A Note on "Secure and Efficient Data Deduplication in JointCloud Storage"

Zhengjun Cao, Lihua Liu

Abstract. We show that the data deduplication scheme [IEEE Trans. Cloud Comput., 11(1), 156-167, 2023] is flawed due to some inconsistent computations. The scheme tries to propose a hybrid encryption and authentication mechanism based on RSA cryptosystem and pairing cryptosystem, but it has confused the different group operations. In the data sharing phase, the cloud service provider cannot determine which stored tag matches the temporary tag, and fails to return the stored data to the requester. To fix, it should explicitly specify which is randomized by modular exponentiation with an RSA modulus, and which is randomized by point multiplication over the underlying elliptic curve.
Keywords: Mutual authentication, Key agreement, RSA cryptosystem, User anonymity.

1 Introduction

Recently, Zhang et al. [1] have proposed a new data deduplication scheme using bilinear pairing and RSA for cloud data sharing. In the considered model, there are two entities, cloud service provider (CSP) and user. Any user having data to be stored in clouds can request data computing services.

The scheme is designed to meet many security requirements, such as data confidentiality, data integrity, tag consistency, resistent to chosen plaintext attack, brute-force attack, poison attack, tampering attack, and collusion attack. In this note, we show that the scheme is flawed because of some inconsistent computations. The cloud service provider cannot determine which stored tag matches the temporary tag. This results in the failure to return the stored data to the requester.

2 Preliminaries

An elliptic curve E over the field K is defined by the equation $y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$, with $\triangle = -d_2^2d_8 - 8d_4^3 - 27d_6^2 + 9d_2d_4d_6$, $d_2 = a_1^2 + 4a_2$, $d_4 = 2a_4 + a_1a_3$, $d_6 = a_3^2 + 4a_6$, $d_8 = a_1^2a_6 + 4a_2a_6 - a_1a_3a_4 + a_2a_3^2 - a_4^2$. The condition $\triangle \neq 0$ ensures the smoothness of the curve, i.e., every point on the curve has at most one tangent line. Suppose that L is an extension

Z. Cao, Department of Mathematics, Shanghai University, Shanghai, 200444, China.
L. Liu, Department of Mathematics, Shanghai Maritime University, Shanghai, 201306, China.
Email: liulh@shmtu.edu.cn

of K. Define the set $E(L) = \{(x, y) \in L \times L : y^2 + a_1xy + a_3y - x^3 - a_2x^2 - a_4x - a_6 = 0\} \cup \{\infty\}$, where ∞ represents the point at infinity.

There is a *chord-and-tangent* rule for adding two points in $E(L)$ to give a third point. Let $P, Q \in E(L), P \neq Q$, ℓ be the line through P and Q, R' be the crossing point of ℓ and E, R be the reflection of R' about x-axis. Define $P + Q = R$. Let ι be the tangent line at P, S' be the another crossing point of ι and E, S be the reflection of S' about x-axis. Define $P + P = S$. The set $E(L)$ is an Abelian group with respect to the operation "+". ∞ is the null element. If $P = (x, y) \in E(L)$, then $-P = (x, -y)$.

Let E be an elliptic curve over K, $p = \text{char}(K)$, the integer $m \geqslant 2$, and $(m, p) = 1$. Then $\sum n_i(P_i)$ is a divisor of some function iff $\sum n_i = 0$ and the additions of n_i points $\sum[n_i]P_i = 0$, Let $E[m] = \{P \in E : [m]P = O\}$. Then $\sharp E[m] = m^2$. If $d \,|\, m$, then $\sharp E[d] = d^2$. Hence, $E[m]$ can be expressed as $\mathbb{Z}_m \times \mathbb{Z}_m$. If $T \in E[m]$, there exists $f \in \overline{K}(E)$ such that $\text{div}(f) = m(T) - m(O)$. Let $T' \in E$ and $[m]T' = T$. Then there exists $g \in \overline{K}(E)$ such that $\text{div}(g) = \sum_{R \in E[m]}(T' + R) - (R)$, which means the composite functions $f \circ [m]$ and g^m have the same divisor. Assume that $f \circ [m] = g^m$. If $S \in E[m]$, for $\forall X \in E$, $g(X + S)^m = f([m]X + [m]S) = f([m]X) = g(X)^m$. Let μ_m be the set of m-th unit roots. Weil pairing is defined as

$$\hat{e}_m : E[m] \times E[m] \to \mu_m, \quad \hat{e}_m(S, T) = g(X + S)/g(X)$$

where $X \in E$ such that $g(X + S) \neq 0, g(X) \neq 0$.

3 Review of the Zhang et al.'s scheme

The scheme makes use of several mathematical tools, including elliptic curve, bilinear pairing, RSA cryptosystem, etc. The involved notations and descriptions are listed below (Table 1). It consists of five phases, System setup, Data outsourcing, Data access, Data update, and Data sharing. The scheme can be briefly restated as follows (Table 2).

Table 1: Notations and descriptions

$\|x\|$	the length of x
ID_j	the identity of user U_j
$\mathbb{G}, \mathbb{G}_T$	cyclic groups with the prime order p and $\mathbb{G} = \langle g \rangle$
$e(\cdot, \cdot)$	bilinear pairing $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$
$f(\cdot)$	key derivation function $f : \mathbb{G}_T \to K$
$H(\cdot)$	collision-resistant hash function $H : \{0, 1\} \to Z_p^*$
τ	the tag of a file m, $\tau = e(a, b)$
$\text{Mat}(\tau_i, \tau_j) = 1$	τ_i matches τ_j such that $e(a^i, b^j) = e(a^j, b^i)$

4 Inconsistent computations

As we see, for a temporary tag $\tau_j = (g^{r_j}, E^{yr_j})$ submitted by the requester, where $E = (g^h)^x \bmod N$, $h = H(m)$ is the hash value of message m, CSPs should find the stored tag $\tau = (g^r, g^{rh})$ such that $\mathrm{Mat}(\tau, \tau_j) = 1$, i.e.,

$$e(g^r, E^{yr_j}) = e(g^{rh}, g^{r_j}) \tag{1}$$

But we find this is a false equation.

Table 2: The Zhang et al.'s scheme

System Setup. Let U_i be the i-th user with the identity ID_i, and RSA public key N_u, e_u, secret key d_u. Let the target cloud service provider be CPS with RSA public key N_c, e_c, secret key d_c. To register, U_i picks a password pw, and sends $A_i = ID_i \| H(pw)$ to CSP. CSP picks ω as a token and returns $\omega^{e_u} \| H(A_i \| \omega)$ to U_i, who checks $H(A_i \| (\omega^{e_u})^{d_u}) = H(A_i \| \omega)$. Then store the valid token. To login, U_i sends $(H(pw) \| \omega)^{e_c} \| H(A_i \| \omega)$ to CSP, who computes $D = ((H(pw \| \omega)^{e_c})^{d_c}$ and checks $H(A_i \| D) = H(A_i \| (H(pw) \| \omega))$. The target CSPs publish the public keys $pk_1 = g^{\alpha_1}, pk_2 = pk_1^{\alpha_2}, \cdots, pk_l = pk_{l-1}^{\alpha_l}$, where $\alpha_\tau \in \mathbb{Z}_p^*, \tau = 1, 2, \cdots, l$.
Data Outsourcing. To outsource data m, user U_j picks a random $r, s \in Z_p^*$ to compute $h = H(m)$, $\tau = (g^r, g^{rh}), K = g^{h \oplus s}$. Pick a random $R \in \mathbb{G}_T$ to encrypt m by the symmetric encryption algorithm AES as $C^1 = SE_{f(R)}(m), C^2 = e(K, pk_l) \cdot R, k = K \oplus g^h$. Generate the deduplication tuple $T = (\tau, k, C^1, C^2)$ and send it to CSPs. Upon receiving a new tuple, CSPs can check if there is a stored tag τ_i such that $\mathrm{Mat}(\tau_i, \tau) = 1$. If holds, the duplicated copy is found.
Data Access. Data Update. Refer to the original description (pp. 160-161, [1]) for the details.
Data Sharing. Assume the user U_i wants to share the data m to the other user U_j. Let (x, N) be U_j's RSA public key, y be the secret key. U_i encrypts the auxiliary information g^h as $E = (g^h)^x \bmod N$, where $h = H(m)$ is the hash value of message m. U_i then sends the sharing request message $share \| ID_i \| ID_j \| \tau$ to CSPs. Meanwhile, CSPs verify ID_i and match tag τ. U_j picks a random $r_j \in \mathbb{Z}_p^*$ to compute a temporary tag $\tau_j = (g^{r_j}, E^{yr_j})$, and submits (ID_j, τ_j) to CSPs. If $\mathrm{Mat}(\tau, \tau_j) = 1$, user U_j can download the target data (k, C) from CSPs, and retrieve the raw data m by computing $K = k \oplus E^y, R = \frac{C^2}{e(K, pk_l)}, m = SD_{f(R)}(C^1)$.

Notice that

$$E = (g^h)^x \bmod N \tag{2}$$

Since bilinear pairings are defined over some elliptic curve groups [3], the point g^h is eventually represented as

$$g^h = (\mu, \nu) \in \mathbb{F}_\rho \times \mathbb{F}_\rho \tag{3}$$

in affine coordinates over the finite field $\mathbb{F}_\rho$. Hence, we have

$$E = (g^h)^x \bmod N = (\mu, \nu)^x \bmod N$$

which can be interpreted as

$$E = (\mu^x \bmod N, \nu^x \bmod N) \tag{4}$$

Under the assumption $|N| > |g^h|$ (see page 161, [1]), we have

$$\begin{aligned}
&E^{yr_j} = (\mu^x \bmod N, \nu^x \bmod N)^{yr_j} \\
&\text{[each coordinate is encrypted and decrypted using RSA]} \\
&= (\mu^{xyr_j} \bmod N, \nu^{xyr_j} \bmod N) \\
&\text{[due to } \ xy \equiv 1 \bmod \phi(N)\text{]} \\
&= (\mu^{r_j} \bmod N, \nu^{r_j} \bmod N) \\
&\text{[each coordinate is randomized by modular exponentiation]} \\
&\neq (g^h)^{r_j} \ \ \text{[point multiplication]}
\end{aligned}$$

which leads to a contradiction.

5 A revision

In the data sharing phase, for the auxiliary information g^h, the data owner U_i represents the point $g^h = (\mu, \nu)$ in affine coordinates and encrypts it as $(E_1, E_2) = (\mu^x \bmod N, \nu^x \bmod N)$, where (N, x) is the requester's RSA public key. Return (E_1, E_2) to the requester U_j.

The user U_j picks a random $r_j \in \mathbb{Z}_p^*$ to compute the point

$$E = (E_1^y \bmod N, E_2^y \bmod N)$$

and the *point multiplication* $F = E^{r_j}$. Generate the temporary tag $\tau_j = (g^{r_j}, F)$ and submit (ID_j, τ_j) to CSPs. The checking equation is

$$e(g^r, F) = e(g^{rh}, g^{r_j}) \tag{1'}$$

Correctness.

$$\begin{aligned}
E &= (E_1^y \bmod N, E_2^y \bmod N) \\
&= ((\mu^x)^y \bmod N, (\nu^x)^y \bmod N) = (\mu, \nu) = g^h \\
F &= E^{r_j} = g^{hr_j}, e(g^r, F) = e(g, g)^{rhr_j} = e(g^{rh}, g^{r_j})
\end{aligned}$$

6 Further discussions

We would like to stress that the conventional point multiplication over an elliptic curve is represented as aP, where P is a point and a is an integer. In the proposed scheme, the point multiplication is represented as g^h, which is easily confused with the general modular exponentiation.

In the scheme, users can be divided into two types. The one is the owner who has the original data. The other is the non-owner who does not possess original data but wants to get the plaintext of the outsourced data from CSPs (see page 158, [1]).

We find a non-owner can impersonate a legitimate owner. For example, the non-owner user U_j can generate a new tag $\tau' = (g^{r_j\xi}, F^{\xi})$, where $\xi \in \mathbb{Z}_p^*$. Then share the tag τ' with another non-owner user U_ϵ. The user U_ϵ sends the tag τ' with its identity ID_ϵ to CSPs. Clearly, $\mathrm{Mat}(\tau', \tau) = 1$. The CSPs will return a logical link about the tag of the duplication data to U_ϵ. Finally, CSPs add ID_ϵ into an ownership list L of the data m (see page 160, [1]).

7 Conclusion

We show that the Zhang et al.'s data deduplication scheme should be revised due to some false computations. We want to stress that the structure of a bilinear pairing is still unfamiliar to some newcomers. We hope the findings in this note could be helpful for the future work on designing such schemes.

References

[1] D. Zhang, J. Le, N. Mu, J. Wu, X. Liao: Secure and efficient data deduplication in jointcloud storage, *IEEE Trans. Cloud Comput.*, vol. 11, no. 1, pp. 156-167, 2023.

[2] D. Hankerson, A. Menezes, S. Vanstone: Guide to Elliptic Curve Cryptography, Springer, USA, 2003.

[3] D. Boneh and M. Franklin, Identity-based encryption from the Weil pairing, *SIAM J. Comput.*, vol. 32, no. 3, pp. 213-229, 2003.