doi:10.11959/j.issn.1000-436x.2016238

云环境中数据安全去重研究进展

熊金波[1]，张媛媛[1]，李凤华[2,3]，李素萍[1]，任君[1]，姚志强[1,3]

(1. 福建师范大学软件学院，福建 福州 350117；2. 中国科学院信息工程研究所信息安全国家重点实验室，北京 100093；
3. 福建省公共服务大数据挖掘与应用工程技术研究中心，福建 福州 350117)

摘 要：为了提高云存储效率和节约网络通信带宽，需要对云端同一数据的多个副本执行重复性检测与去重操作，而云环境下的密文数据阻碍了数据安全去重的实施，这一问题迅速引起学术界和产业界的广泛关注，成为研究热点。从安全性角度出发，分析云环境中数据安全去重的原因及面临的主要挑战，给出云数据安全去重的系统模型和威胁模型，面向云数据安全去重技术的实现机制从基于内容加密的安全去重、基于所有权证明的安全去重和隐私保护的安全去重 3 个方面对近年来相关研究工作进行深入分析和评述，并指出各种关键技术与方法的优势及存在的共性问题；最后给出云数据安全去重领域未来的研究方向与发展趋势。

关键词：安全去重；基于内容的加密；所有权证明；隐私保护；重复数据删除

中图分类号：TP309.2 **文献标识码**：A

Research progress on secure data deduplication in cloud

XIONG Jin-bo[1], ZHANG Yuan-yuan[1], LI Feng-hua[2,3], LI Su-ping[1], REN Jun[1], YAO Zhi-qiang[1,3]

(1. Faculty of Software, Fujian Normal University, Fuzhou 350117, China;
2. State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
3. Fujian Engineering Research Center of Public Service Big Data Mining and Application, Fuzhou 350117, China)

Abstract: In order to improve the efficiency of cloud storage and save the communication bandwidth, a deduplication mechanism for multi-duplicate of the same data in cloud environment was needed. However, the implement of the secure data deduplication was seriously hindered by the ciphertext in cloud. This issue has quickly aroused wide attention of academia and industry, and became a research hotspot. From a security standpoint, firstly the primary cause and the main challenges of secure data deduplication in cloud environment was analyzed, and then the deduplication system model as well as its security model was described. Furthermore, focusing on the realization mechanism of secure data deduplication, the thorough analyses were carried on and reviews for the related research works in recent years from content-based encryption, proof of ownership and privacy protection for secure deduplication, then the advantages and common problems of various key technologies and methods were summed up. Finally, the future research directions and development trends on secure data deduplication in cloud was given.

Key words: secure deduplication, content-based encryption, proof of ownership, privacy protection, data deduplication

1 引言

随着云计算、大数据技术的不断进步以及新型信息传播方式和个性化服务模式不断发展，越来越多的用户选择将数据外包给云端进行存储和管理[1]，使云端数据量将呈爆炸式增长，人们即将进入一个

收稿日期：2016-08-10；**修回日期**：2016-10-18
通信作者：李凤华, lfh@iie.ac.cn
基金项目：国家自然科学基金资助项目(No.61402109, No.61370078, No.61502103)；福建省自然科学基金资助项目(No.2015J05120)；福建省网络安全与密码技术重点实验室(福建师范大学)开放课题基金资助项目(No.15008)；福建省高校杰出青年科研人才培育计划基金资助项目(No.2015)

Foundation Items: The National Natural Science Foundation of China (No.61402109, No.61370078, No.61502103), The Natural Science Foundation of Fujian Province (No.2015J05120), Fujian Provincial Key Laboratory of Network Security and Cryptology Research Fund (No.15008), Distinguished Young Scientific Research Talents Plan in Universities of Fujian Province(No.2015)

由数据驱动的大数据时代，一切都将被数字化。据小米云存储服务统计，2015 年末的客户量已达 9 700 万人，并且已为用户存储 405 亿张照片、5.04 亿视频，存储量超过 100 PB；另据 Gartner 数据统计，在 2013 年，全球数据量为 4.4 ZB，到 2016 年底，将会有 36%的数字内容和个人数据存储到云端服务器，预计到 2020 年，全球数据量将达到 44 ZB。面对如此庞大规模的数据量，如何经济、高效地存储成为云服务提供商最大的挑战之一。

为了提高云存储服务提供商的存储效率，节约用户带宽消耗，最直接的方法是采用数据压缩技术对原始数据压缩处理后再上传到云端。然而，对于同一份数据文件，不同用户可能单独采用不同的压缩技术，如霍夫曼编码、字典编码等，生成不同的压缩文件，产生多数据副本共存，反而使云端存储压力更大，也增加了用户带宽的消耗[2]。因此，迫切需要数据副本重复性检测与删除机制，即数据去重(deduplication)机制，也称重复数据删除机制，与传统利用字典模型识别和消除文件内冗余的压缩技术不同，该技术旨在消除数据集合中文件内部和文件之间的冗余数据，通过仅保留一份数据副本来提高云服务效率与服务质量。

数据去重是一种高效的数据缩减方法，可以减少存储空间、降低传输带宽消耗。对于明文数据，服务器可以采用随机抽样或提取散列值，然后匹配的方法检测用户新上传数据与原存储数据是否相同，相同则删除重复的数据，且无需该用户再次上传，并告知该用户数据的访问链接以便下次访问。敖等[3]综述了通用存储系统中的数据去重的相关技术，从完全文件检测、固定分块检测、可变分块检测、滑动块检测等 4 个方面分析了相同数据的重复性检测技术，从基于 Shingle 的检测、基于布隆过滤器(Bloom filter)的检测、基于模式匹配的检测等 3 个方面介绍了相似数据的重复性检测技术。付等[4]从数据划分方法、I/O 优化技术、高可靠数据配置策略、系统可扩展性等方面综述了网络存储系统中数据去重的关键技术。Xia 等[5]总结了数据去重的关键特征，从数据去重工作流的角度综述了相关技术的研究进展。研究发现，通过采用数据去重技术，可以为备份系统节约 83%的存储空间，为主存系统节约 68%的存储空间，为固态硬盘节约 28%的存储空间，为云虚拟机中通用数据的存储节约高达 80%的空间[6]。

在云环境中，同一数据存储多个冗余副本，在一定程度上可以增加服务的可用性和可靠性，但将大量浪费用户的通信带宽和云存储空间，尤其在用户量与数据量均呈爆炸式增长的大数据时代，这就要求云服务提供商能够提供安全的数据去重服务。而云环境中实施安全去重，除数据机密性之外，还同时存在密钥泄露、非授权访问、用户身份隐私泄露等诸多安全问题，严重影响云服务的健康发展[7]。为了解决上述问题，云数据安全去重技术的研究迅速得到学术界和产业界的广泛关注，成为云数据安全存储领域的研究热点，并取得了一定的研究成果。与已有面向存储系统中明文数据的去重技术综述不同，本文以云计算环境为背景，面向云数据安全去重的技术实现机制，从基于内容加密的安全去重、基于所有权证明(PoW, proof of ownership)的安全去重和隐私保护的安全去重 3 个方面展开，对相关技术与实现机制进行归纳与述评，并提出进一步的研究方向，为科研人员准确把握存储安全领域最新研究动态和未来发展趋势提供借鉴。

2 问题描述

本节首先描述云环境下数据安全去重所面临的挑战，然后给出云数据安全去重问题的系统模型和威胁模型。

2.1 面临的挑战

在传统的信息系统中，数据均以明文形式存储，用户完全拥有其所有权与管理权。为了提高存储效率，需要对其进行重复删除，相关技术比较成熟，已经取得了一定的研究成果，如相同数据的重复性检测和相似数据的重复性检测[3]。

而在云存储服务中，数据的管理权与所有权分离，为了保护数据安全，云数据常以密文形式存储，如何实现对密文数据的安全去重成为学术界的研究焦点。云环境下数据外包与虚拟化等特征也给云数据的安全去重带来如下诸多挑战。

1) 数据外包与机密性保护。用户数据的所有权与管理权分离，为了保护其敏感信息安全，必须对外包数据先进行加密处理。因随机性加密算法采用不同密钥加密相同明文后得到不同密文，云端无法检验这些密文是否对应相同的明文，因而密文数据不能简单采用明文数据的去重方法，这给云数据安全去重带来巨大挑战。

2) 虚拟化与隐私泄露。云计算使用虚拟化技术，建立多用户之间的逻辑隔离，从而实现对多用

户计算资源、存储资源等的按需分配。当云服务器检测到多用户之间具有重复数据后，执行去重操作，则其很容易识别多用户之间的重复数据量，这本身就泄露了用户和数据的部分隐私，如何既实现数据去重又保护用户隐私是亟需解决的一个难题。

3) 侧信道攻击。在跨用户安全去重的任务执行过程中，可能因文件的大小、类型、散列值等信息而产生侧信道攻击，通过识别文件、试图学习文件内容和建立隐蔽通道[8]而揭露用户的身份、职业、敏感文件等隐私信息。如何在云数据安全去重过程中，避免侧信道攻击以保护参与用户的隐私成为亟需解决的另一个关键问题。

2.2　系统模型

本节将介绍云数据安全去重机制的通用系统模型，如图 1 所示，主要实体包含用户、云服务器、第三方服务器。

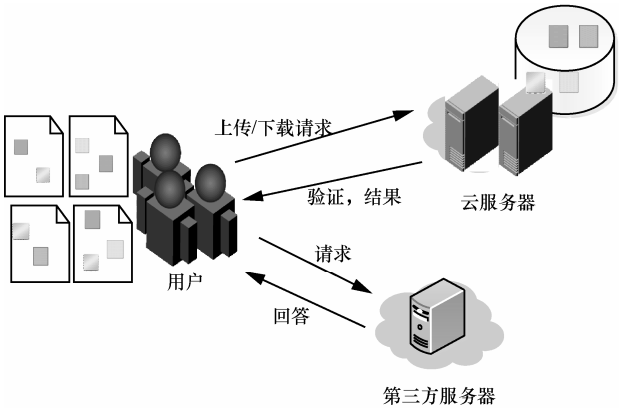

图 1　云数据安全去重的系统模型

用户将文件进行预处理后上传到云服务器，该服务器负责存储文件与相关文件标识，当其他用户再次上传相同文件时，云服务器执行数据安全去重工作。

通常情况下，数据安全去重方案只包含用户与云服务器 2 个实体，如文献[9~13]等。但为了安全需要或实现对密钥的有效管理，许多方案引入第三方服务器，如文献[14~17]等引入密钥服务器来专门存储和管理密钥；文献[18, 19]等引入文件索引服务器提供安全的文件索引，此外，文献[18]还引入身份验证服务器来验证用户身份和抵抗 Sybil 攻击。

2.3　威胁模型

本节主要描述云数据安全去重机制中常见的攻击类型和攻击行为[20]，以及抵抗这些攻击的相关方案，如表 1 所示。

蛮力攻击。收敛加密(CE, convergent encryption)的密钥由原始文件计算而来，因此，知道密文的敌手可对猜测的明文进行加密并与之进行对比，则可能猜测出原始数据。针对该攻击，DupLESS[14]提出使用不经意伪随机函数(OPRF, oblivious pseudo-random function)的密钥服务器(KS, key server)来产生密钥，即密钥是由数据本身和一个系统层面的密钥共同决定，实现了数据的保密性并能抵抗蛮力攻击。针对 KS 和云服务器合谋导致敌手可以获取密文和密钥的问题，Miao 等[17]提出一种基于门限盲签名与可校验秘密共享机制的多密钥服务器数据去重方案，即密钥由多个 KS 合作产生，每个 KS 只有密钥分量，无法得到完整密钥，有效防止单个 KS 与云服务器的合谋。

侧信道攻击。主要分为 3 种：1) 识别文件，攻击者上传特定的文件到云服务器，根据数据去重是否发生来判断其是否拥有该文件；2) 学习文件内容，识别特定文件是否存储在服务器之后，攻击者可能为了确定文件的内容而进行穷举攻击；3) 建立隐蔽信道，攻击者设法在用户的电脑上安装恶意软件，利用数据去重建立隐蔽信道与外部通信。

针对上述攻击，Chen 等[10]提出一种将秘密共享机制 AONT-RS 与 CE 结合的 CDStore 方案，进行用户本地和全局 2 个阶段的去重，有效解决了侧信道攻击。Puzio 等[16]在 CE 加密的基础上增加额外的语义安全加密方案和访问控制机制，提出 ClouDedup 抵抗侧信道攻击。此外，文献[21]提出利用差分隐私(DP, differential privacy)技术[22]来抵抗上述攻击。

字典攻击。在基于 CE 的数据去重方案中，敌手跟云服务器合谋，将明文进行加密，与已知的密文字典进行对比，则可以猜测到目标文件。即使数据加密密钥由用户的私钥加密并且存储在安全的服务器上，只要敌手可以得到加密密文，就可以实施字典攻击。文献[16]方案可以抵抗字典攻击，而文献[19]则提出对不同安全级别的文件进行不同级别的加密来抵抗这类攻击。

伪造攻击。敌手利用云服务器无法区分明文和密文的漏洞，上传一个加密文件以及与其不一致的伪造指纹信息，使另一可信用户上传伪造指纹对应的文件时，服务端回复已有该文件，导致无法上传，而下载的却是敌手上传的伪造文件。为了抵抗这种攻击，Bellare 等[13]在消息锁加密(MLE, message-locked encryption)[11]的基础上提出交互 MLE (iMLE,

interactive MLE)，即文件指纹由数据本身和服务器提供的系统参数共同决定，敌手无法伪造系统参数，从而也无法伪造成文件的拥有者。

Sybil 攻击。在点对点网络系统中，单一节点具有多个身份标识，Sybil 攻击可以控制系统的大部分节点。使用可信证书中心来验证通信实体的身份可以防止这种攻击，如 Stanek 等[18]使用一种身份验证服务器来严格进行身份控制。

表 1 不同方案抵制攻击类型汇总

攻击类型	抵抗攻击的方案
蛮力攻击	DupLESS[14]、REED[15]、multi-server-aided[17]
侧信道攻击	CDStore[10]、ClouDedup[16]、Shin 方案[22]
字典攻击	DupLESS[14]、ClouDedup[16]、PerfectDedup[19]
伪造攻击	MLE[11]、iMLE[13]
Sybil 攻击	Stanek 方案[18]

3 国内外研究进展与分析

对云环境中数据安全去重方面的研究，国内外已经取得了一定的成果。依据研究对象粒度的不同，可以分为文件级安全去重[9, 12, 14, 18]和块级安全去重[10~12, 15]；依据由谁来主导去重任务，可以分为基于目标的安全去重、基于文件源的安全去重和跨用户的安全去重。

本文另辟蹊径，从安全性角度出发，着重关注云数据安全去重技术的实现机制，对相关技术和原理进行系统梳理和归纳。整体而言，可以分为基于内容加密的安全去重、基于 PoW 的安全去重和隐私保护的安全去重 3 个方面。

3.1 基于内容加密的安全去重

基于内容的加密算法属于对称加密算法，由数据内容计算得到的加密密钥保证了相同的数据内容得到相同的密钥和密文，从而可用于对密文数据进行重复性检测，被广泛应用于云数据安全去重机制中。主要包括 2 种算法：收敛加密(CE, convergent encryption)算法和消息锁加密(MLE, message-locked encryption)算法。

3.1.1 基于 CE 实现云数据安全去重

针对传统随机加密算法无法进行重复性检测，使相同文件在云端同时存储多份副本，严重浪费存储空间的问题，Douceur 等提出 CE 算法[9]，其中，密钥生成算法为确定性算法，通常由原数据经过散列运算得到，确保相同的数据得到相同的密钥。

基于 CE 的数据去重交互过程如图 2 所示，用户上传文件 F，首先初始化散列函数 H，计算密钥 $k = H(F)$，同时作为文件 F 的 id。然后，用户端将 $H(F)$上传给云服务器，服务器验证是否存储该文件 id。若没有，则要求用户端上传文件 F，用户端需要对 F 进行加密

$$C = E_{H(F)}(F)$$

并将 C 与 $H(F)$上传给云服务器存储；若已存储，则无需用户端上传文件 H。

当用户端需要访问或下载文件时，只需上传 $H(F)$，服务器根据 $H(F)$将对应的密文 C 传给用户端，用户端进行解密操作

$$F = D_{H(F)}(C)$$

即可实现对密文的去重，大幅节省云服务器存储空间。

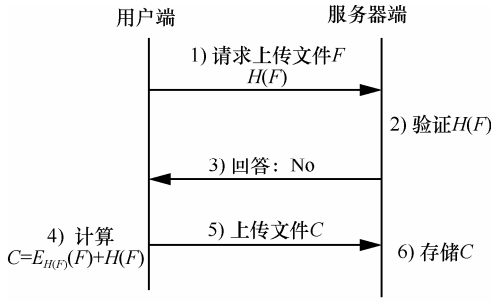

图 2 基于 CE 的数据去重交互过程

CE 加密算法的密文可校验性，使其迅速在云数据去重领域得到广泛应用，许多研究成果都将 CE 结合各种不同机制来实现对密文数据的去重。

CE 结合秘密共享机制实现云数据安全去重。包含随机信息输入的秘密共享机制使相同数据秘密共享后的秘密分量不同，造成服务器需存储多份秘密分量，浪费存储空间。Li 等[10]将秘密共享 AONT-RS(all-or-nothing transform reed-solomon)与 CE 结合，采用收敛扩散(CD, convergent dispersal)机制并提出一种 CDStore 方案，把扩散算法中的随机信息替换为数据的散列指纹，保证该算法的确定性以实现数据的安全去重，实验表明该方案可节省 70%的云存储空间。

针对不同隐私级别的数据需要不同保护程度的需求，Stanek 等[18]结合 CE 提出一种为数据提供不同安全等级加密的方案，该方案将文件分为流行文件和非流行文件。对保密程度不高的流行文件，使用基本的 CE 加密算法；对隐私级别较高的非流行文件，采用语义安全的加密算法。针对用户自定

义数据块重要程度过程中出现的安全隐患，Puzio 等[19]提出 PerfectDedup 方案，一种完美安全的获得数据块重要程度的机制，用户使用完美散列函数对数据块进行操作，该散列值作为一种查找存储在云服务器中数据块重要程度的索引。

针对云数据安全去重方案面临的安全和隐私问题，Puzio 等[16]提出一种安全有效的存储系统 ClouDedup，该系统提供数据块级的去重和保密性，在 CE 的基础上，增加额外的语义安全加密方案和访问控制机制来抵抗现有的 CE 攻击，此外，该系统利用密钥服务器来管理由数据块级别加密而产生的大量密钥，以减少客户端的存储消耗。

然而，CE 没有明确的安全目标，也缺乏形式化描述的安全模型。因此，Bellare 等[11]提出 MLE 算法，一种具有明确安全目标和形式化定义的安全去重方案。

3.1.2　基于 MLE 实现云数据安全去重

为提供严格的形式化安全机制，Bellare 等[11]基于 CE 提出一种新的消息锁加密(MLE)算法，加密密钥 k 由数据和系统参数共同计算 $k=K(p,m)$，其中，p 为系统参数，m 为原始文件；接着使用密钥 k 对数据进行对称加密得到密文 $C=E_k(m)$；t 作为检验消息数据是否相同的标志，实现对密文数据的重复性校验，即 $t=T(C)$；由于对称加密的特性，使用密钥 k 解密密文 C 即可得到明文：$m=D_k(C)$。

为了现实在大规模加密文件中现实高效的去重，Chen 等[12]提出 BL-MLE，只需要少量的元数据就能实现文件级和块级的数据去重、消息块的密钥管理和所有权证明关系，并易拓展为支持可验证的存储证明。除此之外，Bellare 等[13]在 MLE 的基础上提出交互 MLE(iMLE)，相同的文件以系统参数 p 和文件作为输入，输出相同的密钥，实现关联文件的去重。

对于 MLE 密钥更新问题，Li 等[15]结合 MLE 与 AONT-RS 秘密共享机制，提出 REED (rekeying-aware encrypted deduplication)方案。文件密钥由密钥状态进行散列运算得到，通过更新密钥状态来更新文件密钥。实验表明，该方案安全有效地实现数据安全去重。

下面从主要算法、抗攻击类型、是否第三方服务器、数据去重级别等方面对现有基于内容加密的安全去重主流方案进行对比分析，如表 2 所示。

表 3 从客户端开销、服务器端开销和通信带宽消耗等方面重点比较了几种经典数据安全去重方案的计算复杂度。

表 3　基于内容加密的安全去重方案的计算开销比较

方案	客户端开销	服务器端开销	初始通信带宽消耗	常规通信带宽消耗
CE[9]	$O(f)Hash$	$O(f)$	$O(f)$	$O(g)$
MLE[11]	$O(f)Hash{\cdot}Hash$	$O(f)$	$O(f)$	$O(g)$
DupLESS[14]	$O(f)Hash{\cdot}Hash$	$O(f)OPRF$	$O(f)$	$O(g)$
BL-MLE[12]	$O(b)Hash{\cdot}Hash$	$O(b)$ PoW	$O(f)$	$O(g\lambda)$
iMLE[13]	$O(f)Hash$	$O(fp)$	$O(f)$	$O(gp)$

注：f 表示文件的长度，b 表示文件块的长度，p 表示系统参数，λ 表示安全参数，g 表示文件指纹的长度，$Hash$ 表示执行一次散列函数的开销，PoW 表示执行一次所有权证明的开销，$OPRF$ 表示执行一次不经意伪随机函数的开销。

由表 3 可以看出，CE 与 iMLE 只需对生成文件的指纹进行重复性检验，客户端开销相对于其他方案较小，但 iMLE 方案因上传或下载文件均需被服

表 2　基于内容加密的安全去重方案的对比分析

方案	采用的主要算法	抗攻击类型	第三方服务器	数据去重级别	谁主导去重实施
CE[9]	CE	—	无	文件级去重	客户端
CDStore[10]	CE+AONT-RS	侧信道攻击	无	块级去重	客户端
PerfectDedup[19]	CE+PHF	目录攻击	索引服务器	块级去重	客户端
ClouDedup[16]	CE+访问控制策略	目录攻击，侧信道攻击	密钥服务器	块级去重	跨用户
Stanek[18]	门限+CE	Sybil 攻击	索引、身份验证服务器	文件级去重	跨用户
MLE[11]	MLE	—	无	文件级去重	客户端
DupLESS[14]	MLE	蛮力攻击	密钥服务器	文件级去重	客户端
BL-MLE[12]	BL-MLE+PoW	文件分发攻击	无	块级去重	客户端
multi-server-aided[17]	门限盲签名+可校验秘密共享	蛮力攻击	多密钥服务器	文件级去重	客户端
iMLE[13]	iMLE	伪造攻击	无	文件级去重	跨用户
REED[15]	MLE+AONT-RS	蛮力攻击	密钥服务器	块级去重	客户端

务器的参数列验证，因此，服务器开销与常规通信带宽消耗较大，BL-MLE 在进行去重的基础上增加了 PoW 验证，因此服务器端与常规通信消耗较大。

3.2 基于 PoW 的安全去重

针对敌手通过获取用户文件的指纹信息，利用客户端去重机制从服务器得到完整文件的攻击，学者们提出了所有权证明的概念。

PoW 协议采用通用的挑战—应答模型，主要包含 3 个阶段：1) 文件上传阶段，客户端向云服务器发送上传文件的请求，云服务器收到该请求后，在数据库中检索文件是否存在，如果不存在，则要求客户端上传文件，接收并存储该文件，如果已存在，则需要客户端证明拥有者的身份；2) 云服务器挑战阶段，作为验证者的云服务器根据文件索引找到相关文件，生成挑战数组发送给客户端；3) 客户端应答阶段，客户端接收到挑战后，根据拥有的文件生成应答，返回验证，服务器将接收到的结果与正确的应答比较，若匹配，表明客户端确实拥有该文件，返回文件指针给客户端，反之，表明客户端不是该文件的拥有者，返回失败。PoW 协议可分为如下 3 类实现机制。

3.2.1 基于 MHT 的 PoW

PoW 最初由 Halevi 等[23]基于 Merkle Hash Tree (MHT)实现，服务器和客户端都依据原文件内容建立 MHT，通过类似挑战—应答模型由服务器挑战客户端对于给定叶子节点的子集能否正确提供有效的 MHT 路径。该方案提出 3 种不同方法构造 MHT：1) 对原始文件分块后再采用纠删码来构造 MHT；2) 采用两两独立的散列函数构建 MHT；3) 服务器产生稀疏线性文件，利用一种高效的散列函数计算文件指纹来构建 MHT。然而，Halevi 的 PoW 方案仍然存在如下局限性：1) 均对数据明文进行处理，未考虑对敏感数据的加密保护；2) 后 2 种方法本质仍然是基于文件散列来实现数据重复性校验，存在蛮力攻击的风险。

针对 PoW 方案无法保证在客户端去重中证明所有权的同时保证用户数据隐私的问题，Xu 等[24]改进 Halevi 的方案，先对原文件进行加密，再构造 MHT 进行 PoW 校验，在随机预言机模型下证明该方案是安全的。此外，文献[25]面向密文数据提出基于 MHT 的去重方案和基于同态 MAC 的去重方案，通过对密文的拥有证明实现文件级的重复性检测和本地数据块级的重复性检测。

3.2.2 基于随机抽样的 PoW

为了减少计算开销和降低 I/O 读写操作次数，Di Pietro 等[26]提出了一种 PoW 优化方案(s-PoW)，实现高效、信息论安全的文件所有权证明。该方案采用四元组数据结构存储文件信息，通过伪随机数生成器和相关加密算法产生 *seed*，该 *seed* 可作为挑战发送给客户端，服务器保存对应挑战的应答，与接收到客户端发来的应答进行比较，返回结果。随后对该方案进行优化，提出 s-PoW1 和 s-PoW2，提高算法的执行效率。

为了进一步提高现有 PoW 方案中服务器端的工作效率，Blasco 等[27]提出了一种基于 Bloom filter 的灵活、可扩展、可证明安全的云数据去重方案(BF-PoW)。服务器将接收到的文件分成大小相同的 *chunk*，得到对应的 *token*，把经过 PRF 函数处理后的 *token* 插入 Bloom filter 中，服务器采用三元组数据结构存储文件信息。在挑战阶段，服务器要求客户端上传一定数量的 *token* 证明文件的所有权。实验表明该方案能够大幅减少客户端和服务器端的开销。

González-Manzano 等[28]提出一种无可信第三方、没有复杂密钥管理开销的 CE-PoW 方案，服务器储存一个由密文块、挑战、应答和客户端身份列表组成的四元组。在随机预言机模型下该方案是可证明安全的，实验结果表明，与其他 PoW 方案相比，该方案具有较小开销。但采用 CE 算法不能实现语义安全，易导致文件内容猜测攻击[11]。

除了上述几种典型的 PoW 方案外，Xu 等[29]提出了一种支持用户数据隐私保护的去重方案 privacy-preserving PoW，利用随机提取和数据可恢复性证明(PoR, proof of retrievability)[30]实现有限泄露信息下的所有权证明，设计出一种新的大规模输入的随机提取，减少随机种子的长度和熵损失的隐私保护的去重方案。

表 4 对上述 PoW、s-PoW、bf-PoW、ce-PoW 等 4 种典型的 PoW 方案的计算开销进行了综合对比分析。在客户端的开销方面，CE-PoW 除了散列操作，还要对文件进行 CE 加密，因此直接对明文操作的 PoW、s-PoW、BF-PoW 方案计算开销较小；在服务器的开销方面，BF-PoW 方案使用 BF 加快查询效率，更具优势；在通信带宽消耗方面，s-PoW 方案直接将明文上传，通信带宽消耗只和系统安全参数相关，相比之下更有优势。

表 4　　经典基于 PoW 的去重方案的开销对比分析

方案	客户端计算开销	客户端读写开销	服务器初始化计算开销	服务器去重计算开销	服务器初始化读写开销	服务器常规读写开销	服务器内存开销	通信带宽消耗
PoW[23]	$O(f)Hash$	$O(f)$	$O(f)$ $Hash$	$O(1)$	$O(f)$	$O(0)$	$O(1)$	$O(\lambda\log\lambda)$
s-PoW[26]	$O(f)Hash$	$O(f)$	$O(f)$ $Hash$	$O(n\lambda)PRF$	$O(f)$	$O(n\lambda)$	$O(n\lambda)$	$O(\lambda)$
BF-PoW[27]	$O(f)Hash$	$O(f)$	$O(f)$ $Hash$	$O(\Omega)Hash$	$O(f)$	$O(0)$	$O(\Psi)$	$O\left(\frac{l\lambda}{p_f}\right)$
CE-PoW[28]	$O(b)$ $CE{\cdot}Hash{\cdot}Hash$	$O(f)$	$O(b){\cdot}Hash{\cdot}Hash$	$O(nl\lambda)$ $PRNG$	$O(f)$	$O(0)$	$O(nl\lambda)$	$O(l\lambda)$

注：其中，f 表示文件长度，n 表示预设的挑战数目，λ 表示安全参数，p_f 表示 BF 的误判率，l 表示 $token$ 的长度，$Hash$ 表示执行一次散列函数的开销，CE 表示执行一次收敛加密的开销，PRF 表示执行一次伪随机函数的开销，$PRNG$(pseudo-random number generator)表示执行一次伪随机序列发生器的开销，$\Omega=\frac{l\lambda\left(\log\frac{1}{p_f}\right)}{p_f}$，$\psi=\frac{\log\frac{1}{p_f}}{l}$。

3.2.3　基于广义散列函数的 PoW

针对客户端去重中攻击者只需获得原始文件的指纹信息即可从服务器端获得全部原始文件的安全问题，Yang 等[31, 32]提出了 2 种加密的、客户端根据完整源文件而非部分文件信息向服务器证明所有权的安全方案，POF(provable ownership of file)和 POEF(provable ownership of the encrypted file)，分别利用抽样检测、动态系数和随机选择原始文件检索实现所有权证明，安全性分析表明该方案是可证明安全的，并能有效减少客户端开销。

和 PoW 相近的还有数据存储证明(PoS, proof of storage)[33, 34]，PoS 方案允许客户端在本地不保留任何数据副本的前提下，校验存储在云端服务器上的数据仍然是原始文件，包含 2 种基本的方法：可证明数据持有(PDP, provable data possession)[35, 36]和 POR[30, 37]。针对现有 PDP 和 POR 均无法满足不同群组的特殊性要求，尚未实现群组应用中的重复数据删除问题，王等[38]基于矩阵计算和伪随机函数，提出了一种支持数据去重的群组 PDP(GPDP, group PDP)方案，高效地完成数据所有权证明，并支持数据去重，在群组中抵抗恶意方选择成员攻击，在标准模型下证明了 GPDP 的安全性，能够有效减小存储开销和通信开销。

针对现有所有权证明方案的性能弱点以及在智能移动终端缺乏考虑数据去重中实现所有权证明的问题，Yu 等[39]回顾了现有所有权证明方案，分析了各方案性能弱点，并提出一种在去重中实现所有权证明的提高移动终端效率的方案。

3.3　隐私保护的安全去重

云服务提供商在采用数据去重技术来控制单个文件副本数量的同时，攻击者可能利用数据去重过程作为侧信道来对用户隐私信息进行攻击。用户的隐私信息遭到威胁，将严重阻碍云服务的健康发展。因此，在数据去重的过程中保护数据隐私尤为重要，目前面向数据内容隐私保护的安全去重方法主要分为 2 种类型：基于随机化方法实现隐私保护的安全去重和基于差分隐私实现隐私保护的安全去重。

3.3.1　基于随机化方法的隐私保护安全去重

该方法通过增加数据去重发生的随机性来改变随机事件发生的概率，从而达到混淆数据去重事件的效果。

Harnik 等[40]提出随机化的解决方案，在数据去重过程中设置单个文件的数量，当单个文件的上传数量达到此阈值时执行数据去重。Lee 等[41]在 Harnik 方案的基础上，对随机化算法进行了改进，服务器将上传文件 F 的实际数量减去集合中的随机数后，再进行阈值和数量的判断，使数据去重事件的发生有一定的随机性，可以进一步降低隐私泄露的概率。

通过对上述方案的分析可知，修改阈值可以降低隐私泄露的概率，但这种随机化方法可能会上传一些不必要文件，导致增加网络带宽。同时，无法抵御攻击者利用文件之间的相关性推断文件 F 是否存在关联文件攻击，也无法完全抵抗敌手的侧信道攻击。

3.3.2　基于差分隐私的隐私保护安全去重

该方法主要是在数据去重的过程中采用差分隐私保护机制[21, 22]，保证在数据特征不变的前提下添加适量虚拟数据即噪声数据，达到数据失真的效

果来保护数据隐私。数据失真隐私保护技术除了随机化、阻塞、交换之外，Dwork[42]首次提出差分隐私概念，通过添加噪声数据使敏感数据失真但同时保证处理后的数据仍然可以保持某些统计特性不变。采用差分隐私技术保护数据隐私，即使在攻击者已知 N 条记录中的 N–1 条的情况下，依然无法推断出剩下那条记录的信息，并且可以根据数据集的敏感度灵活地实施不同程度的隐私保护。

Shin 等[22]提出了一种面向私人客户端的数据去重协议，利用存储网关作为用户和服务器之间的交互实体来实现高效的数据去重，同时引入差分隐私保护机制来减少信息泄露的风险，该方案的核心是使用虚拟数据的上传混淆数据去重的发生。使用文件的散列验证文件 F 是否存储在服务器，如果不存在，文件 F 需全部上传，文件 F 一个子集的字节 $\{b_1,b_2,\cdots,b_\tau\}$ 从本地磁盘队列 Y 中传送到缓存区，剩余的字节 $\{b_{\tau+1},b_{\tau+1},\cdots,b_{|F|}\}$ 排到队列 Y 中；如果存在，文件 F 不上传，τ byte 的数据从队列 Y 中移除，将 Y 中原有的数据 y 传送到缓冲区。如果 Y 中没有足够的数据，虚拟字节随机生成并填充缓冲区，要求总大小为 τ。利用随机生成的虚拟数据上传到服务器，使攻击者无法判断数据去重是否发生，从而起到保护文件 F 的目的。

实验结果表明，随着时间的变化和参数的不同，Harnik 等[40]、Lee 等[41]通过设定随机阈值的保护方案依然会使文件存在多个数据副本，而 Shin 等[22]的方案可保证文件只有一个副本，大幅减少客户端的通信开销，且采用 DP 机制使用户数据的安全性得到较大提高。

在数据安全去重过程中，不仅要关注隐私保护程度的高低，更要考虑数据安全去重的效率、带宽的消耗以及数据缺损等问题，表 5 给出了上述方案的性能对比分析。

由表 5 可知，Harnik 等[40]与 Lee 等[41]的方案开销较大，在数据去重的过程中需要不断判断服务器所包含文件 F 的副本数量是否已经达到所设定的阈值，因此，通信带宽和服务器内存都会随着文件 F 上传次数的增加而增加，而在 Shin 等[22]的方案中，服务器只需要存储文件 F 的一个副本即可，所以通信带宽及服务器内存消耗大大减少。

3.4 研究进展小结

通过对现有云数据安全去重工作的梳理和分析，可以得到如下结论。

表 5 隐私保护的相关安全去重方法的性能评估

隐私保护方案	隐私保护程度	数据缺损	数据依赖性	通信带宽消耗	服务器内存消耗
Harnik 方案[40]	低	低	低	$O(i)$	$O(if)$
Lee 方案[41]	中	低	低	$O(i)$	$O(if)$
Shin 方案[22]	高	高	高	$O(i)$	$O(f)$

注：i 代表上传文件的次数，f 代表文件 F 的长度。

1) 基于内容的加密实现安全去重能够解决用户间数据重复性检测与加密之间的矛盾，但存在如下局限性：①基于 CE 的方案将文件的散列指纹值作为数据去重的校验依据，容易遭受离线的蛮力攻击，使其只能在特定条件下提供有限的安全与隐私保障；②DupLESS[14]部署专门的密钥服务器来保管系统密钥结合 MLE 以抵抗蛮力攻击，但当系统密钥被泄露时仍存在被蛮力攻击的风险，且密钥服务器生成密钥的计算开销大，在执行海量块级数据去重时，容易导致服务器过载；③CDStore[10]采用的收敛扩散技术仍可能遭受离线的蛮力攻击。

2) 基于 PoW 的安全去重要求用户在从服务器获取有关数据存在性信息之前，先证明其确实拥有该数据。PoW 能够解决恶意用户通过文件的散列指纹信息从服务器获得原始文件的侧信道攻击问题，但也存在如下局限性：①基于 MHT 构造的方案需要在客户端进行多次散列操作，为提高数据的可靠性需对数据块进行纠删码编码，将增加客户端的计算开销；②执行挑战—应答模式的云服务器需要维护三元、四元数组，计算挑战和相应的应答等，均需要较大的计算开销，采用布隆过滤器可以缓解服务器端开销；③执行挑战—应答模式均需要客户端和云服务器进行多次交互，不可避免地消耗传输带宽资源。

3) 面向隐私保护的安全去重采取适当的机制能够有效抵抗侧信道攻击，保护数据隐私安全，但仍存在如下缺陷：①随机化方法通过随机设定上传阈值来随机化数据去重事件的发生，以抵抗敌手攻击，但云服务器端需存储预先设定的随机值数量的数据副本，将带来额外的存储开销；②差分隐私保护方法通过引入噪声数据以保护数据隐私，但同时降低了数据可用性，如何在隐私保护效果和数据可用性之间取得折中还需要深入研究。

表 6 对上述实现云数据安全去重的相关技术和原理进行系统梳理和归纳。

表 6　　**云数据安全去重的相关技术和原理的对比分析**

分类	相关技术	典型方案	主要算法	优点	局限性
基于内容加密的安全去重	基于 CE 实现云数据安全去重	CE[9]	CE	密文的重复性检测	不能达到语义安全，易遭受蛮力攻击
	基于 MLE 实现云数据安全去重	BL-MLE[12]	BL-MLE +PoW	明确安全目标和形式化定义，块级去重	计算开销较大，遭受蛮力攻击
基于PoW的安全去重	基于 MHT 的 PoW	Halevi[23]方案	MHT	提出 PoW 概念，抵抗侧信道攻击	计算开销较大，未考虑敏感数据的加密保护
	基于随机抽样的 PoW	CE-PoW[28]	CE+随机抽样	减少计算开销，高效，无可信第三方	不能实现语义安全，易导致内容猜测攻击
	基于广义散列函数的 PoW	Yang[31] 方案	PoF+抽样检测+动态系数	根据完整源文件的所有权证明	计算开销较大，遭受蛮力攻击
基于隐私保护的安全去重	基于随机化方法的隐私保护安全去重	Harnik[40]方案 Lee[41]方案	随机化方法	数据真实无缺损，减少去重中的隐私泄露的概率	无法抵抗关联文件攻击，易遭受蛮力攻击
	基于差分隐私的隐私保护安全去重	Shin[22]方案	差分隐私	隐私保护程度较高，可抵抗侧信道攻击及关联文件攻击	添加噪声容易导致数据失真，数据依赖性高，需要根据不同的数据计算所添加噪声大小

4　发展趋势与未来研究方向

从已有研究成果看，云数据安全去重的研究当前还处于发展初期，虽取得了一定的成果，但还远不够完善。尤其是大数据、“万物互连”时代给云数据安全去重提出更大挑战，这些挑战已经引起当今学术界和产业界的重点关注。从发展趋势分析，未来云数据安全去重的研究主要关注如下几方面。

1) 协作数据的共享所有权证明机制。云计算环境下的协作创新已成为新时代的新型信息服务模式，如重大项目申请书的协作撰写、疑难病症的多医院联合会诊、联合商业计划书的形成等过程中不可避免地产生协作文件。与传统单一文件由用户单独所有不同，协作文件属于参与方共同所有，不能由某个参与方单方面设置策略与访问权限，对其进行操作需要获得各参与方的共同授权[43]。如何扩展现有 PoW 的概念，对协作文件实施共享所有权证明，以支持协作文件的多副本匹配、校验等操作，并保护各参与方的隐私信息不泄露，以实现对协作云数据的安全去重是一个全新挑战。

2) 大数据安全去重、隐私保护与效率的平衡机理。据 Gartner 预测，2020 年全球数据使用量将达到 44 ZB。如此大规模的数据不可避免地产生大量数据冗余，给抽样、匹配、决策等带来巨大挑战，海量数据的关联分析还将严重威胁用户隐私。半同态加密算法和全同态加密算法虽支持对数据密文进行部分操作，但计算开销大、效率偏低[44]，显然不适合大数据环境。如何对大数据执行密文的安全匹配、重复校验等操作，在保护数据和用户隐私信息的有限泄露的前提下，取得安全去重、隐私保护和效率之间的平衡，仍然是一个开放课题。

3) 数据压缩与安全去重的内在编码机制。数据压缩旨在保护数据有用信息的前提下，对数据进行重新编码组织，以减少冗余空间，提高其传输、存储和处理效率，可以分为有损压缩和无损压缩[45]。数据压缩的编码算法有很多，如霍夫曼编码、BWT (burrows Wheeler transform)变换、PPM (prediction by partial matching)、算术编码、差量压缩等，每种算法压缩数据后均得到不同的压缩文件，相互之间无法直接匹配和校验[46]。Halevi 等[23]利用纠删编码实现重复数据检测，但不能压缩数据。因此，如何构造一种合适的数据编码方法，在数据高效压缩的同时，实现对压缩数据的重复检测与删除是未来重要的研究课题。

4) 融合创新与新领域延伸发展。近期部分研究团队尝试将相关技术和安全去重进行融合创新，还有学者向图像等多媒体领域延伸发展，提出了如下新的研究方法。Li 等[47]提出了一种混合云架构，引入访问控制机制，实现对云数据授权访问的安全去重。Li 等[48]针对外包云存储不完全可信的问题，构造了 2 个安全系统，SecCloud 和 SecCloud+，在云数据安全去重的同时能够实现对数据的完整性审计。阎等[49]分析了 OpenXML 复合文件的属性，为非结构化数据提供细粒度的去重指导。Liu 等[50]结合短散列、口令认证密钥协商、加法同态等算法，实现了一种新的跨用户安全去重方案，但要求用户保持在线且由此引入了额外的交互开销。Armknecht 等[51]提出一种适应于云数据去重的新型

定价模型和新型存储系统 ClearBox。Zheng 等[52]针对多媒体数据提出一种新的安全去重机制。Li 等[53]针对图像文件提出一种隐私保护的模糊去重方法。张等[54]针对云桌面环境提出一种用户感知的新型数据去重方法。熊等[55]系统总结了基于密码学的云数据确定性删除方法，相关方法可以为未来云数据安全去重的研究提供借鉴。李等[56]针对新形势下隐私泄露威胁，提出隐私计算的概念，如何利用隐私计算相关理论，探索云数据安全去重新技术与新方法，并挖掘云数据安全去重在更多新领域的应用延伸也是未来的发展趋势。

5 结束语

如今，信息通信技术飞速发展，用户所产生的数据和通信流量均呈指数级增长，导致云服务提供商需要存储大量的冗余数据，这不仅加重云端的存储负担，而且严重阻碍云服务的推广和发展。为了提高云端存储效率和减少对带宽的消耗，迫切需要对云数据进行重复性检测和去重服务。云环境下的数据安全去重是一个非常活跃的研究方向，目前还处于起步阶段，尚未建立一套完整的理论体系，从关键技术与理论的完善到算法的实际应用还有很大的差距。

本文首先分析了云环境中数据安全去重面临的主要挑战，包括数据外包与机密性保护、虚拟化与隐私泄露、侧信道攻击，抽象出云数据安全去重的系统模型和威胁模型；然后，站在云数据安全去重的技术实现角度，从基于内容加密的安全去重、基于 PoW 的安全去重和面向隐私保护的安全去重 3 个方面对相关研究工作的基本思想、工作原理等进行了深入分析、归纳与总结，分别指出了各种技术方法的优缺点及存在的共性问题；最后，从协作云数据的安全处理、对大数据的支持、对压缩数据的安全处理、新技术和新领域等方面预测了该领域的发展趋势与未来研究方向。

参考文献：

[1] XIONG J, LI F, MA J, et al. A full lifecycle privacy protection scheme for sensitive data in cloud computing [J]. Peer-to-Peer Networking and Applications, 2014, 8(6): 1-13.

[2] MITTAL S, VETTER J. A survey of architectural approaches for data compression in cache and main memory systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 27(5): 1524-1536.

[3] 敖莉, 舒继武, 李明强. 重复数据删除技术[J]. 软件学报, 2010, 21(5): 916-929.
AO L, SHU J W, LI M Q. Data deduplication techniques [J].Journal of Software, 2010, 21(5): 916-929.

[4] 付印金, 肖侬, 刘芳. 重复数据删除关键技术研究进展[J]. 计算机研究与发展, 2012, 49(1): 12-20.
FU Y J, XIAO N, LIU F. Research and development on key techniques of data deduplication [J]. Journal of Computer Research and Development, 2012, 49(1): 12-20.

[5] XIA W, JIANG H, FENG D, et al. A comprehensive study of the past, present, and future of data deduplication [J]. Proceedings of the IEEE, 2016, 104(9): 1681-1710.

[6] PAULO J, PEREIRA J. A survey and classification of storage deduplication systems [J]. ACM Computing Surveys (CSUR), 2014, 47(1): 1-30.

[7] YU S. Big privacy: challenges and opportunities of privacy study in the age of big data [J]. IEEE Access, 2016, 4: 2751-2763.

[8] RABOTKA V, MANNAN M. An evaluation of recent secure deduplication proposals [J]. Journal of Information Security and Applications, 2016, 27: 3-18.

[9] DOUCEUR J, ADYA A, BOLOSKY W, et al. Reclaiming space from duplicate files in a serverless distributed file system[C]//International Conference on Distributed Computing Systems. 2002: 617-624.

[10] LI M, QIN C, LEE P. CDStore: toward reliable, secure, and cost-efficient cloud storage via convergent dispersal [C]//USENIX Annual Technical Conference (USENIX ATC 15). Santa, Clara, 2015: 111-124.

[11] BELLARE M, KEELVEEDHI S, RISTENPART T. Message-locked encryption and secure deduplication [M]//Advances in Cryptology–EUROCRYPT 2013. Springer Berlin Heidelberg, 2013: 296-312.

[12] CHEN R, MU Y, YANG G, et al. BL-MLE: block-level message-locked encryption for secure large file deduplication [J].IEEE Transactions on Information Forensics and Security, 2015, 10(12): 2643-2652.

[13] BELLARE M, KEELVEEDHI S. Interactive message-locked encryption and secure deduplication [M]. Public-Key Cryptography--PKC 2015. Springer Berlin Heidelberg, 2015: 516-538.

[14] KEELVEEDHI S, BELLARE M, RISTENPART T. DupLESS: server-aided encryption for deduplicated storage [C]//22nd USENIX Security Symposium (USENIX Security 13). Washington, 2013: 179-194.

[15] LI J, QIN C, LEE P, et al. Rekeying for encrypted deduplication storage [C]//The 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2016), Toulouse, France, 2016.

[16] PUZIO P, MOLVA R, ONEN M, et al. ClouDedup: secure deduplication with encrypted data for cloud storage [C]//Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on. IEEE, Bristol, UK, 2013: 363-370.

[17] MIAO M, WANG J, LI H, et al. Secure multi-server-aided data deduplication in cloud computing [J]. Pervasive and Mobile Computing, 2015, 24: 129-137.

[18] STANEK J, SORNIOTTI A, ANDROULAKI E, et al. A secure data deduplication scheme for cloud storage [M]. Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2014: 99-118.

[19] PUZIO P, MOLVA R, ÖNEN M, et al. PerfectDedup: secure data deduplication[C]//International Workshop on Data Privacy Management. Springer International Publishing, Atlanta, 2015: 150-166.

[20] RABOTKA V, MANNAN M. An evaluation of recent secure deduplication proposals[J]. Journal of Information Security and Applications,

2016, 27: 3-18.

[21] SHIN Y, KIM K. Differentially private client-side data deduplication protocol for cloud storage services[J]. Security and Communication Networks, 2015, 8(12): 2114-2123.

[22] DWORK C, LEI J. Differential privacy and robust statistics[C]//The forty-first annual ACM symposium on Theory of computing. ACM, Bethesda, 2009: 371-380.

[23] HALEVI S, HARNIK D, PINKAS B et al. Proofs of ownership in remote storage systems [C]//The 18th ACM conference on Computer and Communications Security. ACM, Chicago, 2011: 491-500.

[24] XU J, CHANG E, ZHOU J. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage [C]//8th ACM SIGSAC Symposium on Information, Computer and Communications Security, ASIA CCS '13, ACM. Hangzhou, China, 2013: 195-206.

[25] 陈越, 李超零, 兰巨龙, 等. 基于确定/概率性文件拥有证明的机密数据安全去重方案[J]. 通信学报, 2015, 36(9): 1-12.
CHEN Y, LI C L, LAN J L, et al. Secure sensitive data deduplication schemes based on deterministic/probabilistic proof of file ownership[J]. Journal on Communications, 2015, 36(9): 1-12.

[26] DI PIETRO R, SORNIOTTI A. Boosting efficiency and security in proof of ownership for deduplication [C]//The 7th ACM Symposium on Information, Computer and Communications Security. ACM, Seoul, 2012: 81-82.

[27] BLASCO J, DI PIETRO R, ORFILA A, et al. A tunable proof of ownership scheme for deduplication using bloom filters[C]// Communications and Network Security (CNS), 2014 IEEE Conference on. IEEE, San Francisco, California, 2014: 481-489.

[28] GONZÁLEZ-MANZANO L, ORFILA A. An efficient confidentiality-preserving proof of ownership for deduplication [J]. Journal of Network and Computer Applications, 2015, 50: 49-59.

[29] XU J, ZHOU J. Leakage resilient proofs of ownership in cloud storage, revisited[C]//Applied Cryptography and Network Security. Springer International Publishing, New York, 2014: 97-115.

[30] JUELS A, KALISKI J. PoRs: proofs of retrievability for large files[C]//14th ACM conference on Computer and Communications Security, CCS '07. New York, 2007: 584-597.

[31] YANG C, REN J, MA J. Provable ownership of files in deduplication cloud storage [J]. Security and Communication Networks, 2015, 8(14): 2457-2468.

[32] 杨超, 张俊伟, 董学文, 等. 云存储加密数据去重删除所有权证明方法[J]. 计算机研究与发展, 2015, 52(1): 248-268.
YANG C, ZHANG J W, DONG X W, et al. Proving method of ownership of encrypted files in cloud de-duplication deletion[J]. Journal of Computer Research and Development, 2015, 52(1): 248-268.

[33] ZHENG Q, XU S. Secure and efficient proof of storage with deduplication[C]//The 2nd ACM Conference on Data and Application Security and Privacy. ACM, San Antonio, 2012: 1-12.

[34] ATEBIESE G, DAGDELEN Ö, DAMGÅRD I, et al. Entangled cloud storage [J]. Future Generation Computer Systems, 2016, 62: 104-118.

[35] ATENIESE G, BURNS R, CURTMOLA R, et al. Provable data possession at untrusted stores[C]//The 14th ACM Conference on Computer and Communications Security. ACM, New York, USA, 2007: 598-609.

[36] REN Y, SHEN J, WANG J, et al. Mutual verifiable provable data auditing in public cloud storage [J]. Journal of Internet Technology, 2015, 16(2): 317-323.

[37] WANG B, CHOW S, LI M, et al. Storing shared data on the cloud via security-mediator[C]//Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on. IEEE, Macau, China, 2013: 124-133.

[38] 王宏远, 祝烈煌, 李龙一佳. 云存储中支持数据去重的群组数据持有性证明[J]. 软件学报, 2016, 27(6): 1417-1431.
WANG H Y, ZHU L H, LI L Y J. Group provable data possession with deduplication in cloud storage[J]. Journal of Software, 2016, 27(6): 1417-1431.

[39] YU C, CHEN C, CHAO H. Proof of ownership in deduplicated cloud storage with mobile device efficiency [J]. Network, IEEE, 2015, 29(2): 51-55.

[40] HARNIK D, PINKAS B, SHULMAN-PELEG A. Side channels in cloud services: deduplication in cloud storage [J]. IEEE Security & Privacy, 2010, 8(6):40-47.

[41] LEE S, CHOI D. Privacy-preserving cross-user source-based data deduplication in cloud storage[C]//2012 International Conference on ICT Convergence (ICTC). IEEE, Jeju, Korea, 2012: 329-330.

[42] DWORK C. Differential privacy: a survey of results[C]//International Conference on Theory and Applications of Models of Computation. Springer Berlin Heidelberg. Xi'an, China, 2008, 4978: 1-19.

[43] SORIENTE C, KARAME G, RITZDORF H, et al. Commune: shared ownership in an agnostic cloud[C]//The 20th ACM Symposium on Access Control Models and Technologies. ACM, Austria, 2015: 39-50.

[44] CHENG H, RONG C, HWANG K, et al. Secure big data storage and sharing scheme for cloud tenants [J]. China Communications, 2015, 12(6): 106-115.

[45] SINGH A, SINGH G. A survey on different text data compression techniques [J]. International Journal of Science and Research, 2014, 3.

[46] KAVITHA S, ANANDHI R. A survey of image compression methods for low depth-of-field images and image sequences [J]. Multimedia Tools and Applications, 2015, 74(18): 7943-7956.

[47] LI J, LI Y, CHEN X, et al. A hybrid cloud approach for secure authorized deduplication [J]. Parallel and Distributed Systems, IEEE Transactions on, 2015, 26(5): 1206-1216.

[48] LI J, LI J, XIE D, et al. Secure auditing and deduplicating data in cloud[J]. IEEE Transactions on Computers, 2016, 65(8): 2386-2396.

[49] 阎芳, 李元章, 张全新, 等. 基于对象的 OpenXML 复合文件去重方法研究[J]. 计算机研究与发展, 2015, 52(7): 1546-1557.
YAN F, LI Y Z, ZHANG Q X, et al. Object-based data de-duplication method for openXML [J]. Journal of Computer Research and Development, 2015, 52(7): 1546-1557.

[50] LIU J, ASOKAN N, PINKAS B. Secure deduplication of encrypted data without additional independent servers[C]//The 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, Denver, USA, 2015: 874-885.

[51] ARMKNECHT F, BOHLI J, KARAME G, et al. Transparent data deduplication in the cloud[C]//The 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, Denver, USA, 2015: 886-900.

[52] ZHENG Y, YUAN X, WANG X, et al. Enabling encrypted cloud media center with secure deduplication[C]//The 10th ACM Symposium on Information, Computer and Communications Security. ACM, Singapore, 2015: 63-72.

[53] LI X, LI J, HUANG F. A secure cloud storage system supporting privacy-preserving fuzzy deduplication [J]. Soft Computing, 2016, 20(4): 1437-1448.

[54] 张沪寅, 周景才, 陈毅波, 等. 用户感知的重复数据删除算法[J]. 软件学报, 2015, 26(10): 2581-2595.
ZHANG H Y, ZHOU J C, CHEN Y B, et al. User-aware de-duplication algorithm [J].Journal of Software, 2015, 26(10): 2581-2595.

[55] 熊金波, 李凤华, 王彦超, 等. 基于密码学的云数据确定性删除研究进展[J]. 通信学报, 2016, 37(8): 167-184.
XIONG J B, LI F H, WANG Y C, et al. Research progress on cloud data assured deletion based on cryptography [J]. Journal on Communications, 2016, 37(8): 167-184.

[56] 李凤华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4): 1-11.
LI F H, LI H, JIA Y, et al. Privacy computing: concept, connotation and its research trend [J]. Journal on Communications, 2016, 37(4): 1-11.

作者简介：

熊金波（1981-），男，湖南益阳人，福建师范大学副教授、硕士生导师，中国科学院信息工程研究所博士后，主要研究方向为云数据安全与隐私保护技术。

张媛媛（1992-），女，河南南阳人，福建师范大学硕士生，主要研究方向为云数据安全与隐私保护技术。

李凤华（1966-），男，湖北浠水人，博士，中国科学院信息工程研究所副总工、研究员、博士生导师，主要研究方向为网络与系统安全、信息保护、隐私计算。

李素萍（1991-），女，福建三明人，福建师范大学硕士生，主要研究方向为云数据安全与隐私保护技术。

任君（1993-），女，山西临汾人，福建师范大学硕士生，主要研究方向为云计算与安全服务。

姚志强（1967-），男，福建莆田人，博士，福建师范大学教授、硕士生导师，主要研究方向为信息安全。