上述算法如图 5 所示.

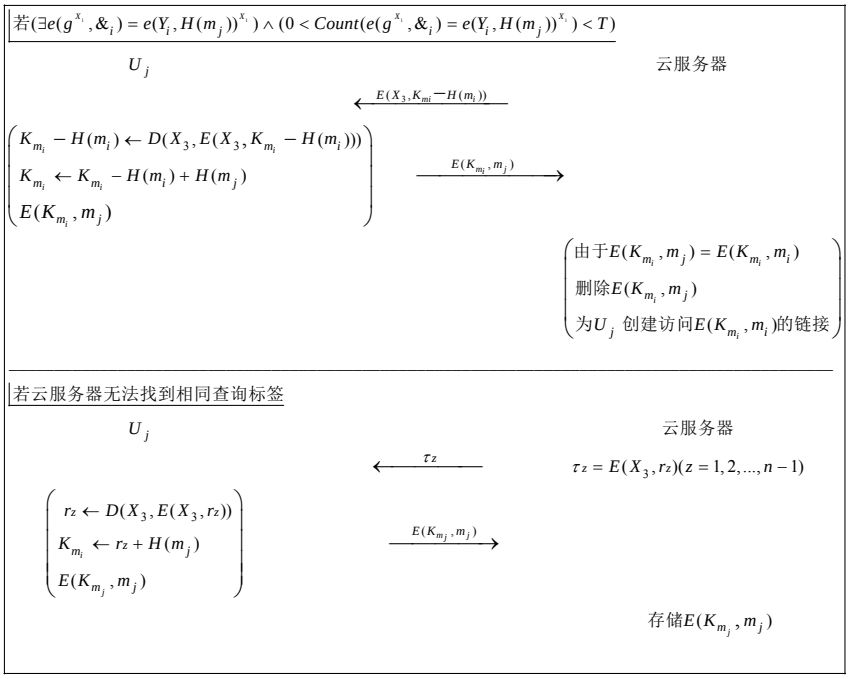

Fig.5 UnpopularDedup

图 5 非流行数据重复删除

4.5 PopularDedup

由于流行数据块的隐私度较低,因此可采用改进后的收敛加密算法对其加密.U_j 查询得知 m_j 为流行数据块,云服务器计算拥有 m_j 的用户数量 $Count(m_j)$.

如图 6 所示,其中,a 为当 $Count(m_j)=T$ 时,U_j 使用 $X_j=H(m_j)+X_2$ 对 m_j 加密得到 $C=E(X_j,m_j)$,将密文 C 上传至云服务器.b 为若 $Count(m_j)>T$,则改用效率更高的客户端重复数据删除(client-side deduplication),U_j 无需上传密文,云服务器为 U_j 创建访问此加密数据块的链接.

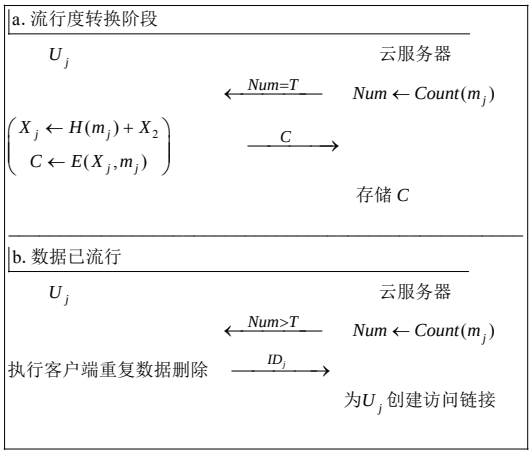

Fig.6 PopularUpload

图 6 流行数据块上传

5 安全性分析与证明

本节从以下 4 个方面详细分析该方案的安全性.

(1) 数据块验证的安全性证明

通过直接比较散列值判断数据块是否相同的方法,容易遭受云服务器的离线穷举攻击.本方案可有效避免此类威胁,并且能够将云服务器中存储的加密数据块与初始上传者的身份绑定.安全性证明如下文所述.

① 数据块验证结果的唯一性

本方案的安全性建立在特殊散列函数 H 安全性的基础上,G_1 表示阶为大素数 P 的加法循环群,g 表示 G_1 的生成元.由 H 的安全性假设,可得以下引理.

引理 1. 对于安全的特殊散列函数 H:$\{0,1\}^*\rightarrow G_1$,不同的数据块 m_i 与 m_j 拥有相同散列值的概率是可忽略的.我们采用ε表示可忽略值.

$$Prob\left[H(m_i)=H(m_j)\mid m_i\neq m_j\right]<\varepsilon.$$

定理 1. 在验证数据块是否相同时,初始上传者 U_i 的查询标签为 $e(g^{X_1},\&_i)$,当前上传者 U_j 的查询标签为 $e(Y_i,H(m_j))^{X_1}$.当 $m_i\neq m_j$ 时,$e(g^{X_1},\&_i)$与 $e(Y_i,H(m_j))^{X_1}$ 相同的概率是可忽略的.

$$Prob\left[e(g^{X_1},\&_i)=e(Y_i,H(m_j))^{X_1}\mid m_i\neq m_j\right]<\varepsilon.$$

证明:不失一般性,由双线性映射的性质可得以下推论:

$$e(Y_i,H(m_j))^{X_1}=e(g^{y_i},H(m_j))^{X_1}=e(g,H(m_j))^{y_i\cdot X_1}=e(g^{X_1},H(m_j)^{y_i}).$$

由引理 1 可得,若 $m_i\neq m_j$, 则 $H(m_i)\neq H(m_j)$, 故 $e(g^{X_1},\&_i)=e(g^{X_1},H(m_i)^{y_i})\neq e(g^{X_1},H(m_j)^{y_i})$, 即:

$$Prob\left[e(g^{X_1},\&_i)=e(Y_i,H(m_j))^{X_1}\mid m_i\neq m_j\right]<\varepsilon.$$

换言之,当且仅当 m_i=m_j 时,$e(g^{X_1},\&_i)$=$e(Y_i,H(m_j))^{X_1}$ 才会成立.因此,数据块的验证结果是唯一的. □

② 数据块验证结果的正确性

定理 2. 若等式 $e(g^{X_1},\&_i)$=$e(Y_i,H(m_j))^{X_1}$ 成立,m_i 与 m_j 不同的概率是可忽略的.

$$Prob\left[m_i\neq m_j\mid e(g^{X_1},\&_i)=e(Y_i,H(m_j))^{X_1}\right]<\varepsilon.$$

证明:不失一般性,假设 U_i 的查询标签为 $e(g^{X_1},\&_i)$,由双线性映射性质可得以下等式:

$$e(g^{X_1},\&_i)=e(g^{X_1},H(m_i)^{y_i})=e(g,H(m_i))^{X_1\cdot y_i}=e(g^{y_i},H(m_i))^{X_1}=e(Y_i,H(m_i))^{X_1}.$$

由引理 1 可得: $m_i=m_j\leftarrow e(Y_i,H(m_i))^{X_1}=e(Y_i,H(m_j))^{X_1}\leftarrow e(g^{X_1},\&_i)=e(Y_i,H(m_j))^{X_1}$.

因此: $Prob\left[m_i\neq m_j\mid e(g^{X_1},\&_i)=e(Y_i,H(m_j))^{X_1}\right]<\varepsilon.$ □

(2) 防止查询标签泄露隐私数据块的明文信息

引理 2. G_1 表示阶为大素数 P 的乘法循环群,给定 $g, g^a, h\in G_1$, 其中, $a\in Z_P^*$, 计算 $h^a\in G_1$ 是困难的(计算 Diffie-Hellman 问题的困难性).

定理 3. 在用户不与云服务器合谋的情况下,云服务器无法以离线穷举攻击的方式从查询标签中获取数据块的任何明文信息.

证明:云服务器对用户发来的查询标签 $e(Y_i,H(m_j))^{X_1}$ 进行离线穷举攻击.

穷举大量数据块 $\{m_r\}, r\in(1,2,3,...,n)$, 试图找到 m_r=m_j.为了验证 m_r 与 m_j 是否相等,云服务器需要计算 $e(Y_i,H(m_j))^{X_1}$ 并将其与 $e(Y_i,H(m_j))^{X_1}$ 进行比较.云服务器容易计算 $e(Y_i,H(m_r))$,但由引理 2 可知,即使持有 $e(Y_i,H(m_j))$ 和 $e(Y_i,H(m_j))^{X_1}$,计算 $e(Y_i,H(m_r))^{X_1}$ 也是困难的.因此,云服务器无法以离线穷举攻击的方式从查询标签中获取数据块明文的任何信息. □

(3) 防止用户进行在线穷举攻击

定理 4. 在本方案中,用户 U_D 在持有辅助密钥 X_1 的情况下,无法对存储在云服务器中的非流行数据块进行

在线穷举攻击.

证明:

① U_D 穷举数据块 $\{m_r\}, r \in (1,2,3,...,n)$, 构造集合 $\{e(Y_i, H(m_r))^{X_1}\}$.

② U_D 将集合中的元素逐一发送至云服务器.

③ 云服务器根据是否存在等式 $e(g^{X_1},\&_i)=e(Y_i,H(m_j))^{X_1}$,回复 U_D 相应的信息.

④ U_D 根据回复的信息判断哪些数据块存储在云服务器.

由算法 UnpopularDedup 可知:

情况(a).当 $m_i=m_r$ 时,云服务器将 $E(X_3, K_{m_i} - H(m_i))$ 回复给 U_D.

情况(b).若云服务器中不存在 m_r,云服务器在密钥池中随机选择 $\tau = E(X_3, r_z)(z = 1,2,...,n-1)$, 并回复给 U_D.

由于两种情况下 U_D 获得的 $K_{m_i} - H(m_i)$ 和 r_z 是由相同方法得到伪随机数,U_D 无法区分情况(a)和情况(b),故无法对存储在云服务器的非流行数据块进行在线穷举攻击. □

(4) 防止恶意用户截取信息

由于恶意用户 U_D 是广播中心 BC 的授权用户,因此 U_D 拥有广播信息 $M=(X_1,X_2)$.假设 U_D 截获了用户 U_i 上传的查询标签 $e(g^{X_1},\&_i)$,并对 $e(g^{X_1},\&_i)$采取离线穷举攻击.

① 穷举数据块 $\{m_r\}, r \in (1,2,3,...,n)$.

② 构造集合 $\{e(Y_i, H(m_r))^{X_1}\}, r \in (1,2,3,...,n)$.

③ 查看是否存在以下等式 $e(g^{X_1},\&_i)=e(Y_i,H(m_r))^{X_1}$.

若 $e(g^{X_1},\&_i)=e(Y_i,H(m_r))^{X_1}$,则 $m_i=m_r$,数据块 m_i 的明文信息便遭到泄露.

解决方法:

① 系统初期,云服务器随机生成密钥对$\langle PK_{CSP}, SK_{CSP}\rangle$.

② 云服务器将 PK_{CSP} 发送至 U_i.

③ U_i 使用 PK_{CSP} 加密 $e(g^{X_1},\&_i)$得到密文 $Enc(PK_{CSP}, e(g^{X_1},\&_i))$,并将密文发送至云服务器.

如此,即使 U_D 截取查询标签也无法对其造成任何安全威胁.

6 仿真与实验分析

实验采用 PBC[27]、GMP[28]、PBC_bce[29]和 OPENSSL[30]函数库,使用 C++语言编程实现了客户端与服务器软件.选用腾讯云的云服务器,其配置为 4GB 内存,4 核 CPU,1Mbps 带宽,1T 存储盘.设定大小为 512bit 的基域,其中每个元素 $element \in Z_P^*$ 的大小为$|P|$=160bit.为了模拟真实情景,我们在云服务器中存储了超过 2 000 个不同的文件,且随机设定拥有每个文件的用户数量 $Count_F$.建立用户数量表 $CountTab$,记录 $Count_F$.设定流行度阈值 T=7,使非流行数据与流行数据的比例大致为 2:3.

实验共分 3 部分:首先,上传一个大小为 30MB 的文件 F_A,记录方案中各阶段所需的时间开销.然后,上传大小为 10MB 的文件 F_B,计算方案所需的总时间开销,并与 perfectDedup 方案对比.最后,以上传 100MB、500MB 的文件为例,分别计算在本方案、perfectDedup 方案与不进行重复数据删除方案(NoDedup)中的存储开销,以此验证本方案在重复数据删除中的高效性.每部分操作重复进行 10 次,取平均值作为最终结果.

(1) 各阶段所需时间开销

由于广播加密仅在系统建立初期执行一次,因此,其所需时间开销不进行计算.非流行数据上传实验结果如图 7 所示.由于方案将大部分的计算外包给云服务器,用户端数据分块、标签生成与对称加密所需时间开销非常小.相对而言,发生在云服务器端的流行度识别与密文上传所需时间开销较大.流行数据上传实验结果如图 8 所示,各阶段所需时间开销与非流行数据上传大致相同.如图 9 所示,当 $Count_{FA}>T$ 时,只需要进行客户端重复数据删除,不再上传收敛加密密文,因此,显著减少了计算开销并节约了网络带宽.

(2) 更少的总时间开销

本文的方案与 perfectDedup 方案的对比结果如图 10~图 12 所示.在数据的流行度查询阶段(包括标签生成

与流行度识别),本方案所需时间开销明显低于 perfectDedup.此外,本方案摆脱了实时在线第三方 IS.因此,本方案在总时间开销上具有较明显的优势.

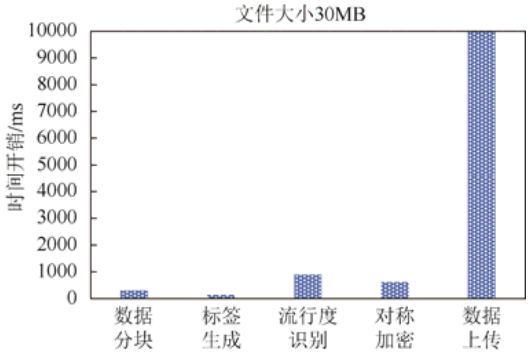

Fig.7 The time span for each phase of experiment ($Count_{FA}<T$)

图 7 实验中各阶段所需时间开销($Count_{FA}<T$)

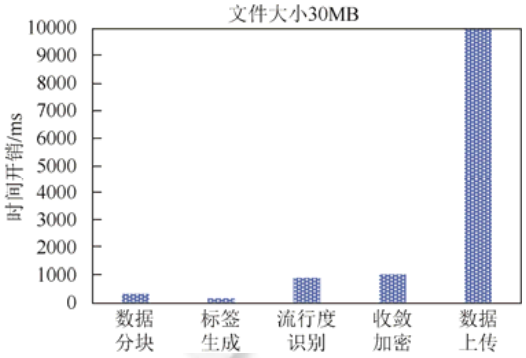

Fig.8 The time span for each phase of experiment ($Count_{FA}=T$)

图 8 实验中各阶段所需时间开销($Count_{FA}=T$)

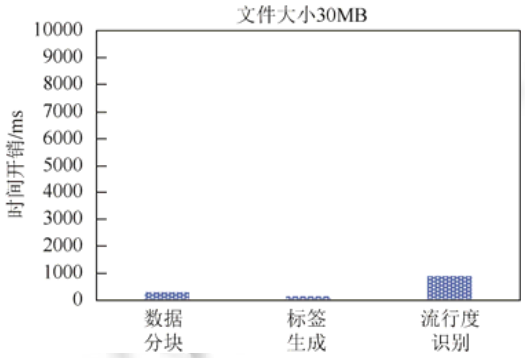

Fig.9 The time span for each phase of experiment ($Count_{FA}>T$)

图 9 实验中各阶段所需时间开销($Count_{FA}>T$)

Fig.10 Comparison of the time span between our scheme and the perfectDedup scheme ($Count_{FB}<T$)

图 10 本文方案与 perfectDedup 方案时间开销对比($Count_{FB}<T$)

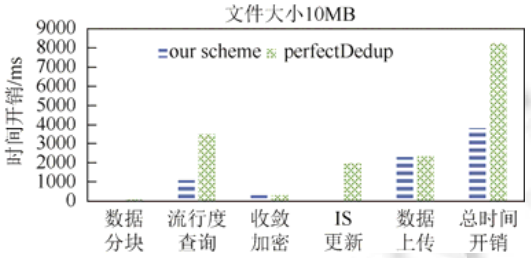

Fig.11 Comparison of the time span between our scheme and the perfectDedup scheme ($Count_{FB}=T$)

图 11 本文方案与 perfectDedup 方案时间开销对比($Count_{FB}=T$)

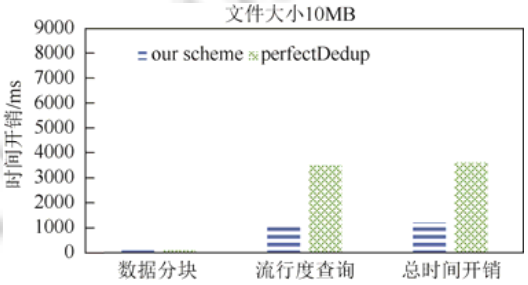

Fig.12 Comparison of the time span between our scheme and the perfectDedup scheme ($Count_{FB}>T$)

图 12 本文方案与 perfectDedup 方案时间开销对比($Count_{FB}>T$)

(3) 更少的存储空间开销

如图 13、图 14 所示,NoDedup 方案不执行重复数据删除,perfectDedup 方案无法删除非流行加密数据.本文方案的存储空间开销与持有数据的用户数量无关.且文件越大,本文方案的优势越明显.

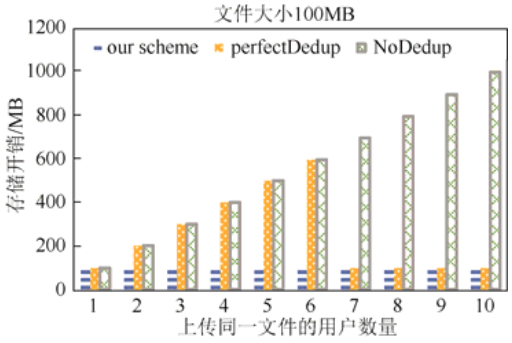

Fig.13 Cloud server storage costs of different schemes (Data size 100M)

图 13 3 种方案中云服务器存储开销对比(每个文件 100M)

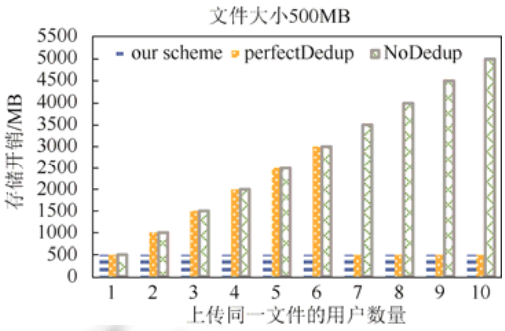

Fig.14 Cloud server storage costs of different schemes (Data size 500M)

图 14 3 种方案中云服务器存储开销对比(每个文件 500M)

(4) 性能分析比较

由以上实验结果可知,划分数据流行度、摆脱实时在线可信第三方能够显著提升重复数据删除方案的执行效率.表 1 给出了本文方案与其他代表性方案是否具有以上两个优点的分析和比较.

Table 1 Comparison of schemes characteristics

表 1 方案特点对比

方案	[7]	[8]	[9]	[15]	[17]	Our
划分数据流行度	×	√	√	×	×	√
摆脱实时在线可信第三方	×	×	×	×	√	√

7 总结与展望

本文研究了云存储环境下加密数据的重复删除问题,提出了一种基于离线密钥分发的加密数据重复删除方案.此方案通过构造语义安全的双线性映射,能够实现在不泄露数据任何明文信息的情况下完成流行度查询.通过广播加密为授权用户生成辅助密钥,保证非流行数据加密密钥的存储与传递的安全.持有相同非流行数据的不同用户能够获取相同的加密密钥,得到相同的加密数据,进而使云服务器能够对非流行数据进行重复数据删除.采用改进后的收敛加密算法保护隐私度较低的流行数据,用户能够自行生成加密密钥,进一步提高了方案的执行效率.通过安全分析与仿真实验,证明本方案具有较高的安全性与实用性.

如何摆脱广播中心,实现只有用户与云服务器两方交互的重复数据删除方案,是下一步的研究重点.

References:

[1] Fu YX, Luo SM, Shu JW. Survey of secure cloud storage system and key technologies. Journal of Computer Research and Development, 2013,50(1):136−145 (in Chinese with English abstract).

[2] Fu YJ, Xiao N, Liu F. Research and development on key techniques of data deduplication. Journal of Computer Research and Development, 2012,49(1):12−20 (in Chinese with English abstract).

[3] Ao L, Shu JW, Li MQ. Data deduplication techniques. Ruan Jian Xue Bao/Journal of Software, 2010,21(5):916−929 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3761.htm [doi: 10. 3724/SP.J.1001.2010.03761]

[4] Jeramiah B. Opendedup: Open-Source deduplication put to the test. Belltown Media, 2013. http://opendedup.org/

[5] Meyer DT, Bolosky WJ. A study of practical deduplication. ACM Trans. on Storage (TOS), 2012,7(4):14.

[6] Douceur JR, Adya A, Bolosky WJ, *et al*. Reclaiming space from duplicate files in aserverless distributed file system. In: Proc. of the ICDCS. IEEE, 2002. 617–624.

[7] Puzio P, Molva R, Onen M. Cloudedup: Secure deduplication with encrypted data for cloud storage. In: Proc. of the CloudCom. IEEE Computer Society, 2013. 363–370.

[8] Puzio P, Molva R, Önen M. PerfectDedup: Secure data deduplication. In: Proc. of the Int'l Workshop on Data Privacy Management. Springer Int'l Publishing, 2015. 150–166.

[9] Stanek J, Sorniotti A, Androulak E, *et al*. A secure data deduplication scheme for cloud storage. In: Christin N, Safavi-Naini R, eds. LNCS 8437. Springer-Verlag, 2014. 99–118.

[10] Xu J, Chang E C, Zhou J. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In: Proc. of the ACM SIGSAC Symp. on Information, Computer and Communications Security. ACM, 2013. 195–206.

[11] Adya A, Bolosky WJ, Castro M, *et al*. Farsite: Federated, available, and reliable storage for an incompletely trusted environment. ACM SIGOPS Operating Systems Review, 2002,36(SI):1–14.

[12] Hur J, Koo D, Shin Y, *et al*. Secure data deduplication with dynamic ownership management in cloud storage. IEEE Trans. on Knowledge and Data Engineering, 2016,28(11):1.

[13] Perttula. Attacks on convergent encryption. 2008. https://tahoe-lafs.org/hacktahoelafs/drew_perttula.html

[14] Bellare M, Keelveedhi S, Ristenpart T. Message-Locked encryption and secure deduplication. In: Proc. of the EUROCRYPT. LNCS 7881, Springer-Verlag, 2013. 296–312.

[15] Mihir B, Keelveedhi S, Ristenpart T. DupLESS: Server-Aided encryption for deduplicated storage. In: Proc. of the 22nd USENIX Conf. on Security. USENIX Association, 2013. 179–194.

[16] Douceur JR. The Sybil attack. In: Proc. of the Peer-to-Peer Systems. Springer-Verlag, 2002. 251–260.

[17] Liu J, Asokan N, Pinkas B. Secure deduplication of encrypted data without additional servers. Technical Report, 455, ePrint archive, 2015. https://eprint.iacr.org/2015/455

[18] Li L, Xue R, Zhang HG, Feng DG, Wang L. Security analysis of authenticated key exchange protocol based on password. ACTA ELECTRONICA SINICA, 2005,33(1):166–170 (in Chinese with English abstract).

[19] Hu XX, Zhang ZF, Liu WF. Universal composable password authenticated key exchange protocol in the standard model. Ruan Jian Xue Bao/Journal of Software, 2011,22(11):2820–2832 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3910.htm [doi: 10.3724/SP.J.1001.2011.03910]

[20] Cui H, Deng RH, Li Y. Attribute-Based storage supporting secure deduplication of encrypted data in cloud. IEEE Trans. on Big Data, 2016, 1–13.

[21] Zhang XS. The construction and calculation of bilinear pairs in cryptography [Ph.D. Thesis]. Beijing: The Chinese Academy of Sciences, 2012 (in Chinese with English abstract).

[22] Chen YM, Cheng XG, Wang S. Pairing certificateless signature scheme based on information network security. Netinfo Security, 2017,(3):53–58 (in Chinese with English abstract).

[23] Sakai R, Furukawa J. Identity-Based broadcast encryption. Journal of Electronics & Information Technology, 2007,33(4): 1047–1050.

[24] Delerablée C. Identity-Based broadcast encryption with constant size ciphertexts and private keys. In: Proc. of the Advances in Crypotology, Int'l Conf. on Theory and Application of Cryptology and Information Security. Springer-Verlag, 2007. 200–215.

[25] Tan ZW, Liu ZJ, Xiao HG. A fully public key tracing and revocation scheme provably secure against adaptive adversary. Ruan Jian Xue Bao/Journal of Software, 2005,16(7):1333–1343 (in Chinese with English abstract). http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=20050716&journal_id=jos [doi: 10.1360/jos161333]

[26] Pang LJ, Li HX, Jiao LC. Design and analysis of a provable secure multi-recipient public key encryption scheme. Ruan Jian Xue Bao/Journal of Software, 2009,20(10):2907–2914 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3552.htm [doi: 10.3724/SP.J.1001.2009.03552]

[27] Lynn B. The pairing-based cryptographic library. 2015. http://crypto.Stanford.edu/pbc/

[28] Loukides M, Oram A. Programming with GNU SoftWare. O'Reilly & Associates, 1997,86(3):350–359.

[29] Steiner M. The PBC_bce broadcast encryption library. 2006. https://crypto.stanford.edu/pbc/bce/

[30] Hu XT, Qin ZP, Zhang H, Hao GS. Research and improved implementation of AES algorithm in OpenSSL. Control & Automation, 2009,25(12):83–85.

附中文参考文献:

[1] 傅颖勋,罗圣美,舒继武.安全云存储系统与关键技术综述.计算机研究与发展,2013,50(1):136–145.

[2] 付印金,肖侬,刘芳.重复数据删除关键技术研究进展.计算机研究与发展,2012,49(1):12–20.

[3] 敖莉,舒继武,李明强.重复数据删除技术.软件学报,2010,21(5):916–929. http://www.jos.org.cn/1000-9825/3761.htm [doi: 10.3724/SP.J.1001.2010.03761]

[18] 李莉,薛锐,张焕国,冯登国,王丽娜.基于口令认证的密钥交换协议的安全性分析.电子学报,2005,33(1):166–170.

[19] 胡学先,张振峰,刘文芬.标准模型下通用可组合的口令认证密钥交换协议.软件学报,2011,22(11):2820–2832. http://www.jos.org.cn/1000-9825/3910.htm [doi: 10.3724/SP.J.1001.2011.03910]

[21] 张旭升,林东岱.密码学中双线性对的构造与计算[博士学位论文].北京:中国科学院大学,2012.

[22] 陈亚萌,程相国,王硕.基于双线性对的无证书群签名方案研究.信息网络安全,2017,(3):53–58.

[25] 谭作文,刘卓军,肖红光.一个安全公钥广播加密方案.软件学报,2005,16(7):1333–1343. http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=20050716&journal_id=jos [doi: 10.1360/jos161333]

[26] 庞辽军,李慧贤,焦李成,王育民.可证明安全的多接收者公钥加密方案设计与分析.软件学报,2009,20(10):2907–2914. http://www.jos.org.cn/1000-9825/3552.htm [doi: 10.3724/SP.J.1001.2009.03552]

张曙光(1991－),男,山东曲阜人,硕士,主要研究领域为密码学,云计算安全.

刘红燕(1994－),女,硕士,主要研究领域为云中重复数据删除.

咸鹤群(1979－),男,博士,副教授,CCF 高级会员,主要研究领域为密码学,云计算安全,系统安全.

侯瑞涛(1993－),男,学士,主要研究领域为数据库数字水印.

王雅哲(1979－),男,博士,副研究员,主要研究领域为物联网安全,智能信息设备安全.