计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140073
52(2):265-281,2015

大数据隐私管理

孟小峰[1]　　张啸剑[2]
[1](中国人民大学信息学院　北京　100872)
[2](河南财经政法大学计算机与信息工程学院　郑州　450002)
(xfmeng@ruc.edu.cn)

Big Data Privacy Management

Meng Xiaofeng[1] and Zhang Xiaojian[2]
[1](*Information School*, *Renmin University of China*, *Beijing* 100872)
[2](*School of Computer & Information Engineering*, *Henan University of Economics and Law*, *Zhengzhou* 450002)

Abstract With the high-speed development of information and network, big data has become a hot topic in both the academic and industrial research, which is regarded as a new revolution in the field of information technology. However, it brings about not only significant economic and social benefits, but also great risks and challenges on individuals' privacy protection and data security. Currently, privacy related with big data has been considered as one of the greatest problems in many applications. This paper analyzes and summarizes the categories generated by big data, the privacy properties and types in terms of difference reasons, the challenges in technologies and laws and regulations on managing privacy, and describes the differences of the current technologies which handle those challenges. Finally, this paper provides an active framework for managing big data privacy on the actual private problems. Under this framework, we illustrate some privacy-preserving technology challenges on big data.

Key words big data; privacy risk; privacy active management; privacy attack; privacy leakage

摘　要　信息化和网络化的高速发展使得大数据成为当前学术界和工业界的研究热点,是IT业正在发生的深刻技术变革.但它在提高经济和社会效益的同时,也为个人和团体的隐私保护以及数据安全带来极大风险与挑战.当前,隐私成为大数据应用领域亟待突破的重要问题,其紧迫性已不容忽视.描述了大数据的分类、隐私特征与隐私类别,分析了大数据管理中存在的隐私风险和隐私管理关键技术;提出大数据隐私主动式管理建议框架以及该框架下关于隐私管理技术的主要研究内容,并指出相应的技术挑战.

关键词　大数据;隐私风险;隐私主动式管理;隐私攻击;隐私泄露

中图法分类号　TP392

大数据正在改变着世界,它是IT业正在发生的深刻技术变革.大数据中那些巨大的数字痕迹已经成为当前工业界与学术界的研究热点.然而,大数据技术发展无法避开的事实是隐私问题.实际上,隐私与新技术变革之间的冲突贯穿着整个信息技术的发展史.19世纪以报纸为代表的新型媒体是最早

收稿日期:2014-09-23;**修回日期**:2014-10-20
基金项目:国家自然科学基金项目(61379050,91224008);国家"八六三"高技术研究发展计划基金项目(2013AA013204);高等学校博士学科点专项科研基金项目(20130004130001)

披露个人隐私的信息技术，这类隐私泄露通常利用法律进行保护；20 世纪 60 年代，信息技术的革新使得大型计算机开始挑战人们对隐私的传统观念，针对这类隐私威胁常采用密码技术进行保护；21 世纪前 10 年，网络技术和社交媒体的蓬勃发展使得个人隐私无处可藏，这类隐私泄露通常利用匿名化技术(anonymization)和模糊化技术(de-identification)进行保护. 过去这些隐私与新技术之间的冲突往往集中于单一的小数据(small data). 模糊化、匿名化、加密、密码学等是防止小数据上隐私泄露的常用技术. 然而，这些技术是基于某些特定的攻击假设和背景知识才能够生效，是对隐私的被动保护(passive protection). 例如，利用背景知识攻击可以推理出 k-匿名[1]之后的敏感数据.

大数据的大规模性、高速性和多样性等特征，使得它不同于小数据. 上述提到的针对小数据的隐私保护方法在大数据上存在着很大的局限性：大数据的多样性带来的多源数据融合使得传统的匿名化和模糊化技术几乎无法生效；大数据的大规模性与高速性带来的实时性分析使得传统的加密和密码学技术遇到了极大的瓶颈. 此外，大规模性数据采集技术、新型存储技术以及高级分析技术使得大数据的隐私保护面临更大的挑战.

1) 在大数据环境下，移动轨迹通常蕴含着丰富的个人敏感信息，例如家庭住址与行为模式等. 文献[2]指出在 1 500 000 条匿名后的个人移动轨迹数据中，在不依赖外部背景知识的前提下，随机给出 2 个时空数据点，可以甄别出 50% 的个人敏感移动轨迹；随机给出 4 个时空点，被甄别出的敏感轨迹数据可达到 95%.

2) 在基因大数据中，基因序列隐含着个人疾病情况. 文献[3]指出虽然对基因工程中 100 000 个自愿者的邮政编码、出生日期与性别进行了匿名化操作，然而通过把基因序列数据与公共选民信息融合后，却重新甄别出 84%～87% 自愿者的身份.

由上述 2 个例子可以看出，大数据独有的隐私问题使得那些传统的被动式保护技术束手无策. 因此，急需新型的隐私保护技术的出现. 基于此，本文提出了基于主动式保护(active protection)思想的隐私管理框架，对公开的大数据进行隐私管理. 以前针对小数据的被动式保护方式仅仅考虑了当前攻击模式下的隐私保护效果，而并不关心将来的某个攻击模式下的隐私泄露. 而主动式保护方式是考虑数据的整个生命周期内的隐私泄露情况，该方式对隐私有着绝对的保护力，并主动参与到整个大数据隐私处理流程中去. 然而，由被动式的隐私保护技术到主动式的隐私管理技术，将是一次巨大的技术进步，同时也面临着巨大的技术挑战.

目前，大数据贯穿 7 大行业：教育、交通、商业、电力、石油天然气、卫生保健以及金融业. 根据麦肯锡公司分析，如果这 7 大行业之间公开数据，将带来 3 万亿美元的经济利益. 然而，公开数据带来巨大经济利益的同时，也给个人和团体的隐私带来威胁.

1) 医学领域中基因研究的快速发展，使得全球超过百万人在不知情的情况下向研究人员公开了他们的 DNA 数据，这些研究可以解决心脏病、糖尿病的问题，却不可避免地会涉及到个人隐私问题. 例如，通过 DNA 序列分析，可以推理出某个人是否是癌症患者.

2) 在社会科学领域，通过分析社交媒体服务(例如 Facebook，Twitter)所产生的大数据，可以捕获社会人群的情感、话题、认知趋势以及发掘有共同兴趣的社区等. 然而这些分析可能泄露个人的敏感信息. 例如，通过分析基于位置的社交网络，可以泄露某个人的敏感位置等.

由此可见，阻碍大数据公开的主要因素是数据隐私问题. 实际上，现实中与个人和团体相关的数据确实处于风险之中. 2013 年 6 月发生的“棱镜门”事件提醒人们如果数据的隐私没有得到充分保护，将会带来非常严重的后果. 当前，很多研究机构同样认识到大数据的隐私问题，并积极关注讨论大数据隐私问题. 2014 年 3 月美国白宫科学与技术政策办公室联合麻省理工大学、纽约大学与加州伯克利大学举办了大数据隐私保护研讨会[4]，主要研讨了大数据带来的机遇和风险、当前隐私保护技术；2014 年 5 月美国白宫发布了《大数据与隐私保护：一种技术视角》白皮书[5]，主要探讨个人隐私存在的风险与保护技术；2014 年中国工业和信息化部电信研究院发布了《大数据白皮书》[6]，主要阐述我国大数据技术发展所面临的挑战.

因此，在大数据时代下，保护数据中隐私信息有着独特的意义，传统的隐私保护理论和技术已经无法涵盖大数据隐私的内涵，有必要对大数据隐私保护问题进行重新思考与定位. 本文在整理大数据研究现状的基础上，重点分析了大数据在收集、集成融合以及分析时存在的隐私泄露问题，详细分析了当前大数据保护关键技术，提出了大数据隐私管理框架，并讨论了该框架是如何主动发现隐私泄露隐患、

如何主动地进行隐私保护,本文在第4节对其进行了初步分析和探讨.

1 大数据类型、隐私特征与类别

1.1 大数据类型

大数据增长速度快,数据格式多样,数据源广泛.根据这些特征,大数据的类型可以分为2种:

1) 天生数字化数据(born digital data).这类数据自然产生出来就适合计算机的存储和处理系统.例如电子邮件与文本信息、GPS位置数据、关联电话呼叫的元数据、商业事务数据、移动用于连接网络的元数据、网页数据以及物联网(Internet of Things)数据等.天生数字化数据的隐私担忧来自于该类数据的过分收集(over-collection)和数据融合.过分收集往往与收集者的初衷相违背.例如,爬虫收集网页数据初衷可能是为了提升网络的访问速度,过分收集数据后进行分析,可以挖掘网络用户的行为模式进而泄露其隐私信息;"最亮手电筒应用(brightest flashlight free app)[5]"可以打开基于Android平台手机中所有可用的灯源.然而,美国联邦贸易委员会却揭露了该免费应用所蕴含的阴谋:该应用能够在用户不知情的情况下,在后台过分收集用户的位置信息,并且卖给第三方,这样用户的位置隐私就被出卖了.相对于过分收集,基于多个数字化数据源的融合所带来的隐私担忧更大.单一的数据源通常对实体简单描述,然而,通过新型数学计算方法(例如基于隐马尔可夫模型的贝叶斯分析方法[7])与模式识别技术对多个数据源进行融合之后,可以得到更加丰富的个人描述信息,进而识别出不同的实体,以至于泄露用户的隐私信息.

2) 天生模拟化数据(born analog data).这类数据是由物理世界特征演化而来的,通过碰撞传感器最终成为可以访问的数字化格式.例如,手机呼叫的音频与视频、个人健康数据(例如心跳、呼吸与步速等)、环境监测视频、超声波、医疗影像、化学与生物样本、合成孔径雷达、可佩戴设备的监控等模拟化数据.天生模拟化数据的隐私需求源自于产生该类数据的物理世界特征.例如,通过分辨率、对比度、测光精度3个参数可以提高视频监控的清晰度,能够清晰地识别几英里之外的门窗结构,然而,个人的活动也不可避免地被监视.手机用户通过GPS向基于位置的服务方(location-based service, LBS)发出请求,该用户的位置信息很有可能被非可信的LBS泄露.一旦模拟化数据转化为数字化,即可与现有数据进行融合,对实体进行识别.

1.2 大数据的隐私特征与类别

普遍的观点认为,隐私具有3种特征:隐私的主体是人、隐私的客体是个人事务与个人信息、隐私的内容是主体不愿意泄露的事实或者行为.由于大数据具有大规模性、多样性与高速性的独有特征,大数据隐私主体可能是人或者组织团体、客体可能是人或者团体的信息.此外,大数据隐私还具有边界难以鉴定的特征.

根据来源的不同,大数据的隐私类别大致分为以下3类:

1) 监视(surveillance)带来的隐私.这里的监视是指通过非法的手段跟踪、收集个人或者团体的敏感信息.例如,网站利用Cookie技术跟踪用户的搜索记录、利用视频监视系统窥视他人的行为等.这类隐私常利用问责系统或者法律手段来保护.

2) 披露(disclosure)带来的隐私.数据披露是指故意或无意中向不可信的第三方透露或遗失数据.该类隐私通常利用匿名化、差分隐私、加密、访问控制等技术来保护.

3) 歧视(discrimination)带来的隐私.这里的歧视是指由于大数据处理技术的不透明性,普通人无法感知和应用,会在有意或无意中产生歧视结果,进而泄露个人或者团体的隐私.该类隐私通常利用法律法规手段来保护.

此外,根据对象的不同,大数据隐私类别可以分为数据隐私(例如关系数据隐私、位置数据隐私等)、查询隐私(例如k近邻查询等)、发布隐私等.

2 大数据的隐私风险

2012年1月,奥巴马在消费者隐私条例草案发布会说"隐私从一开始一直是我们民主制度的心脏,而目前比以往任何时候更需要它,大数据时代更加如此"[8].先前有文献从信息安全的角度阐述大数据管理问题[9],实际上,隐私和安全存在一定的区别.

2.1 数据隐私与信息安全的区别

2.1.1 二者定义的区别

数据隐私是指个人、组织机构等实体不愿意被外部知道的信息.比如,个人的行为模式、位置信息、兴趣爱好、健康状况、公司的财务状况等,本文描述了与个人相关的所有隐私信息,如图1所示.数据隐私主要涉及数据的模糊性、隐私性、可用性.

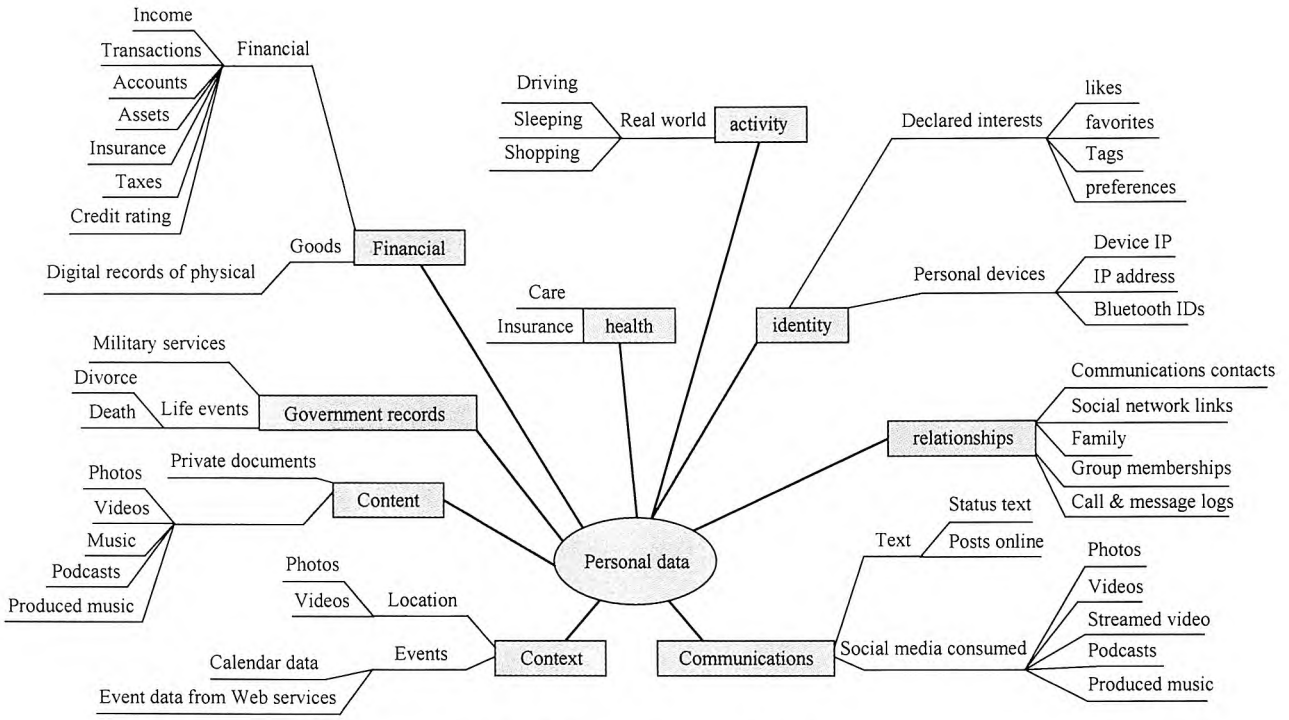

Fig. 1 Privacy related to individual.

图 1 与个人相关的隐私信息

信息安全是指信息及信息系统免受未经授权的访问. 未经授权的操作包括非法使用、披露、破坏、修改、记录及销毁等. 信息安全主要涉及数据的机密性、完整性、可用性.

2.1.2 二者实施技术的区别

信息安全的实施技术包括访问控制和密码学; 而数据隐私的实施技术包括模糊化、匿名化、差分隐私(differential privacy)以及加密等. 虽然信息安全技术能够保证基础设施、通信与访问过程数据的安全性, 但是数据的隐私还有可能被泄露. 例如, 一个被授权的恶意用户可以误用 Alice 的数据并与其他数据融合, 这些操作可能会泄露 Alice 的隐私. 虽然数据隐私和信息安全存在以上区别, 但是二者的最终目的是一致的, 都是为了数据能够被私密地、安全地访问和分析.

本文从数据隐私的角度来讨论大数据隐私管理问题, 而非关注信息安全.

2.2 大数据带来的隐私风险

文献[10]给出了大数据的处理框架, 该框架包括数据收集、数据集成与融合、数据分析以及数据解释 4 个部分. 其中, 数据收集包括公开数据(例如 data.gov 网站①)和私有数据的收集; 数据集成与融合主要处理数据之间的冗余、不一致、相互拷贝关系等问题; 数据分析的目的是从数字化数据与模拟化数据中抽取或者学习到有价值的模型和规则; 而数据解释主要是通过可视化、数据溯源等技术来展示大数据的分析结果. 然而, 在大数据的整个处理框架和生命周期中, 每个步骤均存在披露和破坏数据隐私的风险. 1) 数据收集步骤, 如果个人数据被不可信的第三方服务(untrusted third-party service)收集, 则个人隐私很有可能被泄露或者卖给恶意攻击者. 例如, 不可信的位置服务恶意收集用户的位置信息, 则用户的敏感位置可能会被披露; 2) 数据集成融合以及存储步骤中, 存在着不可信外包服务攻击、无加密索引、记录连接攻击等; 3) 数据分析过程中存在频繁模式支持度攻击、分类与聚类攻击、特征攻击等; 4) 数据解释过程中可能存在前景知识攻击(foreground knowledge attack)[11]、通过数据溯源图挖掘元数据之间的依赖关系等. 本文着重介绍数据收集、集成与融合以及分析这 3 个步骤中的隐私风险.

2.2.1 数据肆意收集带来的风险

在大数据环境中, 可以通过医疗就医记录、购物及服务记录、网站搜索记录、手机通话记录、手机位置轨迹记录等来获取用户的信息. 而收集这些用户

① 该网站 2009 年 3 月上线, 已拥有超过 37 万个数据集, 数据源来自 171 个机构.

个人信息时,通常是未经用户同意,或者用户很少有机会去思考、去认同自己的数据被用作干什么?是谁收集了自己的数据?是谁二次使用了自己的数据?如果自己的数据出现误用,将由谁负责?自己的数据是否在网上被恶意传播?自己的数据什么时候被销毁?2011 年 4 月,《纽约时报》报道,Apple 公司通过 iPhone 手机上的 iOS4 系统无线跟踪并收集用户的地理位置信息,而位置信息通常蕴含着用户的敏感信息,例如距离 Alice 最近的皮肤病医院.地理位置信息的跟踪与收集是在 iPhone 的后台运行,用户根本无法察觉.而位置信息一旦被泄露,通过位置的序列关系可以推理出用户的疾病情况、家庭住址、轨迹模式等私密信息.此外,Google 公司通过 Cookie 跟踪用户的搜索记录,进而披露用户的网上行为模式、政治倾向以及消费习惯等.Google 公司也得到了美国联邦贸易委员会给出的 2 250 万美元的判罚.

因此,通过上述的实例可知,在用户无"知情同意"权[①]的情况下,隐私风险非常巨大.而这类风险主要是缺乏规范与法律法规监管,在收集数据时,为了不危害用户的隐私,通常依靠收集者的自律和自觉遵守一些规范.而在商业化的应用场景中,用户有权利选择自己数据的用途,在收集个人数据之前必须得到用户的许可;用户有权知道自己的数据是否被共享、误用、恶意传播、销毁等.而这些权利的实施,需要政府出台或者加强相关的法律法规建设,为保护用户的个人隐私起到约束与监管作用.

2.2.2 集成融合带来的风险

集成和融合通常采用链接操作使多个异构数据源汇聚在一起,并且识别出相应的实体.小数据源通常能够反映出用户的某个活动,比如接受的医疗、购买的商品、搜索的网站、手机留下的位置特征、与社交网络互动信息、政治活动等.融合不同的小数据可以更好地服务于数据分析与管理.零售商通过集成线上、线下以及销售目录数据库,可以获得更多消费者的个人描述信息、预测消费者的购物偏好等;GPS 服务商通过集成路网不同路段上的传感器数据,可以得到更好的道路规划与交通路线.然而,多个数据源的集成与融合几乎能够推理出个人所有的敏感信息,无形中给个人隐私的保护带来严峻挑战.匿名和模糊化是集成中常用的隐私保护技术,该技术通常比较适用于小型且单一的数据源,保护的效果比较理想.然而,针对于复杂的大数据,即使利用匿名或者模糊化技术将个人敏感信息保护起来,但是当攻击者拥有其他公共的或者隐私的数据源时,可以利用链接攻击(link attack)对匿名之后的数据源进行攻击,极有可能重新识别(re-identify)出匿名后的个人敏感信息,这样造成个人隐私泄露.例如,美国在线公司(AOL)虽然删除了搜索用户的显性标示,用随机数代替名字和 ID 号,然而,纽约时报记者还是通过背景知识识别出 4417749 号是佐治亚州的一名寡妇;Netflix 公司所发布的 Netflix 大奖赛匿名数据,被攻击者通过集成方法甄别出一些用户的身份导致用户的隐私泄露,这一结果直接导致第二次 Netflix 大奖赛的取消[12].

以图 2 中的例子说明集成融合带来的隐私泄露.数据源 1 是满足 k-匿名的医疗发布数据,在属性 ZIP,Birth Date 与 Sex 上做了匿名化处理;数据源 2 是公开的选民注册数据,同样具有 ZIP,Birth Date 与 Sex 属性.攻击者通过集成数据源 2 与数据源 1,可以推理出数据源 1 中用户的身份,并披露其隐私信息,比如个人的政治倾向与医疗记录等.

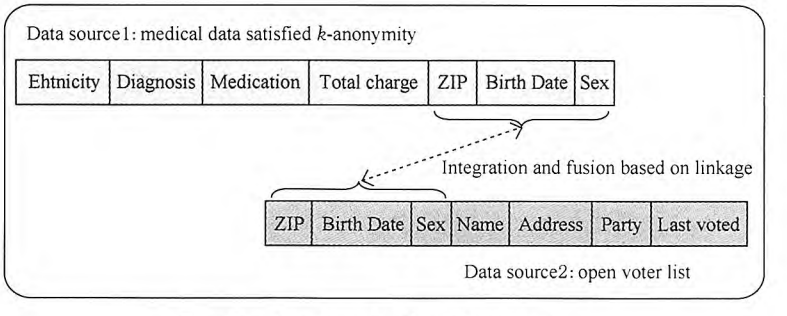

Fig. 2 An example of re-identification.

图 2 个人身份重新甄别例子

① 知情同意(informed consent)通常在医疗方面表示医生和患者之间的关系,是指患者有获知病情并对医生所采取的治疗方案决定取舍的权利.

2.2.3 数据分析带来的风险

目前,基于大数据的计算框架,其计算分析能力能够达到"大海捞针".数据科学家通过分析,可以挖掘出大数据中的异常点、频繁模式、分类模式、数据之间的相关性以及用户行为规律等信息.然而,大数据分析的最大障碍是数据隐私问题.在某种程度上,隐私不可怕,可怕的是用户的行为可以通过大数据分析被预测出来.例如,Facebook 就曾因跟踪用户的数据,并通过分析这些数据来评估 Facebook 的广告效果,而引发了隐私维权机构的质疑;Google 的 Analytics 是最受欢迎的分析工具,企业和政府通常利用该工具分析网站流量.然而,在用户使用该工具时并不能保证自己数据隐私的不被泄露.Analytics 不仅知道用户本身网站所有访客信息,也可以通过关联分析获悉其他网站中的访客信息;大数据下的个性化推荐系统是电子商务网站根据用户的兴趣特点和购买行为,向用户推荐感兴趣的信息和商品.然而,用户的商品购买信息以及行为模式很有可能被商务网站挖掘出来,进而导致隐私信息泄露.

大数据分析带来的隐私问题主要源自于 3 个方面:新型计算框架、高性能算法、更加复杂的分析模型.在大数据环境下,以 Hadoop+MapReduce,Storm,Dremel 以及 R+Hadoop 为代表的强大计算框架,能够以批处理或者流式处理的方式并行处理大规模数据;以前传统的数据挖掘、机器学习与 OLAP 算法不再适应这些新型计算框架,需要重新改写并提高其分析性能.比如,基于 MapReduce 的快速聚类方法 k-center[13] 与 k-median[13]、多维聚类方法 BoW[14]、关联聚类方法 Co-Cluster[15] 等.这些高性能算法不但能够深层分析大数据中那些细小的、彼此之间毫无关联的数据碎片,同时也为恶意分析者提供了确凿的攻击背景知识,进而通过分析泄露大数据中的隐私信息;先前单一的分类、回归分析等模型无法应对大数据的大规模性和多样性,进而出现了更为复杂高效的分析模型,比如基于随机优化(stochastic optimization)的分类方法 SDCA[16] 与回归分析方法 SAG[17] 等.

大数据分析带来的直接风险是泄露数据的隐私信息,间接风险是导致隐私保护方法失效、分析结果的不可擦除性等.因此,需要更具有鲁棒性、可扩展性以及隐私性的数据挖掘和机器学习方法的出现.

3 大数据隐私管理框架

解决大数据隐私问题的当务之急是,针对不同的风险,建立混合式与综合性的隐私管理框架,并积极拓展隐私管理的关键技术研究.

3.1 隐私管理的目标

隐私管理的总体目标是利用我们自己的管理理念和方法,像管理 Web 数据、XML 数据与移动数据一样管理大数据隐私.具体目标包括如下 3 点:

1) 为大数据的应用提供技术支撑.隐私是大数据应用的前提,若隐私问题不能得到很好的解决,则相应的应用很有可能成为空谈.防止数据收集者、数据分析者、分析结果的使用者恶意泄露隐私信息,防止大数据生命周期中收集、处理、存储、转换、销毁各个阶段中隐私的泄露;

2) 为那些悬而未决的隐私挑战寻找方法.目前许多领域仍未找到合适的隐私保护策略,比如,医疗保障和研究领域中,如何挖掘个人临床数据而又不存在保险歧视的风险,如何配送人性化基因药物而不存在医疗数据的误用等;市场营销领域中,如何确保消费者的信息在雇用或保险决策时没有被滥用;

3) 给打算公开数据的企业和个人一个定心丸.对于想公开和共享数据的人来说,数据隐私是第一位的.在不泄露数据隐私的前提下,可以公开数据并允许其他用户访问.比如,为科学研究公开自己的位置信息而不存在被恶意跟踪的风险;公开自己的社交网络信息而不存在丢掉工作的风险等.

3.2 主动式隐私管理框架

本节我们提出一种大数据隐私主动式管理建议框架,如图 3 所示.

该框架包括隐私主动监控体系、隐私主动评估体系、隐私主动管理技术体系、问责系统体系以及法律法规体系 5 大部分,为实现大数据隐私管理提供技术支持.

3.2.1 隐私风险主动监测

隐私风险是指基于个人或者团体数据上的构成隐私泄露的操作.比如,一个恶意攻击者在网站中植入意外查询;挖掘社交网络数据中人与人之间的链接关系等,这些操作均有可能披露隐私.隐私风险主动监测(privacy risk active monitor)体系是为了在处理大数据时,能够主动侦测到那些不正当的或者存有恶意的操作.不同操作的目的不同,比如,过分收集数据是为了挖掘更有价值的知识;Spam、免费 App、广告投放是为了获取更高的商业利益;窃取身份信息、泄露病人病情、黑客入侵、投放计算机病毒等恶意行为是为了窃取财物或者伤及别人.隐私风险主动监测是上层隐私管理技术与法律法规的基础.

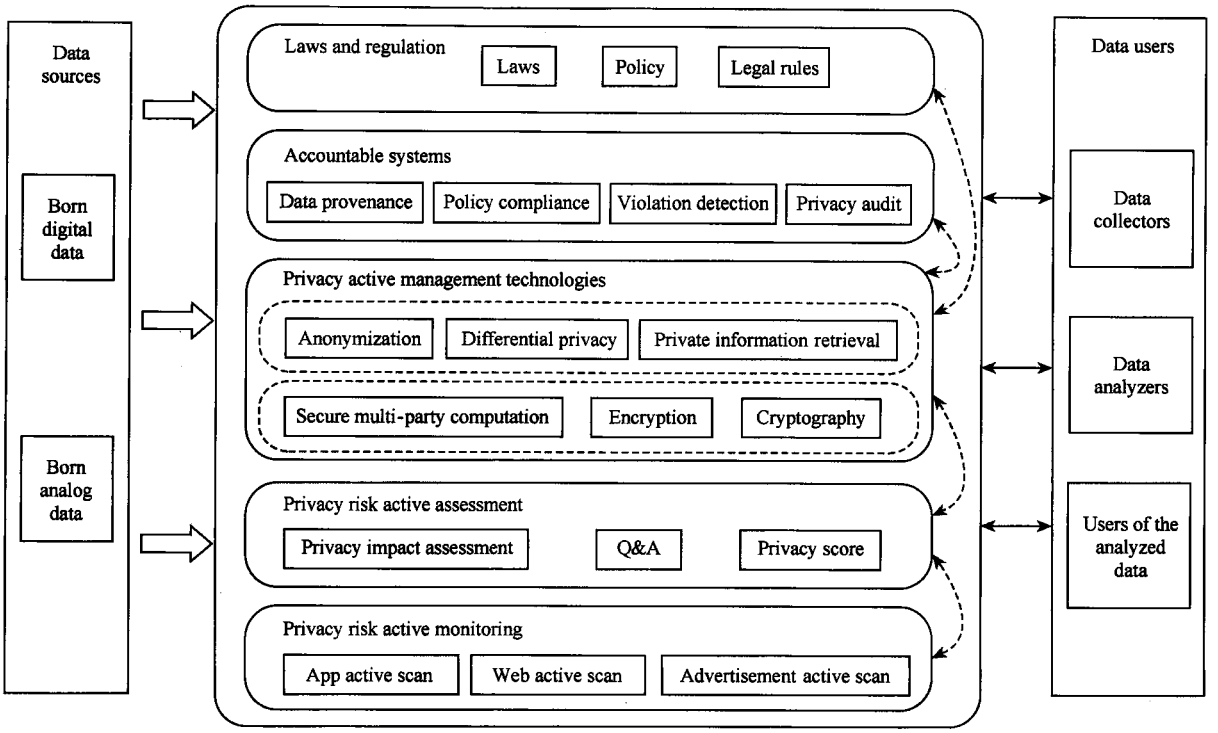

Fig. 3 Active privacy management framework of big data.

图 3 大数据主动式隐私管理框架

风险主动监测包含两个层面的含义:1)在缺乏诚信的应用环境中主动扫描到外部恶意攻击的能力.例如,免费 App 是否扫描自己的手机数据;手机中投放过来的移动广告是否记录自己的地理位置;Web 搜索服务是否利用 Cookies 技术记录自己的会话记录等;2)为上层管理体系主动发布隐私风险的能力.目前常用的隐私风险监测技术是基于成本最优博弈理论(cost-optimal game-theoretical)的方法[18].

3.2.2 隐私风险主动评估

隐私风险主动评估(privacy risk active assessment)是继隐私风险主动监测之后的管理体系,为大数据应用提供基础性服务,是支撑大数据应用的重要手段.风险主动评估同样应具有两层含义:1)在某个大数据应用的初级阶段能够主动分析出隐私风险大小的能力;2)具有指导上层隐私管理技术体系如何选择相应技术的能力.一方面可以通过简单的问答方式(Q&A)进行隐私风险评估,例如,用户数据在服务于一些大数据应用时,这些应用是否与用户本人相关?如果用户数据不含敏感信息,则个人隐私风险可能是轻微的;如果涉及到用户本人,应该给出什么是影响隐私泄露的原因、哪些额外操作甄别了用户数据?涉及应用的所有操作是否可信?另一方面,通过技术手段进行隐私风险主动评估.PIA (privacy impact assessment)[19]与 EBIOS (expression of needs and identification of security)[20]是常用的风险评估技术,其中 PIA 采用阈值技术评估隐私风险;而 EBIOS 使用风险严重程度与发生的可能性来衡量隐私风险的大小.

在进行风险评估时,为了避免触及原始数据,应该在隐私保护下做隐私风险评估,常用的方法是安全多方计算[21].此外,也可以根据隐私风险的不同等级,采用概率模型对操作的敏感性和可见性进行评估,利用隐私风险打分(privacy risk score)机制,自动为相应操作给出分值并起到预警作用[22-23].

3.2.3 隐私主动管理技术

图 3 中的隐私管理技术体系为整个大数据隐私管理框架提供了重要的技术和管理支撑,其核心涵盖以下 4 方面的应用需求:

1) 支持不同类型的查询需求.在隐私管理过程中,查询通常是数据使用者通过交互式环境①提交的,是大数据最常用的应用之一.例如,聚集查询、top-k 查询、workload 查询、范围计数查询、直方图查询等;

2) 支持不同数据类型的发布需求.无论是天生

① 交互式查询也可以称之为在线查询,是通过查询接口提交查询需求.

数字化数据或者天生模拟化数据转换之后的数据均可以表示成不同的数据类型，比如，关系数据、图数据、流数据、字符序列数据等. 而在非交互式环境[①]下发布这些隐私数据，将有利于行业内科技的发展；

3) 支持数据挖掘与机器学习的分析需求. 数据分析是整个大数据处理的核心，是发掘大数据真实价值具体过程. 例如，top-k 频繁模式挖掘、线性与逻辑回归、支持向量机分类、深度学习等；

4) 支持主动或者自适应选择隐私管理技术的需求. 在大数据管理环境中，不同类型的数据所需隐私保护层度不同，使用的技术也不相同. 目前，隐私管理技术包括匿名化技术、差分隐私保护技术、隐私信息检索技术、安全多方计算技术、数据加密技术等. 隐私管理技术体系应能够根据不同的数据类型与隐私风险评估结果，自适应或者主动选择相应的隐私管理技术来实现大数据隐私的管理. 为了利用上述提到的隐私管理技术，本文设计了一种主动式隐私保护框架，如图 4 所示，该框架可以实现隐私管理技术的自适应选择. 有关隐私管理技术的具体细节，本文会在第 4 节给出详细描述.

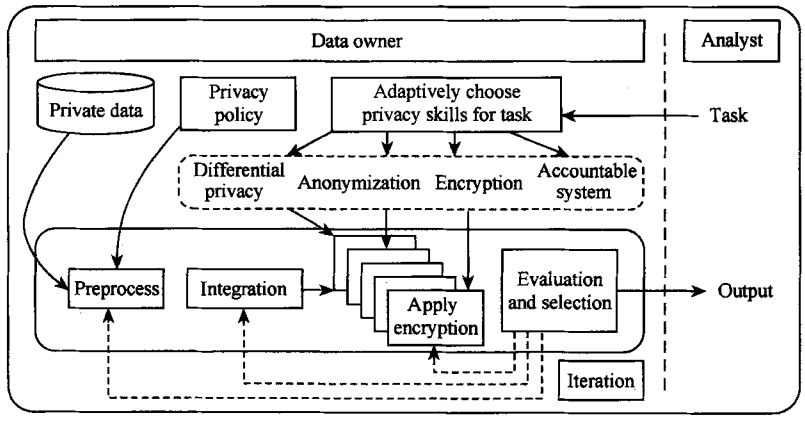

Fig. 4 Active privacy-preserving framework.

图 4 主动式隐私保护框架

3.2.4 问责系统

问责[24-25]是指当一个实体(例如项目负责人)的行为违反了某一策略和规则，则该实体应当受到惩罚. 问责系统[②](accountable system)是隐私管理技术体系与法律法规体系之间的桥梁，与隐私管理技术体系是相辅相成的. 问责系统在整个隐私管理框架中起到的作用犹如法律法规在社会中起到作用一样，对违反操作策略和规定的人起到追究其责任的作用. 隐私管理技术通过模糊化或加密来控制数据的访问，并且在特定的攻击模型下才能生效. 当隐私管理技术不能生效时，问责系统起着问责和追究责任的作用.

问责系统结合计算机技术、社会科学与法律法规对整个大数据操作起到监管作用，其功能应包含 3 点：具有标记不妥当操作的能力；利用策略语言标准(比如 AIR 语言[③])检验是否违反了策略与规定的能力；给出相应惩罚的能力. 此外，实施问责系统需要数据溯源、策略违反检测、隐私审计等技术的支持.

3.2.5 法律法规

由于大数据隐私管理的法律法规的特殊性，本文仅是简单的讨论. 法律法规是隐私保护技术之外的隐私保障手段. 因此，在管理隐私过程中，仅依靠技术是不够的，纯技术代替不了法律和社会道德对侵害隐私的制裁和约束. 美国和欧盟相继颁发了隐私法案，来规范个人数据在收集、使用与传播等方面的行为；2013 年 6 月中国工信部发布的《电信和互联网用户个人信息保护规定》，该规定为互联网个人信息的收集、使用提供了安全与法律法规保障. 由此看来，在大数据隐私管理过程中，政府应制定、改进和完善相应的隐私权法案，从法律法规角度为用户提供强大的隐私保护屏障.

4 现有隐私管理关键技术分析

大数据隐私管理的核心部分是隐私管理关键技

① 非交互式发布也称之为离线发布，是通过发布算法来公开相关信息.

② 在政府和企业中，问责系统被称为问责制度，是指负责人由于故意或者过失给企业和政府造成不良影响和后果的行为，进行内部监督和责任追究的制度.

③ http://dig.csail.mit.edu/2009/AIR/.

术.由于大数据隐私本身的特殊意义,传统的隐私保护理论和技术已经无法涵盖其内涵.目前没有一个万能的方法能够解决所有的隐私问题,每一种方法均有自己的优缺点.本文针对大数据管理过程中面临的隐私风险和挑战,展开大数据隐私管理关键技术的分析.本节选取部分重点隐私管理技术给予介绍.

4.1 匿名化技术

匿名化是指隐藏或者模糊数据以及数据源.该技术一般采用抑制[26]、泛化[27]、剖析[28]、切片[29]、分离[30]等操作匿名数据.k-匿名[1]是该技术的早期代表方法,该方法在发布关系数据时要求每一个泛化后的等价类(equivalence class)至少包含 k 条相互不能区分的数据,即是要求一条数据表示的个人信息至少和其他 $k-1$ 条数据不能区分.然而,k-匿名的缺陷是未对等价类中的敏感属性进行约束进而导致该技术失效,例如,某等价类中任意一个敏感属性取值相同,则攻击者可以推理出该敏感值.与 k-匿名不同,l-diversity[31]方法在匿名关系数据时确保每个等价类至少包含 l 个不同的敏感属性值.虽然 l-多样化保证了敏感属性的多样性,却忽视了敏感属性的全局分布,进而攻击者可能以很高的概率确认出敏感值.为弥补 l-diversity 方法的不足,t-closeness[32]方法要求所有等价类中敏感属性值的分布与该属性的全局分布保持一致.此外,m-invariance[33]与 HD-composition[34]填补了 k-匿名、l-diversity 与 t-closeness 方法仅适用于静态关系数据的不足,确保数据在动态或者增量发布数据时隐私不被泄露.

上述研究是针对关系数据的,而另一部分匿名化研究是着眼于社交网络数据的发布和查询.社交网络中包含大量的敏感信息,例如链接关系、节点属性、节点标记、图结构特征等,攻击者可以借助主动攻击[35]与被动攻击[35]模型推理和披露相关的敏感信息.社交网络数据隐私保护技术分为 2 类:基于聚类泛化法与图结构修改法.基于聚类泛化法[36-39]是指通过聚类的方法把图中的节点和边分成超级节点和超级边,节点和边的敏感信息可以隐藏在它们的超类中.常用方法包括节点聚类法、边聚类法和节点边聚类法;图结构修改法[40-44]是指通过节点和边的插入删除操作改变图的结构,保护边和节点的身份识别以及重新识别.这类方法主要采用类似于 k-匿名思想,防止攻击者借助网络结构作为背景知识进行攻击,例如度攻击、子图攻击、l-近邻攻击等.

相对于关系数据与社交网络而言,大数据的匿名化更为复杂.大数据中多源数据之间的集成融合以及相关性分析使得上述那些针对小数据的被动式保护方法失效.与图 3 中的主动式隐私管理框架相比,传统匿名技术存在缺陷是被动式地防止隐私泄露,结合单一数据集上的攻击假设来制定相应的匿名化策略.然而,大数据的大规模性、多样性使得传统匿名化技术顾此失彼.

4.2 数据加密技术

大数据隐私管理通常以云平台为依托,在云平台下实现隐私管理的首要问题是存储、加密数据上的计算以及通信的安全性,数据加密技术正好满足这一需求.云平台下具体应用通常依赖于数据的存储、索引与检索以及云平台提供的可信度.同态加密(homomorphic encryption)[45-49]、功能加密(functional encryption)[50]、安全多方计算[51-52]等是常用的加密方法.文献[48-49]利用同态加密技术分别提出了 key-value 隐私存储方式以及多级索引处理技术,确保数据拥有者和云平台都不能在用户查询的结点检索过程中识别出结点.密文检索处理技术分为对称加密[53]和公钥加密方法[54-55].其中,文献[53]提出了一种支持动态检索的对称加密方法,具有较高的安全性和检索效率;文献[54]提出了可搜索公钥加密技术并支持关键字检索;而文献[55]针对文献[54]的安全通道和一致性问题,提出了基于随机预言模型与无安全通道的公钥加密方案.此外,功能加密允许在处理密钥时学习密文所隐含的信息.

安全多方计算是另外一类数据加密技术,其核心操作是在分布式环境下基于多方参与者提供的数据计算出相应的函数值,并确保除了参与者的输入以及输出信息外,不会额外地暴露参与者的任何信息.该技术常用于分布式环境下隐私保护的数据挖掘领域[51],并逐渐扩展到无向积[56]与添加矢量[57]等领域.

尽管上述研究为大数据隐私管理提供了一定的思路,但是该技术的缺陷比较明显.类似于匿名化技术,该类技术也是针对某类数据的隐私泄露而被动式的保护.而在大数据环境下,其大规模性、多样性等特点使得该类技术陷入循环怪圈,面对新型应用的隐私泄露,必须新的加密方法才能保护.

4.3 差分隐私技术

无论是匿名技术还是加密技术,二者都是针对当前的外部攻击来设计启发式保护方法,面对新的攻击需要重新制定保护方法.在大数据环境中,这 2 类

方法均由于缺乏很强的数学基础来定义数据隐私性与损失性而不具有普遍应用性. 差分隐私[58-62]的出现弥补了这一空白,该模型是一种由数学理论支撑的、新型的、强健的隐私保护技术. 根据差分隐私形式化定义① 可知,该方法由隐私参数 ε 控制着隐私保护程度与隐私损失的大小,可以确保在某一数据集中插入或者删除一条记录的操作不会影响任何计算的输出结果. 另外,该方法不关心攻击者所具有的背景知识,即使攻击者已经掌握除某一条记录之外的所有记录的信息,该记录的隐私也无法被披露[63],这一特点使得差分隐私技术具有很好的扩展性. 要实现差分隐私保护需要借助于噪音机制和查询敏感性[64]. 常用的噪音机制包括拉普拉斯噪音[64]与指数噪音[65],噪音的大小与函数相关 $f(\Delta/\varepsilon)$,其中,$f(\cdot)$ 表示拉普拉斯分布或者指数分布的分布函数,Δ 表示查询敏感性. 目前,差分隐私技术的研究主要集中在数据发布、数据挖掘与学习、查询处理等方面. 数据发布典型的工作包括:一维和多维直方图发布方法[66-74]、流数据发布[75-77]、图数据发布[78-80]以及空间数据发布[81-82]等;数据挖掘和机器学习近期研究包括:频繁模式挖掘[83-85]、回归分析[86-87]、分类[88-89]等;而查询处理工作包括:范围计数查询[90]、基于矩阵机制的批量查询[91-92]、基于低秩机制的批量查询[93]等.

从上述的研究可以看出,差分隐私已经成为目前隐私保护技术研究热点. 学术界认为差分隐私与大数据具有天然的匹配性[94],其原因是大数据的大规模性和多样性使得在数据集中添加或者删除某个数据点对整体数据的影响非常小,这一特质与差分隐私定义的内涵相吻合.

尽管如此,相对于本文提出的大数据隐私主动式管理框架,差分隐私保护技术仍存在的缺陷包括:无法主动式地控制隐私参数 ε. 小的 ε 导致低可用性与高的隐私性,反之,导致高可用性与低的隐私性. 因此,该参数很难控制. 大数据之间的关联性有可能弱化差分隐私保护效果.

4.4 隐私信息检索技术

隐私信息检索(private information retrieval)[95]技术通常被用于外包数据时的查询安全,用户可以在不可信的服务平台上查询任意数据而不泄露被查询数据的敏感信息. 被查询的数据可以是公开的、匿名的,但是服务平台却无法甄别这些数据的具体内容. 尽管 4.2 节提到的同态加密技术也可以实现对查询的控制,然而,由于查询的复杂性与计算开销使得这类技术不具有实用性. 实现隐私检索的技术包括 2 类:1)基于信息论的检索方法[95],该方法通常是把所有的数据传递给客户端并允许其在本地解码,然而由于传输代价问题,这种技术不太适合大数据;2)基于硬件的可计算检索方法,该方法是目前比较常用的,通常用于 DNA 序列匹配、基于内容的图像检索以及位置隐私查询等领域. 文献[96-97]基于可计算框架分别依据二次剩余假设问题的难解性与伪随机函数的可实现性设计了不同的隐私信息检索方法[98];文献[99]提出了一种单一服务方可计算检索协议,该协议利用 Paillier 加密系统[100]实现了低通信开销的字符传输. 然而,该方法却存在效率低以及信息泄露的危险[101];文献[102]借助于 ORAM (Oblivious RAM)[103]计算提出了一种更加有效的检索协议,该协议不但能够降低通信和计算代价,更能够防止信息泄露. 尽管隐私信息检索技术促进了安全软硬件的发展,但在大数据环境中,这项技术的应用会更加困难和复杂.

4.5 问责系统

本节主要阐述问责系统中所涉及的计算机技术. 问责系统应能够记录用户的数据是如何管理的、哪些人访问过他们的数据、数据什么时候被修改和误用过等,该系统的核心包括数据追踪、违规检测、数据溯源等. 为了实现问责系统,工作流技术至关重要. 工作流经过的途径都有可能要问责,例如,非授权进入安全系统、非授权检索安全数据等. 标记和追踪工作流中所有数据行为是问责系统的关键. 数据溯源是追踪数据流经途径的常用方法,其类别包括标记方法[104-105]、数据驱动的追踪方法[106-107]、集成式追踪方法[108]与分布式追踪方法[109-110]. 文献[104-105]利用标记方法记录数据在数据仓库中的传播和查询历史;文献[106-107]提出了 Flogger 与 S2Logger 方法,分别追踪云框架下与端对端下的数据溯源,可以记录文件系统上数据创建、读写行为. 文献[106]结合网格环境提出了一种溯源集成方法,聚合不同的工作流来记录数据派生史;文献[109-110]提出一种分布式溯源追踪方法,利用网络传播标记所有操作. 此外,事件追踪[111]也是捕捉数据行为的常用方

① **定义 1**. ε-差分隐私. 给定数据集 D 和 D',一个隐私算法 A,$Range(A)$ 为 A 的取值范围. 若算法 A 在数据集 D 和 D' 上任意输出结果 O ($O\in Range(A)$)满足不等式 $\Pr[A(D)=O]\leqslant e^{\varepsilon}\times\Pr[A(D')=O]$,则 A 满足 ε-差分隐私. 其中,概率 $\Pr[\cdot]$ 由算法 A 的随机性控制;隐私预算参数 ε 表示隐私保护程度,ε 越小隐私保护程度越高. D 和 D' 之间至多相差一条记录,则 D 和 D' 为近邻数据集,二者为近邻关系.

法.上述数据溯源研究大都是基于小规模数据集,在应用于大数据集时应注意数据集多变与复杂性以及数据质量等问题.违规检测是问责系统的另一个核心技术.当数据被误用时,问责系统应能够检测出何处出现误用行为和误用行为的制造者.常用的检测技术包括入侵检测[112-113]与统计匹配[113].文献[112-113]提出了网络协议层上的 Backtracker 检测方法,该方法能够实现多主机之间的入侵检测与入侵源追踪;文献[114]利用数据的统计可追溯性可检测出误用行为.

目前,问责系统在大数据环境中管理隐私存在的缺陷包括:缺乏底层风险监测与评估的支持;缺乏可靠的法律法规制度确保问责系统的执行.

5 隐私管理技术面临的挑战

本文所提出的隐私管理技术框架为大数据隐私管理提供了重要的技术支撑.然而,该框架以及框架中所集成的现有隐私保护技术都存在一定的挑战.

5.1 隐私管理框架带来的挑战

第 3 节主要描述了主动式隐私管理框架中各个体系的功能以及实现每个功能模块对应的方法和技术,而该框架同时也面临着诸多挑战与问题.

1) 隐私风险主动监视与评估体系面临的挑战

丰富的数据资源是大数据技术发展的基础,目前我国数据源开放的程度比较低.一旦政府、企业和行业之间突破制约而公开与共享数据,隐私风险主动监视与主动评估将面临着巨大挑战.简单的隐私监视方法有可能无法满足多数据源共享的需求.例如,如何监视到过分收集数据、恶意分析数据的行为与操作等.同时,Q&A 系统与 PIA 技术这些简单的评估方法可能无法应对多数据源公开带来的挑战.因此,针对不同的数据源带来的隐私风险,如何制定新的隐私评估与分析策略、如何对新的隐私风险进行分类以及风险等级划分等都存在很大的问题.

2) 隐私主动管理技术体系面临的挑战

隐私主动管理技术体系的终极目标是在大数据整个生命周期中保护其隐私,并且能够依据隐私风险评估结果主动选择相应的保护技术.然而,大数据资源的连续性公开,使得隐私管理技术面临新型的隐私攻击与隐私泄露,例如多源数据融合带来的隐私威胁等.因此,如何设计应对新型攻击模型的管理技术是个大的挑战,如何把相应的技术集成到我们的主动式隐私保护框架中也是个很大挑战.为了避免主动式管理技术陷入与传统技术相同的怪圈,可以利用机器学习方法对相应的隐私管理技术与隐私泄露原因进行训练与学习,进而达到自适应地选择与应对隐私风险的效果,然而,如何设计学习方法也是个挑战性问题.

3) 问责系统体系面临的挑战

在问责系统体系中,数据溯源是追踪数据操作行为的理想技术,然而,大数据的高速性与多样性等特点使该技术变得更加复杂.如何跨平台、跨领域追踪那些明显发生改变的数据非常困难.虽然数据溯源技术可以描述整个数据的脉络,而在使用该技术时可能导致数据隐私泄露.其原因是数据溯源本身可能蕴含敏感的元数据,在追踪过程中可能会泄露其他的信息等.此外,在大数据环境中,依靠人工来标记数据的使用和误用目的不太现实,如何利用统计方法设计出自动甄别和检测误用行为的方法是个大的挑战.

4) 法律法规体系面临的挑战

在保护大数据隐私方面,我国现有的法律法规主要面临着 3 个方面的挑战:①现有的法律以保护"个人可识别信息"为主,而在大数据环境中,个人可识别信息的界限越来越难界定;②以往的隐私保护制度,例如"操作目的明确,征得个人事先同意、限制信息使用范围"等,越来越难控制;③无法对数据跨境流动带来的隐私危害给予保护.因此,针对上述挑战,如何完善与改进目前的法律法规是我们所面临的问题.

5) 文中 2.2 节描述了当前大数据处理平台的各个层次都存在相应的隐私泄露威胁,而我们提出的主动式隐私管理框架主要是应对这些问题.然而,如何把我们设计的管理框架嵌入到现有的大数据管理框架中是个很大挑战.

5.2 现有隐私保护技术存在的挑战

第 4 节主要分析了当前隐私管理技术的优缺点.在本文提出的隐私管理框架中,现有隐私保护技术面对大数据面临诸多挑战:

1) 匿名化技术面临的挑战

在大数据集成融合的过程中,模式定位(schema alignment)是其核心操作,而在模式定位时,数据源的多样性和动态性会涉及到数据的多种属性,并且这些数据之间彼此存在相关性,甚至导致模式语义发生演化.而传统的匿名方法无法保护模式演化的敏感属性.因此,如何设计兼顾大数据的模式多样性、模式演化与相关性的匿名方法是个挑战性问题.

数据源之间的相关性来自于数据之间拷贝关系,拷贝关系通常导致混乱的数据溯源,进而可能导致虚假信息的存在.在集成分析过程中,需要完整与可信的数据源,若匿名方法保护了虚假的拷贝数据,则真值的隐私可能会泄露.因此,如何设计兼顾拷贝关系与追踪数据溯源的匿名方法是个很大挑战.

此外,现存的匿名方法通常存在可扩展性差、计算代价高、匿名后数据可用性度量不规范等缺陷,因此,如何把现有的方法扩展到目前新型的计算框架中,例如 MapReduce, Storm 与 Spark 等,是个很大的挑战.同时,制定新的适用于大数据匿名的信息损失度量方法也是个大的挑战.

2) 数据加密技术面临的挑战

由于可以从多渠道获得大数据,进而在加密过程中如何保护其私密性非常关键.在处理大数据安全查询时,通常假设云平台是可信的,然而,现实应用中不可信或者半可信的云平台确实存在.在这类云平台上,数据拥有者的数据、用户查询隐私均有可能被披露.例如,某零售公司把人群的个体信息暴露给不可信的云平台,该平台有可能把个体的隐私信息卖给该公司的竞争对手.因此,如何利用可搜索公钥加密技术、同态加密技术、功能加密技术以及安全多方技术来设计一种既能保护用户的查询隐私、数据隐私以及三方交互隐私的方法是个很大的挑战.在处理大数据实时计算时,常采用同态加密技术,然而,目前的同态加密技术效率比较低,因此,如何设计高效的实时的同态加密方法是一个大的挑战.另外,在大数据环境下,协同作业是常用的技术.安全多方计算技术是确保协同作业时彼此不泄露隐私的主要方法,然而,许多基于安全多方技术的方法都是常驻内存的,要求数据全部驻留在内存中,而这类方法不能直接被用于有几千万条记录的大数据上.因此,如何利用安全多方技术来设计非内存的且满足大规模数据协同作业需求的方法是个很大的挑战.

3) 差分隐私技术面临的挑战

差分隐私保护的前提是要求数据集的数据是相互独立的[①],而大数据的多样性与大规模性造成了多源数据之间的相关性,进而很难保证差分隐私技术有效.并且,如果数据以相关性分组的形势存在,特别组比较大时,差分隐私保护效果比较差,进而导致现有的分析和查询方法不能很好地移植到大数据环境.因此,如何设计支持相关性数据发布与查询的差分隐私算法是个挑战性问题.

大数据的高速性要求以流的方式对其进行分析和发布.发布数据要连续更新,否则无法概要全部的统计信息,此外,必须采取在线高效的处理方式并保证发布结果的准确性.因此,如何依据滑动窗技术、抽样技术以及约束推理技术,设计出处理高速的、实时变化的大数据差分隐私管理框架变得尤为重要.

在采用差分隐私保护大数据时,如何调节和分配隐私参数 ε 非常关键,因为 ε 直接决定着数据隐私性与可用性.较大的 ε 值弱化隐私性而增强可用性,反之,减弱可用性而增强隐私性.而实际应用中,ε 还不能充分平衡隐私性与可用性.因此,如何设计合理的隐私参数 ε 分配策略是个很具有挑战性的问题.

4) 隐私信息检索技术面临的挑战

目前,隐私信息检索技术主要针对不可信服务器上的隐私信息查询,例如 KNN 查询、关键字查询、最短路径查询等.在大数据环境中,用户可以利用该技术向不可信服务器发送查询并获得响应结果.该方法通常利用信息冗余技术针对一次提交的查询而给出相应的隐私保护.然而,大数据的多样性与大规模性使得仅依靠一次查询所获得的结果非常不准确.用户有可能针对某个查询向不可信服务器提交多次或者不同形式的查询,例如查询距离自己最近的医院,用户需要多次基于隐私信息检索技术的访问才可能获得准确的查询结果.在大数据环境下,利用隐私信息检索技术对用户多次或者多样化查询进行保护时,需要大量的冗余数据,这样导致很高的计算代价和响应代价.因此,如何设计同时兼顾降低上述两种开销的方法很具有挑战性.

此外,目前隐私信息检索技术通常对用户查询的内容进行保护,没有考虑如何保护服务器端数据的情况.而此种情况要求用户只能查询被授权的数据,同时服务提供方不知道用户具体查询哪些数据.因此,如何将隐私信息检索技术与现有数据加密技术相结合来实现用户查询隐私与服务器数据隐私的保护是个大的挑战.

6 结束语

大数据在当前 IT 业发展十分迅速,具有广阔的发展前景,但同时其所面临的隐私挑战和风险也是空前的,需要隐私保护研究者共同探求管理之道.

① ε-差分隐私的定义,参见脚注定义 1.

本文打破了传统被动式保护技术的约束,提出了主动式隐私管理框架,并讨论了该框架面临的主要技术挑战.大数据隐私管理不仅仅是技术方面的问题,它还涉及到法律法规、监管模式、宗教等诸多方面.因此,仅从技术层面探讨大数据隐私管理问题是远远不够的,需要学术界、企业界以及政府相关部门共同努力才能实现.

参 考 文 献

[1] Sweeney L. K-anonymity: A model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

[2] Montjoye D, Hidalgo C A, Verleysen M, et al. Unique in the crowd: The privacy bounds of human mobility [J]. Nature, Scientific Reports, 2013, 3(2): 1-5

[3] Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name [R/OL]. Cambridge, MA: Harvard University Data Privacy Lab. [2013-04-24]. http://dataprivacylab. org/projects/pgp/1021-1. pdf

[4] Weitzner D J, Bruce E J. Big data privacy workshop: Advancing the state of the art in technology and practice [R]. [2014-03-03]. http://web. mit. edu/bigdata-priv/index. html

[5] Holdren J P, Lander E S. Big data privacy: A technological perspective [R/OL]. [2014-05-01]. http://www. whitehouse. gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014. pdf

[6] China Academy of Telecommunication Research of MIIT. Big data white paper [R/OL]. [2014-07]. China Academy of Telecommunication Research of MIIT (in Chinese)
(工业和信息化部电信研究院. 大数据白皮书[R/OL]. [2014-07]. 工业和信息化部电信研究院, 2014)

[7] Dong X, Laure B E, Srivastava D. Truth discovery and copying detection in a dynamic world [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 562-573

[8] Podesta J, Pritzker P, Moniz E J, et al. Big data: seizing opportunities preserving values [R/OL]. Washington: Executive Office of the President, The White House Washington. [2014-05-01]. http://www. whitehouse. gov/sites/default/files/docs/big_data_privacy_report_may_1_2014. pdf

[9] Feng Dengguo, Zhang Min, Li Hao. Big data security and privacy protection [J]. Chinese Journal of Computers, 2014, 37(1): 246-258
(冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258)

[10] 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)

[11] Wong R C W, Fu A, Wang K, et al. Can the utility of anonymized data be used for privacy breaches [J]. ACM Trans on Knowledge Discovery from Data, 2011, 5(3): 1-16

[12] Narayanan A, Shmatikov V. Roust de-anonymization of large spare datasets [C] //Proc of the 29th IEEE Symp on Security and Privacy (S&P 2008). New York: IEEE, 2008: 111-125

[13] Alina E, Sungjin Im, Moseley B. Fast clustering using MapReduce [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2011). New York: ACM, 2011: 681-689

[14] Caetano T J, Traina A J M, López J, et al. Clustering very large multi-dimensional datasets with MapReduce [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2011). New York: ACM, 2011: 690-698

[15] Flavio C, Nilesh D, Ravi K. Correlation Clustering in MapReduce [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2014). New York: ACM, 2014: 641-650

[16] Hsieh C J, Chang K W, Lin C J, et al. A dual coordinate descent method for large-scale linear SVM [C] //Proc of the 25th Int Conf on Machine Learning (ICML 2008). Menlo Park, CA: AAAI, 2008: 408-415

[17] Schmidt M, Roux N L, Bach F. Convergence rates of inexact proximal-gradient methods for convex optimization [C] // Proc of the 25th Annual Conf on Neural Information Processing Systems (NIPS 2011). Berlin: Springer, 2011: 1458-1466

[18] Abbe E A, Khandani A E, Lo A W. Privacy-preserving methods for sharing financial risk exposures [J]. American Economic Review: Papers & Proceedings, 2012, 102(3): 65-70

[19] Office of the Privacy Commissioner. Privacy impact assessment guide. Australian Government [R/OL]. [2008-07-16]. http://www. privacy. org. nz/news-and-publications/guidance-notes/privacy-impact-assessment-handbook

[20] Methodology for Privacy Risk Management: How to Implement the Data Protection Act [R/OL]. [2012-05-09]. http://www. piawatch. eu/node/1539

[21] Clifton C, Kantarcioglou M, Lin X, et al. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations, 2002, 4(2): 28-34

[22] Liu K, Terzi E. A framework for computing privacy scores of users in online social networks [C] //Proc of the 9th IEEE Int Conf on Data Mining (ICDM 2009). Piscataway, NJ: IEEE, 2009: 235-242

[23] Mislove A, Viswanath B, Gummadi K, et al. You are who you know: Inferring user profiles in online social networks [C] //Proc of the 3rd Int Conf on Web Search and Web Data Mining (WSDM 2010). New York: ACM, 2010: 243-252

[24] Feigenbaum J, Jaggard A D, Wright R A. Towards a formal model of accountability [C] //Proc of the 19th Workshop on New Security Paradigms Workshop (NSPW 2011). New York: ACM, 2011: 45-56

[25] Weitzner D J. Information accountability [J]. Communication of the ACM, 2008, 51(6): 82-87

[26] Wang K, Fung B C M, Yu P S. Handicapping attacker's confidence: An alternative to *k*-anonymization [J]. Knowledge and Information Systems, 2007, 11(3): 345-368

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

[28] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation [C] //Proc of the 32nd Int Conf on Very Large Data Bases (VLDB 2006). New York: ACM, 2006: 139-150

[29] Li T, Li N, Zhang J, et al. Slicing: A new approach for privacy preserving data publishing [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(3): 561-574

[30] Terrovitis M, Liagouris J, Mamoulis N, et al. Privacy preservation by disassociation [J]. Proceedings of the VLDB Endowment, 2012, 5(10): 944-955

[31] Machanavajjhala A, Kifer D, Gehrke J, et al. *l*-diversity: Privacy beyond k-anonymity [J]. ACM Trans on Knowledge Discovery from Data, 2007, 1(1): 1-47

[32] Li N, Li T, Venkatasubramanian S. Closeness: A new privacy measure for data publishing [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(7): 943-956

[33] Xiao X, Tao Y. *m*-invariance: Towards privacy preserving republication of dynamic datasets [C] //Proc of the 27th ACM Int Conf on Management of Data (SIGMOD 2007). New York: ACM, 2007: 689-700

[34] Bu Y, Ada W C F, Wong R C W, et al. Privacy preserving serial data publishing by role composition [J]. Proceedings of the VLDB Endowment, 2008, 1(1): 845-856

[35] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural teganography [C] //Proc of the 16th Int World Wide Web Conf (WWW 2007). New York: ACM, 2007: 122-132

[36] Cormode G, Srivastava D, Bhagat S, et al. Class-based graph anonymization for social network data [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 810-811

[37] Zheleva E, Getoor L. Preserving the Privacy of Sensitive Relationships in Graph Data [C] //Proc of the 1st KDD Workshop on Privacy, Security, and Trust in KDD (PinKDD 2007). Berlin: Spinger, 2007: 53-171

[38] Cormode G, Srivastava D, Yu T, et al. Anonymizing bipartite graph data using safe groupings [J]. Proceedings of the VLDB Endowment, 2008, 1(1): 833-844

[39] Zou L, Chen L, Ozsu M T A. *k*-automorphism: A general framework for privacy preserving network publication [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 946-957

[40] Cheng J, Fu A C W, Liu J. *k*-isomorphism: Privacy preserving network publication against structural attacks [C] //Proc of the 30th ACM Int Conf on Management of Data (SIGMOD 2010). New York: ACM, 2010: 459-470

[41] Wu W, Xiao Y, Wang W, et al. *k*-symmetry model for identity anonymization in social networks [C] //Proc of the 13th Int Conf on Extending Database Technology (EDBT 2010). New York: ACM, 2010: 111-122

[42] Liu K, Terzi T. Towards identity anoymization on graphs [C] //Proc of the 28th ACM Int Conf on Management of Data (SIGMOD 2008). New York: ACM, 2008: 93-106

[43] Ying X, Wu X. Randomizing social networks: A spectrum preserving approach [C] //Proc of the 8th SIAM Conf on Data Mining (SDM 2008). Philadelphia, PA: SIAM, 2008: 739-750

[44] Zhou B, Pei J. A brief survey on anonymization techniques for privacy preserving publishing of social network data [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2008). New York: ACM, 2008: 12-22

[45] Stehlé D, Steinfeld R. Faster fully homomorphic encryption [C] //Proc of the 16th Int Conf on the Theory and Application of Cryptology and Information Security (ASIACRYPT 2010). Berlin: Springer, 2010: 377-394

[46] Josep D F. A provably secure additive and multiplicative privacy homomorphism [C] //Proc of the 5th Int Conf on Information Security (ISC 2002). Berlin: Springer, 2002: 471-483

[47] Gentry C. Fully homomorphic encryption using ideal lattices [C] //Proc of the 1st ACM Symp on Theory of Computing. New York: ACM, 2009: 169-178

[48] Hu H, Xu J, Xu X, et al. Private search on key-value stores with hierarchical indexes [C] //Proc of the 30th IEEE Int Conf on Data Engineering (ICDE 2014). Piscataway, NJ: IEEE, 2014: 628-639

[49] Hu H, Xu J, Ren C, et al. Processing private queries over untrusted data cloud through privacy homomorphism [C] // Proc of the 27th IEEE Int Conf on Data Engineering (ICDE 2011). Piscataway, NJ: IEEE, 2011: 639-644

[50] Goldreich O. The foundations of cryptography—Volume 2 [M]. Cambridge, UK: Cambridge University Press, 2004

[51] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2002). New York: ACM, 2002: 639-644

[52] Sheikh R, Mishra D K, Kumar B. Secure multiparty computation: From millionaires problem to anonymizer [J]. Information Security Journal: A Global Perspective, 2011, 20(1): 25-33

[53] Kamara S, Papamanthou C, Roeder T. Dynamic Searchable Symmetric Encryption [C] //Proc of the 19th ACM Conf on Computer and Communications Security (CCS 2012). New York: ACM, 2012: 965-976

[54] Abdalla M, Chevassut O, Fouque P A, et al. Searchable encryption revisited: Consistency properties, relation to anonymous IBE, and extensions [C] //Proc of the 25th Annual Int Cryptology Conf (CRYPTO 2005). Berlin: Springer, 2005: 205-222

[55] Hyun S R, Willy S, Kim H J. Secure searchable public key encryption scheme against keyword guessing attacks [J]. IEICE Electronic Express, 2009, 6(5): 237-243

[56] Du W, Zhan Z. Building decision tree classifier on private data [C] //Proc of the IEEE Int Conf on Privacy, Security and Data Mining. Piscataway, NJ: IEEE, 2002: 121-128

[57] Vaidya J, Clifton C. Privacy-preserving k-means clustering over vertically partitioned data [C] //Proc of the 9th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2003). New York: ACM, 2003: 206-215

[58] Dwork C. Differential Privacy [C] //Proc of the 33rd Int Colloquium on Automata, Languages and Programming (ICALP 2006). Berlin: Springer, 2006: 1-12

[59] Dwork C. Differential privacy: A survey of results [C] // Proc of the 5th Int Conf on Theory and Applications of Models of Computation (TAMC 2008). Berlin: Springer, 2008: 1-19

[60] Dwork C, Lei J. Differential privacy and robust statistics [C] //Proc of the 41st Annual ACM Symp on Theory of Computing (STOC 2009). New York: ACM, 2009: 371-380

[61] Dwork C, Naor M, Reingold O, et al. On the complexity of differentially private data release: Efficient algorithms and hardness results [C] //Proc of the 41st Annual ACM Symp on Theory of Computing (STOC 2009). New York: ACM, 2009: 381-390

[62] Dwork C. The differential privacy frontier (extended abstract)[C] //Proc of the 6th Theory of Cryptography Conf (TCC 2009). Berlin: Springer, 2009: 496-502

[63] ZhangXiaojian, Meng Xiaofeng. Differential privacy in data publication and analysis [J]. Chinese Journal of Computers, 2014, 37(4): 927-949 (in Chinese)
(张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护 [J]. 计算机学报, 2014, 37(4): 927-949)

[64] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Theory of Cryptography Conf (TCC 2006). Berlin: Springer, 2006: 363-385

[65] McSherry F, Talwar K. Mechanism design via differential privacy [C] //Proc of the 48th Annual IEEE Symp on Foundations of Computer Science (FOCS 2007). Piscataway, NJ: IEEE, 2007: 94-103

[66] Xiao X, Xiong L, Yuan C. Differential privacy via wavelet transforms [J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23(8): 1200-1214

[67] Hay M, Rastogi V, Miklau G, et al. Boosting the accuracy of differentially private histograms through consistency [C] // Proc of the 36th Int Conf on Very Large Data Bases (VLDB 2010). New York: ACM, 2010: 1021-1032

[68] Xu J, Zhang Z, Xiao X, et al. Differential private histogram publication [J]. International Journal of Very Large Database, 2013, 22(6): 797-822

[69] Acs G, Chen R. Differentially private histogram publishing through lossy compression [C] //Proc of the 11th IEEE Int Conf on Data Mining (ICDM 2012). Piscataway, NJ: IEEE, 2012: 84-95

[70] Rastogi V, Nath S. Differentially private aggregation of distributed time-series with transformation and encryption [C] //Proc of the 30th ACM Int Conf on Management of Data (SIGMOD 2010). New York: ACM, 2010: 735-746

[71] Zhang X, Chen R, Xu J, et al. Towards Accurate Histogram Publication under Differential Privacy [C] //Proc of the 14th SIAM Int Conf on Data Mining (SDM 2014). Philadelphia, PA: SIAM, 2014: 587-595

[72] Kellaris G, Papadopoulos S. Practical differential privacy via grouping and smoothing [J]. Proceedings of the VLDB Endowment, 2013, 6(5): 301-312

[73] Qardaji W, Yang W, Li N. PriView: Practical differentially private release of marginal contingency tables [C] //Proc of the 34th ACM Int Conf on Management of Data (SIGMOD 2014). New York: ACM, 2014: 1435-1446

[74] Cormode G, Procopiuc M, Srivastava D, et al. Privbayes: Private data release via bayesian networks [C] //Proc of the 34th ACM Int Conf on Management of Data (SIGMOD 2014). New York: ACM, 2014: 1423-1434

[75] Chan T H H, Shi E, Song D. Private and continual release of statistics [J]. ACM Trans on Information and System Security, 2011, 14(3): 1-23

[76] Bolot J, Fawaz N, Muthukrishnan S, et al. Private decayed predicate sums on streams [C] //Proc of the 16th Int Conf on Database Theory (ICDT 2013). New York: ACM, 2013: 284-295

[77] Fan L, Xiong L. An adaptive approach to real-time aggregate monitoring with differential privacy [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(9): 2094-2106

[78] Karwa V, Raskhodnikova S, Smith A, et al. Private analysis of graph structure [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1146-1157

[79] Chen R, Fung B C M, Yu P S, et al. Correlated network data publication via differential privacy [J]. Very Large Data Bases Journal, 2014, 23(4): 653-676

[80] Xiao Q, Chen R, Tan K L. Differentially private network data release via structural inference [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2014). New York: ACM, 2014: 451-462

[81] Chen R, Fung B C M, Desai B C, et al. Differentially private transit data publication: A case study on the Montreal transportation system [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2012). New York: ACM, 2012: 213-221

[82] Hien T, Ghinita G, Shahabi C. A framework for protecting worker location privacy in spatial crowdsourcing [J]. Proceedings of the VLDB Endowment, 2014, 7(10): 919-930

[83] Li N, Qardaji W, Su D, et al. PrivBasis: Frequent itemset mining with differential privacy [J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1340-1351

[84] Zeng C, Naughton J F, Cai J. On differentially private frequent itemset mining [J]. Proceedings of the VLDB Endowment, 2013, 6(1): 25-36

[85] Shen E, Yu T. Mining frequent graph patterns with differential privacy [C] //Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2013). New York: ACM, 2013: 545-553

[86] Zhang J, Zhang Z, Xiao X, et al. Functional mechanism: Regression analysis under differential privacy [J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1364-1375

[87] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization [J]. Journal of Machine Learning Research, 2011, 12: 1069-1109

[88] Mohammed N, Chen R, Fung B C M, et al. Differentially private data release for data mining [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2011). New York: ACM, 2011: 493-501

[89] Smith A. Privacy-preserving statistical estimation with optimal convergence rate [C] //Proc of the 43rd Annual ACM Symp on Theory of Computing (STOC 2011). New York: ACM, 2011: 813-822

[90] Cormode G, Procopiuc C M, Srivastava D, et al. Differentially private spatial decompositions [C] //Proc of the 28th IEEE Int Conf on Data Engineering (ICDE 2012). Piscataway, NJ: IEEE, 2012: 20-31

[91] Li C, Hay M, Rastogi V, et al. Optimizing linear counting queries under differential privacy [C] //Proc of the 41st Annual ACM Symp on Theory of Computing (PODS 2010). New York: ACM, 2010: 123-134

[92] Li C, Hay M, Gerome M. Data- and workload- aware algorithm for range queries under differential privacy [J]. Proceedings of the VLDB Endowment, 2014, 7(5): 341-352

[93] Yuan G, Zhang Z, Winslett M, et al. Low-rank mechanism: Optimizing batch queries under differential privacy [J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1352-1363

[94] Weitzner D, Madden S, Bruce E. Big data privacy: Exploring the future role of technology in protecting privacy [R/OL]. [2013-06-19]. http://bigdata.csail.mit.edu/node/99

[95] Chor B, Goldreich O, Kushilevitz E, et al. Private information retrieval [J]. Journal of the ACM, 1998, 45(6): 965-981

[96] Kushilevitz E, Ostrovsky R. Replication is not needed: Single database, computationally-private information retrieval [C] //Proc of the 35th Annual IEEE Symp on Foundations of Computer Science (FOCS 1997). Piscataway, NJ: IEEE, 1997: 364-373

[97] Goldreich O, Goldwasser S, Micali S. How to construct random functions [J]. Journal of the ACM, 1986, 33(4): 792-807

[98] Wang L, Meng X. Location privacy preservation in big data era: A survey [J]. Journal of Software, 2014, 25(4): 693-712 (in Chinese)
(王璐, 孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报, 2014, 25(4): 693-712)

[99] Chang Y. Single-database private information retrieval with logarithmic communication [C] //Proc of the 9th Australasian Conf on Information Security and Privacy (ACISP 2004). Berlin: Springer, 2004: 50-61

[100] Paillier P. Public-key cryptosystems based on composite degree residuosity classes [C] //Proc of the 17th Int Conf on Theory and Application of Cryptographic Techniques (Eurocrypt 1999). Berlin: Springer, 1999: 223-238

[101] Sion R, Carbunar B. On the practicality of private information retrieval [C] //Proc of the 13th Network and Distributed Systems Security Symp (NDSS 2007). Reston, Virginia: The Internet Society, 2007: 1-8

[102] Williams P, Sion R. Usable private information retrieval [C] //Proc of the 14th Network and Distributed Systems Security Symposium (NDSS 2008). Reston, Virginia: The Internet Society, 2008: 12-19

[103] Goldreich O, Ostrovsky R. Software protection and simulation on oblivious ram [J]. Journal of the ACM, 1996, 45(3): 431-473

[104] Cui Y, Widom J, Janet L. Wiener: Tracing the lineage of view data in a warehousing environment [J]. ACM Trans on Database System, 2000, 25(2): 179-227

[105] Cui Y, Widom J. Lineage tracing for general data warehouse transformations [J]. VLDB J, 2003, 12(1): 41-58

[106] Suen C H, Ko R K L, Tan Y S, et al. S2Logger: End-to-end data tracking mechanism for cloud data provenance [C] //Proc of the 12th IEEE Int Conf on Trust, Security and Privacy in Computing and Communications (TrustCom 2013). Piscataway, NJ: IEEE, 2013: 594-602

[107] Ko R K L, Jagadpramana P, Lee B S. Flogger: A file-centric logger for monitoring file access and transfers with cloud computing environments [C] //Proc of the 3rd IEEE Int Workshop on Security in e-Science and e-Research (ISSR 2011). Piscataway, NJ: IEEE, 2011: 765-771

[108] Zhao J, Suny F, Torniaix C, et al. A provenance-integration framework for distributed workflows in grid environments [C] //Proc of the 1st Workshop on Grid and Utility Computing (WGUC 2008). New York: ACM, 2008: 1-9

[109] Jacobi J. Data Provenance in Distributed Propagator Networks [C] //Proc of the 3rd Int Provenance and Annotation Workshop (IPAW 2010). Berlin: Springer, 2010: 260-264

[110] Cruz D, Mattoso M. Provenance services for distributed workflows [C] //Proc of the 8th IEEE Int Symp on Cluster Computing and the Grid. Piscataway, NJ: IEEE, 2008: 526-533

[111] Feigenbaum J, Jaggard A D, Wright R N. Towards a formal model of accountability [C] //Proc of the 14th New Security Paradigms Workshop. New York, ACM, 2011: 45-56

[112] King S T, Chen P M. Backtracking intrusions [C] //Proc of the 19th ACM Symp on Operating Systems Principles (SOSP 2003). New York, ACM, 2003: 223-236

[113] King S T, Mao Z M, Lucchetti D G, et al. Enriching Intrusion Alerts through Multi-host Causality [C] //Proc of the 11th Network and Distributed Systems Security Symp (NDSS 2005). Reston, Virginia: The Internet Society, 2005: 36-44

[114] Ahmed M, Quercia D, Hailes S. A statistical matching approach to detect privacy violation for trust-based collaborations [C] //Proc of the 1st Int Workshop on Trust, Security and Privacy for Ubiquitous Computing (WoWMoM 2005). Piscataway, NJ: IEEE, 2005: 598-602

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Executive director of China Computer Federation. His main research interests include cloud data management, Web data management, native XML databases, and flash-based databases, privacy-preserving, etc.

Zhang Xiaojian, born in 1980. PhD. Member of China Computer Federation. His main research interests include differential privacy, data mining, and graph data management.