计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131330
52(1):229-238, 2015

大数据时代的个人隐私保护

刘雅辉[1,2] 张铁赢[1] 靳小龙[1] 程学旗[1]
[1](中国科学院计算技术研究所 北京 100190)
[2](石河子大学 新疆石河子 832003)
(liuyahui@software.ict.ac.cn)

Personal Privacy Protection in the Era of Big Data

Liu Yahui[1,2], Zhang Tieying[1], Jin Xiaolong[1], and Cheng Xueqi[1]
[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2](*Shihezi University, Shihezi, Xinjiang* 832003)

Abstract With the development of information technology, emerging services based on Web2.0 technologies such as blog, microblog, social networks, and the Internet of things produce various types of data at an unprecedented rate, while cloud computing provides a basic storage infrastructure for big data. All of these lead to the arrival of the big data era. Big data contains great value. Data become the most valuable wealth of the enterprise, but big data also brings grand challenges. Personal privacy protection is one of the major challenges of big data. People on the Internet leave many data footprint with cumulativity and relevance. Personal privacy information can be found by gathering data footprint in together. Malicious people use this information for fraud. It brings many trouble or economic loss to personal life. Therefore, the issue of personal privacy has caused extensive concern of the industry and academia. However, there is little work on the protection of personal privacy at present. Firstly, the basic concepts of big data privacy protection are introduced, and the challenges and research on personal privacy concern are discussed. Secondly, the related technology of privacy protection is described from the data layer, application layer and data display layer. Thirdly, several important aspects of the personal privacy laws and industry standards are probed in the era of big data. Finally, the further research direction of personal privacy protection is put forward.

Key words personal privacy protection; personal privacy concern; privacy protection technology; big data privacy; big data

摘 要 随着信息技术的发展,以 Web2.0 技术为基础的博客、微博、社交网络等新兴服务和物联网以前所未有的发展速度产生了类型繁多的数据,而云计算为数据的存储提供了基础平台,这一切造就了大数据时代的正式到来.大数据中蕴藏着巨大的价值,是企业的宝贵财富.但大数据同时也带来了巨大的挑战,个人隐私保护问题就是其中之一.迅速发展的互联网已经成为人们生活中不可或缺的一部分,人们在网络上留下了许多数据足迹,这些数据足迹具有累积性和关联性,将多处数据足迹聚集在一起,就可以发现个人的隐私信息.恶意分子利用这些信息进行欺诈等行为,给个人的生活带来了许多麻烦或经济

收稿日期:2013-10-12;修回日期:2014-05-06
基金项目:国家"九七三"重点基础研究发展计划基金项目(2012CB316303,2013CB329602);国家"八六三"高技术研究发展计划基金项目(2012AA011003);国家自然科学基金重点项目(61232010,61173064);国家科技支撑计划基金项目(2012BAH39B04);国家自然科学基金项目(61202214)

损失,因此大数据的个人隐私问题引起了工业界和学术界的广泛关注.首先介绍了大数据时代个人隐私保护的相关概念,讨论了个人隐私保护面临的挑战和研究问题;然后从数据层、应用层以及数据展示层叙述了个人隐私保护所使用的技术,探讨了个人隐私保护的相关法律以及行业规范的几个重要方面;最后提出了大数据个人隐私保护的进一步研究方向.

关键词 个人隐私保护;个人隐私问题;隐私保护技术;大数据隐私;大数据

中图法分类号 TP311

计算机的出现促使各种事务数字化,如过去不方便存储、分析和共享的很多纸质材料都被数字化,计算机逐渐成为不可替代的数据处理工具.随着数据量的不断增加,数据库应运而生,数据库技术的迅速发展以及数据库管理系统的广泛应用使人们积累的数据越来越多,人们迫切需要将这些数据转换成有用的知识并揭示其潜在的价值,广泛地用于各种应用.数据挖掘就是为顺应这种需要而发展起来的数据处理技术,它通过分析企业的数据作出归纳性的推理,从中挖掘出潜在的价值,帮助决策者调整策略,减少风险,作出正确的决策.美国1991年出现了商用互联网服务,商业机构一踏入互联网就发现了它在通信、资料检索、客户服务等方面的巨大潜力.于是,其势一发不可收拾,迎来了互联网发展史上一个新的飞跃,而以Web技术为代表的信息发布系统成为互联网的主要应用.Web2.0技术的出现使得博客和社会网络迅速发展,产生了大量的文字、图像、视频等非结构化数据,随之又迎来了云计算,为用户提供了服务资源的基础平台.2008年"大数据"这一术语开始在技术圈内出现,2008年末,"大数据"得到部分美国知名计算机科学研究人员的认可,由此大数据时代拉开了序幕.大数据成为直接影响国家、社会稳定以及关系到国家安全的战略性问题,带来了许多的科学思考和科学问题,同时也面临着前所未有的挑战[1],如现有的IT架构以及机器处理和计算能力等.

大数据中的大部分数据来源于人和传感器,包括用户上网浏览的资料、社交网络上用户的信息和评论、传感器数据和监视数据等.从浩瀚的半结构或非结构化数据宝藏中获得有价值的信息成为各大企业收集数据的主要目的,大数据的价值不再单纯来源于它的基本用途,而更多源于它的二次利用,以提升企业在市场中的竞争力.因此,数据成为公司有价值的财产、重要的经济投入和新型商业模式的基石.

企业所采集的大部分数据包含了个人信息,虽然有的数据表面上并不是个人数据,但经由大数据处理之后就可以追溯到个人.许多企业或组织基于大数据中数据巨大价值的驱动,无限制地收集、处理、使用和发布个人信息,还有许多大企业之间或企业与第三方间共享用户的信息.

1) 这种用户数据的使用和共享,给企业带来商机的同时,也对个人产生了惊人的影响.如一些购物网站基于用户过去一段时间的购买行为,有针对性的推荐产品或进行个性化广告的推荐;人们在犯罪之前,可以根据他们在互联网上的行为记录,准确地预测犯罪行为的发生.很显然,这些信息是从大数据的分析中获得的.

2) 这种用户数据的使用和共享给用户带来了风险:个人隐私泄露的频繁发生威胁到个人的生活安全,也成为影响社会治安的主要因素.据北京中关村派出所统计,2012年全年接报的电信诈骗占立案的32%,为比例最高的发案类型.诈骗中常采用6种手段:1)个人或交友圈信息泄露后的身份冒充,如犯罪分子冒充公检法机关、邮政、电信、银行、社保的工作人员或者亲友等实施诈骗,占诈骗案件总数的42%;2)购物信息泄露后冒充卖家诈骗;3)电话、QQ或邮箱等通信方式泄露后的中奖诈骗;4)寻求工作信息泄露后收到的虚假招聘信息;5)交友信息泄露后的网络交友诈骗;6)家庭信息泄露后的绑架诈骗.由此可见,许多企业都在不同程度上泄露了用户的个人信息.

3) 个人隐私信息的泄露引发了部分用户的恐慌,他们担心隐私数据丢失或者被恶意窃取.一项民意调查报告显示有72%的人担心他们的在线行为被公司跟踪和分析①.因此,大部分人提高了隐私保护意识,而很多企业对用户隐私保护重视不够,导致企业承受了潜在客户的丢失和经济利益损失[2].

由此可见,大数据时代数据分析存在着多面性,

① Consumer Reports Poll: Americans Extremely Concerned About Internet Privacy (http://markets.financialcontent.com/stocks/news/read/6669586/Consumer_Reports_Poll 2008,9,25)

如果对分析结果合理利用不仅能促进企业的发展,也能为用户提供更好的服务,但是,一旦出现不合理的运用会给个人的生活带来很多的烦恼,甚至是威胁.随着企业拥有数据量的不断增加,如果没有很好的策略解决个人隐私信息的保护问题,将会对企业甚至对整个社会造成不良的影响.当前,对个人隐私信息的保护还没有成熟的技术、成文的法律规定或行业标准,个人隐私保护更应该提上日程,实施各种有效措施保证个人隐私的安全,这也是大数据时代所面临的巨大挑战之一.

本文介绍了个人隐私的基本概念,讨论了大数据时代个人隐私面临的严峻挑战和研究问题,综述了个人隐私的保护技术,提出了企业或组织应遵守的法律和行业规范,最后探索了个人隐私保护的进一步研究方向.

1 个人隐私的概念及在大数据中面临的挑战

1.1 个人隐私的相关概念

1) 个人隐私的概念

隐私的提出要追溯到 Warren 等人[3]在 1890 年发表的《隐私权》,它成为美国传统法律的开创性著作. Warren 和 Brandeis 提出个人隐私权是一项独特的权利,应该受到保护,免遭他人对个人生活中想保守秘密细节的无根据发布.

隐私的概念在社会科学的所有领域(如哲学、心里学、社会学)已被研究大概 100 多年,但是并没有一个明确的既符合时代发展需求又符合实践检验的定义[4]. 隐私的定义主要分为 2 类:基于价值的,把隐私看作一种人权,是社会道德价值体系的一部分,一种商品[5],是人和社会的价值(如用户上网时,担心隐私问题的同时,在许多情况下为了达到自己的一些需求,仍然提交他们的个人信息);基于同源的,把隐私关系到个人的思想、感知和认识,看作一种状态(包含 4 种子状态:匿名、隐匿、保留和隐密),一种控制[6],表示个人和他人之间的交易控制,其最终的目标是增强自治或减少泄密.基于控制的隐私定义曾经是隐私研究的主流,但是也有研究把控制作为隐私的一个要素,两种研究成为学术界争论的焦点之一.

在某种意义上,隐私被描述为多维的、灵活的以及动态的,它随着生活的经验而变化,是机密、秘密、匿名、安全和伦理的概念重叠,同时也依赖特殊的情景(如时间、地点、职业、文化、理由)[7],因此不可能定义出通用的隐私概念.隐私保护随着信息技术的演化过程如表 1 所示[4]:

Table 1 The Evolution of Privacy with IT

表 1 隐私随着 IT 的演化过程

Period	Characteristics
Privacy Baseline 1945—1960	Limited information technology development, high public trust in government and business sector, and general comfort with the information collection.
First Era of Privacy Evolution 1961—1979	Rise of information privacy as an explicit social, political, and legal issue. Formulation of the Fair Information Practices (FIP) Framework and establishing government regulatory mechanisms.
Second Era of Privacy Evolution 1980—1990	Rise of computer and network systems, database capabilities, federal legislation designed to channel the new technologies into FIP, Some nations made data protection laws.
Third Era of Privacy Evolution 1991—2003	Rise of the Internet, data mining and the terrorist attack dramatically changed the landscape of information exchange and caused a user privacy concerns and the attention of the researchers.
Fourth Era of Privacy Evolution 2004—present	Rise of Web 2.0, cloud computing, Internet of things, and big data collected a lot of personal information, privacy concerns rose to new highs.

在特定的情景下,对不同的事,不同的人,隐私是指用户认为是自身敏感的且不愿意公开的信息. Banisar 等人[8]把个人隐私分为 4 类:①信息隐私,即个人数据的管理和使用,包括身份证号、银行账号、收入和财产状况、婚姻和家庭成员、医疗档案、消费和需求信息(如购物、买房、车、保险)、网络活动踪迹(如 IP 地址、浏览踪迹、活动内容)等;②通信隐私,即个人使用各种通信方式和其他人的交流,包括电话、QQ、E-mail、微信等;③空间隐私,即个人出入的特定空间或区域,包括家庭住址、工作单位以及个人出入的公共场所;④身体隐私,即保护个人身体的完整性,防止侵入性操作,如药物测试等.本文所指的个人隐私是公民个人生活中不愿为他人公开或知悉的个人信息,如用户的身份、轨迹、位置等敏感信息.隐私的范围包括私人信息、私人活动和私人空间.

2) 个人隐私的泄露

互联网已经成为我们生活的一部分,留下了我们访问各大网站的数据足迹.在大数据环境下,这使我们的隐私泄露变得更加容易,我们时刻暴露在"第三只眼"下,如淘宝、亚马逊、京东等各大购物网站都在监视着我们的购物习惯;百度、必应、谷歌等监视我们的查询记录;QQ、微博、电话记录等窃听了我们的社交关系网;监视系统监控着我们的E-mail、聊天记录、上网记录等;Flash cookies[9-10]泄露了我们的某些使用习惯或者位置等信息,广告商便跟踪我们的这些信息并推送相关广告等.

我们的日常活动也被监视着,如智能手机监视着我们所在位置;工作单位、各大活动场所、商店、小区等监视我们的出入行为.数字传感器技术的发展使得我们日常情况下的新型数据也可以被收集,如基于射频识别(radio frequency identification, RFID)的自动付款系统和车牌识别系统[11]、可植入的传感器监视病人的健康[12]、监视系统监视着在家的老人[13]等.随着传感器技术的不断成熟,各种类型的传感器将会被广泛地用于我们个人或组织.这些系统的特点是交互变得越来越模糊,因此,需要新的机制来管理个人信息和隐私产生的风险[14].

企业获得了大量的个人数据,他们会利用这些数据挖掘其蕴含的巨大价值,促进企业的发展或者获得更多的经济利益.个人隐私数据的保护面临着内忧外患.内忧主要指的是企业内部,Smith等人[15]指出企业在处理数据的过程中造成隐私泄露问题有4个相关的数据维:信息的收集、误用、二次使用以及未授权访问.此外,业内人可以对外发布数据,无授权地访问或窃取,把个人数据卖给第三方、金融机构或政府机构或者同他们共享数据等[4].外患主要指的是外部人为了获取数据,通过系统的漏洞对数据的窃取.同时,研究者们也发现通过财务奖励补偿用户,可以鼓励他们进行信息发布[16],同样,如果用户想要获得个性化服务,他们可能会提供更多的个人信息.因此,个人隐私的泄露不仅有企业的责任而且也有个人的因素,而个人隐私的泄露可能影响到个人的情感、身体以及财物等多个方面[17].

3) 不同人对个人隐私的担忧

个人的经历和自身特性也影响对隐私问题的不同看待.IBM的调查[18]显示:高管们通常都会低估客户对隐私的担忧;更多精通技术和受过教育的受访者更会意识到且更担心潜在的网上隐私的侵犯;Sheehan等人[19]发现女人比男人更担心她们隐私信息被收集;Culnan[20]发现年轻人、穷人、接受更少教育的人更少担忧个人隐私的泄露.

一些研究者也发现,个人对企业或组织的信任也影响隐私数据的收集.Bowie等人[21]发现企业在对待用户隐私方面值得用户信任,将在竞争中更占据优势.用户对企业信任会更少担心他们的隐私被泄露,也更愿意提供个人信息.

4) 个人隐私与安全的关系

Belanger等人[22]认为人们对隐私与安全的关系缺乏理解.安全对应个人信息保护问题的3个具体目标:①完整性,确保信息在传输和存储过程中不被篡改;②认证,对用户身份以及数据访问资格的验证;③保密,要求数据的使用只限于被授权的人.Culnan等人[23]认为组织可以安全地存储个人信息,但是可能对随后个人信息的使用作出错误的决定,导致隐私信息泄露的问题.Ackerman[24]也表示安全对隐私是必要的,但是安全不足够保证随后的使用,不足够将发布的风险最小化,也不足够使用户放心.由此可见,安全并不能保证个人隐私完全受到保护,必须在确保个人信息安全的基础上,加之对个人信息的正确使用才能确保个人隐私不被泄露的可能.

1.2 大数据时代个人隐私面临的挑战和研究问题

"人、机、物"三元世界在网络空间中交互、融合产生的网络大数据带来了巨大的机遇,同时也给现有的IT架构、机器处理以及计算能力带来许多科学问题和极大挑战[25].此外,大数据具有数据量大、数据类型繁多、数据生成速度快以及价值密度低等特点,加之个人隐私随着诸多因素动态变动的特性,使得保护大数据时代的个人隐私更是难上加难.下面针对大数据的个人隐私保护,阐述相关的6个挑战和研究问题.

1) 个人隐私保护的范围难以确定.根据以上对个人隐私概念的阐述,隐私的概念是随着信息技术的发展而变化的,同时还要考虑不同人的特性和背景,因此,隐私保护哪些敏感数据很难界定.

2) 侵犯个人隐私的行为难以认定.侵犯个人隐私的形式复杂多样,对于界定是否构成侵权行为,根据目前的法律却无法判断.用户在网络上通常使用假名,这种匿名方式使受害人很难收集证据并找到真正的侵权人.即使受害人通过网页备份等手段取得证据,但网页总是处于不断更新之中,只要侵权人不予承认也难以发挥证据的效力.因此,如何判定是谁侵犯了个人隐私面临着极大的挑战.

3) 随着信息和通信技术变得越来越普遍,管理个人隐私信息也变得更加困难. 管理个人隐私信息包括个人隐私信息的收集、存储、使用以及发布. ①在收集个人信息时,如何保证收集到的信息在传输过程中维持其完整性;②在存储个人信息时,使用何种技术保证信息不被窃取或非法访问;③对于个人信息的使用,应该如何设置严格的访问控制策略,使不同的人见到不同访问级别的数据,同时不增加太多的管理工作量;④在发布信息时,控制需要发布什么信息以及谁可以在网络上访问发布的信息已经成为企业越来越关注的问题. 对于将要发布的数据,如何保证数据不会泄露个人的隐私信息,同时保证数据的效用,而不能为了保护隐私就将所有的数据都加以隐藏,这样则不能体现数据的价值所在.

企业的管理者越来越意识到保护个人隐私数据的重要性,因为这些数据将直接关系到企业的利益. 然而,如何管理好数据,即保证数据使用效用的同时保护个人隐私,是大数据时代企业面临的巨大挑战之一.

4) 个人隐私保护的技术挑战. 当人们意识到要保护自己的隐私,试图将自己的行为隐藏起来时,却没有想到自己的行为已经在互联网尤其是社交网络的不同的地点产生了许多数据足迹[26]. 这种数据具有累积性和关联性的特点,单个地点的信息可能不会暴露用户的隐私,但是如果将某个人的很多行为从不同的独立地点聚集在一起时,他的隐私就会暴露,因为有关他的信息已经足够多,这种隐性的数据暴露往往是个人无法预知和控制的. 从技术层面来说,可以通过数据抽取和集成实现用户隐私的获取,而在现实中通过所谓的"人肉搜索"的方式能更快速、准确地得到结果. 服务提供商也可能从授权用户数据的二次使用来获得利益,如目标广告的投放,目前,对数据的二次使用还没有技术障碍. 此外,大数据时代数据具有产生速度快的特点,对动态数据需要怎样的处理技术以迅速地构建隐私保护,而不影响到数据的使用效用,面临着技术和人力层面的双重考验.

5) 为构建良好的大数据生态环境,构建多维的、灵活的个人隐私保护政策面临着极大的挑战. 企业为了提高市场竞争力或为用户提供更好的服务,要求用户注册时提供一些包括个人敏感信息的相关数据,而用户为了得到某些服务也依据要求提供了自己的相关数据,但是在数据的传输或使用过程中,欺诈犯罪和个人隐私泄露频繁发生,威胁到了个人的生活安全. 用户意识到需要保护自己的隐私时,注册的个人信息不再填写真实的数据,而企业为了提供更好的个性化服务,对用户的相关数据进行分析时,由于用户信息的不真实,造成分析的结果与现实存在很大的偏差,达不到企业想为用户提供服务的效果. 在这种情况下,如果没有相关的个人隐私保护政策出台,将引起个人信息不真实与企业提供个性化服务偏差的恶性循环. 因此,提出更好的个人隐私保护策略、构建良好的大数据生态环境,是急需解决的问题.

6) 大数据的数据来源成为研究者的研究障碍. 由于大数据的数据量巨大(如 Web 数据、科学数据、财政数据、移动对象数据等),因此,只有大公司拥有这样的数据,以至于研究者很难得到数据,加之对个人隐私的动态研究紧密关系到用户的行为过程,而不能建立在假设的基础上,导致许多研究无法进行.

总之,大数据的个人隐私保护在人员、管理、生态环境和研究的各个层面上提出了挑战性研究问题. 目前,大数据的个人隐私保护研究刚开始起步,各大企业也在摸索着行业规则,谨慎地处理个人的信息. 当然本文提出的挑战只是个人隐私保护的几个方面,随着技术和观念的不断成熟和演化,会有更多的挑战等待解决.

2 大数据个人隐私保护技术

现有的隐私保护技术分为 3 类:数据扰动技术、数据加密技术和数据匿名化技术,而个人隐私数据经历收集、存储和使用过程(使用包括数据的二次使用、数据共享以及数据发布),因此,应该实施数据的多级安全保护,本节结合大数据的特征从数据层、应用层以及数据展示层对个人隐私保护技术和相关的工作进行叙述.

2.1 数据层的个人隐私保护

通信中的数据可以使用 SSL 协议保证数据的安全,因此,数据层的数据保护主要是指对数据的存储和管理的保护. 保证数据层个人信息的安全是其他一切以数据为基础应用的根本,包括保证数据的机密性、完整性和可用性. 本节主要从数据的加密和访问控制两方面叙述保护个人隐私数据的相关研究.

2.1.1 数据加密的个人隐私保护

数据加密技术已有悠久历史,进入数字化时代之后,它仍然是计算机系统对敏感信息保护的一种

可靠的方法.数据加密的作用是防止入侵者窃取或者篡改重要的数据.按照加密的密钥算法,数据加密可分为对称加密算法和非对称加密算法.

1) 对称加密算法是加密和解密时使用相同的密钥,主要用于保证数据的机密性.最具有代表性的算法是20世纪70年代IBM公司提出的DES(data encryption standard)算法;在此基础上又提出了许多DES的改进算法,如三重DES(triple DES)、随机化DES(RDES)、IDEA(international data encryption algorithm)、广义DES(generalized DES)、NewDES、Blowfish、FEAL以及RC5等.2001年美国国家标准与技术研究院发布高级加密标准(advanced encryption standard, AES)取代了DES,成为对称密钥加密中最流行的算法之一.

对称加密算法的优点是计算开销小、加密速度快,适用于少量或海量数据的加密,是目前用于信息加密的主要算法.其缺点是通信双方使用相同的密钥,很难确保双方密钥的安全性;密钥数据量增长时,密钥管理会给用户带来负担;此外,它仅适用于对数据进行加解密处理,提供数据的机密性,它不适合在分布式网络系统中使用,密钥管理困难,且成本较高.

2) 非对称加密算法也叫公开密钥算法,其加密和解密是相对独立的,使用不同的密钥.它主要用于身份认证、数字签名等信息交换领域.公钥密码体制的算法中最著名的代表是RSA,此外还有背包密码、DSA,McEliece密码、Diffe_Hellman、Rabin、零知识证明、椭圆曲线、EIGamal算法等.

非对称加密算法的优点是可以适应网络的开放性要求,且密钥管理问题也较为简单,可方便地实现数字签名和验证.其缺点是算法复杂、加密数据的速率较低.

然而,无论是对称加密算法还是非对称加密算法都存在密钥泄露的风险.因此,Rivest在1989年开发出MD2算法①,不需要密钥,引发了杂凑算法(也称Hash函数)的研究,即把任意长的输入消息字符串变化成固定长的输出串,不需要密钥,且过程是单向的,不可逆的.比较流行的算法有MD5,sha-1,RIPEMD以及Haval等.杂凑算法不存在密钥保管和分发问题,非常适合在分布式网络系统上使用,但因加密计算复杂,通常只在数据量有限的情形下使用,如广泛应用在注册系统中的口令加密、软件使用期限加密等.

数据加密技术能保证最终数据的准确性和安全性,但计算开销比较大,加密并不能防止数据流向外部,因此,加密自身不能完全解决保护数据隐私的问题.

数据加密算法作为隐私保护的一项关键技术,大数据时代研究重点将集中在对已有算法的完善;综合使用对称加密算法和非对称加密算法.随着新技术的出现会研究出符合新技术发展的新加密算法.

2.1.2 数据库的个人隐私保护

数据库仍然是信息系统的主体,如政府数据库存储的大量个人及家庭信息;金融数据库存储的个人财务信息;医疗数据库存储的个人医疗历史信息等,网络上使用的网上银行、邮件信息以及个人注册信息等.大数据时代虽然MapReduce技术广泛用于相关的数据分析,成为数据库的竞争者,但是MapReduce不能完全替代数据库,它们之间可以相互学习,并且走向集成,形成新生态系统[27].

数据库不但面临入侵者的威胁,而且也面临内部人员的威胁,主要包括未授权的数据查看、不正确的数据修改以及数据的不可用性[28].保证数据库安全要从4个层面考虑[29]:物理安全、操作系统安全、DBMS安全和数据库加密.前3层不足以保证数据的机密性,数据库加密能保证敏感信息以密文的形式存在从而受到保护.为了保护数据库中的敏感数据,采取数据加密和访问控制的双重机制.由于数据加密和访问控制的研究工作已经比较成熟,这里只叙述使用加密和访问控制时注意的事项.

对数据库中的数据进行加密增强了DBMS的安全性,但是对数据操作时的加密和解密操作也带来计算成本的开销,因此应该考虑实际的需求[30]:1)只加密敏感数据;2)在查询期间,只加密或解密感兴趣的数据;3)基于加密属性值建立索引,会导致一些索引特性的丢失,如范围查询;4)加密的数据库不应该增加太多的存储空间.

单纯的数据库加密不能防止各种攻击,还需要通过访问控制来确保数据的安全.访问控制技术起源于20世纪70年代,为了满足当时系统上共享数据授权访问的需要.访问控制是数据库保护资源的关键策略之一,保证合法用户对资源只能进行经过相应授权的合法操作,其内容包括认证、控制策略实现和安全审计,其中安全审计可以审计用户的行为,

① RFC 1319, The MD2 Message-Digest Algorithm (http://tools.ietf.org/html/rfc13191992.4)

并将用户的行为记录在审计日志中,作为一项重要事件追踪的依据,所有的用户都无权修改.数据库的访问控制对象包括数据库、关系、元组以及属性,因此,访问控制级别分为粗粒度(如数据库或表)和细粒度(如元组或属性)两种.访问控制策略包括自主访问控制策略、强制访问控制策略以及基于角色的访问控制策略等.根据大数据对数据访问灵活性的需求,访问控制策略应该根据应用灵活地设置,如非级联权限回收、时间段内的授权以及使用视图支持基于内容的控制策略等.

数据加密确保个人的敏感信息以密文的形式存储,即使攻击者获得受保护的数据,也无法读取和使用.对于内部人员使用细粒度的访问控制策略,确保不同的人或群组拥有不同的访问权限.所有人员的操作都必须记录到审计日志中,通过日志可以跟踪到具体人员的操作行为.

2.1.3 云存储环境下的个人隐私保护

云计算可以看成高速公路,而大数据则是高速公路上的一辆车.云计算为大数据提供了基础存储平台,以一种实惠且容易使用的方式帮助组织存储、管理、共享以及分析大数据.现在许多企业和个人把数据存储在云上,节约了软硬件成本,减轻了本地存储和维护的负担,而且能不限地理位置地随意访问,但是企业和个人失去了对数据的完全控制,云计算也给数据的安全带来了新挑战.

个人数据并非以一种完全加密的形式存储在云服务器中,面临着入侵者和内部人员对数据的威胁.因此,存在个人隐私数据泄露的风险,加之云提供商没有完善的审计和监测技术,不能及时检测到所有入侵和违规操作[31];提供商可以记录用户的服务需求,并且推断用户的隐私信息;管理员的误用导致丢失了用户的隐私数据;员工为了经济利益或者恶意用户突破机器的安全窃取数据;数据被其他有相同服务且没有被授权的用户的访问等.

云计算中通常关系到个人数据的收集、使用、发布、存储、销毁等[32].在云计算方面已经有许多关于隐私问题的研究文章:Chen 等人[32]分析了在云中整个数据生态圈(包括 7 个阶段:数据产生、传输、使用、共享、存储、存档、销毁)的隐私保护问题;Roy 等人[33]把分散信息流控制和差分隐私保护技术应用到云中的数据产生和计算阶段,并提出一个隐私保护系统 Airavat,该系统在 Map-Reduce 计算过程中可以阻止未经许可的隐私泄露;Mowbray 等人[34]提出使用 policy-based 模糊处理(obfuscation)的隐私管家来增强隐私保护,即用户的隐私数据以加密的形式被发送到云上,且处理时也是加密的数据,隐私管家对处理过的输出通过消除模糊处理来显示正确的结果,这种方法不仅减小了一些人员从云上窃取数据的风险,也防止了他们对数据未授权的使用;Zhang 等人[35]针对提供商可能根据用户的需求推断用户的隐私信息的问题,提出噪声产生策略 HPNGS,即根据用户需求历史发生的概率产生需求噪声,使得所有噪声需求和真实需求达到相同的发生概率,这样服务提供商很难辨别哪个是用户的真实需求,从而达到隐私保护的目的;Wang 等人[36]首次提出了云存储的隐私保护公共审计,在云上存储数据的用户或企业可以求助于第三方审计来检测数据的安全性,而不需要数据的本地复制,也不会增加云用户的在线负担.在审计过程中,对用户数据的隐私不会增加新的威胁.隐私保护公共审计被证明是安全的,而且是高效的.

为了确保云平台上数据的安全,从部署和服务对象的范围把云计算分为公有云、私有云以及混合云;从提供服务层次上分为 IaaS,PaaS,SaaS.虽然每种分类中都实施了很多隐私保护技术,但是隐私泄露事件仍然不断出现,尤其公有云上的数据.云存储作为大数据的基础存储平台,应该把隐私贯穿于云计算设计的每个阶段[37],对敏感的信息进行加密处理,并且制定比较细粒度的访问控制策略.

2.2 应用层的个人隐私保护

针对具体的大数据应用,研究相应的个人隐私保护技术是目前企业更加切合实际且满足具体应用需求的做法.本节主要从大数据时代比较流行的应用,即在线社会网络、移动定位以及射频识别 3 个方面讲述个人隐私保护的技术方法.

2.2.1 在线社会网络隐私保护

在社会网络中,当用户参与更多活动且包含更多的社会内容时,隐私问题便更加凸显出来[38].当个性化中有新的内容加入到社会网络中时,这些内容中的隐私信息可以在用户不能预见的多种途径上被共享[39].用户不仅担心他们的隐私信息被使用,而且担心无意中隐私信息流入到社会网络.

在线社会网络(online social networks, OSNs)提供给许多人一种交流、共享兴趣以及更新他们当前活动的途径.现在流行的 OSNs 包括社交网站(如人人网、Facebook)、微博(如新浪微博、Twitter)、博客等.隐私问题最大的威胁是信息的泄露:SNS

提供商采用的内容管理策略允许第三方为不同的目的利用 OSNs 用户信息;另一个泄露的危险应用是信息链接[40],如为了推断一个用户的身份或者行为信息,没有授权的第三方有从不同社会数据中整合数据的可能性. 因此,实际的危险是用户失去了对他们自身信息传播的完整控制,如用户经常无条件同意提供商制定的条款;允许提供商使用和挖掘用户数据;存储在 OSNs 提供商处的数据存在潜在的被盗、内部攻击或执法机构的查看等危险. OSNs 提供商对数据缓存的通常做法和离线存储增加了隐私泄露的风险,并对用户的隐私构成了永久的威胁. 因此,OSNs 面临的一个关键问题是用户隐私保护问题[41]. 处理在 OSNs 中的个人隐私需对数据的拥有者授权对数据的控制,如用户可授权给他的朋友访问他的数据,同时对提供商和其他未授权的实体隐藏数据[42-43].

在 OSNs 中,用户的隐私不仅包括个人信息,而且也包括交流信息. 保护这些信息需要达到的目的是只有被用户直接授权的人才能访问,访问控制需要是细粒度的并且每个属性能分开管理. 然而,在 OSNs 中隐私保护面临的问题是:1)用户不能控制他们的隐私数据,社会网络的提供商可以全权访问用户的数据,潜在着对用户私有数据如简介、通讯录等的访问,或对这些数据进行挖掘,或卖于第三方;2)这些网络中,用户只能定义粗粒度的访问控制,不能设置细粒度的访问控制,如微博发布时权限有密友圈、仅自己可见、分组可见(可以选择已定义好的分组)以及公开(默认权限). 可见,用户的隐私信息易受到供应商的误用以及意外或恶意的泄露,因此,需要更有效的方法来保护用户的隐私,避免用户隐私的泄露.

针对 OSNs 的隐私保护,在研究界已经提出很多方法. 集中式 OSNs 是一个支持高可用性和实时内容传播的社会网络模型. Singh 等人[44]提出了集中设计(centralized designs),信任拥有用户数据的 OSNs 提供商,并允许提供商执行有效的流量分析,通过学习用户的社交接触来决定用户的隐私. Persona[45]通过基于属性加密和传统公钥加密技术的组合,提供灵活的细粒度的访问控制,通过加密技术确保数据的保密性和隐私. De Cristofaro 等人[46]认为微博(如新浪微博、Twitter)的隐私机制应该不同于社交网站(如人人网、Facebook). 微博除了内容隐私外,使用标签标记以及检索内容可能泄露个人的习惯、政治观点、甚至健康状况,因此,需要检测是否符合存储和管理内容的信任或者检索的标准,同时增强用户的访问控制列表. 在集中的 OSNs,他们提出的 Hummingbird 是 Twitter 隐私增强的变体,它保留了 Twitter 的关键特征,同时增加了 2 个隐私敏感的要素,即 1)粉丝的细粒度授权:一个推特用户可以加密一个推文,并选择允许谁访问它;2)粉丝的隐私:他们订阅任意的 hashtags 而不会对任何实体泄露他们的兴趣.

使用 P2P 架构来解决 ONSs 隐私问题. 分散式架构可能隐藏实时消息的可用性,或者需要用户购买云存储来存储他们的数据. 分散设计(decentralized designs)不依靠单方信任或不信任实体,这样的设计不集中数据管理,用户自己或他们信任的联系人存储数据. 在分布式 OSNs 的研究中,同时满足安全、隐私和服务质量需求存在极大的挑战. 和 Persona 相关的是 DECENT[41],DECENT 是建立在分散 OSN 上的加强访问控制的架构,支持访问授权,同时提供极其细粒度的访问控制,通过基于属性的标签验证用户的完整性. 通过加密技术解决存储服务,防止存储节点或第三方干预用户的隐私策略. Safebook[47]是一个分散和隐私保护的 OSN 应用,集中在资源可用性、内容隐私以及端对端交流的保密性. PeerSoN[48]通过加密确保访问控制加上 P2P 的方法来替代传统 OSNs 的集中授权,防止用户、供应商或广告商违反隐私保护的企图. PeerSoN 利用非信任的 P2P 系统,提供的隐私保护比 Safebook 弱. LotusNet[49]是一个基于分布式散列表(distributed Hash table, DHT)的 OSN 架构模式. 在分布式社会网络中,处理安全隐私和服务权衡问题,通过一个灵活细粒度的自主访问控制来控制私有资源,提供给用户调整他们隐私设置的可能性.

大量的社会网络被建模成图,也包含了大量敏感信息,隐私问题关系到图数据的分析和管理,因此隐私保护给图处理带来了更大的挑战. Zheleva 等人[50]研究在图数据中保护隐私的敏感关系的问题,把从匿名图数据中推断敏感关系的问题看作链接重鉴定问题,根据数据移除量和隐私保护量提出了 5 种不同的隐私保护策略. Zhou 等人[51]解决邻域攻击问题,为了保护顶点的隐私,已知社会网络图中的任何顶点在超出给定的自信度阈值时不能在发布的 k-anonymity 图中重新鉴别出来. Liu 等人[52]发现节点的度可能暴露个人身份,研究特定图的 k-degree 匿名化问题,防止有人故意利用确切节点度的先验知识进行个人的重鉴定.

2.2.2 移动定位的隐私保护

随着无线通信和移动定位技术(如 GPS,WiFi)的出现,以及移动数据带宽的增加,定位服务(location-based services, LBSs)变得越来越普及,许多新的应用使用用户的物理位置为商业、社会或信息的目的提供 LBSs. 因此,服务提供商可以持续地跟踪用户的位置,根据对用户精确的物理定位为他们提供服务,如开发新的移动应用、提高个性化搜索结果、提供移动广告服务以及天气信息等. 电子商务服务也根据用户的位置进行差异定价或提供优惠券等. 基于定位服务给各方带来利益的同时,也暴露了移动用户的个人信息,如追踪、暴露家庭位置、被老板跟踪、被政府跟踪以及基于位置广告的打扰等,这成为提供基于位置服务所担忧的关键问题[53].

针对如何提供定位服务的同时保护好移动用户的位置隐私(如用户想查找"据他所在位置最近的购物商场"同时隐藏他的确切位置以及他查询的敏感信息),研究者们已经提出了很多方法. 两种最常用的隐私度量标准是匿名和干扰技术. 当隐藏了用户的身份信息时,一些基于用户身份的在线服务将不可用. 针对这种情况,研究者们提出一种解决方法是保护用户隐私的同时减少位置信息的精确性. 如 Google+允许用户根据不同的好友圈子范围在不同程度上分享自己的地理位置:用户可以和家人分享自己精确的地理位置,而与同事或朋友只分享所在的城市.

为了防止依赖位置攻击,保护位置隐私,Pan 等人[54]提出采用隐私粒度和位置 k-anonymity 作为隐私的测量标准,利用图模型来形式化问题,并转换成在图中寻找 k 节点团的问题,提出了一个基于团的递增的匿名算法 ICliqueCloak,当新的需求达到时能快速识别并产生匿名区域保护位置隐私. Ardagna 等人[55]为了解决对用户位置隐私的不同等级,提出了模糊处理由传感技术测量的位置信息的解决办法,保护用户的位置隐私. 使用相关性来测量位置信息的精确性和隐私,保证解决方案的鲁棒性,并使用自动协商协议对服务提供商需要的定位服务的位置精确性等级和用户要求的位置信息保护进行权衡. Zhu 等人[56]为解决用户通过他的设备发送虚假位置,使他能访问一个受限的资源或者提供虚假不在场证明的问题,提出 APPLAUS,在 APPLAUS 中,协同定位的蓝牙能使移动设备相互地产生位置证明,并更新到一个位置证明服务器. 定期的变换移动设备使用的假名来保护相互的原位置隐私和不信任的位置证明服务器. 个体用户使用以用户为中心的位置隐私模型来实时评估他们的位置隐私等级,并决定是否以及何时接收根据他们的位置隐私等级的需求交换位置证明.

2.2.3 RFID 的安全和隐私

射频识别(radio frequency identification, RFID)[①]是一种无线通信技术,用来识别物体或人. 许多行业都运用了射频识别技术:射频标签附着在一辆正在生产中的汽车上,厂方可以追踪此车在生产线上的进度;射频标签附于产品上可以监控产品在物流管理供应链中准确和实时地移动信息;射频标签也可以附于牲畜或宠物上,方便对牲畜或宠物的识别;射频识别的身份识别卡可以使员工得以进入建筑锁住的部分;汽车上的射频应答器也可以用来征收收费路段与停车场的费用等.

由于 RFID 标签无需直接与收发器接触,当 RFID 的标签序列号和个人信息关联时,可能会在未经本人许可的情况下读取个人信息,这就威胁到个人隐私,包括两个主要的隐私担忧:秘密跟踪和推断[57]. 如消费者使用信用卡购物时,商店可能建立起他的身份和标签序列号之间的联系,卖主可能使用 RFID 阅读器网络识别和分析消费者. 一方面,消费者带着有 RFID 标签的物品可能提供一个秘密的物理跟踪. 另一方面,对个人拥有的 RFID 标签的物品的推断可以获得重要的个人信息:通过个人所带的药物判断得了什么病;通过个人所带的 RFID 优惠卡判断他会在哪里购物;甚至可以推断出个人所穿衣服的尺寸以及饰品的偏好等. 可见,一旦 RFID 普遍应用,RFID 所面临的挑战是如何保护好个人隐私的问题.

为了保护个人隐私,已经提出了很多隐私增强技术:1)当购买商品后,销售点的设备将"杀掉" RFID 标签,使 RFID 标签永久不起作用;2)重命名方法,包括重贴标签、"简约"加密、重加密、通用重加密;3)代理方法,用户不依靠公共的 RFID 阅读器来增强隐私保护,而使用自有的 RFID 隐私增强设备,如一些手机包含了 RFID 功能,他们可能最终会支持隐私保护;4)距离测量,在 RFID 阅读器与射频标签间有一个粗略的距离测量,只有在距离范围内才能获得更多的具体信息;5)阻塞(blocking),在射频标签中加入一个可修正的隐私位,0 表示无限制的

① http://zh.wikipedia.org/wiki/射频识别 2013,6,4

公开扫描,1 表示隐私,此种方法包括软阻塞和信任计算;6)通过立法来保护个人隐私等.

2.3 数据发布的个人隐私保护

政府、企业以及个人可以对收集到的数据进行分析,从而提升服务或者作出决策,在这种利益的驱动下,他们之间需要共享或发布一些数据.如果数据的发布者没有考虑隐私保护而发布数据,将对企业造成经济或名誉损失的严重后果,因此,数据发布面临的挑战是发布的数据既能保证个人隐私信息不泄露,又能最大程度地提高发布数据的效用.

传统的数据库在有数据需求的情况下才发布数据,即使用拉的策略;但是,在 Web 应用环境下,数据在无需求的情况下也会发送给授权的主体,即使用推的策略[28].因此数据的发布不仅要有正确的发布策略,也应该有方法支持发布给第三方的信息架构,保证发布数据的可用性,同时也要保护好个人的隐私信息.

2.3.1 匿名化方法的个人隐私保护

数据发布技术是个人隐私保护的研究热点,目前已有很多数据的发布技术,主要的研究集中于匿名化方法.

匿名化方法是通过隐藏用户的身份和敏感数据达到隐私保护的目的.在数据发布前,主要的匿名操作有泛化[58-59]、压缩、分解[60]、置换[61]以及干扰.其中泛化和压缩是隐藏准标识符(能识别出用户的属性集)的一些细节,使用一个通用的值替换一个具体的值;分解和置换是通过分组和混排敏感属性,解耦准标识符和敏感属性之间的关联;干扰是通过添加噪声(如使用随机化方法)、数据交换以及合成数据生成等来干扰敏感数据.

通常的个人隐私受到的威胁是通过发布的数据记录推断某条记录的相关个人,为了解决这个问题,可以通过匿名化方法避免攻击者使用链接(包括属性链接、记录链接以及表链接)推断出个人的隐私信息.避免记录链接的方法以 k-anonymity[62] 为基础,及其改进方法 (X,Y)-anonymity[63] 和 MultiR k-anonymity[64] 等,它们是把某一记录隐藏在一个大组记录中,达到保护个人隐私的目的.避免属性链接的方法有 l-diversity[65],(α,k)-anonymity[66],t-closeness[67]等,防止攻击者通过发布的数据推断敏感属性值.避免表链接的方法 δ-Presence[68],为防止攻击者推断某一用户的记录是否出现在表中,通过限定表中每个元组在一个指定的概率范围内达到隐私保护的目的.

国内学者针对数据发布提出的比较著名的匿名化隐私保护方法有 Alpha$^+$[69],对不同领域考虑属性权重的数据匿名发布算法 WAK-anonymity[70]等.当然还有许多根据不同的攻击类型而提出的匿名化方法,它们都有各自的优缺点[71],其中匿名化隐私保护的一些方法存在的最大缺点是当攻击者拥有大量的背景知识时,通过结合发布的信息进行关联分析,还是容易推断出某个记录的敏感信息.大数据环境下,为了减少数据共享或发布时无意的数据泄露,数据在传输前应该匿名化,并结合其他技术使接受者对收到数据无法作关联推断,这样既能利用那些数据,又能避免牵扯到具体的个人.利用匿名化共享或发布数据的应用很多,如关系数据、个人的移动数据[72]、时空数据(特定时间个人的位置)[73]、社会网络数据[74-75]、查询日志[76]以及数据挖掘等.

2.3.2 PPDM 的数据发布

数据挖掘可以使我们发现隐藏在数据中有价值的信息.这就驱动人们使用各种挖掘算法挖掘出能支持管理者决策的信息或者带来更多商业利益的信息.与此同时,也带来了安全和隐私问题.IBM Almaden 研究中心的 Agrawal 领导的研究小组,在 2000 年的 ACM SIGMOD 会议上首次提出了"隐私保护数据挖掘(privacy-preserving data mining, PPDM)"的概念[77].隐私保护数据挖掘主要考虑 2 个问题[78]:1)为了数据的接收者不危害他人的隐私,原始数据的敏感信息像标识符、姓名、地址等应该被修改或者从原始数据中去除;2)通过使用数据挖掘算法,从数据中挖掘出的敏感知识也应该被去除,因为这些知识可能同样危害到个人隐私.PPDM 的主要目的是利用算法在一定程度上对原始数据进行修改,使得隐私数据和隐私知识在挖掘过程之后仍然保持隐私.目前,PPDM 主要有 2 种方法:干扰、加密以及匿名化[79].下面主要讲述匿名化方法.

在 PPDM 中,公布带隐私的数值或分类数据时,最常用的匿名方法有 k-anonymity,l-diversity,t-closeness 等.PPDM 中对于动态数据采用顺序发布,Wang 等人[63]提出使用有损连接防止攻击者对两次发布的视图进行连接以识别记录的身份.当对原始的数据增加或者删除属性时,为了避免再次发布时隐私泄露的风险,Xiao 等人[80]提出了 m 不变性的泛化原则.这些方法的主要思想是对于当前和以前的表发布的数据都需要满足 k-anonymity 要求.

大数据中也存在文本和字符串数据类型,PPDM 中对它们采用了其他方法辅助匿名化的方法.

文本数据的特点是高维、稀疏,因此并不能使用标准的 k-anonymization 技术解决 PPDM 问题. Aggarwal 等人[81]利用文本数据稀疏的特征提出了基于草绘的方法构建数据的匿名化表示. 字符串数据的特点是不同记录的字符串长度不同,构建变长的属性匿名化非常困难. Aggarwal 等人[82]提出了基于压缩的方法对字符串数据进行匿名化.

许多 PPDM 方法面临的最大挑战是维度灾难,为了解决这个问题,通过找出定义大多数行为的关键属性来降低数据的维数或者对大量的属性进行压缩.

PPDM 是大数据时代价值发现的主要研究领域,因此还会有更多新的方法来适应大数据应用的发展,如 2.3.3 节所讲的差分隐私保护.

2.3.3 差分隐私保护

随着隐私保护的需求越来越严格,针对匿名方法存在由背景知识推断某些记录敏感信息的缺点,2006 年 Dwork[83]首次提出了一个统计隐私模型,即差分隐私保护(differential privacy)解决了这个问题. 差分隐私保护的优点是它提供了一个更多的语义保证,无论攻击者拥有怎样的背景知识和权力,只能从个人数据中得出有限的结论. 差分隐私保护定义了一个极为严格的攻击模型,并对隐私泄露风险进行了严格的数学证明和定量化表示,攻击者即使知道除一条记录之外所有记录的敏感信息,仍然不能推断出这条记录的任何敏感信息,所以隐私泄露的风险很小. 在数据集中添加或删除一条记录也不会对输出结果产生影响. 差分隐私保护的目的是最小化隐私泄露、最大化数据效用[84],因此,差分隐私保护提出后就在统计数据库领域得到了相当大的支持,它与特定领域无关的特性能与其他领域很好地结合,现已广泛地应用到其他领域,如数据挖掘、机器学习、社交网络、安全通信、决定论、经济学以及密码学等[85].

差分隐私保护是基于数据失真技术,在数据集中加入满足特定分布的随机噪声,从而达到隐私保护的目的,但所加入的噪声量与数据集大小无关,只与全局敏感性密切相关,因此对于大型数据集,仅通过添加少量的噪声就能达到高级别的隐私保护. 常用的添加噪声的机制有拉普拉斯机制[86]、指数机制[87]和数据库访问机制[88].

差分隐私的数据发布技术主要采用非交互式框架发布带敏感数据的信息,且发布的数据满足数据分析者的需求. 常采用的发布技术有直方图[89]、采样和过滤[90]、数据立方体[91]以及划分[92](如树或网格)等. 这些方法中采用不同的添加噪声策略,主要有 2 种:1)对原始数据添加噪声;2)对转换后的原始数据添加噪声.

差分隐私保护在大大降低隐私泄露风险的同时,极大地保证了数据的可用性,成为了现今使用的新的隐私保护模型和各领域的研究焦点,也是大数据时代隐私保护的主要技术,比较适合个人隐私保护的需求,如用户购买商品的信息和行为模式的挖掘、抽取用户兴趣特征的个性化推荐或广告推荐、社交网络中用户社交圈的挖掘、移动终端对用户位置的定位以及发布数据给第三方或与第三方共享数据等. 然而,大数据具有产生快的动态性,如何解决好动态数据的差分隐私保护还有待研究.

2.3.4 数据访问控制的个人隐私保护

现在一些企业也提供了一些机制使个人也可以控制自己的敏感信息是否对外发布或者对哪些人发布,他可以编辑许可约束限制权或指定条件才能访问他的数据. 如在新浪微博发布信息时,可以选择哪些用户能见到你发布的信息,主要权限有“密友圈”、“仅自己可见”、“分组可见”(可以选择你的分组),如果都不选择则默认是公开发布,所有人都可见. 在发表博客时查看权限有:“公开”、“博友”、“私人”,根据自己发表博文的内容选择可见的用户. 在最常用的 QQ 通信中,权限设置包括“所有人可见”、“仅好友可见”、“仅自己可见”,根据你的公布每项个人信息的意愿,选择访问权限. Facebook 有 5 种权限设置:“私人”、“指定人”,“仅朋友”、“朋友的朋友”、“每个人”,默认设置是每个人. Liu 等人[93]针对隐私设置对 200 名 Facebook 的用户作了调查,发现对共享默认的隐私设置有 36%的内容;隐私设置满足用户的期望只有 37%的时间,表明当前的设置在大多数时间都不正确;当用户改变他们默认的隐私设置时,改变的设置只满足用户期望的 39%的时间,表明有更多隐私意识的用户也很难正确地管理和维护他们隐私设置. 和 Facebook 相比,2011 年 Google 推出的 Google+在隐私设置上显示了突出的优点,Google+是社交网站与身份服务,对隐私功能进行了细粒度划分,让用户可以在不同的朋友圈里分享信息.

由用户自己决定哪些自身的信息是他们比较关心的信息,信息可以被哪些人看到,这是大数据时代保护个人隐私发展的一种趋势. 现在企业开发的软件这方面的功能还比较弱,不能满足用户的隐私保护需求,因此,企业应该对现有的软件进行完善或更新,

为用户提供更细粒度的访问控制机制,使用户对自己要保护的信息有更主动的控制权.企业可以根据用户的设置确定信息的保护范围和保护级别,并对他们的信息进行合理的存储、管理、使用和发布,更好地保护个人隐私,提供更人性化的服务.

2.3.5 数据发布的个人隐私保护评估

带敏感信息的发布必须在效用和隐私间做到很好的权衡[94].效用的目标是对每个潜在的用户独立他的辅助信息和偏好,最优化效用.发布完全准确的信息需要最大化效用同时最小化隐私.

对隐私保护技术的度量通常的做法有:1)隐私保护度.通过发布数据的泄露信息的风险来反映,泄露信息的风险越小,隐私保护度越高.如2009年加拿大隐私高级代表办公室对Facebook隐私功能进行评估,要求Facebook对其隐私政策进行升级.Facebook增加了向用户提供有关其隐私功能的信息,以及采取技术调整措施,以加强隐私保护力度.2)数据指标.是对发布数据质量的度量,它反映通过隐私保护技术处理后信息的丢失程度:数据缺损越高信息丢失越多,数据利用率(utility)越低.具体的度量有信息丢失[95]、重构数据与原始数据的相似度[96]等.3)搜索指标.指匿名化算法的每一步最大化信息可用性、最小化信息的失真.

3 大数据个人隐私保护的法律和行业规范

个人隐私保护是一个复杂的社会问题,除了需要先进的保护技术外,还需要结合国家制定的相关政策法规以及行业间形成的行业规范来保护好个人隐私,确保个人免遭人身安全的威胁以及财产损失.

3.1 隐私保护相关法律

到目前为止,我国还没有相关法律条例可以用来规范对个人信息数据的管理与使用.早在2002年12月23日九届人大常委会第31次会议首次审议的民法草案中已有明确界定,私人信息、私人活动和私人空间都属隐私范畴.在《未成年人保护法》规定:"不得披露未成年人的隐私",即隐私权是公民民事权利能力的内容之一.郭瑜2012年2月出版了《个人数据保护法研究》[97]一书,研究了中国应如何建立个人数据保护的法律制度,对中国应如何制定独立的、综合性的个人数据保护法提出了具体建议,并对个人数据使用者应如何正确使用个人数据提供了针对性的意见和建议.

在保护隐私问题上,中国与欧美的差距很大.美国1974年制定《联邦隐私权法》,1986年通过《电子通信隐私法》,1998年出台了第1部关于未成年上网隐私的法律《儿童网上隐私保护法》,还有《公民网络隐私权保护暂行条例》、《个人隐私权与国家信息基础设施》等法律作为业界自律的辅助手段.欧盟在1995年通过了《个人数据保护指令》;1997年通过了《电信事业个人数据处理及隐私保护指令》;之后又制定了《Internet上个人隐私权保护的一般原则》;《信息公路上个人数据收集、处理过程中个人权利保护指南》等相关法令.在欧洲联盟国家,如果数据当事人知道数据处理及其目的,一般只允许个人身份信息被处理,对敏感数据的处理设置了特殊的限制[98].由此可见,国外对数据隐私的保护给予了相当的重视,希望可以通过立法来打击数据隐私侵害行为.

Bansal等人[7]指出所有的声明和法律要求对个人信息必须做到:1)要公平、合法地获得;2)只用作最初规定的目的;3)适当地、相关地并且不过分地使用;4)信息是准确和最新的;5)对主体是可访问的;6)确保安全性;7)完成目的后毁掉.

3.2 个人隐私保护的法律和行业规范

在大数据时代到来之前,一些政策专家就看到了信息化给人们的隐私带来的威胁,社会也已经建立起了庞大的规则体系来保证个人的信息安全.然而在大数据时代,对原有规范进行修修补补已经不能满足个人隐私保护的需求,也不足以抑制大数据所带来的风险,因此,这些规则都不再适用,需要重新定义规则来满足现今的需求.数据提供者、企业以及政府需要提升对隐私保护的高度重视,个人隐私保护应做到数据使用者为其行为承担责任;建立完善的个人隐私保护的法律法规;加强行业的自律性建设及制定行业隐私法.

3.2.1 责任承担

用户如果想在互联网上使用某种服务,如购物、医疗、交友、建立个人主页、免费邮箱、下载资源等,服务商往往要求用户申请注册,并填写登录姓名、年龄、住址、身份证、手机号、工作单位等身份信息,还要同意他们所制定的一些条款,往往在这步操作时,用户不会详细阅读,而直接同意,这使得服务商以合法的形式获得了用户信息的支配和使用权.

在大数据时代需要设立一个不一样的隐私保护模式,该模式应该着重于数据使用者为其行为承担责任,而不是将重心放到收集数据之初取得个人同意上.将责任从用户转移到数据的使用者很有意义,

因为数据使用者比任何人都明白他们想要如何使用数据,他们是数据二次应用的最大受益者,所以应该让他们对自己的行为负责.服务商对收集的个人数据,有义务和责任保守个人的敏感信息,未经授权不得泄露.个人也应该意识到保护好自己的隐私,如果隐私保护机制存在缺陷,个人应该加以区分并拒绝提供敏感数据,尽量避免面临生命和财产威胁的隐患.因此,服务商需要使用正规的评测方法评测数据再利用的行为对个人所造成的影响,且这种影响不能对用户的生活构成威胁.

3.2.2 建立个人隐私数据保护法

大数据时代使用技术手段保护个人隐私远远不够,它不能代替法律体制,必须要建立个人隐私保护的法律法规和基本规则,加大对侵害个人隐私行为的打击力度.

2006年3月8日,民建中央企业委员会在全国政协会议期间向大会提交了《个人信息数据保护法的提案》①,要求通过制定法规对公民个人信息数据的采集、使用、营销等方面进行明确限制,并对触犯法规的行为予以处罚,从而更完善地保护公民权利与安全,保障社会稳定与国家安全,增强经济发展.《个人信息数据保护法》②从数据获得的限制、数据使用的限制、数据营销的限制以及刑事处罚4个方面进行提议.虽然该提议在当时有一定的意义,但是对于进入大数据时代的今天,这些提议已经明显不能满足个人隐私数据保护的需求.因此,应该根据大数据的特点以及个人隐私数据的特征建立通用的大数据《个人隐私数据保护法》.

法律建立的目的是维护大数据时代个人隐私保护的权利,明确大数据时代个人隐私数据保护的范围,如第4节1)中的分类方法;法律管辖的对象是用户及企业或特定的组织;法律的监督机构是建议设立或委托专门的行业协会和行业自律组织辅助相关政府部门监管该项法律的实施,如中国互联网协会.

法律建立的视角是个人数据的收集、使用、发布、共享以及刑事处罚.1)关于数据收集的规定:任何企业或组织不能为某种特定目的以欺骗的手段收集个人的信息,对收集到的用户信息,要保证在传输过程中不会被窃听;不能试图获得某些特定用户群的更详细信息,而对他们进行跟踪;在用户并不知情的情况下,企业或组织收集到个人信息时,不能滥用或者卖于他人.2)关于数据使用的规定:对个人隐私数据进行二次使用时要保证不能丢失、泄露或者滥用个人的隐私信息;数据使用中应该建立严格的等级访问控制策略,保证敏感数据的安全.3)关于数据发布的规定:发布出来的数据信息既有利于数据挖掘研究又能保护到个人的隐私信息;对发布的数据要有非常清晰地权限界定,不能造成个人隐私的泄露.4)关于数据共享的规定:在数据共享的过程中,数据共享的双方需签订一份有法律效力的合同或者协议,能保证用户的数据不被泄露,一旦引起用户隐私数据的泄露,将追究所有参与方的连带刑事责任.5)刑事处罚:对违反上述条款的企业或组织,依据对个人生活或财产造成后果的严重程度,予严厉的刑事处罚.

3.2.3 个人隐私保护的行业规范

客户是企业利益的源泉,企业在遵守《个人隐私数据保护法》的同时,也应该根据企业的应用需求遵守相关的行业规范,避免损失潜在的利益,吸引更多的客户.行业规范应包括以下4个方面.

1) 企业实施隐私保护机制的数据访问系统应该定义3个标准[1]:①灵活性.不同的人有各自的隐私保护需求,因此要为用户提供一个灵活的机制,能根据他们的需求来设置保护策略.②数据质量.在保护用户隐私的同时应保证数据的质量.③简单.政策的建立应该简单并且容易实施.

2) 遵守行业隐私法.一些特殊的行业会涉及到更复杂的隐私数据管理,因此,要制定更精细的行业隐私法来更好地保护个人隐私数据.在美国对特定类型的记录有各自的行业法,如信用报告、视频租用记录以及敏感信息类如健康信息等[99].

3) 数据访问权限的传递控制.数据提供者应该明确数据使用者访问数据的目的、条件、保持时间以及责任.数据提供者也应该注意传递数据的隐私等级,并确保传输的安全,使用内容加密和辅助措施相结合.

4) 建立企业与用户间的信任.Patrick等人[100]强调在人们对系统的接受上,一个重要的因素是人们对系统的信任问题.因此,为了减少用户对自身隐私的担忧,企业应尽量建立有效的个人隐私数据保护机制.用户信任企业就更少地担心他们的隐私被

① 民建中央企业委员会呼吁立法加强保护个人信息数据安全(http://tech.sina.com.cn/i/2006-03-08/1749861507.shtml 2006,3,8)

② 关于出台《个人信息数据保护法》的提案(http://tech.sina.com.cn/i/2006-03-08/1751861508.shtml 2006,3,8)

泄露,更愿意提供个人信息.当企业和用户之间建立起相互的信任关系时,企业便形成了良好的发展环境.

3.3 小 结

个人隐私保护范围的动态性使得企业开发的应用以及创新技术具有特定性,由此,不能建立覆盖企业所有方面的《个人隐私数据保护法》,只能建立各企业通用的《个人隐私数据保护法》加之个人隐私保护的行业规范来达到个人隐私保护的目的.虽然一些国家制定了隐私保护法,但是没有足够的监督和实施机制,法律并没有起到隐私保护的效力;在其他一些国家,法律的制定和实施跟不上技术的发展,在个人隐私保护上出现了严重的脱节.因此,法律和行业规范的制定与实施应该和技术保持同步,相互补充,企业对个人隐私数据的存储、使用和发布也必须严格地按照法律和行业规定执行,构建良好的大数据环境,这样才能更好地保护好个人隐私.

当然,在我国个人隐私保护法的制定还需要时间,希望能在大数据发展的促动下很快出台相关的政策法规,和个人隐私保护技术紧密结合,提高企业的管理、使用以及发布数据的规范性,使得网络用户的违法行为能在法律和技术的融合下有迹可追.

4 进一步的研究方向

根据个人隐私信息在数据层、应用层以及数据展示层实施的保护技术,本节提出了大数据时代个人隐私保护的进一步研究方向,并从不同方面解答了大数据个人隐私保护面临的部分挑战的解决方法.

1) 个人信息作为大数据的重要来源,具有数据量大的特征,应该对个人信息采用分类分级保护的技术方法.个人信息按照保护级别分为个人身份信息、敏感信息、准标识符信息、公开信息以及日志信息[37].个人身份信息指能够定位到个人的信息,如姓名、地址以及身份证号等;敏感信息指个人比较关心且需要额外保护的信息,如工资、健康、宗教或种族、经济财产状况、工作性质等;准标识符信息指几个属性在一起时,可以根据背景知识来识别出个人,如性别、年龄、邮政编码等;日志信息指用户使用互联网服务过程中产生的信息,如用户消费信息、访问信息(如IP地址)、位置信息及网络行为信息(如网页购物记录、搜索内容)等.除公开信息外,其他类型的信息均需纳入个人隐私的保护范围.

保护级别的划分主要考虑4个要素:①是否能依据信息直接识别出特定个人;②信息与个人生活的紧密程度;③是否能通过某些信息获得其他关联信息;④泄露了某些信息对个人产生多大的风险.综合考虑这4个要素保护级别由高到低表示为个人身份信息、准标识符信息、敏感信息、日志信息.

根据保护级别,企业可以在信息流转各个环节(如收集、存储、使用、发布或共享、删除)实施不同的技术保障.个人信息这样的分类分级管理不但有效地保护了个人隐私,而且提高了企业管理信息的效率,是大数据时代数据管理的发展的趋势,能应对大数据带来的更多挑战.

2) 在个人隐私保护中,并不是所有的隐私保护责任都针对企业,个人也应该有权利和责任保护自己的隐私.软件开发商应该寻找更好的方法帮助个人管理他们的敏感信息:开发可重用的个人隐私保护工具、服务或构建隐私敏感用户接口来管理他们的隐私.2002年万维网联盟(W3C)公布的一项隐私保护推荐标准P3P(platform for privacy preferences)[101],它的构想是:Web站点的隐私策略应该告之访问者该站点所收集的信息类型、信息将提供给哪些人、信息将被保留多少时间及其使用信息的方式,用户有权查看站点隐私报告,然后决定是否接受cookie或是否使用该网站.这种构建思想比较适合大数据的发展需求,用户对自身信息保护的决定给企业提供了更准确的个人信息分类范围以及更明确的技术实施目标,有利于企业对信息的管理以及为用户提供更好的服务.

3) 大量的新兴技术如基于位置的服务、RFID以及社交网络等,在被一个社区采用后,在短期内就会很快地流行起来.这样的技术在提高人们生活质量的同时,也产生了许多新的隐私问题.通常的状况是新兴技术带来隐私问题时,再根据暴露的隐私风险情况研究相应的保护技术,造成了隐私保护技术一直跟随新兴技术之后扮演补丁的角色,浪费了大量的人力、物力和财力,因此,应该考虑在新兴技术开发过程中融入隐私保护技术,将个人隐私保护作为新技术开发的一个需求.

1980年经济合作与发展组织(OECD)将保护个人隐私指导方针的公平信息实践(FIPs)成文后,在20世纪90年代基于FIPs首次提出了"从设计着手隐私(privacy by design, PbD)"的概念,反映了在线隐私受到威胁的增大.进入大数据时代后,隐私问

题更加凸显,2009 年 Ann Cavoukian 提出 PbD①,把隐私主动的嵌入特殊技术、商业操作、物理架构以及网络架构中,把隐私看作一个商业问题而不是依从性问题[102],达到隐私保护和承诺功能的双赢. 2011 年她又提出了从设计着手隐私(PbD)的 7 个基本原则[103]:①主动而不是被动,预防而不是补救;②把隐私看作默认需求;③把隐私嵌入到设计中;④全部功能——正和而不是零和;⑤端对端的安全机制——整个生命周期的保护;⑥可见性和透明性;⑦尊重用户的隐私. IBM 实体分析组首席科学家 Jonas 是一个理解大数据的真正远见者,他应用他现实世界的实际经验在软件设计和开发中推进创新,同时提供更好的隐私保护,提出使用高级数据的相关性,同时只使用不可逆加密哈希的突破性创新技术[104],这种隐私增强技术被称为"匿名识别(anonymous resolution)". Jonas 利用 PbD 的理念开发了意会系统,标志着从设计着手保护隐私(PbD)现在已经逐渐的从艺术走入实践应用.

从设计着手隐私把隐私考虑到大数据的整个生命周期中,无论对技术和实施都提出了严峻的挑战,但是,随着大数据的发展,这也是必然的趋势,只有这样才能够解决大数据自身特征所产生的一系列问题,如数据的安全存储、数据的合理使用、数据的发布以及动态数据的保护处理等.

4) 企业或组织最关心的问题是数据或服务质量是否下降、有价值信息是否丢失、成本以及系统复杂性是否增加. 隐私保护技术只解决了和计算机技术相关的问题,除此之外还应该考虑用户的心理、所处的社会环境、当下的法律法规以及政策,因此,应该对个人隐私保护实施跨学科研究,从不同角度对隐私问题有一个更好的理解,促进个人隐私保护技术的成功开发.

总之,传统的一些隐私保护技术难以直接应用于大数据中,发展一套全新的大数据系统的个人隐私保护技术目前并不现实,仍需要时间. 当前,最切合实际的做法是研究新的符合大数据系统要求的个人隐私保护的同时,对于具体的应用,提取大数据中的个人敏感信息,结合现有的个人隐私保护技术,实施分类分级的保护策略.

5 结 论

大数据时代拉开了序幕,带来机遇的同时也带来了巨大的挑战,个人隐私保护就是大数据所面临的挑战之一. 本文首先给出了个人隐私的基本概念以及个人隐私保护所面临的挑战和研究问题,然后从数据层、应用层以及数据展示层叙述了个人隐私的相关研究技术,侧重叙述了大数据时代主要使用的数据加密技术、匿名技术、访问控制以及数据发布技术. 由于大数据所具有的特点,以前的个人隐私保护方法已经不再适合,需要站在大数据的角度重新考虑. 文中从责任承担、建立通用的《个人隐私数据保护法》、以及个人隐私的行业规范作为出发点叙述了大数据时代个人隐私保护所要遵守的法律和行业规范的几个重要方面. 最后提出了大数据时代个人隐私保护的 4 个研究方向.

总的来说,大数据的个人隐私保护还处于起步阶段,尽管隐私保护对用户来说是一个重要的问题,但是企业不愿为了实施隐私保护,而不能充分利用用户信息或者为用户提供更好的服务,以至于限制企业的发展或在市场上的竞争力. 根据本文对个人隐私保护问题的分析,期望将来有一个完整和可理解的安全解决方案来满足个人隐私保护的需求. 对于广大的用户通过实行全民教育与技术防范同步的方式,提高人们对个人信息的自我保护意识. 当然,处于信息化时代,只要我们使用网络,完全保护个人隐私是不现实的,同时,用户的数据足迹遍布互联网,保证所有企业的发布一致信息也是很困难的事情. 因此,应该把立法以及行业规范融入技术实施和企业行为过程中,并保持它们的同步来获得最大化数据的使用效用和最小化隐私的泄露,以满足当下需求,并解决面临的更多挑战. 文中从技术和立法以及行业规范的不同角度回答了大数据时代保护个人隐私所面临的一些挑战的解决办法,希望能给后续的研究提供一些参考.

① http://www.privacybydesign.ca/

参 考 文 献

[1] Li Guojie, Cheng Xueqi. Research status and scientific thinking of big data [J]. 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——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(6): 647-657

[2] Chen M Y, Yang C C, Hwang M S. Privacy protection data access control [J]. International Journal of Network Security, 2013, 15(6): 391-399

[3] Warren S D, Brandeis L D. The right to privacy [J]. Harvard Law Review, 1890, 4(5): 193-220

[4] Smith J, Dinev T, Xu H. Information privacy research: An interdisciplinary review [J]. MIS Quarterly, 2011, 35(4): 989-1016

[5] Bennett C J. The political economy of privacy: A review of the literature[R]. Hackensack, NJ: Center for Social and Legal Research, 1995

[6] Westin A F. Privacy and Freedom [M]. New York: Atheneum, 1968

[7] Bansal G, Zahedi F, Gefen D. The moderating influence of privacy concern on the efficacy of privacy assurance mechanisms for building trust: A multiple-context investigation [C] //Proc of the Int Conf on Information System (ICIS 2008). Australian: AIS, 2008: 14-17

[8] Banisar D, Davies S. Global trends in privacy protection: An international survey of privacy, data protection, and surveillance laws and developments [J]. Journal of Computer & Information Law, 1999, 18(1): 3-111

[9] Soltani A, Canty S, Mayo Q, et al. Flash cookies and privacy [C] //Proc of the AAAI Spring Symp: Intelligent Information Privacy Management 2010. Menlo Park, CA: AAAI, 2010: 1-8

[10] Ayenson M, Wambach D, Soltani A, et al. Flash cookies and privacy II: Now with HTML5 and etag respawning [OL]. (2011-07-29) [2013-02-13]. http://ssrn.com/abstract=1898390

[11] Foresti G L, Mahonen C, Regazzoni C S. Multimedia Video-Based Surveillance Systems: Requirements, Issues, and Solutions [M]. Berlin: Springer, 2000

[12] Maheu M, Whitten P, Allen A. E-Health, Telehealth, and Telemedicine: A Guide to Startup and Success [M]. San Francisco: John Wiley & Sons, 2001

[13] Beckwith R. Designing for ubiquity: The perception of privacy [J]. IEEE Pervasive Computing, 2003, 2(2): 40-46

[14] Iachello G, Hong J. End-user privacy in human-computer interaction [J]. Foundations and Trends in Human Computer Interaction, 2007, 1(1): 1-137

[15] Smith H J, Milberg S J, Burke S J. Information privacy: measuring individuals' concerns about organizational practices [J]. MIS Quarterly, 1996, 20(2): 167-196

[16] Xu H, Teo H H, Tan B C, et al. The role of push-pull technology in privacy calculus: The case of location-based services [J]. Journal of Management Information Systems, 2009, 26(3): 135-174

[17] Moon Y. Intimate exchanges: Using computers to elicit self-disclosure from consumers [J]. Journal of Consumer Research, 2000, 26(4): 323-339

[18] Interactive H. IBM multi-national consumer privacy survey [R]. New York: IBM, 1999

[19] Sheehan K B, Hoy M G. Using e-mail to survey Internet users in the united states: Methodology and assessment [J]. Journal of Computer Mediated Communication, 1999, 4(3): 1-9

[20] Culnan M J. Consumer awareness of name removal procedures: implications for direct marketing [J]. Journal of Direct Marketing, 1995, 9(2): 10-19

[21] Bowie N E, Jamal K. Privacy rights on the internet: Self-regulation or government regulation? [J]. Business Ethics Quarterly, 2006, 16(3): 323-342

[22] Belanger F, Hiller J S, Smith W J. Trustworthiness in electronic commerce: The role of privacy, security, and site attributes [J]. The Journal of Strategic Information Systems, 2002, 11(3): 245-270

[23] Culnan M J, Williams C C. How ethics can enhance organizational privacy: Lessons from the choicepoint and TJX data breaches [J]. MIS Quarterly, 2009, 33(4): 673-687

[24] Ackerman M S. Privacy in pervasive environments: Next generation labeling protocols [J]. Personal and Ubiquitous Computing, 2004, 8(6): 430-439

[25] Wang Yuanzhuo, Jin Xiaolong, Cheng Xueqi. Network big data: Present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138 (in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据:现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138)

[26] Meng Xiaofeng, Ci Xiang. Big data management: Concepts, techniques and challenges [J]. 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[J]. 计算机研究与发展, 2013, 50(1): 146-169)

[27] Tan Xiongpai, Wang Huiju, Du Xiaoyong, et al. Big data analysis—Competition and symbiosis of RDBMS and MapReduce [J]. Journal of Software, 2012, 23(1): 32-45 (in Chinese)
(覃雄派, 王会举, 杜小勇, 等. 大数据分析——RDBMS与MapReduce的竞争与共生[J]. 软件学报, 2012, 23(1): 32-45)

[28] Bertino E, Sandhu R. Database security-concepts, approaches, and challenges [J]. IEEE Trans on Dependable and Secure Computing, 2005, 2(1): 2-19

[29] Fernandez E B, Summers R C, Wood C. Database Security and Integrity [M]. Boston, MA: Addison-Wesley Longman Publishing, 1981

[30] Shmueli E, Vaisenberg R, Elovici Y, et al. Database encryption: An overview of contemporary challenges and design considerations [J]. ACM SIGMOD Record, 2010, 38(3): 29-34

[31] Jansen W, Grance T. Guidelines on security and privacy in public cloud computing [OL]. 2011 [2014-05-05]. http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf

[32] Chen D, Zhao H. Data security and privacy protection issues in cloud computing [C] //Proc of the 7th Int Conf on Computer Science and Electronics Engineering (ICCSEE). Piscataway, NJ: IEEE, 2012: 647-651

[33] Roy I, Setty S T, Kilzer A, et al. Airavat: Security and privacy for MapReduce [C] //Proc of the the 7th USENIX Symp on Network Systems Design and Implementation (NSDI). Berkeley, CA: USENIX Association, 2010: 297-312

[34] Mowbray M, Pearson S, Shen Y. Enhancing privacy in cloud computing via policy-based obfuscation [J]. The Journal of Supercomputing, 2010, 61(2): 267-291

[35] Zhang G, Yang Y, Chen J. A historical probability based noise generation strategy for privacy protection in cloud computing [J]. Journal of Computer and System Sciences, 2012, 78(5): 1374-1381

[36] Wang C, Wang Q, Ren K, et al. Privacy-preserving public auditing for data storage security in cloud computing [C] // Proc of the 29th INFOCOM 2010. Piscataway, NJ: IEEE, 2010: 1-9

[37] Pearson S. Taking account of privacy when designing cloud computing services [C] //Proc of the 31st ICSE Workshop on Software Engineering Challenges of Cloud Computing. Piscataway, NJ: IEEE, 2009: 44-52

[38] Lewis K, Kaufman J, Christakis N. The taste for privacy: An analysis of college student privacy settings in an online social network [J]. Journal of Computer Mediated Communication, 2008, 14(1): 79-100

[39] Toch E, Wang Y, Cranor L F. Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems [J]. User Modeling and User-Adapted Interaction, 2012, 22(1): 203-220

[40] Krishnamurthy B, Wills C E. Characterizing privacy in online social networks [C] //Proc of the 1st Workshop on Online Social Networks. New York: ACM, 2008: 37-42

[41] Jahid S, Nilizadeh S, Mittal P, et al. DECENT: A decentralized architecture for enforcing privacy in online social networks [C] //Proc of the Pervasive Computing and Communications Workshops (PERCOM Workshops). Piscataway, NJ: IEEE, 2012: 326-332

[42] Luo W, Xie Q, Hengartner U. Facecloak: An architecture for user privacy on social networking sites [C] //Proc of the Computational Science and Engineering 2009 (CSE'09). Piscataway, NJ: IEEE, 2009: 26-33

[43] Guha S, Tang K, Francis P. NOYB: Privacy in online social networks [C] //Proc of the 1st Workshop on Online Social Networks. New York: ACM, 2008: 49-54

[44] Singh K, Bhola S, Lee W. xBook: Redesigning privacy control in social networking platforms [C] //Proc of the USENIX Security Symp 2009. Berkeley, CA: USENIX Association, 2009: 249-266

[45] Baden R, Bender A, Spring N, et al. Persona: An online social network with user-defined privacy [C] //Proc of the ACM SIGCOMM 2009. New York: ACM, 2009: 135-146

[46] De Cristofaro E, Soriente C, Tsudik G, et al. Hummingbird: Privacy at the time of twitter [C] //Proc of the 33rd IEEE Symp on Security and Privacy (SP) 2012. Piscataway NJ: IEEE, 2012: 285-299

[47] Cutillo L A, Molva R, Strufe T. Safebook: A privacy-preserving online social network leveraging on real-life trust [J]. IEEE Communications Magazine, 2009, 47(12): 94-101

[48] Buchegger S, Schioberg D, Vu L H, et al. PeerSoN: P2P social networking: early experiences and insights [C] //Proc of the 2nd ACM EuroSys Workshop on Social Network Systems. New York: ACM, 2009: 46-52

[49] Aiello L M, Ruffo G. LotusNet: Tunable privacy for distributed online social network services [J]. Computer Communications, 2012, 35(1): 75-88

[50] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data [G] //Privacy, Security, and Trust in KDD. Berlin: Springer, 2008: 153-171

[51] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks [C] //Proc of the 24th Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2008: 506-515

[52] Liu K, Terzi E. Towards identity anonymization on graphs [C] //Proc of the ACM SIGMOD Int Conf on Management of Data 2008. New York: ACM, 2008: 93-106

[53] Tsai J Y, Kelley P G, Cranor L F, et al. Location-sharing technologies: Privacy risks and controls [J]. A Journal of Law and Policy for the Information Society (ISJLP), 2010, 6(1): 119-144

[54] Pan X, Xu J, Meng X. Protecting location privacy against location-dependent attacks in mobile services [J]. IEEE Trans on Knowledge and Data Engineering (TKDE), 2012, 24(8): 1506-1519

[55] Ardagna C A, Cremonini M, De Capitani Di Vimercati S, et al. An obfuscation-based approach for protecting location privacy [J]. IEEE Trans on Dependable and Secure Computing (TDSC), 2011, 8(1): 13-27

[56] Zhu Z, Cao G. Applaus: A privacy-preserving location proof updating system for location-based services [C] //Proc of the INFOCOM 2011. Piscataway, NJ: IEEE, 2011: 1889-1897

[57] Juels A. RFID security and privacy: A research survey [J]. IEEE Journal on Selected Areas in Communications, 2006, 24(2): 381-394

[58] Fung B C, Wang K, Yu P S. Anonymizing classification data for privacy preservation [J]. IEEE Trans on Knowledge and Data Engineering (TKDE), 2007, 19(5): 711-725

[59] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571-588

[60] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation [C] //Proc of VLDB 2006. New York: ACM, 2006: 139-150

[61] Zhang Q, Koudas N, Srivastava D, et al. Aggregate query answering on anonymized tables [C] //Proc of the 23rd Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2007: 116-125

[62] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, SRI-CSL-98-04 [R]. Palo Alto, California: SRI Computer Science Laboratory, 1998

[63] Wang K, Fung B. Anonymizing sequential releases [C] // Proc of the 12th ACM SIGKDD 2006. New York: ACM, 2006: 414-423

[64] Nergiz M E, Cliftion C, Nergiz A E. Multirelational k-anonymity [J]. IEEE Trans on Knowledge and Data Engineering(TKDE), 2009, 21(8): 1104-1117

[65] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond k-anonymity [J]. ACM Trans on Knowledge Discovery from Data (TKDD), 2007, 1(1): 1-52

[66] Wong R C W, Li J, Fu A W C, et al. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing [C] //Proc of the 12th ACM SIGKDD 2006. New York: ACM, 2006: 754-759

[67] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity [C] //Proc of the 23rd Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2007: 106-115

[68] Nergiz M E, Atzori M, Cliftion C. Hiding the presence of individuals from shared databases [C] //Proc of ACM SIGMOD 2007. New York: ACM, 2007: 665-676

[69] Liu Yubao, Huang Zhilan, Fu Weici, et al. A data privacy preservation method based on lossy decomposition [J]. Journal of Computer Research and Development, 2009, 46(7): 1217-1225 (in Chinese)
(刘玉葆, 黄志兰, 傅慰慈, 等. 基于有损分解的数据隐私保护方法[J]. 计算机研究与发展, 2009, 46(7): 1217-1225)

[70] Xu Yong, Qin Xiaolin, Yang Yitao, et al. A QI weight-aware approach to privacy preserving publishing data set [J]. Journal of Computer Research and Development, 2012, 49(5): 913-924 (in Chinese)
(徐勇, 秦小麟, 杨一涛, 等. 一种考虑属性权重的隐私保护数据发布方法[J]. 计算机研究与发展, 2012, 49(5): 913-924)

[71] Zhou Shuigeng, Li Feng, Tao Yufei, et al. Privacy Preservation in database applications: A survey [J]. Chinese Journal of Computers, 2009, 32(5): 847-861 (in Chinese)
(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[J]. 计算机学报, 2009, 32(5): 847-861)

[72] Bonchi F, Lakshmanan L V, Wang H W. Trajectory anonymity in publishing personal mobility data [J]. ACM SIGKDD Explorations Newsletter, 2011, 13(1): 30-42

[73] Cicek A E, Nergiz M E, Saygin Y. Ensuring location diversity in privacy-preserving spatio-temporal data publishing [J]. The VLDB Journal, 2013, 11(1): 1-17

[74] Tai C H, Yu P S, Yang D N, et al. Privacy-preserving social network publication against friendship attacks [C] // Proc of the 17th ACM SIGKDD 2011. New York: ACM, 2011: 1262-1270

[75] Zhou B, Pei J, Luk W. A brief survey on anonymization techniques for privacy preserving publishing of social network data [J]. ACM SIGKDD Explorations Newsletter, 2008, 10(2): 12-22

[76] Poblete B, Spiliopoulou M, Baeza Y R. Website privacy preservation for query log publishing [G] //Privacy, Security, and Trust in KDD. Berlin: Springer. 2008: 80-96

[77] Agrawal R, Srikant R. Privacy-preserving data mining [J]. ACM Sigmod Record, 2000, 29(2): 439-450

[78] Verykios V S, Bertino E, Fovino I N, et al. State-of-the-art in privacy preserving data mining [J]. ACM Sigmod Record, 2004, 33(1): 50-57

[79] Ilavarasi A, Poorani S. A survey on privacy preserving data mining techniques [J]. Int Journal of Computer Science and Business Informatics, 2013, 7(1): 1-12

[80] Xiao X, Tao Y. M-invariance: Towards privacy preserving re-publication of dynamic datasets [C] //Proc of ACM SIGMOD 2007. New York: ACM, 2007: 689-700

[81] Aggarwal C C, Philip S Y. On privacy-preservation of text and sparse binary data with sketches [C] //Proc of the 7th Int Conf on Data Minin (SDM). New York: SIAM, 2007: 57-67

[82] Aggarwal C C, Philip S Y. On anonymization of string data [C] //Proc of the 7th Int Conf on Data Minin (SDM). New York: SIAM 2007: 419-424

[83] Dwork C. Differential privacy [G] //Automata, Languages and Programming. Berlin: Springer. 2006: 1-12

[84] Dwork C. The promise of differential privacy: A tutorial on algorithmic techniques [C] //Proc of the Foundations of Computer Science (FOCS) 2011. Piscataway, NJ: IEEE, 2011: 1-2

[85] Dwork C. A firm foundation for private data analysis [J]. Communications of the ACM, 2011, 54(1): 86-95

[86] Dwork C, Mcsherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [M]. Theory of Cryptography. Berlin: springer. 2006: 265-284

[87] Mcsherry F, Talwar K. Mechanism design via differential privacy [C] //Proc of the Foundations of Computer Science (FOCS) 2007. Piscataway, NJ: IEEE, 2007: 94-103

[88] Dwork C, Nissim K. Privacy-preserving datamining on vertically partitioned databases [C] //Proc of the Advances in Cryptology (CRYPTO 2004). Berlin: Springer, 2004: 528-544

[89] Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication [C] //Proc of the 28th Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2012: 797-822

[90] Cormode G, Procopiuc C, Srivastava D, et al. Differentially private summaries for sparse data [C] //Proc of the 15th Int Conf on Database Theory. New York: ACM, 2012: 299-311

[91] Ding B, Winslett M, Han J, et al. Differentially private data cubes: optimizing noise sources and consistency [C] // Proc of ACM SIGMOD 2011. New York: ACM, 2011: 217-228

[92] Chen R, Mohammed N, Fung B C, et al. Publishing set-valued data via differential privacy [J]. Proc of the VLDB Endowment, 2011, 4(11): 1087-1098

[93] Liu Y, Gummadi K P, Krishnamurthy B, et al. Analyzing facebook privacy settings: User expectations vs reality [C] //Proc of the ACM SIGCOMM 2011. New York: ACM, 2011: 61-70

[94] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms [J]. SIAM Journal on Computing, 2012, 41(6): 1673-1693

[95] Xu J, Wang W, Pei J, et al. Utility-based anonymization using local recoding [C] //Proc of the 12th ACM SIGKDD 2006. New York: ACM, 2006: 785-790

[96] Kifer D, Gehrke J. Injecting utility into anonymized datasets [C] //Proc of ACM SIGMOD 2006. New York: ACM, 2006: 217-228

[97] Guo Yu. Legal Protection of Rersonal Data [M]. Beijing: Peking University Press, 2012 (in Chinese)
(郭瑜. 个人数据保护法研究[M]. 北京: 北京大学出版社, 2012)

[98] Economico O D C Y D. Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [M]. French: Organisation for Economic Cooperation and Development, 1981

[99] Smith H J. Information privacy and its management [J]. MIS Quarterly Executive, 2004, 3(4): 291-313

[100] Patrick A, Marsh S, Briggs P. Designing systems that people will trust [J]. Security and Usability, 2005, 1(1): 75-99

[101] World Wide Web Consortium. Platform for privacy preferences (P3P)[OL]. 1994 [2007-11-20]. http://www.w3.org/P3P/

[102] Cavoukian A. Privacy by Design—Take the Challenge [M]. Toronto, Canada: Information and Privacy Commissioner of Ontario, 2009

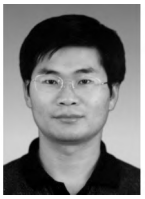
[103] Cavoukian A. Privacy by design: The 7 foundational principles [OL]. [2009-08-20]. http://www.privacybydesign.ca/index.php/paper/privacy-by-design-the-7-foundational-principles-multi-language/

[104] Cavoukian A, Jonas J. Privacy by Design in the Age of Big Data [M]. Toronto, Canada: Information and Privacy Commissioner of Ontario, 2012

Liu Yahui, born in 1979. PhD candidate, lecturer. Her research interests include distributed database, graph, and big data.

Zhang Tieying, born in 1982. PhD, assistant professor. His research interests include computer networks, distributed computing, peer-to-peer system, multimedia networking, and network security.

Jin Xiaolong, born in 1976. Associate professor, PhD supervisor. His research interests include social computing, network performance modelling and evaluation.



Cheng Xueqi, born in 1971. Professor, PhD supervisor. His research interests include information retrieval, social computing, and distributed computing.