DOI:10.3969/j.issn.1009-6868.2016.02.012
网络出版地址:http://www.cnki.net/kcms/detail/34.1228.TN.20160225.1441.002.html

大数据安全与隐私保护态势
Big Data Security and Privacy Protection

中图分类号:TP393 文献标志码:A 文章编号:1009-6868 (2016) 02-0053-004

摘要: 指出安全与隐私防护是大数据面临的两个重要的问题。认为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。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。认为信息安全企业未来的发展前景为: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,通过构建安全大数据,逐步形成大数据安全生态环境。

关键词: 大数据;安全;隐私;认证

Abstract: Security and privacy protection are two important issues with big data. On the one hand, big data creates new security problems and challenges. On the other hand, it creates new opportunities for the development of information security. Big-data-based information security technologies can be used for security and privacy protection. Big-data-based data authenticity analysis is widely considered to be the most effective method. Development prospects for information security are: the data underlying service is the foundation, and enterprises between each other can form the system in which they have mutual dependence, mutual support of information security service. By building up the security big data system, a good environment for information security industry is formed.

Key words: big data; security; privacy; authentication

范渊/FAN Yuan
(杭州安恒信息技术有限公司,浙江 杭州 310051)
(DBAPPSecurity, Hangzhou 310051, China)

- 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
- 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
- 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

1 大数据研究现状

目前,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2.7 ZB,而到2015年这一数值预计会达到8 ZB。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4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:2016年全球在大数据方面的总花费将达到2 320亿美元。Gartner将大数据技术列入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"。在中国,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,为中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。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。多项

收稿时间:2016-02-23
网络出版时间:2016-02-25

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。

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。与其他信息一样，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，较以往其他安全问题（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。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。

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。文章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技术可以反过来用于大数据的安全和隐私保护[1]。

2 大数据安全的挑战

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。而近年爆发的“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。

（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。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。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他数据源结合时，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 top 100、top 500、top 1 000 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攻击的效果更好。

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。包括：数据采集时的隐私保护，如数据精度处理；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；数据分析时的隐私保护；数据生命周期的隐私保护；隐私数据可信销毁等。

（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。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[2]。例如，餐馆电话号码已经变更，但早期的信息已经被其他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）大数据访问控制的实现

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：

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。

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 Web 2.0 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。如何能够统一地描述与表达访问控制需求也是一个极具挑战性的问题[3]。

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，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。

3 基于认证分析的大数据分析技术

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。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、服务资产和其他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 6 类信息资产，因此所能发现的威胁也是有限的。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他企业“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。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，可以有效应对高级持续性威胁（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。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。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 5 000 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。

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。这些技术面临着两个问题：（1）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。(2)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USB Key，一旦忘记口令或者忘记携带USB Key，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。

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。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。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)减轻了用户负担。用户行为和设备行为特征数据的采集、存储等都由认证系统完成。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，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的种种不便。

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。

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。

4 结束语

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。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。文章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，当前全球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] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 36(01): 246–258
[2] 谢邦昌, 姜叶飞. 大数据时代 隐私如何保护[J]. 中国统计, 2013(06): 24–28
[3] 应钦. 大数据安全与隐私保护技术探究[J]. 硅谷, 2014(10): 15–19. doi:10.3969/j.issn.1671–7597.2014.10.044

作者简介

范渊，毕业于美国加州州立大学，现任杭州安恒信息技术有限公司董事长兼CEO；长期从事在线应用安全、数据库安全和审计、Compliance（如SOX、PCI、ISO17799/27001）等方面的研究；并作为项目负责人已承担国家级科技计划项目8项，省、市级科技计划项目16项；先后入选国家“千人计划”特聘专家，国家科技部“科技创新创业人才”，第3届世界浙商大会创业创新奖等；申请专利45项，授权发明专利8项，并在重要学术期刊发表多篇重要论文。