大数据隐私保护技术综述

方滨兴[1,2]，贾焰[2]，李爱平[2]，江荣[2]
1. 北京邮电大学，北京 100876；2. 国防科学技术大学计算机学院，湖南 长沙 410073

摘要
大数据分析带来的隐私泄露问题日趋严重，如何在利用大数据为各行各业服务的同时，保护隐私数据和防止敏感信息泄露成为新的挑战。大数据具有规模大、来源多、动态更新等特点，传统的隐私保护技术大都已不再适用。为此，给出了大数据时代的隐私概念和生命周期保护模型；从大数据生命周期的发布、存储、分析和使用4个阶段出发，对大数据隐私保护中的技术现状进行了分类阐述，并对各技术的优缺点、适用范围等进行分析；对大数据隐私保护技术发展的方向和趋势进行了阐述。

关键词
大数据；隐私保护；数据发布；数据挖掘；数据访问

中图分类号：TP309　　文献标识码：A
doi: 10.11959/j.issn.2096-0271.2016001

Privacy preservation in big data: a survey

FANG Binxing[1,2], JIA Yan[2], LI Aiping[2], JIANG Rong[2]
1. Beijing University of Posts and Telecommunications, Beijing 100876, China
2. School of Computer, National University of Defense Technology, Changsha 410073, China

Abstract
Privacy disclosure issue becomes more and more serious due to big data analysis. Privacy-preserving techniques should be conductive to the big data applications while preserving data privacy. Since big data has the characteristics of huge scale, numerous sources and dynamic update, most traditional privacy preserving technologies are not suitable any more. Therefore, the concept of privacy and life cycle protection model of big data era were introduced firstly. Technical state of big data privacy preservation was elaborated from the points of view of four stages in big data life cycle, i.e. data publishing, storage, analysis and use. The relative merits and scope of application of each technology were investigated as well. Finally, some important direction and tendency of privacy preservation technologies for big data were suggested.

Key words
big data, privacy preservation, data dissemination, data mining, data access

1 引言

随着智慧城市、智慧交通、智能家居、智能电网、智慧医疗、在线社交网络、Web 3.0等数字化技术的发展，人们的衣食住行、健康医疗等信息被数字化，可以随时随地通过海量的传感器、智能处理设备等终端进行收集和使用，实现物与物、物与人、人与人等之间在任何时候、任何地点的有效连接，也促成了大数据时代的到来[1]。

大数据蕴含着巨大的商业价值，目前各行各业都在做大数据分析和挖掘，企业、运营商等在各自拥有的数据或互联网上发布的数据中发掘潜在价值，为提高自己的利润或达到其他目的服务。然而，在享受大数据挖掘得到的各种各样有价值的信息给生产、生活带来便利的同时，也不可避免地泄露了人们的隐私。例如，亚马逊公司推出了“未下单先调货”计划，利用大数据分析技术，基于对网购数据的关联挖掘分析，在用户尚未下单前预测其购物内容，提前发出包裹至转运中心，缩短配送时间，但如果处理不好，很可能会泄露大量用户的隐私；医院在给疾病控制中心等研究部门提供大数据，进行疾病预防和决策时，如果不进行数据处理，则会泄露病人的隐私信息；上市公司在发布自己财务年报或其他新产品信息时，如果不对发布的数据进行适当处理，就会给商业上的竞争者以可乘之机。

如何在不泄露用户隐私的前提下，提高大数据的利用率，挖掘大数据的价值，是目前大数据研究领域的关键问题，将直接关系到大数据的民众接受程度和进一步发展趋势。具体而言，实施大数据环境下的隐私保护，需要在大数据产生的整个生命周期中考虑两个方面：如何从大数据中分析挖掘出更多的价值；如何保证在大数据的分析使用过程中，用户的隐私不被泄露。有时数据发布者恶意挖掘大数据中的隐私信息，此种情况下，更需要加强对数据发布时的隐私保护，以达到数据利用和隐私保护二者之间的折中。

本文的主要贡献为：首先，给出了大数据隐私的概念及隐私保护的生命周期模型；然后，从大数据生命周期的4个阶段（即数据的发布、存储、分析和使用）出发，对大数据隐私保护中的技术现状和发展趋势进行了分类阐述，并对该技术的优缺点、适用范围等进行分析；最后，对大数据隐私保护技术发展的方向和趋势进行了阐述。

2 大数据隐私概念与表示模型

2.1 隐私的概念及量化

在维基百科中，隐私的定义是个人或团体将自己或自己的属性隐藏起来的能力，从而可以选择性地表达自己[1]。具体什么被界定为隐私，不同的文化或个体可能有不同的理解，但主体思想是一致的，即某些数据是某人（或团体）的隐私时，通常意味着这些数据对他们而言是特殊的或敏感的。综上所述认为，隐私是可确认特定个人（或团体）身份或其特征，但个人（或团体）不愿被暴露的敏感信息。在具体应用中，隐私即用户不愿意泄露的敏感信息，包括用户和用户的敏感数据。

例如，病人的患病数据、个人的位置轨迹信息、公司的财务信息等敏感数据都属于隐私。但当针对不同的数据以及数据所有者时，隐私的定义也会存在差别[2]。例如，保守的病人会视疾病信息为隐私，而开放的病人却不视之为隐私；小孩子的定位

1 https://en.wikipedia.org/wiki/Privacy

信息对于父母而言不是隐私，对于其他人而言却是隐私；有些用户的数据现在是隐私，可能几十年后就不是隐私。从隐私的类型划分，隐私可划分为五大类。

- 财务隐私：与银行和金融机构相关的隐私。
- 互联网隐私：使某用户在互联网上暴露该用户自己的信息以及谁能访问这些信息的能力。
- 医疗隐私：患者患病和治疗信息的保护。
- 政治隐私：用户在投票或投票表决时的保密权。
- 信息隐私：数据和信息的保护。

在隐私数据的整个生命周期中，都必须对隐私数据进行准确描述和量化，才能全面地保护隐私数据。隐私可简单描述为：隐私=（信息本体+属性）×时间×地点×使用对象。

可以看出，信息本体就是拥有隐私的用户，隐私以信息本体和属性为基础，包含时间、地点、来源和使用对象等多个因素。为了更好地管理隐私以及进行隐私计算，明确在何种情况下数据发布者、数据存储方以及数据使用者对哪些隐私数据进行保护，需要对隐私数据进行量化。在隐私数据的量化过程中，需要综合考虑用户的属性、行为、数据的属性、传播途径、利用方式等因素，并对隐私数据的计算和变更有很好的支撑。

2.2 大数据生命周期的隐私保护模型

在大数据发布、存储、挖掘和使用的整个生命周期过程中，涉及数据发布者、数据存储方、数据挖掘者和数据使用者等多个数据的用户，如图1所示。在大数据生命周期的各个阶段，大数据隐私保护模型各部分的风险和技术如下所述。

（1）数据发布

数据发布者即采集数据和发布数据的实体，包括政府部门、数据公司、网站或者用户等。与传统针对隐私保护进行的数据发布手段相比，大数据发布面临的风险是大数据的发布是动态的，且针对同一用户的数据来源众多，总量巨大，如何在数据发布时，保证用户数据可用的情况下，高效、可靠地去掉可能泄露用户隐私的内容，是亟待解决的问题。传统针对数据的匿名发布技术，包括k-匿名、l-diversity

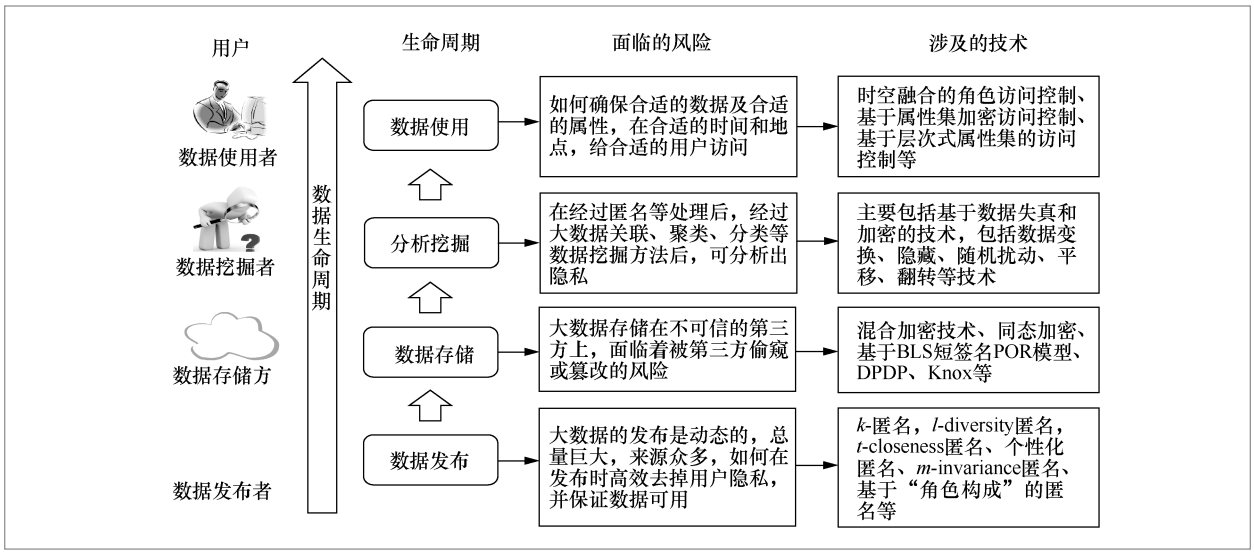


图1 大数据隐私保护生命周期模型

匿名、t-closeness匿名、个性化匿名、m-invariance匿名、基于“角色构成”的匿名等方法，可以实现对发布数据时的匿名保护。在大数据的环境下，如何对这些技术进行改进和发展，以满足大数据发布的隐私保护需求，是需要着重研究的内容。

（2）数据存储

在大数据时代，数据存储方一般为云存储平台，与传统数据的拥有者自己存储数据不同，大数据的存储者和拥有者是分离的，云存储服务提供商并不能保证是完全可信的。用户的数据面临着被不可信的第三方偷窥数据或者篡改数据的风险。加密方法是解决该问题的传统思路，但是，由于大数据的查询、统计、分析和计算等操作也需要在云端进行，为传统加密技术带来了新的挑战。比如，同态加密技术、混合加密技术、基于BLS短签名POR模型、DPDP、Knox等方法，是针对数据存储时防止隐私泄露而采取的一些方法。

（3）数据挖掘

数据挖掘者即从发布的数据中挖掘知识的人或组织，他们往往希望从发布的数据中尽可能多地分析挖掘出有价值的信息，这很可能会分析出用户的隐私信息。在大数据环境下，由于数据存在来源多样性和动态性等特点，在经过匿名等处理后的数据，经过大数据关联分析、聚类、分类等数据挖掘方法后，依然可以分析出用户的隐私。针对数据挖掘的隐私保护技术，就是在尽可能提高大数据可用性的前提下，研究更加合适的数据隐藏技术，以防范利用数据发掘方法引发的隐私泄露。现在的主要技术包括：基于数据失真和加密的方法，比如数据变换、隐藏、随机扰动、平移、翻转等技术。

（4）数据使用

数据使用者是访问和使用大数据以及从大数据中挖掘出信息的用户，通常为企业和个人，通过大数据的价值信息扩大企业利润或提供个人生活质量。在大数据的环境下，如何确保合适的数据及属性能够在合适的时间和地点，给合适的用户访问和利用，是大数据访问和使用阶段面临的主要风险。为了解决大数据访问和使用时的隐私泄露问题，现在的技术主要包括：时空融合的角色访问控制、基于属性集加密访问控制（attribute-based encryption access control，ABE）、基于密文策略属性集的加密（ciphertext policy attribute set based encryption，CP-ASBE）、基于层次式属性集的访问控制（hierarchical attribute set based encryption，HASBE）等技术。

下面针对大数据生命周期中的发布、存储、挖掘和使用4个过程中的隐私保护技术进行阐述。

3 大数据发布隐私保护技术

为了从大数据中获益，数据持有方有时需要公开发布己方数据，这些数据通常会包含一定的用户信息，服务方在数据发布之前需要对数据进行处理，使用户隐私免遭泄露。此时，确保用户隐私信息不被恶意的第三方获取是极为重要的。一般的，用户更希望攻击者无法从数据中识别出自身，更不用说窃取自身的隐私信息，匿名技术就是这种思想的实现之一。

Samarati P和Sweeney L在1998年首次提出了匿名化的概念[3]。数据发布匿名是匿名技术在数据发布中的应用，在确保所发布的信息数据公开可用的前提下，隐藏公开数据记录与特定个人之间的对应联系，从而保护个人隐私。最初，服务方仅仅删除数据表中有关用户身份的属性作为匿名实现方案。但实践表明，这种匿名处理方案是不充分的。攻击者能从其他渠道获

得包含了用户标识符的数据集，并根据准标识符连接多个数据集，重新建立用户标识符与数据记录的对应关系。这种攻击称为链接攻击（linking attack），曾多次造成重大的安全事故。

为了防御链接攻击，常见的静态匿名技术有k-匿名[4]、l-diversity匿名[5]、t-closeness匿名[6]以及以它们的相关变形为代表的匿名策略[7,8]。随着研究的进步，这些匿名策略的效果逐步提高。但是这些匿名策略以信息损失为代价，不利于数据挖掘与分析。为此，研究者随即提出了个性化匿名、带权重的匿名等一系列匿名策略。相对于对所有记录执行相同的匿名保护，这类匿名策略给予每条数据记录以不同程度的匿名保护，减少了非必要的信息损失。下面首先介绍大数据中的静态匿名技术。

3.1 大数据中的静态匿名技术

在静态匿名策略中，数据发布方需要对数据中的准标识码进行处理，使得多条记录具有相同的准标识码组合，这些具有相同准标识码组合的记录集合被称为等价组。k-匿名技术就是每个等价组中的记录个数为k个，即针对大数据的攻击者在进行链接攻击时，对于任意一条记录的攻击都会同时关联到等价组中的其他k-1条记录[4,9]。这种特性使得攻击者无法确定与特定用户相关的记录，从而保护了用户的隐私。攻击者在进行链接攻击时至少将无法区分等价组中的k条数据记录。

若等价类在敏感属性上取值单一，即使无法获取特定用户的记录，攻击者仍然可以获得目标用户的隐私信息。为此，研究者提出了l-diversity匿名策略[5]。l-diversity保证每一个等价类的敏感属性至少有l个不同的值，l-diversity使得攻击者最多以1/l的概率确认某个个体的敏感信息。这使得等价组中敏感属性的取值多样化，从而避免了k-匿名中的敏感属性值取值单一所带来的缺陷。

若等价类中敏感值的分布与整个数据集中敏感值的分布具有明显的差别，攻击者可以以一定概率猜测目标用户的敏感属性值。为此，研究者提出了t-closeness匿名策略[6]。t-closeness匿名以EMD（earth mover's distance）衡量敏感属性值之间的距离，并要求等价组内敏感属性值的分布特性与整个数据集中敏感属性值的分布特性之间的差异尽可能大。即在l-diversity基础上，t-closeness匿名考虑了敏感属性的分布问题，它要求所有等价类中敏感属性值的分布尽量接近该属性的全局分布。

上述匿名策略都会造成较大的信息损失。在进行数据使用时，这些信息损失有可能使得数据使用者做出误判[7]。不同的用户对于自身的隐私信息有着不同程度的保护要求。使用统一的匿名标准显然会造成不必要的信息损失。个性化匿名[7]技术应运而生，即可根据用户的要求对发布数据中的敏感属性值提供不同程度的隐私保护。各级匿名标准提供的匿名效果不同，相应的信息损失也不同。以此避免了不必要的信息损失，从而可显著提高发布数据的可用性。

对于大数据的使用者而言，属性与属性之间的重要程度往往并不相同。例如，对于医学研究者而言，一个患者的住址邮编或者工作单位显然不如他的年龄、家族病史等信息重要[10]。根据这种思想，带权重的匿名策略对记录的属性赋予不同的权重[8]。较为重要的属性具有较大的权重，从而提供较强的隐私保护，其他属性则以较低的标准进行匿名处理，以此尽可能减少重要属性的信息损失。

数据发布匿名最初只考虑了发布后不

再变化的静态数据，但在大数据环境中，数据的动态更新是大数据的重要特点之一。一旦数据集更新，数据发布者便需要重新发布数据，以保证数据的可用性。此时，攻击者可以通过对不同版本的发布数据进行联合分析与推理，上述基于静态数据的匿名策略将会失效，下面介绍大数据中的动态匿名技术。

3.2 大数据中的动态匿名技术

针对大数据的持续更新特性，研究者提出了基于动态数据集的匿名策略，这些匿名策略不但可以保证每一次发布的数据都能满足某种匿名标准，攻击者也将无法联合历史数据进行分析与推理。这些技术包括支持新增的数据重发布匿名技术[11]、m-invariance匿名技术[12]、基于角色构成的匿名[13]等支持数据动态更新匿名保护的策略。

Byun等人最先提出了一种支持新增的数据重发布匿名策略[11]，使得数据集即使因为新增数据而发生改变，但多次发布后不同版本的公开数据仍然能满足l-diversity准则，以保证用户的隐私。在这种匿名策略中，数据发布者需要集中管理不同发布版本中的等价类。若新增的数据集与先前版本的等价类无交集并能满足l-diversity准则，则可作为新版本发布数据中的新等价类出现，否则需要等待；而若新增的数据集与先前版本的等价类有交集，则需要插入最为接近的等价类中；若一个等价类过大，还需要对等价类进行划分，以形成新的较小的等价类。

为了在支持新增操作的同时，支持数据重发布对历史数据集的删除，m-invariance匿名策略被提出[12]。参考文献[12]的作者在研究中发现，对于任意一条记录，只要此记录所在的等价组在前后两个发布版本中具有相同的敏感属性值集合，不同发布版本之间的推理通道就可以被消除。因此，为了保证这种约束，作者在这种匿名策略中引入虚假的用户记录，这些用户记录不对应任何原始数据记录，只是为了消除不同数据版本间的推理通道而存在。在这种匿名策略中，对应于这些虚假的用户记录，作者还引入了额外的辅助表标识等价类中的虚假记录数目，以保证数据使用时的有效性。

为了支持数据重发布对历史数据集的修改，研究者注意到在不同版本的数据发布中，敏感属性可分为常量属性与可变属性两种，并针对这种情况提出HD-composition匿名策略[13]。这种匿名策略同时支持数据重发布的新增、删除与修改操作。为由于数据集的改变而发生的重发布操作提供了有效的匿名保护。

在大数据环境下，海量数据规模使得匿名技术的效率变得至关重要。研究者结合大数据处理技术，实现了一系列传统的数据匿名策略，提高了匿名技术的效率。下面介绍提高大数据匿名处理的效率技术。

3.3 大数据中的匿名并行化处理

在大数据环境下，大数据的巨规模特性使得匿名技术的效率变得至关重要。大数据环境下的数据匿名技术也是大数据环境下的数据处理技术之一，通用的大数据处理技术也能应用于数据匿名发布这一特定目的。当前，大数据环境下数据匿名技术的思想、模型与传统的数据匿名技术一致，主要的不同与问题在于使用大数据环境下的相关技术实现先前的各类数据匿名算法。

研究者结合大数据处理技术，实现了一系列传统的数据匿名策略，提高了匿名技术的效率。分布式多线程是主流

的解决思路，一类实现方案是利用特定的分布式计算框架实施通常的匿名策略，如Zhang等人[14,15]利用MapReduce分布式计算模型成功地实现了大数据集上可扩展的匿名系统；另一类实现方案是Mohammadian等人[16]将匿名算法并行化，使用多线程技术加速匿名算法的计算效率，从而节省了大数据中的匿名并行化处理的计算时间。

使用已有的大数据处理工具与修改匿名算法实现方式是大数据环境下数据匿名技术的主要趋势，这些技术能极大地提高数据匿名处理效率。除此之外，大数据环境为信息的搜集、存储与分析提供了更为强大的支持，攻击者的能力也随之提高，从而匿名保护变得更为困难，研究者需要花费更多的努力确保大数据环境下的匿名安全[17,18]。

此外，数据的多源化为数据发布匿名技术带来了新的挑战[19]。攻击者可以从多个数据源中获得足够的数据信息以对发布数据进行去匿名化。现有的匿名策略还难以防范此类攻击，有待进一步改进。

4 大数据存储隐私保护技术

云计算的出现为大数据的存储提供了基础平台，通过云服务器的计算和存储能力，对大数据的访问将更快速、更便宜、更简单和更标准化。但将敏感的数据存放在不可信的第三方服务器中存在潜在的威胁，因为云服务器提供商可能对用户的数据进行偷窥，也可能出于商业的目的与第三方共享数据或者无法保证数据的完整性。如何安全可靠地将敏感数据交由云平台存储和管理，是大数据隐私保护中必须解决的关键问题之一。

大数据存储给隐私保护带来了新的挑战，主要包括：大数据中更多的隐私信息存储在不可信的第三方中，极易被不可信的存储管理者偷窥；大数据存储的难度增大，存储方有可能无意或有意地丢失数据或篡改数据，从而使得大数据的完整性得不到保证。为解决上述挑战，应用的技术主要包括加密存储和第三方审计技术等，具体介绍如下。

4.1 大数据加密存储技术

对于含有敏感信息的大数据来说，将其加密后存储在云端能够保护用户的隐私，然而若使用传统的DES[20]、AES[21]等对称加密手段，虽能保证对存储的大数据隐私信息的加解密速度，但其密钥管理过程较为复杂，难以适用于有着大量用户的大数据存储系统。而使用传统的RSA[22]、Elgamal[23]等非对称加密手段，虽然其密钥易于管理，但算法计算量太大，不适用于对不断增长的大数据隐私信息进行加解密。数据加密加重了用户和云平台的计算开销，同时限制了数据的使用和共享，造成了高价值数据的浪费。因此，开发适用于大数据平台的快速加解密技术成为大数据隐私信息存储保护的一个重要研究方向。

Lin H Y等人[24]于2012年提出了一种针对HDFS（Hadoop分布式文件系统）的混合加密技术，该技术将对称加密和非对称加密进行了融合。当有新的隐私数据文件需要加密时，先通过非对称加密方法（AES或RC4）对该文件内容进行快速加密，并将其分布式存储于每个HDFS节点上，然后使用对称加密方法对用于加密该文件内容的密钥进行加密，并将结果存储于该数据的头文件中，以此提供对密钥的有效管理。该方法能够很好地实现对大数据隐私信息的存储保护，但是这些

加密后的隐私信息需要先经过解密才能在大数据平台中进行运算，其运算结果在存储到大数据平台时同样需要重新加密，这个加解密过程会造成很大的时间开销。

同态加密算法可以允许人们对密文进行特定的运算，而其运算结果解密后与用明文进行相同运算所得的结果一致。全同态加密算法则能实现对明文所进行的任何运算，都可以转化为对相应密文进行恰当运算后的解密结果[25,26]。将同态加密算法用于大数据隐私存储保护，可以有效避免存储的加密数据在进行分布式处理时的加解密过程，Chen X等人于2013年将全同态加密技术和MapReduce编程模型进行结合[27]，通过在reduce模块之前，增加一个在密文状态下进行计算的转换模块，使得经过全同态加密后的文件可以在不解密的情况下进行MapReduce运算，从而能够大大优化存储的大数据隐私信息的运算效率。Wang等人[28]基于代理重签名的思想，设计了一个可以有效地支持用户撤销的云端群组数据的同态解密验证方案，保护群组用户的身份隐私，且在群组用户的撤销过程中，因维护数据完整性所产生的开销主要由云端而不是用户来承担，减轻了群组在用户撤销过程中的计算和通信开销。

4.2 大数据审计技术

当用户将数据存储在云服务器中时，就丧失了对数据的控制权。如果云服务提供商不可信，其可能对数据进行篡改、丢弃，却对用户声称数据是完好的。为了防止这种危害，云存储中的审计技术则被提出。云存储审计指的是数据拥有者或者第三方机构对云中的数据完整性进行审计。通过对数据进行审计，确保数据不会被云服务提供商篡改、丢弃，并且在审计的过程中用户的隐私不会被泄露。

当前已有很多研究者对云存储中的审计进行了研究。Ateniese等人[29]提出了一种可证明的数据持有（provable data possession，PDP）模型，该模型可以对服务器上的数据进行完整性验证。该模型先从服务器上随机采样相应的数据块，并生成持有数据的概率证据。客户端维持着一定数量的元数据，并利用元数据来对证据进行验证。在该模型中，挑战应答协议传输的数据量非常少，因此所耗费的网络带宽较小。

Juels等人[30]提出可恢复证明（proof of retrievability，POR）模型，该模型主要利用纠错码技术和消息认证机制来保证远程数据文件的完整性和可恢复性。在该模型中，原始文件首先被纠错码编码并产生对应标签，编码后的文件及标签被存储在服务器上。当用户选择服务器上的某个文件块时，可以采用纠错码解码算法来恢复原始文件。POR模型面临的挑战在于需要构建一个高效和安全的系统来应对用户的请求，Shacham等人[31]改进了POR模型。他们的模型构建基于BLS短签名（BLS short signature），即基于双线性对构造的数字签名方案，该模型拥有很短的查询和响应时间。

上述方案都只能适用于静态数据的审计，无法支持对动态数据的审计。Ateniese等人[32]改进了PDP模型，该模型基于对称密钥加密算法，并且支持数据的动态删除和修改。Erway等人[33]改进了PDP模型，提出了DPDP模型。该模型扩展了传统的PDP模型以支持存储数据的更新操作，该操作的时间复杂度为$O(1)$到$O(\lg n)$。Wang Q等人[34]改进了前人的POR模型，通过引入散列树来对文件块标签进行认证。同时，他们的方法也支持对数据的动态操作，但是此方案无法对用户的隐私进行有效的保护。

Wang C等人[35]提出了一种支持隐私保护的审计方案。他们认为第三方审计（third party auditor, TPA）应该满足如下要求：一是第三方审计能够高效地完成对数据的审计，并且不为用户带来多余的负担；二是第三方审计不能为用户隐私带来脆弱性。他们提出的方法基于公钥加密和同态认证，能够在保护用户隐私的情况下完成公开审计。Wang B Y等人[36]首次提出一种用于对云中共享数据进行审计的隐私保护策略。他们在对数据的审计过程中利用环形签名来对数据完整性进行验证。此策略能够很好地对用户的隐私进行保护。其不足之处在于通信开销比较大。Wang B Y等人[37]还提出了一种名为Knox的云中数据的隐私保护策略。该策略利用群组签名来构造同态认证，使得第三方审计机构不需要从云中获取整个数据即能完成对数据完整性的审计。

随着大数据时代的发展，可以预见到，未来存储在云中的数据会越来越多，这也为大数据审计技术带来了巨大的挑战。在未来的研究中，以下几个方向也许值得研究者们关注：一个是云中数据量越来越大、数据种类越来越丰富，如何提供更加高效、安全的审计服务值得关注；另一个是随着人们在线上的交互越来越频繁，云中数据动态操作可能更加频繁，如何应对如此频繁的数据动态操作也值得研究者们关注。

5 大数据挖掘隐私保护技术

随着技术的进步，数据挖掘过程中的隐私保护问题逐渐走进了人们的视线，尤其是在大数据时代，成为数据挖掘界一个新的研究热点。隐私保护数据挖掘，即在保护隐私前提下的数据挖掘，其主要关注点有两个：一是对原始数据集进行必要的修改，使得数据接收者不能侵犯他人隐私；二是保护产生模式，限制对大数据中敏感知识的挖掘。

大数据中的隐私保护数据挖掘依旧处于起步阶段，大数据的种种特性给数据挖掘中的隐私保护提出了不少难题和挑战：对于大规模数据集而言，还没有有效并且可扩展的隐私保护技术[38]；分布式存储环境下，如何有效地对用户信息进行隐藏，还没有合适的解决方法[39]；大数据背景下，如何快速、有效地区分不同数据挖掘应用的领域背景存在一定的困难，而不同应用对于隐私保护的要求也是不同的[40]。下面主要从频繁模式挖掘、分类和聚类3个方面讨论限制敏感信息的知识挖掘技术。

5.1 关联规则的隐私保护

关联规则的隐私保护主要有两类方法[41]：第一类是变换（distortion），即修改支持敏感规则的数据，使得规则的支持度和置信度小于一定的阈值而实现规则的隐藏；第二类是隐藏（blocking），该类方法不修改数据，而是对生成敏感规则的频繁项集进行隐藏。这两类方法都对非敏感规则的挖掘具有一定的负面影响。下面分别对这两类方法进行介绍。

在变换方法中，Atallah等人[42]证明了采用变换方法进行关联规则挖掘是一个NP难问题。他们将敏感规则相关的支持数据进行变换，从而降低敏感规则的支持度和置信度。Oliveira等人[43]提出了一种对于数据进行变换的方法。首先，对于每一条敏感规则rp_i，找到对应的敏感事务$T[rp_i]$；其次，对于每一条敏感规则，将其中对规则支持度最低的项设为牺牲项$Victim_{rpi}$；然后，根据事先设定的暴露阈值ψ，对每一条敏感规则计算其需要隐藏的事务数

量$NumTrans_{rpi}$；最后进行数据重构，对于每一条敏感规则rp_i，对$T[rp_i]$中的事务按照冲突程度升序排序，选取$T[rp_i]$中前$NumTrans_{rpi}$个事务$TransToSanitize$，对于数据集D中的事务t，如果$t \in TransToSanitize$，则将t中牺牲项$Victim_{rpi}$替换之后置入新的数据集D'中。

Chang等人[44]提出了关联规则隐藏的方法，这类方法的特点是不对数据进行修改，而是将敏感规则的相关数据进行隐藏（标记为未知，常用问号替代），保持了数据的真实性。Aggarwal等人[45]研究了如何隐藏一个最小集合，使得对方无法通过数据挖掘的方法预测出敏感信息。他们提出了一种简洁的问题建模方法，并设计了一个有效的启发式算法。首先挖掘出对抗规则，接着推导出隐私集合。在广泛的人工数据集和实际数据集上的测试表面，使用该方法对数据处理后，数据集对数据挖掘算法的各项参数不敏感，从而可以有效保护隐私。

5.2 分类结果的隐私保护

分类方法的结果通常可以发现数据集中的隐私敏感信息，因此需要对敏感的分类结果信息进行保护。这类方法的目标是在降低敏感信息分类准确度的同时，不影响其他应用的性能。

Agrawal等人[46]采用随机扰动的方式对原始数据进行加密，以实现分类结果的隐私保护。算法首先对数据进行随机扰动，对于原始数据$X_1, X_2, \cdots, X_n$，将其看成满足特定分布的随机变量X，为了隐藏原始数据值，在每个原始值上添加一个服从随机分布Y的随机数$Y_1, Y_2, \cdots, Y_n$，则扰动后的数据为X_1+Y_1，$X_2+Y_2, \cdots, X_n+Y_n$的形式，记为Z；然后对数据进行恢复，数据恢复即已知随机变量分布Y、Z以及$X+Y=Z$的关系，用Y、Z的值估计X的过程，应用贝叶斯公式可以得到原始数据估计的迭代方程，从而得到原始数据的近似X'；最后是分类过程，得到了原始数据的模糊近似X'之后即可应用普通的分类方法，如利用决策树对数据进行分类，降低分类的准确度。

Moskowitz L M等人[47]设计了名为“Rational Downgrader”的隐私保护系统，该系统着力于降低信息公开过程中隐私泄露的程度，确保普通用户无法通过已经或将要公开的信息推测出应被保护的隐私信息。该系统主要包括3个部分：其决策部分用于评估哪些分类规则可能被推测出来；示警部分用于测定已经泄露的隐私信息量；降级约束部分降低敏感结果的分类准确度。

Chang等人[48]提出了一种新的范式，以处理由降级（downgrading）引发的隐私信息推测问题。这种新范式包含两大部分：对隐私信息推测问题应该采用决策树进行分析以及对降级问题进行约束、限制。其中，他们使用了一种新的热力学激励的方式来处理对分类规则进行推理的过程，这些被推理的规则来源于部分公开的数据。

5.3 聚类结果的隐私保护

与分类结果的隐私保护类似，保护聚类的隐私敏感结果也是当前研究的重要内容之一。Oliveira等人[49]对发布的数据采用平移、翻转等几何变换的方法进行变换，以保护聚类结果的隐私内容。此方法首先是对原始数据进行几何变换，以对敏感信息进行隐藏，然后是聚类过程，经过几何变换后的数据可以直接应用传统的聚类算法（如K近邻）进行聚类。他们提出的方法在聚类准确度和保护隐私方面达到了较好的平衡。

Vaidya等人[50]提出了一种分布式

K-means聚类方法，该方法专门面向不同站点上存有同一实体集合的不同属性的情况。使用此聚类方法，每个站点可以学习对每个实体进行聚类，但在学习过程中并不会获知其他站点上所存属性的相关信息，从而在信息处理的过程中保障了数据隐私。

6 大数据访问控制技术

大数据访问控制技术主要用于决定哪些用户可以以何种权限访问哪些大数据资源，从而确保合适的数据及合适的属性在合适的时间和地点，给合适的用户访问，其主要目标是解决大数据使用过程中的隐私保护问题。早期的访问控制技术，如自主访问控制（discretionary access control，DAC）[51]、强制访问控制（mandatory access control，MAC）[52]等都面向封闭环境，访问控制的粒度都比较粗，难以满足大数据时代开放式环境下对访问控制的精细化要求。

大数据给传统访问控制技术带来的挑战如下。

- 大数据的时空特性，大数据下的访问控制模型需要在传统访问控制的基础上，充分考虑用户的时间信息和位置信息。
- 在大数据时代的开放式环境下，用户来自于多种组织、机构或部门，单个用户又通常具有多种数据访问需求[53]，如何合理设定角色并为每个用户动态分配角色是新的挑战。
- 大数据面向的应用需求众多，不同的应用需要不同的访问控制策略。以社交网站为例：对于用户个人主页的数据，需要基于用户社交关系的访问控制；对于网站数据，需要基于用户等级的访问控制等。

传统的访问控制方式，包括自主访问控制和强制访问控制技术，难以应对上述挑战。因此，大数据时代的访问控制技术主要包括基于角色的访问控制和基于属性的访问控制方法。

6.1 基于角色的访问控制

基于角色的访问控制（role-based access control，RBAC）[54]中，不同角色的访问控制权限不尽相同。通过为用户分配角色，可实现对数据的访问权限控制。由此，在基于角色的访问控制中，角色挖掘是前提。通常，角色是根据工作能力、职权及责任确定的。大数据场景下的角色挖掘，需要大量人工参与角色定义、角色划分及角色授权等问题，衍生出了所谓角色工程（role engineering）[55]。角色工程的最终目的是根据个体在某一组织内所担当的角色或发挥的作用来实现最佳安全管理。有效的角色工程可以为用户权限提供最优分配、鉴别异常用户、检测并删除冗余或过量的角色、使角色定义及用户权限保持最新、降低随之发生的各类风险等。大数据时代，可用于角色挖掘的数据丰富多样，对角色权限的配置也更加灵活复杂。一方面需要通过挖掘己方数据，合理配置权限，实现数据的访问可控；另一方面，需要挖掘可收集到的对方数据，找出重要目标角色，以便重点关注。因此，大数据下的角色工程需要从攻击和防护的角度综合考虑。

RBAC最初也主要应用于封闭环境之中。针对大数据时空关联性，一些研究者提出将时空信息融合到RBAC当中。如Ray等人提出了LARB（location-aware role-based）访问控制模型，在RBAC的基础之上引入了位置信息，通过考虑用户的位置来确定用户是否具有访问数据的权限[56]。Damiani等人提出的GEO-RBAC，也在分配用户角色时综合考虑了用户的空间位置信息[57]。张颖君等人提出的基于尺度的时空RBAC访问控制模型，引入了尺度的概

念，使得访问控制策略的表达能力得到增强，同时也增强了模型的安全性[58]。

随着大数据环境下角色规模的迅速增长，设计算法自动实现角色的提取与优化逐渐成为近年来的研究热点。参考文献[59]尝试将角色最小化，即找出能满足预定义的用户—授权关系的一组最小角色集合。参考文献[60]提出最小扰动混合角色挖掘方法，首先以自顶向下的方法预先定义部分角色，然后以自底向上的方法挖掘候选角色集合。自动化角色挖掘大大减少了人工工作量，但也面临时间复杂度高的问题，部分问题甚至属于NP完全问题。参考文献[61]提出了一种简单的启发式算法SMA来简化角色求解。参考文献[62]针对大数据及噪声数据场景，提出选择稳定的候选角色，并进一步将角色挖掘问题分解以降低复杂度。

大数据时代的访问控制应用场景广泛，需求也不尽相同。一些研究通过广泛收集研究对象的应用数据，试图挖掘出其中的关键角色，从而有针对性地采取处理措施。参考文献[63]提出在RBAC的基础上增加责任的概念，即responsibility-RBAC，对用户职责进行显式确认，以根据实际应用场景优化角色的数量。

6.2 基于属性的访问控制

基于属性的访问控制（attribute-based access control，ABAC）[64]通过将各类属性，包括用户属性、资源属性、环境属性等组合起来用于用户访问权限的设定。RBAC以用户为中心，而没有将额外的资源信息，如用户和资源之间的关系、资源随时间的动态变化、用户对资源的请求动作（如浏览、编辑、删除等）以及环境上下文信息进行综合考虑。而基于属性的访问控制ABAC通过对全方位属性的考虑，可以实现更加细粒度的访问控制。

大数据环境下，越来越多的信息存储在云平台上。根据云平台的特点，基于属性集加密访问控制[65]、基于密文策略属性集的加密[66]、基于层次式属性集合的加密[67]等相继被提出。这些模型都以数据资源的属性加密作为基本手段，采用不同的策略增加权限访问的灵活性。如HASBE通过层次化的属性加密，可以实现云平台上数据的更加细粒度的访问控制，层次化也使得模型更加灵活，具有更好的可扩展性。除了提供属性加密访问控制之外，ABAC也被当作云基础设施上访问控制中的一项服务[68]。

ABE将属性与密文和用户私钥关联，能够灵活地表示访问控制策略。但对于存储在云端的大数据，当数据拥有者想要改变访问控制策略时，需要先将加密数据从云端取回本地，解密原有数据，之后再使用新的策略重新加密数据，最后将密文传回云端。在这一过程中，密文需要来回传输，会消耗大量带宽，从而引发异常，引起攻击者的注意[69]，对数据的解密和重新加密也会使得计算复杂度显著增大。为此，Yang等人提出了一种高效的访问控制策略动态更新方法[70]。当访问控制策略发生变化时，数据拥有者首先使用密钥更新策略UKeyGen生成更新密钥UK_m，并将其和属性变化情况（如增加、减少特定属性）一起发送到云端。之后，在云端上按照密文更新策略CTUpdate对原有的密文进行更新，而不用对原有密文进行解密。

云端代理重加密将基于属性的加密与代理重加密技术结合，实现云中的安全、细粒度、可扩展的数据访问控制[71-73]。新的用户获取授权或原有用户释放授权时的重加密工作由云端代理，减轻数据拥有者的负担。同时对数据拥有者来说，云端可能并非是完全可信的，在利用云端进行代理重加密的同时还应防止数据被云端窥探。

用户提交给云的是密文，云端无法解密，云端利用重加密算法转换为另一密文，新的密文只能被授权用户解密，而在整个过程中云端服务器看到的始终是密文，看不到明文。云中用户频繁地获取和释放授权，使得数据密文重加密工作繁重，由云端代理重加密工作，可以大大减轻数据拥有者的负担。同时，云端无法解密密文，也就无法窥探数据内容。

Sun等人[74]提出了支持高效用户撤销的属性关键词搜索方案，实现了可扩展且基于用户制定访问策略的高细粒度搜索授权，通过代理重加密和懒惰重加密技术，将用户撤销过程中系统繁重的密钥更新工作交给半可信的云服务器。Wang等人[75]针对多中心云计算环境的数据安全访问特点，将多中心属性加密和外包计算相结合，提出了一种轻量级的安全的访问控制方案。该方案具有解密密钥短、加解密计算开销小等优势，适用于轻量级设备。该方案可以无缝应用到群组隐私信息保护中，实现了群组成员之间的隐私信息定向发布和共享、群组外的隐私信息保护功能。

大数据为访问控制带来了诸多挑战，但也暗藏机遇。随着计算能力的进一步提升，无论是基于角色的访问控制还是基于属性的访问控制，访问控制的效率将得到快速提升。同时，更多的数据将被收集起来用于角色挖掘或者属性识别，从而可以实现更加精准、更加个性化的访问控制。总体而言，目前专门针对大数据的访问控制还处在起步阶段，未来将角色与属性相结合的细粒度权限分配将会有很大的发展空间。

7 结束语

如何在不泄露用户隐私的前提下，提高大数据的利用率，挖掘大数据的价值，是目前大数据研究领域的关键问题。本文首先介绍了大数据带来的隐私保护问题，然后介绍了大数据隐私的概念和大数据生命周期的隐私保护模型，接着从大数据生命周期的发布、存储、分析和使用4个阶段出发，对大数据隐私保护中的技术现状和发展趋势进行了分类阐述，对该技术的优缺点、适用范围等进行分析，探索了大数据隐私保护技术进一步发展的方向。

参考文献:

[1] 方滨兴，刘克，吴曼青,等. 大搜索技术白皮书[R/OL]. (2015-01-06)[2015-05-23]. http://wenku.baidu.com/link?url=gqavgz5O7VROHQgJH4_egRVHB_JtcskcX-vWvRgEdzhfMuyidxhO_kdGemK8Qve-z0z-dBIJRpSqZj7oCYLd0i-2iT1mXE2B1B5p4nPW0TO.
FANG B X, LIU K, WU M Q, et al. White paper on big search[R/OL]. (2015-01-06)[2015-05-23]. http://wenku.baidu.com/link?url=gqavgz5O7VROHQgJH4_egRVHB_JtcskcX-vWvRgEdzhfMuyidxhO_kdGemK8Qve-z0z-dBIJRpSqZj7oCYLd0i-2iT1mXE2B1B5p4nPW0TO.

[2] 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.
ZHOU S G, LI F, TAO Y F, et al. Privacy preservation in database applications: a survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861.

[3] SAMARATI P, SWEENEY L. Generalizing data to provide anonymity when disclosing information[C]// Proceedings of the 17th ACM Sigact-Sigmod-Sigart Symposium on Principles of Database System, June 1-3, 1998, Seattle, Washington, USA. New York: ACM Press, 1998.

[4] SWEENY L. k-anonymity: amodel for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge

Based Systems, 2012, 10(5): 557-570.
[5] BARBARO M, ZELLER T. A face is exposed for AOL searcher No. 4417749[N/OL]. New York Times, (2006-08-09)[2013-09-10]. http://www.nytimes.com/2006/08/09/technology/09aol.html.
[6] NARAYANAN A, SHMATIKOV V. How to break anonymity of the netflix prize dataset[J]. Eprint Arxiv Cs, 2006, arXiv:cs/0610105.
[7] MACHANAVAJJHALA A, GEHRKE J, KIFER D, et al. l-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 24.
[8] LI N, LI T, VENKATASUBRAMANIAN S. t-closeness: privacy beyond k-anonymity and l-diversity[C]// Proceedings of IEEE 23rd International Conference on Data Engineering, April 11-15, 2007, Istanbul, Turkey. Piscataway: IEEE Press, 2007: 106-115.
[9] NIU B, LI Q H, ZHU X Y, et al. Enhancing privacy through caching in location-based services[C]//Proceedings of IEEE INFOCOM, April 26-May 1, 2015, Hong Kong, China. Piscataway: IEEE Press, 2015: 1017-1025.
[10] LI A, JIN S, ZHANG L, et al. A sequential decision-theoretic model for medical diagnostic system [J]. Technology and Health Care, 2015, 23(s1): S37-S42.
[11] BYUN J W, SOHN Y, BERTINO E, et al. Secure anonymization for incremental dataset[C]//Proceedings of the 3rd VLDB Workshop on Secure Data Management (SDM), September 10-11, 2006, Seoul, Korea. [S.l.: s.n.], 2006.
[12] XIAO X K, TAO Y F. m-invariance: towards privacy preserving re-publication of dynamic datasets[C]// Proceedings of the 2007, ACM SIGMOD International Conference on Management of Data, June 12-14, 2007, Beijing, China. New York: ACM Press, 2007: 689-700.
[13] BU Y Y, FU A W C, WONG R C W, et al. Privacy preserving serial data publishing by role composition[C]// Proceedings of the 34th International Conference on Very Large Data Bases, August 23-28, 2008, Auckland, New Zealand. [S.l.: s.n.], 2008: 845-856.
[14] ZHANG X, LIU C, NEPAL S, et al. A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud [J]. Journal of Computer & System Sciences, 2014, 80(5): 1008-1020.
[15] ZHANG X, LIU C, NEPAL S, et al. Combining top-down and bottom-up: scalable sub-tree anonymization over big data using MapReduce on cloud [J]. IEEE International Conference on Trust, 2013, 52(1): 501-508.
[16] MOHAMMADIAN E, NOFERESTI M, JALILI R. FAST: fast anonymization of big data streams[C]// Proceedings of the 2014 International Conference on Big Data Science and Computing, Aug 4-7, 2014,Beijing, China. [S.l.:s.n.], 2014.
[17] SEDAYAO J, BHARDWAJ R, GORADE N. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues[C]// Proceedings of the 3rd International Congress on Big Data, June 27-July 2, 2014, Anchorage, Alaska, USA. Piscataway: IEEE Press, 2014.
[18] SUN G Z, WEI S, XIE X. De-anonymization technology and applications in the age of big data [J]. Information &Communications Technologies, 2013(6): 52-57.
[19] NARAYANAN A, SHMATIKOV V. Robust de-anonymization of large sparse datasets[C]//Proceedings of the 2008 IEEE Symposium on Security and Privacy, May 18-21, 2008, Oakland, USA. Piscataway: IEEE Press, 2008: 111-122.
[20] National Bureau of Standards. Proposed federal information processing data encryption standard [J]. Creptologia, 1977, 1(3): 292-306.
[21] FIPS. Advanced encryption standard (AES): FIPS PUB 197[S/OL]. [2010-10-16]. http://wenku.baidu.com/link?url=dqgVVuI1EvKAh4fSiHu7mSAgObQji-LiI6C1_KlYWtuiIUFIZaJUZOpkcOWQMPy9U91SHgPcPrt7UWmAQmT3b8WJZ80idSjZ-qLVileRY3a.
[22] RIVEST R L, SHAMIR A, ADLERNAN L M. A method for obtaining digital signatures and public-key cryptosystems [J].

Communications of the ACM, 1978, 21(6): 120-126.
[23] ELGAMAL T. A public key cryptosystem and a signature scheme based on discrete logarithms[J]. IEEE Transactions on Information Theory, 1985, 31(4): 469-472.
[24] LIN H Y, SHEN S T, TZENG W G, et al. Toward data confidentiality via integrating hybrid encryption schemes and Hadoop distributed file system[C]// Proceedings of IEEE 26th International Conference on Advanced Information Networking and Applications (AINA), March 26-29, 2012, Fukuoka, Japan. Washington DC: IEEE Computer Society Press, 2012: 740-747.
[25] GENTRY C. A fully homomorphic encryption scheme [D]. Palo Alto: Stanford University, 2009.
[26] VAN DIJK M, GENTRY C, HALEVI S, et al. Fully homomorphic encryption over the integers[C]// Proceedings of the 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, May 30-June 3, 2010, Riviera, French. New York: Springer Berlin Heidelberg, 2010: 24-43.
[27] CHEN X, HUANG Q. The data protection of MapReduce using homomorphic encryption[C]// Proceedings of the 4th IEEE International Conference on Software Engineering and Service Science (ICSESS), May 23-25, 2013, Beijing, China. Piscataway: IEEE Press, 2013: 419-421.
[28] WANG B Y, LI B C, LI H. Public auditing for shared data with efficient user revocation in the cloud[C]// Proceedings of IEEE INFOCOM, April 26-May 1, 2015, Hong Kong, China. Piscataway: IEEE Press, 2015: 2904-2912.
[29] ATENIESE G, BURNS R, CURMOLA R, et al. Provable data possession at untrusted stores[J]. ACM Conference on Computer & Communications Security, 2007, 14(1): 598-609.
[30] JUELS A, KALISKI B S. PORs: proofs of retrievability for large files[C]// Proceedings of the 14th ACM Conference on Computer and Communications Security, October 29-November 2, 2007, Alexandria, VA, USA. New York: ACM Press, 2007: 584-597.
[31] SHACHAM H, WATERS B. Compact proofs of retrievability[J]. Journal of Cryptology, 2013, 26(3): 442-483.
[32] ATENIESE G, PIETRO R, MANCIN L V, et al. Scalable and efficient provable data possession[C]// Proceedings of International Conference on Security & Privacy in Communication Networks, September 22-25, 2008, Istanbul, Turkey. New York: ACM Press, 2008.
[33] ERWAY C, KüPçü A, PAPAMANTHOU C, et al. Dynamic provable data possession[C]// Proceedings of the 16th ACM Conference on Computer and Communications Security, November 9-13, 2009, Chicago, IL, USA. New York: ACM Press, 2009: 213-222.
[34] WANG Q, WANG C, LI J, et al. Enabling public verifiability and data dynamics for storage security in cloud computing[C]// Proceedings of ESORICS, September 21-25, 2009, Saint Malo, France. [S.l.:s.n.], 2009: 355-370.
[35] WANG C, WANG Q, REN K, et al. Privacy-preserving public auditing for data storage security in cloud computing[C]// Proceedings of IEEE INFOCOM, March 15-19, 2010, San Diego, CA, USA. Piscataway: IEEE Press, 2010: 525-533.
[36] WANG B Y, LI B C, LI H. Oruta: privacy preserving public auditing for shared data in the cloud[C]//Proceedings of IEEE 5th International Conference on Cloud Computing, November 22-24, 2012, Honolulu, Hawaii, USA. Piscataway: IEEE Press, 2012: 295-302.
[37] WANG B Y, LI B C, LI H. Knox: privacy preserving auditing for shared data with large groups in the cloud[C]// Proceedings of the 10th International Conference on Applied Cryptography and Network Security, June 26-29, 2012, Singapore. Berlin: Springer, 2012.
[38] THURAISINGHAM B. Big data security and privacy[C]//Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, March 2-4, 2015, San Antonio, TX, USA. New York: ACM

Press, 2015: 279-280.
[39] WONG R. Big data privacy[J]. J Inform Tech SoftwEng, 2012(2): e114.
[40] WU X, ZHU X, WU G Q, et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97-107.
[41] AGGARWAL C C, PHILIP S Y. A General Survey of Privacy-Preserving Data Mining Models and Algorithms[M]. New York: Springer US, 2008.
[42] ATALLAH M, BERTINO E, ELMAGARMID A, et al. Disclosure limitation of sensitive rules[C]// Proceedings of Workshop on Knowledge and Data Engineering Exchange, November 7, 1999, Chicago, IL, USA. Piscataway: IEEE Press, 1999: 45-52.
[43] OLIVEIRA S R M, ZAIANE O R. Privacy preserving frequent itemset mining[C]// Proceedings of IEEE International Conference on Data Mining, Japan, December 9-12, 2002, Maebashi City. Piscataway: IEEE Press, 2002: 43-54.
[44] CHANG L W, MOSKOWITZ I S. An Integrated Framework for Database Inference and Privacy Protection[M]. Ifip Tc11/ Wg113 Fourteenth Working Conference on Database Security: Data & Application Security. New York: Springer US, 2000: 161-172.
[45] AGGARWAL C, PEI J, ZHANG B. A framework for privacy preservation against adversarial data mining[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, Philadelphia, USA. New York: ACM Press, 2006.
[46] AGRAWAL R, SRIKANT R. Privacy-preserving data mining[J]. ACM SIGMOD Record, 2000, 29(2): 439-450.
[47] MOSKOWITZ L W, CHANG I S. A Decision Theoretical Based System for Information Downgrading[R/OL]. (2011-08-27)[2015-11-20]. http://wenku.baidu.com/link?url=JAg4rujC4hcwRVbIulvyqgkMJaP fMQ41JAr8v4zfRmZwXWwBNndmDUm10WAIvXYEvlCWb2m34GnIBkADnLpgm8za3iyAHiDnChiaPZwthAW.
[48] CHANG L W, MOSKOWITZ I S. Parsimonious downgrading and decision trees applied to the inference problem[C]// Proceedings of the 1998 Workshop on New Security Paradigms, Charlottesville, Virginia, USA, 1998. New York: ACM Press, 1998: 82-89.
[49] OLIVEIRA S R M, ZAIANE O R. Privacy preserving clustering by data transformation[J]. Journal of Information and Data Management, 2010, 1(1): 37.
[50] VAIDYA J, CLIFTON C. Privacy-preserving k-means clustering over vertically partitioned data[C]// Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington DC, USA. New York: ACM Press, 2003: 206-215.
[51] SANDHU R S, SAMARATI P. Access control: principle and practice[J]. IEEE Communications Magazine, 1994, 32(9): 40-48.
[52] SANDHU R S. Lattice-based access control models[J]. Computer, 1993, 26(11): 9-19.
[53] ZHANG W, LI A, CHEEMA M, et al. Probabilistic n-of-N skyline computation over uncertain data streams[J]. World Wide Web, 2015, 18(5): 1331-1350.
[54] SANDHU R S, COYNE E J, FEINSTEIN H L, et al. Role-based access control models[J]. Computer, 1996(2): 38-47.
[55] KUHLMANN M, SHOHAT D, SCHIMPF G. Role mining-revealing business roles for security administration using data mining technology[C]// Proceedings of the 8th ACM Symposium on Access Control Models and Technologies, June 2-3, 2003, Como, Italy. New York: ACM Press, 2003: 179-186.
[56] RAY I, KUMAR M, YU L J. LRBAC: a location-aware role-based access control model[C]// Proceedings of the 2nd International Conference on Information Systems Security, December 19-21, 2006, Kolkata, India. New York: Springer US, 2006: 147-161.
[57] DAMIANI M L, BERTINO E, CATANIA B, et al. Geo-rbac: a spatially aware rbac[J]. ACM Transactions on Information and System Security (TISSEC), 2007, 10(1): 2.

[58] 张颖君, 冯登国. 基于尺度的时空RBAC模型[J]. 计算机研究与发展, 2015, 47(7): 1252-1260.
ZHANG Y J, FENG D G. A role-based access control model based on space, time and scale[J]. Journal of Computer Research and Development, 2010, 47(7): 1252-1260.

[59] ENE A, HORNE W, MILOSAVLJEVIC N, et al. Fast exact and heuristic methods for role minimization problems[C]// Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, June 11-13, 2008, Estes Park, CO, USA. New York: ACM Press, 2008: 1-10.

[60] 翟志刚, 王建东, 曹子宁, 等. 最小扰动混合角色挖掘方法研究[J]. 计算机研究与发展, 2015, 50(5): 951-960.
ZHAI Z G, WANG J D, CAO Z N, et al. Hybrid role mining methods with minimal perturbation[J]. Journal of Computer Research and Development, 2015, 50(5): 951-960.

[61] BLUNDO C, CIMATO S. A simple role mining algorithm[C]//Proceedings of the 2010 ACM Symposium on Applied Computing, March 22-26, 2010, Sierre, Switzerland. New York: ACM Press, 2010: 1958-1962.

[62] NINO V V. Role mining over big and noisy data theory and some applications[D]. Roma: Roma Tre University, 2011.

[63] FELTUS C, PETIT M, SLOMAN M. Enhancement of business it alignment by including responsibility components in RBAC[C]//Proceedings of the 5th International Workshop on Business/IT Alignment and Interoperability BUSITAL, June 2010, Hammamet, Tunisia. [S.l.:s.n.], 2010.

[64] Attribute-based access control[EB/OL]. [2015-12-08]. https://en.wikipedia.org/wiki/Attribute-based_access_control.

[65] GOYAL V, PANDEY O, SAHAI A, et al. Attribute-based encryption for fine-grained access control of encrypted data[C]// Proceedings of the 13th ACM Conference on Computer and Communications Security, October 30-November 3, 2006, Alexandria, Virginia, USA. New York: ACM Press, 2006: 89-98.

[66] BOBBA R, KHURANA H, PRABHAKARAN M. Attribute-sets: a practically motivated enhancement to attribute-based encryption[C]// Proceedings of the 14th European Symposium on Research in Computer Security, September 21-25, 2009, Saint-Malo, France. [S.l.: s.t.], 2009: 587-604.

[67] WAN Z, LIU J E, DENG R H. HASBE: a hierarchical attribute-based solution for flexible and scalable access control in cloud computing[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 743-754.

[68] JIN X. Attribute-based access control models and implementation in cloud infrastructure as a service[D]. San Antonio: The University of Texas at San Antonio, 2014.

[69] LI A, HAN Y, ZHOU B, et al. Detecting hidden anomalies using sketch for high-speed network data stream monitoring[J]. Applied Mathematics and Information Sciences, 2012, 6(3): 759-765.

[70] YANG K, JIA X, REN K, et al. Enabling efficient access control with dynamic policy updating for big data in the cloud[C]// Proceedings of IEEE INFOCOM, April 27-May 2, 2014, Toronto, Canada. Piscataway: IEEE Press, 2014: 2013-2021.

[71] BLAZE M, BLEUMER G, STRAUSS M. Divertible protocols and atomic proxy cryptography[C]// Proceedings of International Conference on the Theory and Application of Cryptographic Techniques Espoo, May 13, 1998, Finland. Berlin: Springer, 1998: 127-144.

[72] LI A, XU J, GAN L, et al. An efficient approach on answering top-k queries with grid dominant graph index[C]// Proceedings of the 15th Asia-Pacific Web Conference, April 4-6, 2013, Sydney, Australia. Berlin: Springer, 2013: 804-814.

[73] ZHANG W M, CHEN B, YU N H. Improving various reversible data hiding schemes via optimal codes for binary covers[J]. IEEE Transactions on Image Processing, 2012, 21(6): 2991-3003.

[74] SUN W H, YU S C, LOU W J, et al.

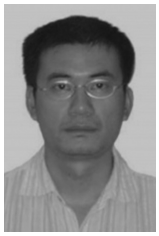
Protecting your right: attribute-based keyword search with fine-grained owner-enforced search authorization in the cloud[C]//Proceedings of IEEE Conference on Computer Communications, April 27-May 2, 2014, Toronto, Ontario, Canada. Piscataway: IEEE Press, 2014.

[75] WANG Y C, LI F H, XIONG J B, et al. Achieving lightweight and secure access control in multi-authority cloud[C]//Proceedings of the 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, August 20-22, 2015, Helsinki, Finland. Piscataway: IEEE Press, 2015: 459-466.

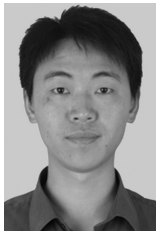
作者简介

方滨兴(1960-),男,博士,中国工程院院士,主要研究方向为大数据、计算机网络和信息安全。

贾焰(1960-),女,博士,国防科学技术大学教授,主要研究方向为大数据、网络信息安全和社交网络。



李爱平(1974-),男,博士,国防科学技术大学研究员,主要研究方向为大数据分析、数据挖掘和网络信息安全。



江荣(1984-),男,博士,国防科学技术大学助理研究员,主要研究方向为隐私保护和网络信息安全。

收稿日期:2015-12-24