女,1975年生,博士,副研究员,主要研究方向为数据隐私保护、可信计算与云存储安全.李 昊,男,1983年生,博士,助理研究员,主要研究方向为数据隐私保护与可信计算.

在价值. 目前，大数据已成为继云计算之后信息技术领域的另一个信息产业增长点. 据 Gartner 预测，2013 年大数据将带动全球 IT 支出 340 亿美元，到 2016 年全球在大数据方面的总花费将达到 2320 亿美元. Gartner 将“大数据”技术列入 2012 年对众多公司和组织机构具有战略意义的十大技术与趋势之一. 不仅如此，作为国家和社会的主要管理者，各国政府也是大数据技术推广的主要推动者. 2009 年 3 月美国政府上线了 data. gov 网站，向公众开放政府所拥有的公共数据. 随后，英国、澳大利亚等政府也开始了大数据开放的进程，截至目前，全世界已经正式有 35 个国家和地区构建了自己的数据开放门户网站①. 美国政府联合 6 个部门宣布了 2 亿美元的“大数据研究与发展计划”. 在我国，2012 年中国通信学会、中国计算机学会等重要学术组织先后成立了大数据专家委员会，为我国大数据应用和发展提供学术咨询.

目前大数据的发展仍然面临着许多问题，安全与隐私问题是人们公认的关键问题之一[1-2]. 当前，人们在互联网上的一言一行都掌握在互联网商家手中，包括购物习惯、好友联络情况、阅读习惯、检索习惯等等. 多项实际案例说明，即使无害的数据被大量收集后，也会暴露个人隐私[1]. 事实上，大数据安全含义更为广泛，人们面临的威胁并不仅限于个人隐私泄漏. 与其它信息一样，大数据在存储、处理、传输等过程中面临诸多安全风险，具有数据安全与隐私保护需求. 而实现大数据安全与隐私保护，较以往其它安全问题（如云计算中的数据安全等）更为棘手. 这是因为在云计算中，虽然服务提供商控制了数据的存储与运行环境，但是用户仍然有些办法保护自己的数据，例如通过密码学的技术手段实现数据安全存储与安全计算，或者通过可信计算方式实现运行环境安全等. 而在大数据的背景下，Facebook 等商家既是数据的生产者，又是数据的存储、管理者和使用者，因此，单纯通过技术手段限制商家对用户信息的使用，实现用户隐私保护是极其困难的事[1].

当前很多组织都认识到大数据的安全问题，并积极行动起来关注大数据安全问题. 2012 年云安全联盟 CSA 组建了大数据工作组，旨在寻找针对数据中心安全和隐私问题的解决方案. 本文在梳理大数据研究现状的基础上，重点分析了当前大数据所带来的安全挑战，详细阐述了当前大数据安全与隐私保护的关键技术. 需要指出的是，大数据在引入新的安全问题和挑战的同时，也为信息安全领域带来了新的发展契机，即基于大数据的信息安全相关技术可以反过来用于大数据的安全和隐私保护. 本文在第 5 节对其进行了初步分析与探讨.

2　大数据研究概述

2.1　大数据来源与特征

普遍的观点认为，大数据是指规模大且复杂、以至于很难用现有数据库管理工具或数据处理应用来处理的数据集②. 大数据的常见特点包括大规模（volume）、高速性（velocity）和多样性（variety）.

根据来源的不同，大数据大致可分为如下几类[3]：

（1）来自于人. 人们在互联网活动以及使用移动互联网过程中所产生的各类数据，包括文字、图片、视频等信息；

（2）来自于机. 各类计算机信息系统产生的数据，以文件、数据库、多媒体等形式存在，也包括审计、日志等自动生成的信息；

（3）来自于物. 各类数字设备所采集的数据. 如摄像头产生的数字信号、医疗物联网中产生的人的各项特征值、天文望远镜所产生的大量数据等.

2.2　大数据分析目标

目前大数据分析应用于科学、医药、商业等各个领域，用途差异巨大. 但其目标可以归纳为如下几类：

（1）获得知识与推测趋势.

人们进行数据分析由来已久，最初且最重要的目的就是获得知识、利用知识. 由于大数据包含大量原始、真实信息，大数据分析能够有效地摒弃个体差异，帮助人们透过现象、更准确地把握事物背后的规律. 基于挖掘出的知识，可以更准确地对自然或社会现象进行预测. 典型的案例是 Google 公司的 Google Flu Trends 网站. 它通过统计人们对流感信息的搜索，查询 Google 服务器日志的 IP 地址判定搜索来源，从而发布对世界各地流感情况的预测③. 又如，人们可以根据 Twitter 信息预测股票行情④等.

（2）分析掌握个性化特征.

个体活动在满足某些群体特征的同时，也具有

① http：//www. chinaeg. gov. cn/show-4150. html
② 维基百科 http：//en. wikipedia. org/wiki/Big_data
③ http：//www. google. org/flutrends/
④ http：//tech2ipo. com/6322/

鲜明的个性化特征. 正如“长尾理论”中那条细长的尾巴那样,这些特征可能千差万别. 企业通过长时间、多维度的数据积累,可以分析用户行为规律,更准确地描绘其个体轮廓,为用户提供更好的个性化产品和服务,以及更准确的广告推荐. 例如 Google 通过其大数据产品对用户的习惯和爱好进行分析,帮助广告商评估广告活动效率,预估在未来可能存在高达到数千亿美元的市场规模①.

(3)通过分析辨识真相.

错误信息不如没有信息. 由于网络中信息的传播更加便利,所以网络虚假信息造成的危害也更大. 例如,2013 年 4 月 24 日,美联社 Twitter 帐号被盗,发布虚假消息称总统奥巴马遭受恐怖袭击受伤. 虽然虚假消息在几分钟内被禁止,但是仍然引发了美国股市短暂跳水. 由于大数据来源广泛及其多样性,在一定程度上它可以帮助实现信息的去伪存真. 目前人们开始尝试利用大数据进行虚假信息识别. 例如,社交点评类网站 Yelp 利用大数据对虚假评论进行过滤,为用户提供更为真实的评论信息②; Yahoo③和 Thinkmail④等利用大数据分析技术来过滤垃圾邮件.

2.3 大数据技术框架

大数据处理涉及数据的采集、管理、分析与展示等. 图 1 是相关技术示意图.

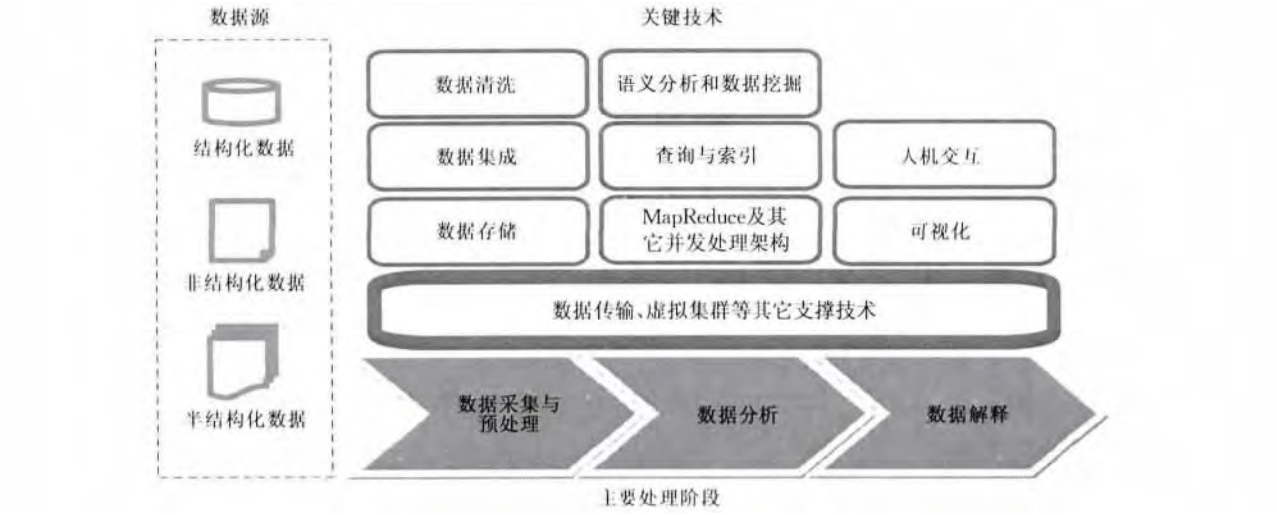

图 1 大数据技术架构

(1)**数据采集与预处理**(Data Acquisition & Preparation).

大数据的数据源多样化,包括数据库、文本、图片、视频、网页等各类结构化、非结构化及半结构化数据. 因此,大数据处理的第一步是从数据源采集数据并进行预处理操作,为后继流程提供统一的高质量的数据集.

由于大数据的来源不一,可能存在不同模式的描述,甚至存在矛盾. 因此,在数据集成过程中对数据进行清洗,以消除相似、重复或不一致的数据是非常必要的. 文献[4-7]中数据清洗和集成技术针对大数据的特点,提出非结构化或半结构化数据的清洗以及超大规模数据的集成.

数据存储与大数据应用密切相关. 某些实时性要求较高的应用,如状态监控,更适合采用流处理模式,直接在清洗和集成后的数据源上进行分析. 而大多数其它应用则需要存储,以支持后继更深度的数据分析流程. 为了提高数据吞吐量,降低存储成本,通常采用分布式架构来存储大数据. 这方面有代表性的研究包括:文件系统 GFS[8]、HDFS[9]和 Haystack[10]等;NoSQL数据库 Mongodb、CouchDB、HBase、Redis、Neo4j 等.

(2)**数据分析**(Data Analysis).

数据分析是大数据应用的核心流程. 根据不同层次大致可分为 3 类:计算架构、查询与索引以及数据分析和处理.

在计算架构方面,MapReduce[11]是当前广泛采用的大数据集计算模型和框架. 为了适应一些对任务完成时间要求较高的分析需求,文献[12]对其性

① http://server.yesky.com/datacenter/172/34705172.shtml
② http://adage.com/article/digital/fake-reviews-rise-yelp-crack-fraudsters/237486/
③ http://readwrite.com/2010/05/24/map-reduce-yahoo-mail#awesm=~obIr1Wwi9X9dMN
④ http://cloud.yesky.com/20/34984520.shtml

能进行了优化；文献[13]提出了一种基于 MapReduce 架构的数据流分析解决方案 MARISSA，使其能够支持实时分析任务；文献[14]则提出了基于时间的大数据分析方案 Mastiff；文献[15]也针对广告推送等实时性要求较高的应用，提出了基于 MapReduce 的 TiMR 框架来进行实时流处理.

在查询与索引方面，由于大数据中包含了大量的非结构化或半结构化数据，传统关系型数据库的查询和索引技术受到限制，而 NoSQL 类数据库技术得到更多关注. 例如，文献[16]提出了一个混合的数据访问架构 HyDB 以及一种并发数据查询及优化方法. 文献[17]对 key-value 类型数据库的查询进行了性能优化.

在数据分析与处理方面，主要涉及的技术包括语义分析与数据挖掘等. 由于大数据环境下数据呈现多样化特点，所以对数据进行语义分析时，就较难统一术语进而挖掘信息. 文献[18]针对大数据环境，提出了一种解决术语变异问题的高效术语标准化方法. 文献[19]对语义分析中语义本体的异质性展开了研究. 传统数据挖掘技术主要针对结构化数据，因此迫切需要对非结构化或半结构化的数据挖掘技术展开研究. 文献[20]提出了一种针对图片文件的挖掘技术，文献[21]提出了一种大规模 TEXT 文件的检索和挖掘技术.

(3) 数据解释(Data Interpretation).

数据解释旨在更好地支持用户对数据分析结果的使用，涉及的主要技术为可视化和人机交互.

目前已经有了一些针对大规模数据的可视化研究[22-23]，通过数据投影、维度降解或显示墙等方法来解决大规模数据的显示问题. 由于人类的视觉敏感度限制了更大屏幕显示的有效性，以人为中心的人机交互设计也将是解决大数据分析结果展示的一种重要技术.

(4) 其它支撑技术(Data Transmission & Virtual Cluster).

虽然大数据应用强调以数据为中心，将计算推送到数据上执行，但是在整个处理过程中，数据的传输仍然是必不可少的，例如一些科学观测数据从观测点向数据中心的传输等. 文献[24-25]针对大数据特征研究高效传输架构和协议.

此外，由于虚拟集群具有成本低、搭建灵活、便于管理等优点，人们在大数据分析时可以选择更加方便的虚拟集群来完成各项处理任务. 因此需要针对大数据应用展开的虚拟机集群优化研究[26].

3　大数据带来的安全挑战

科学技术是一把双刃剑. 大数据所引发的安全问题与其带来的价值同样引人注目. 而最近爆发的“棱镜门”事件更加剧了人们对大数据安全的担忧. 与传统的信息安全问题相比，大数据安全面临的挑战性问题主要体现在以下几个方面.

3.1　大数据中的用户隐私保护

大量事实表明，大数据未被妥善处理会对用户的隐私造成极大的侵害. 根据需要保护的内容不同，隐私保护又可以进一步细分为位置隐私保护、标识符匿名保护、连接关系匿名保护等.

人们面临的威胁并不仅限于个人隐私泄漏，还在于基于大数据对人们状态和行为的预测. 一个典型的例子是某零售商通过历史记录分析，比家长更早知道其女儿已经怀孕的事实，并向其邮寄相关广告信息①. 而社交网络分析研究也表明，可以通过其中的群组特性发现用户的属性. 例如通过分析用户的 Twitter 信息，可以发现用户的政治倾向、消费习惯以及喜好的球队等[27-28].

当前企业常常认为经过匿名处理后，信息不包含用户的标识符，就可以公开发布了. 但事实上，仅通过匿名保护并不能很好地达到隐私保护目标. 例如，AOL 公司曾公布了匿名处理后的 3 个月内部分搜索历史，供人们分析使用. 虽然个人相关的标识信息被精心处理过，但其中的某些记录项还是可以被准确地定位到具体的个人. 纽约时报随即公布了其识别出的 1 位用户. 编号为 4 417 749 的用户是 1 位 62 岁的寡居妇人，家里养了 3 条狗，患有某种疾病，等等. 另一个相似的例子是，著名的 DVD 租赁商 Netflix 曾公布了约 50 万用户的租赁信息，悬赏 100 万美元征集算法，以期提高电影推荐系统的准确度. 但是当上述信息与其它数据源结合时，部分用户还是被识别出来了. 研究者发现，Netflix 中的用户有很大概率对非 top100、top500、top1000 的影片进行过评分，而根据对非 top 影片的评分结果进行去匿名化(de-anonymizing)攻击的效果更好[29].

目前用户数据的收集、存储、管理与使用等均缺乏规范，更缺乏监管，主要依靠企业的自律. 用户无法确定自己隐私信息的用途. 而在商业化场景中，用

① http://news.xinhuanet.com/info/2013-04/11/c_132300013.htm

户应有权决定自己的信息如何被利用，实现用户可控的隐私保护. 例如用户可以决定自己的信息何时以何种形式披露，何时被销毁. 包括：(1) 数据采集时的隐私保护，如数据精度处理；(2) 数据共享、发布时的隐私保护，如数据的匿名处理、人工加扰等；(3) 数据分析时的隐私保护；(4) 数据生命周期的隐私保护；(5) 隐私数据可信销毁等.

3.2 大数据的可信性

关于大数据的一个普遍的观点是，数据自己可以说明一切，数据自身就是事实①. 但实际情况是，如果不仔细甄别，数据也会欺骗，就像人们有时会被自己的双眼欺骗一样.

大数据可信性的威胁之一是伪造或刻意制造的数据，而错误的数据往往会导致错误的结论. 若数据应用场景明确，就可能有人刻意制造数据、营造某种“假象”，诱导分析者得出对其有利的结论. 由于虚假信息往往隐藏于大量信息中，使得人们无法鉴别真伪，从而做出错误判断. 例如，一些点评网站上的虚假评论，混杂在真实评论中使得用户无法分辨，可能误导用户去选择某些劣质商品或服务. 由于当前网络社区中虚假信息的产生和传播变得越来越容易，其所产生的影响不可低估. 用信息安全技术手段鉴别所有来源的真实性是不可能的.

大数据可信性的威胁之二是数据在传播中的逐步失真. 原因之一是人工干预的数据采集过程可能引入误差，由于失误导致数据失真与偏差，最终影响数据分析结果的准确性. 此外，数据失真还有数据的版本变更的因素. 在传播过程中，现实情况发生了变化，早期采集的数据已经不能反映真实情况. 例如，餐馆电话号码已经变更，但早期的信息已经被其它搜索引擎或应用收录，所以用户可能看到矛盾的信息而影响其判断.

因此，大数据的使用者应该有能力基于数据来源的真实性、数据传播途径、数据加工处理过程等，了解各项数据可信度，防止分析得出无意义或者错误的结果.

密码学中的数字签名、消息鉴别码等技术可以用于验证数据的完整性，但应用于大数据的真实性时面临很大困难，主要根源在于数据粒度的差异. 例如，数据的发源方可以对整个信息签名，但是当信息分解成若干组成部分时，该签名无法验证每个部分的完整性. 而数据的发源方无法事先预知哪些部分被利用、如何被利用，难以事先为其生成验证对象.

3.3 如何实现大数据访问控制

访问控制是实现数据受控共享的有效手段. 由于大数据可能被用于多种不同场景，其访问控制需求十分突出.

大数据访问控制的特点与难点在于：

(1) 难以预设角色，实现角色划分. 由于大数据应用范围广泛，它通常要为来自不同组织或部门、不同身份与目的的用户所访问，实施访问控制是基本需求. 然而，在大数据的场景下，有大量的用户需要实施权限管理，且用户具体的权限要求未知. 面对未知的大量数据和用户，预先设置角色十分困难.

(2) 难以预知每个角色的实际权限. 由于大数据场景中包含海量数据，安全管理员可能缺乏足够的专业知识，无法准确地为用户指定其所可以访问的数据范围. 而且从效率角度讲，定义用户所有授权规则也不是理想的方式. 以医疗领域应用为例，医生为了完成其工作可能需要访问大量信息，但对于数据能否访问应该由医生来决定，不应该需要管理员对每个医生做特别的配置. 但同时又应该能够提供对医生访问行为的检测与控制，限制医生对病患数据的过度访问.

此外，不同类型的大数据中可能存在多样化的访问控制需求. 例如，在 Web2.0 个人用户数据中，存在基于历史记录的访问控制；在地理地图数据中，存在基于尺度以及数据精度的访问控制需求；在流数据处理中，存在数据时间区间的访问控制需求，等等. 如何统一地描述与表达访问控制需求也是一个挑战性问题.

4 大数据安全与隐私保护关键技术

当前亟需针对前述大数据面临的用户隐私保护、数据内容可信验证、访问控制等安全挑战，展开大数据安全关键技术研究. 本节选取部分重点相关研究领域予以介绍.

4.1 数据发布匿名保护技术

对于大数据中的结构化数据(或称关系数据)而言，数据发布匿名保护是实现其隐私保护的核心关键技术与基本手段，目前仍处于不断发展与完善阶段.

以典型的 k 匿名方案为例. 早期的方案[30-31]及其优化方案[32-34]通过元组泛化、抑制等数据处理，将

① http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

准标识符分组. 每个分组中的准标识符相同且至少包含 k 个元组，因而每个元组至少与 $k-1$ 个其它元组不可区分. 由于 k 匿名模型是针对所有属性集合而言，对于具体的某个属性则未加定义，容易出现某个属性匿名处理不足的情况. 若某等价类中某个敏感属性上取值一致，则攻击者可以有效地确定该属性值. 针对该问题研究者提出 l 多样化(l-diversity)[35] 匿名. 其特点是在每一个匿名属性组里敏感数据的多样性满足要大于或等于 l. 实现方法包括基于裁剪算法的方案[36] 以及基于数据置换的方案[37] 等. 此外，还有一些介于 k 匿名与 l 多样化之间的方案. 进一步的，由于 l-diversity 只是能够尽量使敏感数据出现的频率平均化. 当同一等价类中数据范围很小时，攻击者可猜测其值. t 贴近性(t-closeness)方案[38] 要求等价类中敏感数据的分布与整个数据表中数据的分布保持一致. 其它工作包括 (k,e) 匿名模型[39]、(X,Y)匿名模型[40] 等. 上述研究是针对静态、一次性发布情况. 而现实中，数据发布常面临数据连续、多次发布的场景. 需要防止攻击者对多次发布的数据联合进行分析，破坏数据原有的匿名特性[41-42].

在大数据场景中，数据发布匿名保护问题较之更为复杂：攻击者可以从多种渠道获得数据，而不仅仅是同一发布源. 例如，在前所提及的 Netflix 应用中，人们[43] 发现攻击者可通过将数据与公开可获得的 imdb 相对比，从而识别出目标在 Netflix 的账号. 并据此获取用户的政治倾向与宗教信仰等(通过用户的观看历史和对某些电影的评论和打分分析获得). 此类问题有待更深入的研究.

4.2　社交网络匿名保护技术

社交网络产生的数据是大数据的重要来源之一，同时这些数据中包含大量用户隐私数据. 截至 2012 年 10 月 Facebook 的用户成员就已达 10 亿. 由于社交网络具有图结构特征，其匿名保护技术与结构化数据有很大不同.

社交网络中的典型匿名保护需求为用户标识匿名与属性匿名(又称点匿名)，在数据发布时隐藏了用户的标识与属性信息；以及用户间关系匿名(又称边匿名)，在数据发布时隐藏用户间的关系. 而攻击者试图利用节点的各种属性(度数、标签、某些具体连接信息等)，重新识别出图中节点的身份信息.

目前的边匿名方案大多是基于边的增删. 随机增删交换边的方法可以有效地实现边匿名. 其中文献[44]在匿名过程中保持邻接矩阵的特征值和对应的拉普拉斯矩阵第二特征值不变，文献[45]根据节点的度数分组，从度数相同的节点中选择符合要求的进行边的交换，类似的还有文献[46-47]. 这类方法的问题是随机增加的噪音过于分散稀少，存在匿名边保护不足问题.

另一个重要思路是基于超级节点对图结构进行分割和集聚操作. 如基于节点聚集的匿名方案[48]、基于基因算法的实现方案[49]、基于模拟退火算法的实现方案[50] 以及先填充再分割超级节点的方案[51]. 文献[52]所提出的 k-security 概念，通过 k 个同构子图实现图匿名保护. 基于超级节点的匿名方案虽然能够实现边的匿名，但是与原始社交结构图存在较大区别，以牺牲数据的可用性为代价.

社交网络匿名方案面临的重要问题是，攻击者可能通过其它公开的信息推测出匿名用户，尤其是用户之间是否存在连接关系. 例如，可以基于弱连接对用户可能存在的连接进行预测[53]，适用于用户关系较为稀疏的网络；根据现有社交结构对人群中的等级关系进行恢复和推测[54]；针对微博型的复合社交网络进行分析与关系预测[55]；基于限制随机游走方法，推测不同连接关系存在的概率[56]，等等. 研究表明[57]，社交网络的集聚特性对于关系预测方法的准确性具有重要影响，社交网络局部连接密度增长，集聚系数增大，则连接预测算法的准确性进一步增强. 因此，未来的匿名保护技术应可以有效抵抗此类推测攻击.

4.3　数据水印技术

数字水印是指将标识信息以难以察觉的方式嵌入在数据载体内部且不影响其使用的方法，多见于多媒体数据版权保护. 也有部分针对数据库和文本文件的水印方案.

由数据的无序性、动态性等特点所决定，在数据库、文档中添加水印的方法与多媒体载体上有很大不同. 其基本前提是上述数据中存在冗余信息或可容忍一定精度误差. 例如，Agrawal 等人[58-59] 基于数据库中数值型数据存在误差容忍范围，将少量水印信息嵌入到这些数据中随机选取的最不重要位上. 而 Sion 等人[60-61] 提出一种基于数据集合统计特征的方案，将一比特水印信息嵌入在一组属性数据中，防止攻击者破坏水印. 此外，通过将数据库指纹信息嵌入水印中[62]，可以识别出信息的所有者以及被分发的对象，有利于在分布式环境下追踪泄密者；通过采用独立分量分析技术(简称 ICA)，可以实现无需密钥的水印公开验证[63]. 其它相关工作包括文献[64-65]. 若在数据库表中嵌入脆弱性水印，可以帮

助及时发现数据项的变化[66].

文本水印的生成方法种类很多,可大致分为基于文档结构微调的水印[67],依赖字符间距与行间距等格式上的微小差异;基于文本内容的水印[68],依赖于修改文档内容,如增加空格、修改标点等;以及基于自然语言的水印[69],通过理解语义实现变化,如同义词替换或句式变化等.

上述水印方案中有些可用于部分数据的验证.例如在文献[58-59]中,残余元组数量达到阈值就可以成功验证出水印.该特性在大数据应用场景下具有广阔的发展前景,例如:强健水印类(Robust Watermark)可用于大数据的起源证明,而脆弱水印类(Fragile Watermark)可用于大数据的真实性证明.存在问题之一是当前的方案多基于静态数据集,针对大数据的高速产生与更新的特性考虑不足,这是未来亟待提高的方向.

4.4 数据溯源①技术

如前所述,数据集成是大数据前期处理的步骤之一.由于数据的来源多样化,所以有必要记录数据的来源及其传播、计算过程,为后期的挖掘与决策提供辅助支持.

早在大数据概念出现之前,数据溯源(Data Provenance)技术就在数据库领域得到广泛研究.其基本出发点是帮助人们确定数据仓库中各项数据的来源,例如了解它们是由哪些表中的哪些数据项运算而成,据此可以方便地验算结果的正确性,或者以极小的代价进行数据更新.数据溯源的基本方法是标记法,如在[70-72]中通过对数据进行标记来记录数据在数据仓库中的查询与传播历史.后来概念进一步细化为 why-和 where-两类[73],分别侧重数据的计算方法以及数据的出处.除数据库以外,它还包括 XML 数据、流数据与不确定数据的溯源技术[77].数据溯源技术也可用于文件的溯源与恢复.例如文献[74]通过扩展 Linux 内核与文件系统,创建了一个数据起源存储系统原型系统,可以自动搜集起源数据.此外也有其在云存储场景中的应用[75].

未来数据溯源技术将在信息安全领域发挥重要作用.在 2009 年呈报美国国土安全部的"国家网络空间安全"的报告中,将其列为未来确保国家关键基础设施安全的 3 项关键技术之一[76].然而,数据溯源技术应用于大数据安全与隐私保护中还面临如下挑战:

(1)数据溯源与隐私保护之间的平衡.一方面,基于数据溯源对大数据进行安全保护首先要通过分析技术获得大数据的来源,然后才能更好地支持安全策略和安全机制的工作;另一方面,数据来源往往本身就是隐私敏感数据.用户不希望这方面的数据被分析者获得.因此,如何平衡这两者的关系是值得研究的问题之一.

(2)数据溯源技术自身的安全性保护.当前数据溯源技术并没有充分考虑安全问题,例如标记自身是否正确、标记信息与数据内容之间是否安全绑定等等.而在大数据环境下,其大规模、高速性、多样性等特点使该问题更加突出.

4.5 角色挖掘

基于角色的访问控制(RBAC)是当前广泛使用的一种访问控制模型.通过为用户指派角色、将角色关联至权限集合,实现用户授权、简化权限管理.早期的 RBAC 权限管理多采用"自顶向下"的模式:即根据企业的职位设立角色分工.当其应用于大数据场景时,面临需大量人工参与角色划分、授权的问题(又称为角色工程).

后来研究者们开始关注"自底向上"模式,即根据现有"用户-对象"授权情况,设计算法自动实现角色的提取与优化,称为角色挖掘[78-82].简单来说,就是如何设置合理的角色.典型的工作包括:以可视化的形式,通过用户权限二维图的排序归并的方式实现角色提取[83];通过子集枚举以及聚类的方法提取角色[84]等非形式化方法;也有基于形式化语义分析、通过层次化挖掘来更准确提取角色的方法[85].总体来说,挖掘生成最小角色集合的最优算法时间复杂度高,多属于 NP-完全问题.因而也有研究者关注在多项式时间内完成的启发式算法[86].在大数据场景下,采用角色挖掘技术可根据用户的访问记录自动生成角色,高效地为海量用户提供个性化数据服务.同时也可用于及时发现用户偏离日常行为所隐藏的潜在危险.但当前角色挖掘技术大都基于精确、封闭的数据集,在应用于大数据场景时还需要解决数据集动态变更以及质量不高等特殊问题.

4.6 风险自适应的访问控制

在大数据场景中,安全管理员可能缺乏足够的专业知识,无法准确地为用户指定其可以访问的数据.风险自适应的访问控制是针对这种场景讨论较多的一种访问控制方法.Jason 的报告[87]描述了风险量化和访问配额的概念.随后,Cheng 等人[88]提

① 也被译成"数据世系",英文有的称作 Data Lineage 或 Data Pedigree,含义略有区别

出了一个基于多级别安全模型的风险自适应访问控制解决方案. Ni 等人[89]提出了另一个基于模糊推理的解决方案，将信息的数目和用户以及信息的安全等级作为进行风险量化的主要参考参数. 当用户访问的资源的风险数值高于某个预定的门限时，则限制用户继续访问. 文献[90]提出了一种针对医疗数据提供用户隐私保护的可量化风险自适应访问控制. 通过利用统计学和信息论的方法，定义了量化算法，从而实现基于风险的访问控制. 但同时，在大数据应用环境中，风险的定义和量化都较之以往更加困难.

5　大数据服务与信息安全

5.1　基于大数据的威胁发现技术

由于大数据分析技术的出现，企业可以超越以往的“保护-检测-响应-恢复”(PDRR)模式，更主动地发现潜在的安全威胁. 例如，IBM 推出了名为 IBM 大数据安全智能的新型安全工具①，可以利用大数据来侦测来自企业内外部的安全威胁，包括扫描电子邮件和社交网络，标示出明显心存不满的员工，提醒企业注意，预防其泄露企业机密.

“棱镜”计划也可以被理解为应用大数据方法进行安全分析的成功故事. 通过收集各个国家各种类型的数据，利用安全威胁数据和安全分析形成系统方法发现潜在危险局势，在攻击发生之前识别威胁.

相比于传统技术方案，基于大数据的威胁发现技术具有以下优点.

(1) 分析内容的范围更大.

传统的威胁分析主要针对的内容为各类安全事件. 而一个企业的信息资产则包括数据资产、软件资产、实物资产、人员资产、服务资产和其它为业务提供支持的无形资产. 由于传统威胁检测技术的局限性，其并不能覆盖这六类信息资产，因此所能发现的威胁也是有限的. 而通过在威胁检测方面引入大数据分析技术，可以更全面地发现针对这些信息资产的攻击. 例如通过分析企业员工的即时通信数据、Email 数据等可以及时发现人员资产是否面临其它企业“挖墙脚”的攻击威胁. 再比如通过对企业的客户部订单数据的分析，也能够发现一些异常的操作行为，进而判断是否危害公司利益. 可以看出，分析内容范围的扩大使得基于大数据的威胁检测更加全面.

(2) 分析内容的时间跨度更长.

现有的许多威胁分析技术都是内存关联性的，也就是说实时收集数据，采用分析技术发现攻击. 分析窗口通常受限于内存大小，无法应对持续性和潜伏性攻击. 而引入大数据分析技术后，威胁分析窗口可以横跨若干年的数据，因此威胁发现能力更强，可以有效应对 APT 类攻击.

(3) 攻击威胁的预测性.

传统的安全防护技术或工具大多是在攻击发生后对攻击行为进行分析和归类，并做出响应. 而基于大数据的威胁分析，可进行超前的预判. 它能够寻找潜在的安全威胁，对未发生的攻击行为进行预防.

(4) 对未知威胁的检测.

传统的威胁分析通常是由经验丰富的专业人员根据企业需求和实际情况展开，然而这种威胁分析的结果很大程度上依赖于个人经验. 同时，分析所发现的威胁也是已知的. 而大数据分析的特点是侧重于普通的关联分析，而不侧重因果分析，因此通过采用恰当的分析模型，可发现未知威胁.

虽然基于大数据的威胁发现技术具有上述的优点，但是该技术目前也存在一些问题和挑战，主要集中在分析结果的准确程度上. 一方面，大数据的收集很难做到全面，而数据又是分析的基础，它的片面性往往会导致分析出的结果的偏差. 为了分析企业信息资产面临的威胁，不但要全面收集企业内部的数据，还要对一些企业外的数据进行收集，这些在某种程度上是一个大问题. 另一方面，大数据分析能力的不足影响威胁分析的准确性. 例如，纽约投资银行每秒会有 5000 次网络事件，每天会从中捕捉 25 TB 数据. 如果没有足够的分析能力，要从如此庞大的数据中准确地发现极少数预示潜在攻击的事件，进而分析出威胁是几乎不可能完成的任务.

5.2　基于大数据的认证技术

身份认证是信息系统或网络中确认操作者身份的过程. 传统的认证技术主要通过用户所知的秘密，例如口令，或者持有的凭证，例如数字证书，来鉴别用户. 这些技术面临着如下两个问题.

首先，攻击者总是能够找到方法来骗取用户所知的秘密，或窃取用户持有的凭证，从而通过认证机制的认证. 例如攻击者利用钓鱼网站窃取用户口令，或者通过社会工程学方式接近用户，直接骗取用户所知秘密或持有的凭证.

其次，传统认证技术中认证方式越安全往往意味着用户负担越重. 例如，为了加强认证安全，而采

① http://www.36kr.com/p/201176.html

用的多因素认证.用户往往需要同时记忆复杂的口令,还要随身携带硬件 USBKey.一旦忘记口令或者忘记携带 USBKey,就无法完成身份认证.为了减轻用户负担,一些生物认证方式出现,利用用户具有的生物特征,例如指纹等,来确认其身份.然而,这些认证技术要求设备必须具有生物特征识别功能,例如指纹识别.因此很大程度上限制了这些认证技术的广泛应用.

而在认证技术中引入大数据分析则能够有效地解决这两个问题.基于大数据的认证技术指的是收集用户行为和设备行为数据,并对这些数据进行分析,获得用户行为和设备行为的特征,进而通过鉴别操作者行为及其设备行为来确定其身份.这与传统认证技术利用用户所知秘密,所持有凭证,或具有的生物特征来确认其身份有很大不同.具体地,这种新的认证技术具有如下优点.

(1)攻击者很难模拟用户行为特征来通过认证,因此更加安全.利用大数据技术所能收集的用户行为和设备行为数据是多样的,可以包括用户使用系统的时间、经常采用的设备、设备所处物理位置,甚至是用户的操作习惯数据.通过这些数据的分析能够为用户勾画一个行为特征的轮廓.而攻击者很难在方方面面都模仿到用户行为,因此其与真正用户的行为特征轮廓必然存在一个较大偏差,无法通过认证.

(2)减小了用户负担.用户行为和设备行为特征数据的采集、存储和分析都由认证系统完成.相比于传统认证技术,极大地减轻了用户负担.

(3)可以更好地支持各系统认证机制的统一.基于大数据的认证技术可以让用户在整个网络空间采用相同的行为特征进行身份认证,而避免不同系统采用不同认证方式,且用户所知秘密或所持有凭证也各不相同而带来了种种不便.

虽然基于大数据的认证技术具有上述优点,但同时也存在一些问题和挑战亟待解决.

(1)初始阶段的认证问题.基于大数据的认证技术是建立在大量用户行为和设备行为数据分析的基础上,而初始阶段不具备大量数据.因此,无法分析出用户行为特征,或者分析的结果不够准确.

(2)用户隐私问题.基于大数据的认证技术为了能够获得用户的行为习惯,必然要长期持续地收集大量的用户数据.那么如何在收集和分析这些数据的同时,确保用户隐私也是亟待解决的问题.它是影响这种新的认证技术是否能够推广的主要因素.

5.3 基于大数据的数据真实性分析

目前,基于大数据的数据真实性分析被广泛认为是最为有效的方法.许多企业已经开始了这方面的研究工作,例如 Yahoo 和 Thinkmail 等利用大数据分析技术来过滤垃圾邮件;Yelp 等社交点评网络用大数据分析来识别虚假评论;新浪微博等社交媒体利用大数据分析来鉴别各类垃圾信息等.

基于大数据的数据真实性分析技术能够提高垃圾信息的鉴别能力.一方面,引入大数据分析可以获得更高的识别准确率.例如,对于点评网站的虚假评论,可以通过收集评论者的大量位置信息、评论内容、评论时间等进行分析,鉴别其评论的可靠性.如果某评论者为某品牌多个同类产品都发表了恶意评论,则其评论的真实性就值得怀疑;另一方面,在进行大数据分析时,通过机器学习技术,可以发现更多具有新特征的垃圾信息.然而该技术仍然面临一些困难,主要是虚假信息的定义、分析模型的构建等.

5.4 大数据与“安全-即-服务(Security-as-a-Service)”

前面列举了部分当前基于大数据的信息安全技术,未来必将涌现出更多、更丰富的安全应用和安全服务.由于此类技术以大数据分析为基础,因此如何收集、存储和管理大数据就是相关企业或组织所面临的核心问题.除了极少数企业有能力做到之外,对于绝大多数信息安全企业来说,更为现实的方式是通过某种方式获得大数据服务,结合自己的技术特色领域,对外提供安全服务.一种未来的发展前景是,以底层大数据服务为基础,各个企业之间组成相互依赖、相互支撑的信息安全服务体系,总体上形成信息安全产业界的良好生态环境.

6 小 结

大数据带来了新的安全问题,但它自身也是解决问题的重要手段.本文从大数据的隐私保护、信任、访问控制等角度出发,梳理了当前大数据安全与隐私保护相关关键技术.但总体上来说,当前国内外针对大数据安全与隐私保护的相关研究还不充分.只有通过技术手段与相关政策法规等相结合,才能更好地解决大数据安全与隐私保护问题.

参 考 文 献

[1] Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution that Will Transform How We Live, Work and Think. Boston: Houghton Mifflin Harcourt, 2013

[2] Meng Xiao-Feng, Ci Xiang. Big data management: Concepts, techniques and challenges. Journal of Computer Research and Development, 2013, 50(1): 146-169(in Chinese)
(孟小峰，慈祥. 大数据管理：概念、技术与挑战. 计算机研究与发展，2013，50(1)：146-169)

[3] Li Guo-Jie, Cheng Xue-Qi. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chinese)
(李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域. 中国科学院院刊，2012，27(6)：647-657)

[4] Li Xian, Dong Xin Luna, Lyons Kenneth, et al. Truth finding on the deep web: Is the problem solved?//Proceedings of the 39th International Conference on Very Large Data Bases (VLDB'2013). Trento, Italy, 2013: 97-108

[5] Arasu A, Chaudhuri S, Chen Z, et al. Experiences with using data cleaning technology for bing services. IEEE Data Engineering Bulletin, 2012, 35(2): 14-23

[6] Liu Xuan, Dong Xin Luna, Ooi Beng Chin, Srivastava Divesh. Online data fusion//Proceedings of the 37th International Conference on Very Large Data Bases (VLDB'2011). Seattle, USA, 2011: 932-943

[7] Sarma Anish Das, Dong Xin Luna, Halevy Alon. Data integration with dependent sources //Proceedings of the 14th International Conference on Extending Database Technology. Uppsala, Sweden, 2011: 401-412

[8] Ghemawat Sanjay, Gobioff Howard, Leung Shun-Tak. The Google file system//Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York, USA, 2003: 29-43

[9] HDFS Architecture Guide. http://hadoop. apache. org/docs/stable/hdfs_design. html. 2013-05-12

[10] Beaver D, Kumar S, et al. Finding a needle in haystack: Facebook's photo storage//Proceedings of the 9th USENIX Symposium on Operating Systems Design (OSDI'2010). Vancouver, Canada, 2010: 1-8

[11] Dean Jeffrey, Ghemawat Sanjay. MapReduce: Simplified data processing on large clusters//Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation. San Francisco, USA, 2004: 107-113

[12] Verma A, Cherkasova L, Kumar V, Campbell R. Deadline-based workload management for mapreduce environments: Pieces of the performance puzzle//Proceedings of the Network Operations and Management Symposium (NOMS'2012). Hawaii, USA, 2012: 900-905

[13] Dede E, Fadika Z, Hartog J, et al. MARISSA: MApReduce implementation for streaming science applications//Proceedings of the IEEE 8th International Conference on E-Science. Chicago, USA, 2012: 1-8

[14] Guo Si-Jie, Xiong Jin, Wang Wei-Ping, Lee Ru-Bao. Mastiff: A MapReduce-based system for time-based big data analytics //Proceedings of the 2012 IEEE International Conference on Cluster Computing (CLUSTER). Beijing, China, 2012: 72-80

[15] Chandramouli B, Goldstein J, Duan S. Temporal analytics on big data for Web advertising//Proceedings of the 28th IEEE International Conference on Data Engineering (ICDE). Washington DC, USA, 2012: 90-101

[16] Zhu Qing, Qin Zuo-Yan. HyDB: Access optimization for data-intensive service//Proceedings of the 14th International Conference on High Performance Computing and Communications (HPCC). Liverpool, UK, 2012: 580-587

[17] Wang Yao-Guang, Lu Wei-Ming, Wei Bao-Gang. Transactional multi-row access guarantee in the key-value store//Proceedings of the 2012 IEEE International Conference on Cluster Computing (CLUSTER). Beijing, China, 2012: 572-575

[18] Hwang M, Jeong D H, Kim J, et al. A term normalization method for better performance of terminology construction//Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing. Zakopane, Poland, 2012: 682-690

[19] Ketata I, Mokadem R, Morvan F. Biomedical resource discovery considering semantic heterogeneity in data grid environments//Proceedings of the International Conference on Innovative Computing Technology. Sao Carlos, Brazil, 2011: 12-24

[20] Kang U, Chau D H, Faloutsos C. Pegasus: Mining billion-scale graphs in the cloud//Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, Japan, 2012: 5341-5344

[21] Gubanov M, Pyayt A. MEDREADFAST: A structural information retrieval engine for big clinical text//Proceedings of the 13th International Conference on Information Reuse and Integration (IRI). Las Vegas, USA, 2012: 371-376

[22] Ahrens J, Brislawn K, Martin K, et al. Large-scale data visualization using parallel data streaming. IEEE Computer Graphics and Applications, 2001, 21(4): 34-41

[23] Scheidegger Luiz, Vo Huy T, Krüger Jens, et al. Parallel large data visualization with display walls//Proceedings of the 2012 Conference on Visualization and Data Analysis (VDA). Burlingame, USA, 2012: 1-8

[24] Narayanan S, Madden T J, Sandy A R, et al. GridFTP based real-time data movement architecture for x-ray photon correlation spectroscopy at the Advanced Photon Source//Proceedings of the IEEE 8th International Conference on E-Science. Chicago, USA, 2012: 1-8

[25] Tierney B, Kissel E, Swany M, Pouyoul E. Efficient data transfer protocols for big data//Proceedings of the IEEE 8th International Conference on E-Science. Chicago, USA, 2012: 1-9

[26] Yan Cai-Rong, Zhu Ming, Yang Xin, et al. Affinity-aware virtual cluster optimization for MapReduce applications//Proceedings of the 2012 IEEE International Conference on Cluster Computing (CLUSTER). Beijing, China, 2012: 63-71

[27] Ye Mao, Yin Pei-Feng, Lee Wang-Chien, Lee Dik-Lun. Exploiting geographical influence for collaborative point-of-interest recommendation//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'11). Beijing, China, 2011: 325-334

[28] Goel S, Hofman J M, Lahaie S, et al. Predicting consumer behavior with Web search. National Academy of Sciences, 2010, 7(41): 17486-17490

[29] Narayanan A, Shmatikov V. How to break anonymity of the netflix prize dataset. ArXiv Computer Science e-prints, 2006, arXiv:cs/0610105: 1-10

[30] Sweeney L. *k*-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10 (5): 557-570

[31] Sweeney L. *k*-Anonymity: Achieving *k*-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588

[32] Bayardo R J, Agrawal R. Data privacy through optimal *k*-anonymization//Proceedings of the 21st International Conference on Data Engineering. Tokyo, Japan, 2005: 217-228

[33] Kristen LeFevre, David J DeWitt, Raghu Ramakrishnan. Incognito: Efficient full-domain *K*-anonymity//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, USA, 2005: 49-60

[34] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional *K*-anonymity//Proceedings of the 22nd International Conference on Data Engineering. Georgia, USA, 2006: 25-35

[35] Machanavajjhala Ashwin, Kifer Daniel, Gehrke Johannes, Venkitasubramaniam Muthuramakrishnan. *L*-diversity: Privacy beyond *k*-anonymity. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1-52

[36] Xiao Xiao-Kui, Tao Yu-Fei. Anatomy: Simple and effective privacy preservation//Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06). Seoul, Korea, 2006: 139-150

[37] Qing Zhang, Koudas N, Srivastava D, Ting Yu. Aggregate query answering on anonymized tables//Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE'2007). Istanbul, Turkey, 2007: 116-125

[38] Li Ning-Hui, Li Tian-Cheng, Venkatasubramanian S. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity//Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE'2007). Istanbul, Turkey, 2007: 106-115

[39] Zeng K. Publicly verifiable remote data integrity//Proceedings of the 10th International Conference on Information and Communications Security (ICICS'2008). Birmingham, UK, 2008: 419-434

[40] Wang Ke, Fung B C M. Anonymizing sequential releases//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06). Philadelphia, USA, 2006: 414-423

[41] Xiao Xiao-Kui, Tao Yu-Fei. *M*-invariance: Towards privacy preserving re-publication of dynamic datasets//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD'07). Beijing, China, 2007: 689-700

[42] Bu Ying-Yi, Fu Ada Wai Chee, Wong Raymond Chi Wing, et al. Privacy preserving serial data publishing by role composition//Proceedings of the 34th International Conference on Very Large Data Bases (VLDB'2008). Auckland, New Zealand, 2008: 845-856

[43] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets//Proceedings of the 2008 IEEE Symposium on Security and Privacy (S&P'2008). Oakland, USA, 2008: 111-125

[44] Ying X, Wu X. Randomizing social networks: A spectrum preserving approach//Proceedings of the SIAM International Conference on Data Mining (SDM'08). Georgia, USA, 2008: 739-750

[45] Zhang Li-Jie, Zhang Wei-Ning. Edge anonymity in social network graphs//Proceedings of the International Conference on Computational Science and Engineering (CSE'09). Vancouver, Canada, 2009: 1-8

[46] Li Na, Das Sajal K. Applications of *k*-anonymity and ℓ-diversity in publishing online social networks//Proceedings of the IEEE Social Computing Conference. Cambridge, USA, 2012: 153-180

[47] Zou Lei, Chen Lei, Özsu M T. *k*-automorphism: A general framework for privacy preserving network publication//Proceedings of the 35th International Conference on Very Large Data Bases (VLDB'2009). Lyon, France, 2009: 946-957

[48] Campan Alina, Truta Traian Marius. Data and structural *k*-anonymity in social networks//Proceedings of the 2nd ACM SIGKDD International Workshop (PinKDD2008). Las Vegas, USA, 2008: 1-10

[49] Sihag Vikas Kumar. A clustering approach for structural *k*-anonymity in social networks using genetic algorithm//Proceedings of the CUBE International Information Technology Conference(CUBE'12). Pune, India, 2012: 701-706

[50] Hay Michael, Miklau Gerome, Jensen David, et al. Resisting structural re-identification in anonymized social networks//Proceedings of the 34th International Conference on Very Large Data Bases (VLDB'2008). Auckland, New Zealand, 2008: 102-114

[51] Zhang Li-Jie, Zhang Wei-Ning. Efficient edge anonymization of large social graphs. http://venom. cs. utsa. edu/dmz/techrep/2011/CS-TR-2011-004. pdf. 2013-06-10

[52] Cheng James, Fu Ada Wai-Chee, Liu Jia. *K*-isomorphism: Privacy preserving network publication against structural

attacks//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis, USA, 2010: 459-470

[53] Lü L, Zhou T. Link prediction in weighted networks: The role of weak ties. Europhysics Letters, 2010, 89(1): 18001-18006

[54] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks. Nature, 2008, 453(7191): 98-101

[55] Yin Da-Wei, Hong Liang-Jie, Xiong Xiong, Davison B D. Link formation analysis in microblogs//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11). Beijing, China, 2011: 1235-1236

[56] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10). Washington DC, USA, 2010: 243-252

[57] Feng X, Zhao J C, Xu K. Link prediction in complex networks: A clustering perspective. The European Physical Journal B, 2012, 85(3): 1-9

[58] Agrawal R, Haas P J, Kiernan J. Watermarking relational data: Framework, algorithms and analysis. The International Journal on Very Large Data Bases, 2003, 12(2): 157-169

[59] Agrawal R, Kiernan J. Watermarking relational databases//Proceedings of the 28th International Conference on Very Large Data Bases (VLDB'02). Hong Kong, China, 2002: 155-166

[60] Sion R, Atallab M, Prabhakar S. On watermarking numeric sets//Proceedings of the 1st International Workshop on Digital Watermarking (IWDW'2002). Seoul, Korea, 2002: 130-146

[61] Sion R, Atallah M, Prabhakar S. Rights protection for relational data//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD'2003). San Diego, USA, 2003: 98-109

[62] Guo Fei, Wang Jian-Min, Li De-Yi. Fingerprinting relational databases//Proceedings of the 2006 ACM Symposium on Applied Computing (SAC'06). Dijon, France, 2006: 487-492

[63] Jiang Chuan-Xian, Sun Xing-Ming, Yi Ye-Qing, Yang Heng-Fu. Study of database public watermarking based on JADE algorithm. Journal of System Simulation, 2006, 18(7): 1781-1785(in Chinese)
(姜传贤，孙星明，易叶青，杨恒伏. 基于JADE算法的数据库公开水印算法的研究. 系统仿真学报, 2006, 18(7): 1781-1785)

[64] Zhao Y, Niu X M, Zhao D N. A method of protecting relational databases copyright with cloud watermark. International Journal of Information Technology, 2004, 1(1): 206-210

[65] Liu Yu-Chao, Ma Yu-Tao, Zhang Hai-Su, et al. A method for trust management in cloud computing: Data coloring by cloud watermarking. International Journal of Automation and Computing, 2011, 8(3): 280-285

[66] Guo H, Li Y, Liu A, Jajodia S. A fragile watermarking scheme for detecting malicious modifications of database relations. Information Sciences, 2006, 176(10): 1350-1378

[67] Low S H, Maxemchuk N F, Brassil J T, O'Gorman L. Document marking and identification using both line and word shifting//Proceedings of the Conference on Computer Communications, 14th Annual Joint Conference of the IEEE Computer and Communications Societies, Bringing Information to People (INFOCOM'95). Boston, USA, 1995: 853-860

[68] Pease A, Niles I, Li J. The suggested upper merged ontology: A large ontology for the semantic web and its applications//Proceedings of the AAAI-2002 Workshop on Ontologies and the Semantic Web. Edmonton, Canada, 2002: 1-4

[69] Atallah M J, Raskin V, Hempelmann C F, et al. Natural language watermarking and tamperproofing//Proceedings of the 5th International Workshop on Information Hiding (IH'2002). Noordwijkerhout, Netherlands, 2002: 196-212

[70] Cui Ying-Wei, Widom Jennifer, Wiener J L. Tracing the lineage of view data in a warehousing environment. ACM Transactions on Database Systems (TODS), 2000, 25(2): 179-227

[71] Cui Y, Widom J. Practical lineage tracing in data warehouses//Proceedings of the 16th International Conference on Data Engineering (ICDE'2000). San Diego, USA, 2000: 367-378

[72] Cui Y, Widom J. Lineage tracing for general data warehouse transformations. The International Journal on Very Large Data Bases, 2003, 12(1): 41-58

[73] Buneman P, Khanna S, Wang-Chiew T. Why and where: A characterization of data provenance//Proceedings of the 8th International Conference on Database Theory (ICDT2001). London, UK, 2001: 316-330

[74] Muniswamy-Reddy K K, Holland D A, Braun U, Seltzer M. Provenance-aware storage systems//Proceedings of the 2006 USENIX Annual Technical Conference. Boston, USA, 2006: 43-56

[75] Muniswamy-Reddy K K, Macko P, Seltzer M. Provenance for the cloud//Proceedings of the 8th USENIX Conference on File and Storage Technologies. San Jose, USA, 2010: 15-28

[76] Wybourne M N, Austin M F, Palmer C C. National cyber security research and development challenges. Institute for Information Infrastructure Protection, 2009. http://www.thei3p.org/docs/publications/i3pnational cybersecurity.pdf

[77] Gao Ming, Jin Che-Qing, Wang Xiao-Ling, et al. A survey on management of data provenance. Chinese Journal of Computers, 2010, 33(3): 374-389(in Chinese)
(高明，金澈清，王晓玲等. 数据世系管理技术研究综述. 计算机学报, 2010, 33(3): 374-389)

[78] Ene A, Horne W, Milosavljevic N, et al. Fast exact and heuristic methods for role minimization problems//Proceedings of the 13th ACM Symposium on Access Control Models and Technologies. Estes Park, USA, 2008: 1-10

[79] Frank M, Basin D, Buhmann J M. A class of probabilistic models for role engineering//Proceedings of the 15th ACM Conference on Computer and Communications Security (CCS'2008). Alexandria, USA, 2008: 299-310

[80] Lu H, Vaidya J, Atluri V. Optimal Boolean matrix decomposition: Application to role engineering//Proceedings of the IEEE 24th International Conference on Data Engineering (ICDE'2008). Cancun, Mexico, 2008: 297-306

[81] Molloy I, Chen H, Li T, et al. Mining roles with semantic meanings//Proceedings of the 13th ACM Symposium on Access Control Models and Technologies (SACMAT'2008). Estes Park, USA, 2008: 21-30

[82] Streich A P, Frank M, Basin D, Buhmann J M. Multi-assignment clustering for Boolean data//Proceedings of the 26th Annual International Conference on Machine Learning (ICML'2009). Montreal, Canada, 2009: 969-976

[83] Colantonio A, Di Pietro R, Ocello A, Verde N V. Visual role mining: A picture is worth a thousand roles. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(6): 1120-1133

[84] Vaidya J, Atluri V, Warner J, Qi Guo. Role engineering via prioritized subset enumeration. IEEE Transactions on Dependable and Secure Computing, 2010, 7(3): 300-314

[85] Molloy I, Park Y, Chari S. Generative models for access control policies: Applications to role mining over logs with attribution//Proceedings of the 17th ACM symposium on Access Control Models and Technologies (SACMAT'12). Newark, USA, 2012: 45-56

[86] Carlo Blundo, Stelvio Cimato. A simple role mining algorithm//Proceedings of the 2010 ACM Symposium on Applied Computing(SAC'10). Sierre, Switzerland, 2010: 1958-1962

[87] The MITRE Corporation. Horizontal integration: Broader access models for realizing information dominance. http://www.fas.org/irp/agency/dod/jason/classpol.pdf. 2013-06-13

[88] Cheng P C, Rohatgi P, Keser C, et al. Fuzzy multi-level security: An experiment on quantified risk-adaptive access control//Proceedings of the 2007 IEEE Symposium on Security and Privacy (S&P'2007). Oakland, USA, 2007: 222-230

[89] Ni Q, Bertino E, Lobo J. Risk-based access control systems built on fuzzy inferences//Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security (ASIACCS'2010). Beijing, China, 2010: 250-260

[90] Wang Q, Jin H. Quantified risk-adaptive access control for patient privacy protection in health information systems//Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS'2011). Hong Kong, China, 2011: 406-410

FENG Deng-Guo, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include information security and cryptology, trusted computing and information assurance.

ZHANG Min, born in 1975, Ph. D., associate professor. Her research interests include data privacy protection, trusted computing and cloud storage security.

LI Hao, born in 1983, Ph. D., research assistant. His research interests include data privacy protection and trusted computing.

Background

Currently big data is increasingly attracting the attention of both academic and industrial researchers. However, there are many unknown security risks in its collection, storage and processing. To make things worse, privacy issues related with big data analysis spell trouble for individuals. This paper summarizes and analyzes the security challenges brought by big data, and then describes the key technologies which can be exploited to deal with these challenges. This paper also argues that big data brings not only challenges, but also technical revolution in the field of information security.

This work is partly supported by the National Natural Science Foundation of China (Nos. 91118006, 61232005, 61100237), and "863" Project (No. 2011AA0123824001). These projects aim to provide security and privacy protection for massive data storage and processing. The group has been working on the trust computing, cloud storage security and database security. Many papers have been published in respectable international conferences, such as AsiaCCS, TrustCom, CloudCom etc.