开题报告

1.1 研究工作的背景与意义

1.1.1 研究背景

随着信息技术产业的高速发展，数字信息量呈爆炸式增长。Gartner 研究表明[1]，仅 2015 年的移动数据流量就较 2014 年增长 59%；并且，这一增长率将持至 2018 年末，移动数据流量水平达 1.73 亿 TB。数据的快速增长导致企业面临的存储和管理成本越来越高[2]。另一方面，在存储系统所保存的数据中，高达 60%的数据都是冗余的，随着时间的推移，这些冗余数据的比例将进一步上升[3]。近年来，存储系统中数据高冗余的特点得到越来越多研究人员的关注，利用该特点来节省存储容量是一个热门研究课题。

数据重删技术（data deduplication）是指通过识别数据流中的冗余，只传输或存储唯一数据（unique data），而使用指向已存储数据的指针替换重复副本，以达到节省带宽或存储空间的目的[4]。由于能够有效地降低存储开销，数据重删技术非常适合为管理日益增长的海量数据节省成本。在工业界，EMC Data Domain[5] 和 Avamar[6]、Veritas 的 NetBackup Appliances[7] 以及 Commvault 的开放数据平台[8]都是比较知名的数据重删应用产品；此外，各大云存储厂商（例如 Dropbox、Google Drive、Bitcasa、Moza 等）也纷纷将数据重删技术应用于各自的云服务产品中，以提升经济效益[9]。

如图1-1所示，在支持数据重删的存储系统（统称为数据重删系统）中，重删后的任何数据块都被一个或多个文件引用，而文件则以指向这些数据块的指针的集合形式存储。这种文件共用数据块的存储模式强调了数据块的敏感性，因为一个数据块的泄漏可能扩散影响到共用这个数据块的所有文件。如何保护重删后的数据的隐私，成为信息安全领域的一个研究热点。

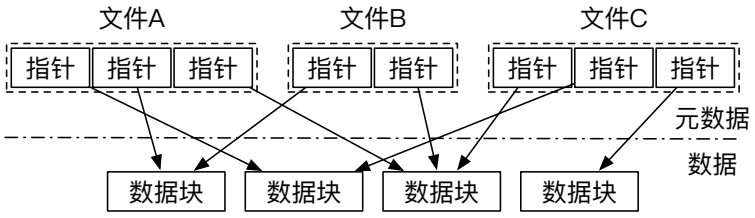

图 1-1 数据重删系统的存储模式

为了保护数据隐私，加密重复数据删除（encrypted deduplication）增加了一层

作用于逻辑数据块的加密操作。如图1-2所示，该加密层基于数据内容来产生加密密钥[10]（例如将数据块的哈希值作为密钥[11]），从而将相同的明文数据块加密为相同的密文数据块。系统计算每个密文数据块的哈希值(称为指纹，fingerprint)，查询指纹索引（fingerprint index）确定该数据块是否已经存储，最后保存仅具有唯一指纹的密文数据块。需要指出的是，部分加密重复数据删除方案[10]采用随机加密算法，但基于明文数据块产生指纹，因此仍然可以通过检查指纹来识别重复数据。

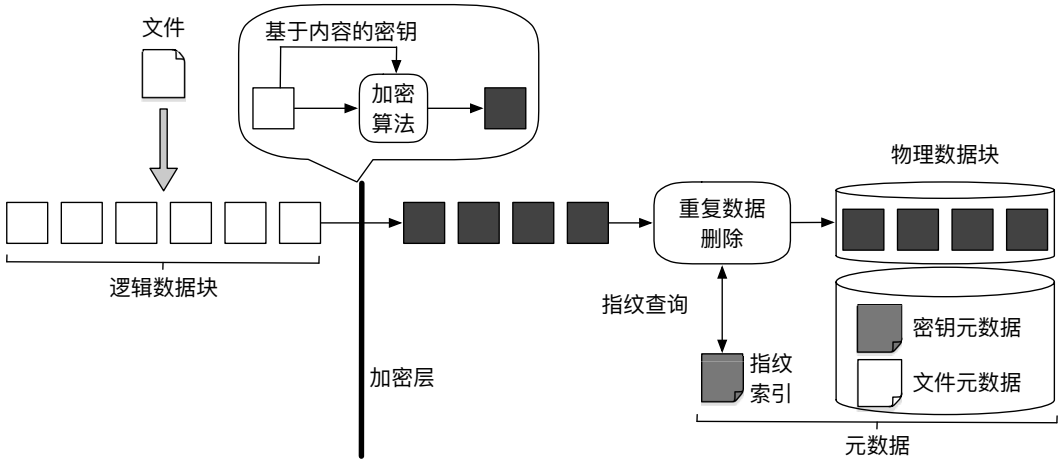

图 1-2 加密重复数据删除系统逻辑视图

除了指纹索引以外，加密重复数据删除系统须存储额外的元数据（metadata），包括:

1. 文件元数据记录了文件内逻辑数据块与相应物理数据块的映射关系，用于重构完整文件。
2. 密钥元数据记录了文件内逻辑数据块的解密密钥，用于恢复相应的明文内容，由于密钥元数据包含密钥信息，需由文件属主的主密钥（master key）加密后以密文形式存储。

1.1.2 问题和动机

数据块频率泄漏问题

由于基于数据块内容产生密钥，加密重复数据删除泄漏了数据块的频率信息，即如果一个明文数据块出现了 n 次，则它对应的密文数据块也将出现 n 次。另一方面，真实数据集中数据块的出现频率往往呈非均匀分布，调研了 FSL 和 VM 备份数据集的数据块频率分布特征，发现三种数据集有超过 97% 的数据块的频率低于 100 次，而至多只有 0.04% 的数据块的频率高于 10,000 次（图1-3），这种非均匀的分布特点使攻击者可以利用频率来确定相应数据块。

基于以上原因，认为加密重复数据删除可能受到频率分析[12]的威胁，拟通过

本课题，深入研究频率分析攻击对加密重复数据删除安全性的影响，以及提高频率分析攻击效果的方法。

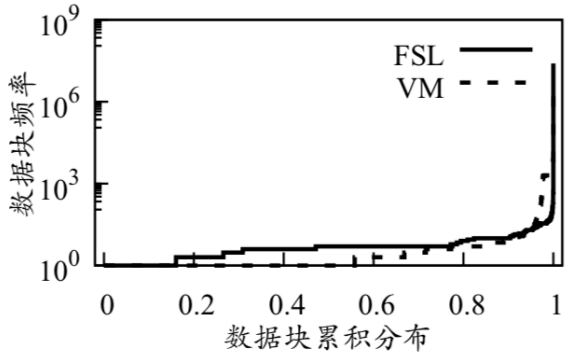

图 1-3 两种真实数据集的数据块频率分布

1.1.3 研究意义

本课题研究将填补频率分析攻击研究空白，对理解加密重复数据删除的实际安全性，并降低其在非适合场景下的误用风险具有重要作用。

尽管加密重复数据删除的频率泄漏已是学术界公认的安全问题，但针对性的频率分析攻击研究仍为空白 (即利用频率泄漏来获取隐私数据仍是一个开放性问题)，致使部分厂商盲目地将加密重复数据删除技术应用于商业产品[13,14] 和开源系统[15–18] 中。本项目将研究加密重复数据删除技术在频率分析攻击下的实际安全性，以指导其在适合场景下被正确使用。

1.2 国内外研究历史与现状

1.2.1 加密重复数据删除

在传统对称加密方式下，每个用户具有不同的密钥，不同用户之间的相同明文会被加密为不同密文，难以执行 (密文) 重复数据删除操作。

消息锁定加密（message-locked encryption，MLE）确立了加密重复数据删除的密码学基础[10]: 基于数据内容产生密钥 (称为 MLE 密钥)，从而将相同明文加密为相同密文。最流行的 MLE 实例是收敛加密（convergent encryption，CE）[11]，它使用明文的哈希值作为 MLE 密钥，并基于密文哈希值计算指纹，以识别重复数据（如图1-2）。基于 CE 的加密重复数据删除方案还包括:

1. 哈希收敛加密（hash convergent encryption，HCE）[11] 与 CE 具有相同的 MLE 密钥产生规则，但基于明文哈希值计算指纹。
2. 随机收敛加密（random convergent encryption，RCE）[11] 使用随机密钥加密

以产生非确定的密文，同时也基于明文哈希值来进行重复检查。

3. 收敛扩散（convergent dispersal，CD）[19] 使用明文哈希值作为秘密共享（secret sharing）的输入种子，在兼容重复数据删除的基础上提高了密文存储的可靠性。

上述 MLE 实例基于明文产生 MLE 密钥（CE、HCE 和 CD）或指纹（HCE 和 RCE），如果明文是可预测的 (即所有可能的明文的数量有限)，这些方案易于受到离线暴力破解攻击[10,20]。为了抵御该攻击，DupLESS[20] 基于第三方密钥服务器实现了服务器辅助 MLE（server-aided MLE），确保无法从离线明文派生出相应的 MLE 密钥。以服务器辅助 MLE 为基础，现有研究进一步解决了加密重复数据删除的故障容错[21,22]、透明价格模型[23]、点对点密钥管理[24]、层次密钥管理[25] 等问题。围绕 MLE 扩展功能的一系列研究包括: 兼容加密重复数据删除的数据完整性审计协议[26]；支持动态访问控制的加密重复数据删除系统 REED[27,28]。

无论是 MLE 还是服务器辅助 MLE，均须为相同的明文产生相同的密文（CE、HCE、CD 和 DupLESS）或指纹（HCE 和 RCE），泄漏了数据的出现频率。一些理论研究[29–31] 基于零知识证明、全同态加密、双线性对运算等底层密码技术实现了随机加密，但这些方案存在计算复杂性高、依赖多轮信息交互等问题，难以应用到系统实践中。本文关注可实际应用的加密重复数据删除系统/方案的频率泄漏问题，研究其安全性影响和防御对策。

1.2.2 频率分析攻击

频率分析[32] 是一种针对确定性加密（deterministic encryption）的密码分析技术，被应用于破解匿名查询日志[33]、破坏关键词隐私[34–38]、重构密文数据库记录[12,39–43] 等实际攻击。本项目研究针对数据块的频率分析攻击，与现有攻击目标（查询日志条目、关键词、数据库记录等）相比，数据块数量极其庞大（呈千万级），并且大量数据块具有相同频率，致使当前的频率分析攻击算法难以适用。

面向加密重复数据删除，已有工作[44] 提出了基于数据块局部性（chunk locality）[45–47] 的频率分析攻击。本课题的部分技术路线（1.4）就是在该工作[44] 的基础上提出来的，并进一步研究频率分析攻击的准确率和依赖条件，以及面向真实系统的攻击原型。除了频率分析和暴力破解（1.2.1）之外，加密重复数据删除还可能遭受边信道攻击[9,48,49]、副本伪造攻击[10]、基于数据块长度的攻击[50] 等威胁，但这些攻击可通过所有权证明[48]、守卫解密 (guarded decryption)[10]、固定长度分块等措施进行防御，而本文所研究的频率分析攻击超出了现有保护措施的防御范畴。

1.3 课题的研究内容、研究目标、以及拟解决的关键问题

1.3.1 研究内容

加密重复数据删除的频率分析攻击

在传统频率分析模式下，攻击者能够访问明文逻辑数据块集合 M 和密文逻辑数据块集合 C（M 和 C 包含重复的明文和密文数据块）。攻击者根据出现频率分别对 M 和 C 中的数据块进行排序，然后将 C 中的密文数据块映射为 M 中与其具有相同排名的明文数据块。但是，传统频率分析在加密重复数据删除中难以形成有效的攻击，主要原因是：M 和 C 的原始内容可能存在差异 (例如 M 和 C 来源于同一个文件系统在不同时间点的备份镜像)，将打乱数据块频率排序的对应关系；并且，在频率排序过程中可能存在大量明文和密文数据块具有相同的频率，频率分析难以排序这些数据块来形成正确的对应关系。

为了提高传统频率分析的攻击效果，首先研究基于数据特征的新型频率分析攻击技术。然后，分别从抵抗频率排序干扰和降低攻击发生条件两方面改进攻击技术。最后，实现针对真实系统的频率分析攻击原型，并分析该攻击对各类数据安全性的影响。

1.3.2 研究目标

针对以上研究内容，预期实现如下研究目标:

1. 在理论上，构造针对加密重复数据删除的频率分析攻击，揭示实践中的安全隐患。
2. 在技术上，以理论研究为支撑，设计并实现针对加密重复数据删除系统/方案的频率分析攻击工具，并在真实系统中进行理论验证和攻击效果测试。

1.3.3 拟解决的关键问题

本课题致力于解决传统频率分析攻击方法在针对加密重复数据删除方案/系统进行攻击时难以解决的如下问题:

1. 明密文数据块的的原始内容可能存在差异 (例如一组明密文集合分别来源于同一个文件系统在不同时间点的备份镜像)，将打乱数据块频率排序的对应关系。
2. 在频率排序过程中可能存在大量明文和密文数据块具有相同的频率，基本频率分析攻击难以排序这些数据块来形成正确的对应关系。

1.4 拟采取的研究方案

如图1-4所示，在支撑研究（基于数据块局部性的频率分析攻击方案[44]）的基础上，根据相对频率分布特性，设计抗排序干扰的攻击方法；在数据相似性的基础上，设计低依赖条件的攻击方法。最终设计出针对加密重复数据删除方案/系统的新型频率分析攻击方法及对应的原型软件工具。

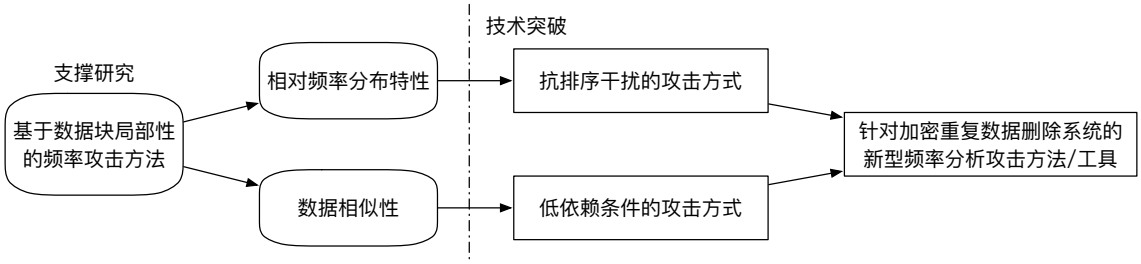

图 1-4 技术路线图

1.4.1 支撑研究

基于数据块局部性的频率分析攻击方案（locality-based attack）[44] 作为支撑研究，该方案主要用于破译加密数据备份，即已知的明文数据块集合 M 和目标密文数据块集合 C 源于同一个系统在两个不同时间点的备份镜像。攻击利用了数据块的局部性特征：在不同备份之间，绝大多数数据块保持了相同的局部顺序；例如，每天备份工作项目的进度快照，若一天内的改动较小，则在两次备份之间未被改动的大部分数据块之间的相对顺序保持不变。因此，得出一个关键推论：如果明文数据块 M 是密文数据块 C 的原始明文，那么 M 左边和右边相邻的明文数据块有较大可能也是 C 左边和右边相邻密文数据块的原始明文。

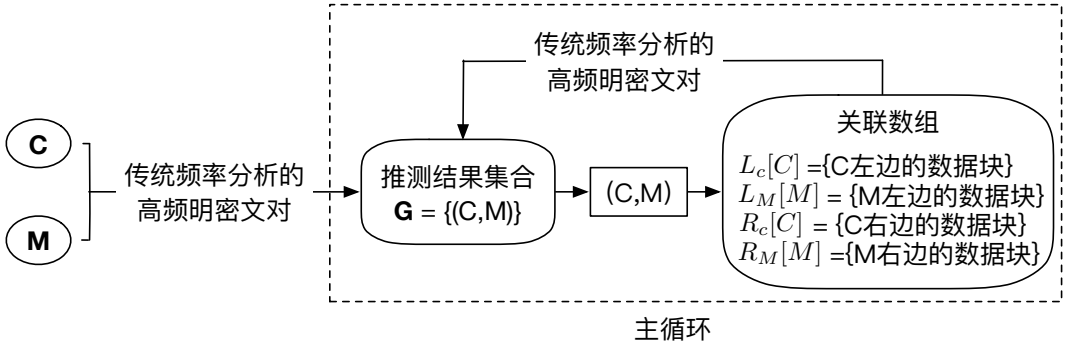

图 1-5 基于数据块局部性的频率分析攻击

基于此，攻击流程如图1-5所示：首先，对 C 和 M 应用传统频率分析（1.2.2），将获得的若干组高频明密文对加入推测结果集合 G；然后，每次从 G 中选取一组

明密文对 (M, C)，分别对其左右相邻的明文和密文数据块集合 $L_M[M]$ 和 $L_C[C]$，以及 $R_M[M]$ 和 $R_C[C]$ 实施频率分析，并将获得的高频明密文对也加入 G；继续对 G 中的明密文对进行基于相邻数据块的频率分析，直至所有明密文对都被处理。

为了验证攻击效果，定义推测率为正确推测出原始明文的 (不同) 密文数据块个数与 C 中 (不同) 密文数据块总个数的比率。在基于真实数据集的实验验证中，攻击方案能够达到 17.8% 推测率，远远高于传统频率分析方法的 0.0001% 推测率。

1.4.2 技术突破

1.4.2.1 基于分布的频率分析攻击方法

基于分布的频率分析攻击利用密文和明文的相对顺序信息来增强频率分析的有效性。该攻击方法建立在数据块局部性 [45–47] 的基础上。通过以下三个关键特性构造出基于分布的频率分析攻击方法。

- 数据块局部性指出数据的原始排序可能会在各种备份文件中得以保留，因此可以用于在备份文件中推理出类似的与位置相关的密文-明文对。
- 相邻数据的共现频率的相对频率分布可以通过分析得到。
- 明文及其相应的密文具有相似的相对频率分布的特性。

1.4.2.2 基于聚类的频率分析攻击方法

在基于分布的频率分析攻击方法的基础上，通过引入相似性 [51] 这一属性来消除基于分布的频率分析攻击方法对明文数据的细粒度顺序信息的需求。基于此提出基于聚类的频率分析攻击方法。

1.5 可行性分析

1.5.1 研究方法可行

本课题通过引入加密重复数据删除方案/系统的相关特性来设计新型频率分析攻击方法，针对性提高频率分析攻击方法的有效性。该研究问题具有很高的实际价值且在面向加密重复数据删除的攻击中，已有工作提出了基于数据块局部性（chunk locality） [45–47] 的频率分析攻击 [44]，在此基础上改进攻击方案以提高频率分析攻击实际效果具有明确的可行性。

1.5.2 研究条件可行

前期基于数据块局部性的频率分析攻击方案给出了本课题研究两种新型频率分析攻击方法的基本理论和工具基础。在原有方案上改进即可用于本课题研究。

研究者在本科阶段深入钻研了 CDStore、REED、SDFS 等加密重复数据删除系统，以及前期基于数据块局部性的频率分析攻击方案的理论和工具实现。这些经验可以帮助设计新型频率分析方法以及其在真实系统中的实践。

1.6 本文特色与创新之处

本研究的特色与创新之处有以下几点：

1. 研究针对加密重复数据删除提出了两种新型的频率分析攻击方法。除了利用由于加密重复数据删除具有确定性导致的频率泄漏之外，两种攻击都利用重复数据删除工作负载的特性来增加频率分析攻击的有效性。
2. 本文研究使用多个真实数据集（包括长期备份[52,53]，Windows 文件系统快照[54]和 VM 磁盘映像[19,28]）以及开源重复数据删除系统 SDFS[?]、Destor[?]，评估两种新型频率分析攻击方法和基本频率分析攻击方法的实际效果。通过评估频率攻击方法的攻击结果，进一步分析本频率分析攻击方法对实际加密重复数据删除带来的安全性影响。
3. 本研究讨论了减少加密重复数据删除中信息泄漏的可能方案。

参考文献

[1] Gartner. Gartner forecasts 59 percent mobile data growth worldwide in 2015[EB/OL]. http://www.gartner.com/newsroom/id/3098617, Dec 12, 2018

[2] 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

[3] J. McKnight, T. Asaro, B. Babineau. Digital archiving: End-user survey and market forecast 2006–2010[J]. The Enterprise Strategy Group, 2006,

[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

[5] EMC. Emc data domain[EB/OL]. http://www.emc.com/en-us/data-protection/data-domain.htm, Dec 12, 2018

[6] EMC. Avamar deduplication backup software and system[EB/OL]. http://www.emc.com/data-protection/avamar.htm, Dec 12, 2018

[7] veritas. Netbackup appliances[EB/OL]. https://www.veritas.com/product/backup-and-recovery/netbackup-appliances.html, Dec 12, 2018

[8] CommVault. Commvault solutions for data protection backup and recovery[EB/OL]. http://www.commvault.com/solutions/by-function/data-protection-backup-and-recovery, Dec 12, 2018

[9] D. Harnik, B. Pinkas, A. Shulman-Peleg. Side channels in cloud services: Deduplication in cloud storage[J]. IEEE Security & Privacy, 2010, 6): 40-47

[10] M. Bellare, S. Keelveedhi, T. Ristenpart. Message-locked encryption and secure deduplication[C]. Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2013, 296-312

[11] J. R. Douceur, A. Adya, W. J. Bolosky, et al. Reclaiming space from duplicate files in a serverless distributed file system[C]. Distributed Computing Systems, 2002. Proceedings. 22nd International Conference on, 2002, 617-624

[12] M. Naveed, S. Kamara, C. V. Wright. Inference attacks on property-preserving encrypted databases[C]. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, 644-655

[13] MEGA. Mega[EB/OL]. https://mega.nz, Dec 12, 2018

[14] ElephantDrive. Elephantdrive[EB/OL]. https://www.elephantdrive.com, Dec 12, 2018

[15] Cryptosphere. Cryptosphere[EB/OL]. https://cryptosphere.io, Dec 12, 2018

[16] Freenet. Freenet[EB/OL]. https://freenetproject.org, Dec 12, 2018

[17] GNU. Gnu's framework for secure peer-to-peer networking[EB/OL]. https://gnunet.org, Dec 12, 2018

[18] Tahoe-LAFS. Tahoe-lafs[EB/OL]. https://tahoe-lafs.org/trac/tahoe-lafs., Dec 12, 2018

[19] M. Li, C. Qin, J. Li, et al. Cdstore: Toward reliable, secure, and cost-efficient cloud storage via convergent dispersal[J]. IEEE Internet Computing, 2016, 3): 45-53

[20] S. Keelveedhi, M. Bellare, T. Ristenpart. Dupless: server-aided encryption for deduplicated storage[C]. USENIX Security 13, 2013, 179-194

[21] M. Li, C. Qin, P. P. Lee. Cdstore: Toward reliable, secure, and cost-efficient cloud storage via convergent dispersal.[C]. USENIX Annual Technical Conference, 2015, 111-124

[22] Y. Duan. Distributed key generation for encrypted deduplication: Achieving the strongest privacy[C]. Proceedings of the 6th edition of the ACM Workshop on Cloud Computing Security, 2014, 57-68

[23] F. Armknecht, J.-M. Bohli, G. O. Karame, et al. Transparent data deduplication in the cloud[C]. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, 886-900

[24] J. Liu, N. Asokan, B. Pinkas. Secure deduplication of encrypted data without additional independent servers[C]. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, 874-885

[25] Y. Zhou, D. Feng, W. Xia, et al. Secdep: A user-aware efficient fine-grained secure deduplication scheme with multi-level key management[C]. Mass Storage Systems and Technologies (MSST), 2015 31st Symposium on, 2015, 1-14

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

[27] J. Li, C. Qin, P. P. Lee, et al. Rekeying for encrypted deduplication storage[C]. Dependable Systems and Networks (DSN), 2016 46th Annual IEEE/IFIP International Conference on, 2016, 618-629

[28] C. Qin, J. Li, P. P. Lee. The design and implementation of a rekeying-aware encrypted deduplication storage system[J]. ACM Transactions on Storage (TOS), 2017, 13(1): 9

[29] M. Abadi, D. Boneh, I. Mironov, et al. Message-locked encryption for lock-dependent messages[M]. Springer, 2013, 374-391

[30] J. Stanek, A. Sorniotti, E. Androulaki, et al. A secure data deduplication scheme for cloud storage[C]. International Conference on Financial Cryptography and Data Security, 2014, 99-118

[31] M. Bellare, S. Keelveedhi. Interactive message-locked encryption and secure deduplication[C]. IACR International Workshop on Public Key Cryptography, 2015, 516-538

[32] J. Katz, A. J. Menezes, P. C. Van Oorschot, et al. Handbook of applied cryptography[M]. CRC press, 1996

[33] R. Kumar, J. Novak, B. Pang, et al. On anonymizing query logs via token-based hashing[C]. Proceedings of the 16th international conference on World Wide Web, 2007, 629-638

[34] D. Cash, P. Grubbs, J. Perry, et al. Leakage-abuse attacks against searchable encryption[C]. Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015, 668-679

[35] P. Grubbs, R. McPherson, M. Naveed, et al. Breaking web applications built on top of encrypted data[C]. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, 1353-1364

[36] M. S. Islam, M. Kuzu, M. Kantarcioglu. Access pattern disclosure on searchable encryption: Ramification, attack and mitigation.[C]. Ndss, 2012, 12

[37] D. Pouliot, C. V. Wright. The shadow nemesis: Inference attacks on efficiently deployable, efficiently searchable encryption[C]. Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, 1341-1352

[38] Y. Zhang, J. Katz, C. Papamanthou. All your queries are belong to us: The power of file-injection attacks on searchable encryption.[C]. USENIX Security Symposium, 2016, 707-720

[39] V. Bindschaedler, P. Grubbs, D. Cash, et al. The tao of inference in privacy-protected databases[J]. Proceedings of the VLDB Endowment, 2018, 11(11): 1715-1728

[40] F. B. Durak, T. M. DuBuisson, D. Cash. What else is revealed by order-revealing encryption?[C]. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, 1155-1166

[41] P. Grubbs, K. Sekniqi, V. Bindschaedler, et al. Leakage-abuse attacks against order-revealing encryption[C]. Security and Privacy (SP), 2017 IEEE Symposium on, 2017, 655-672

[42] G. Kellaris, G. Kollios, K. Nissim, et al. Generic attacks on secure outsourced databases[C]. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, 1329-1340

[43] M.-S. Lacharité, B. Minaud, K. G. Paterson. Improved reconstruction attacks on encrypted data using range query leakage[C]. 2018 IEEE Symposium on Security and Privacy (SP), 2018, 297-314

[44] J. Li, C. Qin, P. P. Lee, et al. Information leakage in encrypted deduplication via frequency analysis[C]. Dependable Systems and Networks (DSN), 2017 47th Annual IEEE/IFIP International Conference on, 2017, 1-12

[45] B. Zhu, K. Li, R. H. Patterson. Avoiding the disk bottleneck in the data domain deduplication file system.[C]. Fast, 2008, 1-14

[46] M. Lillibridge, K. Eshghi, D. Bhagwat, et al. Sparse indexing: Large scale, inline deduplication using sampling and locality.[C]. Fast, 2009, 111-123

[47] W. Xia, H. Jiang, D. Feng, et al. Silo: A similarity-locality based near-exact deduplication scheme with low ram overhead and high throughput.[C]. USENIX annual technical conference, 2011, 26-30

[48] S. Halevi, D. Harnik, B. Pinkas, et al. Proofs of ownership in remote storage systems[C]. Proceedings of the 18th ACM conference on Computer and communications security, 2011, 491-500

[49] M. Mulazzani, S. Schrittwieser, M. Leithner, et al. Dark clouds on the horizon: Using cloud storage as attack vector and online slack space.[C]. USENIX security symposium, 2011, 65-76

[50] H. Ritzdorf, G. Karame, C. Soriente, et al. On information leakage in deduplicated storage systems[C]. Proceedings of the 2016 ACM on Cloud Computing Security Workshop, 2016, 61-72

[51] D. Bhagwat, K. Eshghi, D. D. Long, et al. Extreme binning: Scalable, parallel deduplication for chunk-based file backup[C]. Modeling, Analysis & Simulation of Computer and Telecommunication Systems, 2009. MASCOTS'09. IEEE International Symposium on, 2009, 1-9

[52] Z. Sun, G. Kuenning, S. Mandal, et al. A long-term user-centric analysis of deduplication patterns[C]. Mass Storage Systems and Technologies (MSST), 2016 32nd Symposium on, 2016, 1-7

[53] FSL. FSL traces and snapshots public archive[EB/OL]. http://tracer.filesystems.org/, Dec 12, 2018

[54] D. T. Meyer, W. J. Bolosky. A study of practical deduplication[J]. ACM Transactions on Storage (TOS), 2012, 7(4): 14