**INTRODUCTION**

WITH the rapid development of cloud storage, more and more individuals and enterprises tend to outsource their sensitive data to remote cloud service providers in a pay-per-use manner [1], [2], [3], [4], [5]. According to the study from Internet Data Center (IDC) sponsored by Dell EMC, the digital universe is doubling in size every two years and the volume of the universe data is expected to reach 44 zettabytes (ZB) or 44 trillion gigabytes (GB) in 2020 (more than 5200 gigabytes for each man, woman, and child) [6]. However, the growth of data puts heavy pressures on cloud service providers. To cope with it, a straightforward method is to require cloud service providers continuously increasing the capacity of storage space, so as to meet users’ requirements for high-quality storage services.

However, cloud service providers may store plentiful and repetitive data (such as movies, music and genome data), which inevitably incurs a mass of redundant storage and backup space, consequently to cost a vast amount of computing and management overhead during its whole life cycle. To solve this problem, Bolosky *et al*. first proposed the technique of data de duplication [7], which decreases the redundant storage space and bandwidth by eliminating duplicate copies and only storing one copy of them. Nowadays, data deduplication techniques have been widely deployed by cloud service providers, such as Drop box [8], Google Drive [9] and Memopal [10]. Researches [11], [12] have shown that most of the genome data (*≥* 83%) and disk (*≥* 90%) of business applications can be reduced by exploiting data de duplication technique. While the technique of data de duplication has plentiful advantages, there are still some security challenges that need to be addressed. In particular, the cloud service providers are often assumed to be not fully trusted, which may try to infer and analyze the outsourced data [13], [14], [15], [16]. To protect the confidentiality of their sensitive data, cloud users generally encrypt their data before outsourcing them to cloud service providers. However, different users encrypt the same data with their private keys, which leads the same data to output different cipher texts, and makes the function of data de duplication unachievable. Douceur *et al*. [17] proposed the first feasible solution to protect the confidentiality of data and achieve de duplication on cipher texts. However, the cloud user encrypts sensitive data with a convergent key, which is derived from the hash value of the data and unchanged. It will lead the revoked cloud user to access the sensitive data through the reserved convergent key.

User revocation is a severe security problem in the cloud environment. We take the case in genome research as an example to illustrate this point. Considering the enormous volume of genome datasets, genome researchers tend to use the cloud to store the genome data [18]. Google Genomics [19] and Amazon [20] have deployed specific platforms for managing and analyzing genome data. However, some sensitive genome data produced by disease sequencing projects must be protected. For example, when a researcher is no longer a member of the genome project, he will be prohibited from accessing the genome datasets. This problem has been addressed by using techniques such as re-encryption and group key distribution [21], [22], [23]. By using symmetric encryption (such as AES-128 or AES-256) to re-encrypt the sensitive data and distribute group key for group users, those schemes can support user joining and user revocation. Although the re-encryption scheme uses a new encryption key to encrypt the entire message to protect the data confidentiality, it will result in a waste of excessive computation overhead. William *et al*. [24] provided evidence that, in situations involving even a minimal amount of policy dynamism, the cryptographic enforcement of access controls is likely to carry prohibitive costs.

Recently, Li *et al*. [25], [26] proposed a rekeying-aware encrypted deduplication storage system (REED), which supports a lightweight re-encryption. Instead of re-encrypting the entire package, data owners are only required to reencrypt a small part of it through the CAONT, thereby effectively reducing the computation overhead of the system. However, we point that the REED is vulnerable to the stub-reserved attack, which will be described in Section 3.2. In short, if a revoked cloud user keeps the last bytes of a package as *stub* package, he can use the reserved *stub* package and *trimmed* package (downloaded from the cloud service providers) to recover the plaintext. Therefore, existing proposed schemes cannot well support secure dynamic ownership management of cloud users and efficient

re-encryption.

**Our Contribution.** In this paper, we further study the above problems of secure and efficient re-encryption for de duplication storage. Our contributions are three folds:

*•* We point out a security weakness of the enhanced encryption of REED scheme [25], [26]. That is, this scheme is vulnerable to the so-called stub-reserved attack proposed in this paper.

*•* We propose a location selection method based on Bloom filter and a secure data de duplication scheme with efficient re-encryption. By using the symmetric encryption and this new location selection method, the revoked cloud user cannot obtain the sensitive data from the data owner. Thus data privacy is ensured. Moreover, instead of re-encrypting the entire package, data owners are only required to re-encrypt a small part of it through the CAONT, thereby effectively reducing the computation overhead of the scheme.

*•* We provide security analysis and performance evaluation of our scheme, and the results show that our scheme is secure and efficient.