IJNS
International Journal of Network Security

ISSN 1816-353X (Print)
ISSN 1816-3548 (Online)

Vol. 26, No. 4 (July 2024)

INTERNATIONAL JOURNAL OF NETWORK SECURITY

PUBLISHING OFFICE

Min-Shiang Hwang
Department of Computer Science & Information Engineering, Asia University, Taichung 41354, Taiwan, R.O.C.
Email: mshwang@asia.edu.tw

International Journal of Network Security is published both in traditional paper form (ISSN 1816-353X) and in Internet (ISSN 1816-3548) at http://ijns.jalaxy.com.tw

PUBLISHER: Candy C. H. Lin

Volume: 26, No: 4 (July 1, 2024)

International Journal of Network Security

Policy Extraction and Optimization with Top-down and Bottom-up Approaches for Attribute-based Access Control

Wei Sun, Jun Lu, and Mengzhao Wang
(Corresponding author: Wei Sun)

School of Computer and Information Technology, Xinyang Normal University
No. 237, Nanhu Road, Xinyang 464000, P. R. China
Email: sunny810715@xynu.edu.cn

(Received Apr. 7, 2023; Revised and Accepted Nov. 7, 2023; First Online June 22, 2024)

Abstract

Attribute-based access control (ABAC) has received much attention and has emerged as a desired access control mechanism due to its flexibility and identity-independent property. The policy engineering technology seems to be very efficient for constructing ABAC systems, and it mainly comprises two construction approaches: The top-down and bottom-up. However, the former is costly and error-prone, while the engineering scale is large, and the policy result lacks interpretability for the latter. Furthermore, many redundant and inaccurate rules increase the complexity of policy engineering. This study proposes a novel policy-engineering method combining top-down and bottom-up approaches to address these issues. Firstly, to accurately extract policy elements while alleviating manual workloads, natural language processing techniques such as access policy sentence identification and semantic role labeling are utilized to automatically handle the top-level policy specification documents in natural language. Secondly, to reduce the engineering scale while enhancing the policy interpretability, an unsupervised learning method is adopted to develop an initial policy set according to both the bottom-level and the top-level access information, and then an algorithm is presented to mine effective rules. Lastly, to further optimize the mined rules while improving the policy conciseness, a pruning algorithm is proposed to eliminate the redundant policy rules, and the experiments demonstrate the efficiency and effectiveness of the proposal and show encouraging results.

Keywords: Attribute-based Access Control; Policy Engineering; Policy Optimization; Top-down; Bottom-up

1 Introduction

With the rapid development and widespread application of mobile communication technology and high-performance computing, security has become a basic constraint requirement in emerging fields such as the Internet of Things, smart contracts, block chain, and industrial information integration systems [21, 25]. In large-scale, distributed collaborative management systems, there is a large amount of information storage and resource sharing, and enterprises adopt corresponding access control mechanisms to ensure the system security. However, traditional access control mechanisms such as discretionary access control (DAC), mandatory access control (MAC), and role-based access control (RBAC) rely on identity and lack flexibility and scalability [23], which cannot meet the fine-grained and dynamic functional requirements in actual application scenarios. As an alternative mechanism, Attribute-based Access Control (ABAC) overcomes the shortcomings of the existing mechanisms and has high flexibility and expressiveness, especially for fine-grained and dynamic access requests in large-scale distributed and collaborative working environments [12]. It has received much attention and in-depth research in both academia and industry in recent few years [3].

Similar to construction of the role-engineering system, there are also two construction approaches for policy engineering based on ABAC: Top-down [17] and bottom-up [14, 15, 19]. Usually, the former accurately evaluates business processes and decomposes them into several independent and small units, which are then associated with access permissions. Researchers have proposed various top-down methods for policy engineering. To obtain access control policies (ACPs) from the given document resources in natural language, Xia *et al.* [26] proposed a new approach towards extracting policies from natural language documents. To automatically extract ACPs from unlabeled datasets, Xiao *et al.* [27] utilized the deductive reasoning and proposed a novel top-down approach called Text2Policy to construct ACPs. To construct suitable ACPs from labeled datasets, Slankas *et al.* [20] adopted the supervised learning and inductive reasoning and proposed a new approach called ACRE

to extract rules and develop an access control system. Actually, many enterprises have high-level requirement specifications for accessing resources, which state who can do what under which circumstances and conditions in human understandable ways. These specifications are referred to as natural language access control policies (NLACPs) [17], which usually cannot be determined or directly enforced in ABAC systems. To automatically derive policies from unrestricted natural language documents, Narouei *et al.* [18] proposed a top-down policy engineering methodology using the deep neural network. Alohaly *et al.* [1] used the current popular natural language processing and machine learning techniques, proposed a new method to automatically extract ABAC properties from natural language access control policies and then generated a set of natural language access control strategies to evaluate the proposed method. Subsequently, to further bridge the gap between the natural language and formalized representation for specifying ABAC constraints, Alohaly *et al.* [2] proposed an automated framework to infer the policy constraints from the natural language documents. However, the top-down approach manually selects policy engineering elements from the natural language specification documents to construct an ABAC system while determining access control policies, which is time-consuming, labor-intensive and error-prone.

On the other hand, the bottom-up approach extracts rules from the existing patterns of access permissions and automatically constructs a desire system framework, which overcomes the limitations of the top-down approach and is often called policy mining. Xu and Stoller [28] discovered ABAC policies from access control lists (ACLs) and their corresponding attribute data, and they first proposed a bottom-up mining method. This method first cycled through the given ACL to select access patterns and construct a candidate rule set, and then selected rules that could cover more patterns in the ACL from the candidate rule set. Das *et al.* [8] proposed a scheme to solve the policy mining problem using Gini impurity, which considered environmental attributes while mining policies. Talukdar *et al.* [24] pointed out that policy mining was similar to identifying functional dependencies in relational tables and proposed a policy mining algorithm that could accurately enumerate all possible subject-object pairs. Iyer *et al.* [10] proposed a policy mining method that could construct positive and negative ABAC authorization rules. Das *et al.* [7] believed that the problem of the policy engineering was similar to the role engineering problem. They considered that constructing the corresponding access control systems are equally important for solving such two problems, and then presented a detailed survey for both the engineering techniques. However, the mining scales are very large and the existing bottom-up approaches do not consider the actual functional requirements of the top-level organizations. To reduce the mining scale while enhancing the policy interpretability, Das *et al.* [9] rearranged the authorization matrix with the visual representation and proposed a visually mining method (VisMAP). To reduce the engineering scale while ensuring the system security, Sun *et al.* [22] used the partition and compressing technologies and proposed a novel optimization method with separation-of-duty constraints (PEO_VR&SOD). To guarantee the correctness of the policy rules, Cotrini *et al.* [6] presented a new criterion for evaluating the policy quality and proposed a novel mining method. Furthermore, the mined policy should be as concise as possible. Nevertheless, there are numerous same or similar policy rules using the existing methods, which increase the complexity of the policy engineering problem and influence policy quality.

To address these existing problems, this study proposes a novel policy-engineering method, which is called policy extraction and optimization that combines the top-down and bottom-up approaches. To sum up, our main work contributions are as follows:

1) To automatically and accurately extract the access control policy elements from the top-level policy specification documents in natural language, we utilize natural language processing techniques such as the syntax parsing, access policy sentences identification, semantic role labeling, named-entity recognition and argument extension to extract the policy elements and define the data dictionary.

2) To reduce the mining scale and complexity of the policy engineering, we adopt an unsupervised leaning method to develop an initial policy set according to both the bottom-level and the top-level access information, and then present an algorithm to derive effective rules based on the initial rule clusters and the constructed generalized Cartesian product among entities and operations. We also present the performance evaluation of policy mining through experiments.

3) To improve the policy quality, we propose a pruning algorithm to eliminate redundant or similar policy rules and take the weighted structural complexity as the evaluation criterion, in order to further optimize the mined rules while improving the policy conciseness. We also present the performance evaluation of policy optimization through experiments.

The rest of the article is structured as follows. The preliminaries used in our work are briefly discussed in Section 2. Section 3 proposes a novel policy-engineering method, which includes three phases: The top-down extraction of policy-engineering elements, bottom-up policy mining, and policy optimization. We implement experiments and comprehensively present the experimental analysis in Section 4. Section 5 concludes the article and discusses future work.

2 Preliminaries

Before proposing our methodology, some preliminaries are briefly discussed, which involve the basic ABAC elements and the natural language processing and clustering techniques.

2.1 Basic Elements of the ABAC Model

The basic ABAC model [22] mainly consists of the following sets, relationships, and functions:

1) U, O, S, and OP represent the finite sets of requesting subjects (or users), requested objects, environments, and operations, respectively;

2) A_u, A_o, and A_s represent the attribute sets of user u, object o, and session s, respectively;

3) E and A represent the finite sets of all the entities and entity attributes in the system, where $E = U \cup O \cup S$, $A = A_U \cup A_O \cup A_S$;

4) $Val\_Att(e, a)$ represents a function that returns the values that attribute a of entity e takes;

5) ρ represents an authorization rule, which is a two-tuple form $< AC, op >$, where AC indicates the set of attribute conditions, and op indicates an operation. The set of all the authorization rules is referred to as an policy, denoted as $P = \{\rho_1, \rho_2, \cdots\}$.

6) UUA, OOA, and AL are Boolean matrices that correspond to the many-to-many assignment relationships of users to attribute values, objects to attribute values, and access logs of users to objects, respectively. They can be formalized as:

$$UUA[i][j] = \begin{cases} 1 & \text{if user } u_i \text{ has the attrubute value in the } j^{th} \text{ column;} \\ 0 & \text{otherwise} \end{cases}$$

$$OOA[i][j] = \begin{cases} 1 & \text{if object } o_i \text{ has the attrubute value in the } j^{th} \text{ column;} \\ 0 & \text{otherwise} \end{cases}$$

$$AL_{op}[i][j] = \begin{cases} 1 & \text{if user } u_i \text{ is permitted to perform operation } op \text{ on object } o_j \\ 0 & \text{otherwise} \end{cases}$$

where u_i indicates the user in the i^{th} row, o_j indicates the object in the j^{th} column, and op indicates an operating action.

2.2 Natural Language Processing Techniques

Generally, the natural language processing parser [16], which is used to analyze NLACPs, consists of tokenization, sentence segmentation, part-of-speech tagging (POS), lemmatization, named-entity recognition (NER), co-reference resolution, and semantic role labeling (SRL). Specifically, the tokenization aims to detect individual words, punctuations, and other items; the sentence segmentation is to identify the sentence boundary, the POS determines the token type, the lemmatization generates the common root prototype, the NER categorizes the nouns that are present in a sentence, the co-reference resolution determines whether or not two expressions correspond to the same entity or event, and the SRL identifies predicates within a sentence and implements annotations with the semantic arguments.

2.3 Unsupervised Learning Method

The clustering technique with an unsupervised learning method can specify the structure of sample data, particularly for the categorical data with no labels. Essentially, the process of the policy mining can be regarded as the mapping from the log set to a set of clusters that indicate ideal ABAC rules. Such mapping relationship can be represented as a function $f : X \to Y$, where X is a set of authorization-tuple records, and Y is a set of cluster labels.

3 Methodology

In this section, a novel policy-engineering method is proposed, which mainly involves three phases: (1) Top-down extraction of policy elements, (2) bottom-up policy mining, which includes identification of initial set of clustering rules and mining of effective rules, and (3) policy optimization. Specifically, given the top-level requirement documents described in natural language, syntax parsing, access policy sentence recognition, semantic role labeling, and other natural language processing techniques are used to extract the policy elements. Next, for the given bottom-level access-log documents, the tuples of access permissions that involve various entities with similar characteristics are gathered together using the clustering technique. According to the attribute values of the cluster center and other cluster numbers, the data dictionary is adopted to determine an effective attribute set that is related to the ABAC policy, and then an initial set of ABAC rules is developed; subsequently, a new matrix is constructed through the generalized Cartesian product among subjects, objects, attributes and operations, and then effective rules are further derived based on the initial rule clusters. However, it is possible to derive the similar rules from different rule clusters during the policy-mining stage. Last, the rule pruning technique is adopted to identify and eliminate redundant rules from the candidate rule set, in order to optimize and improve the policy quality. The flow chart of the proposal is presented in Figure 1.

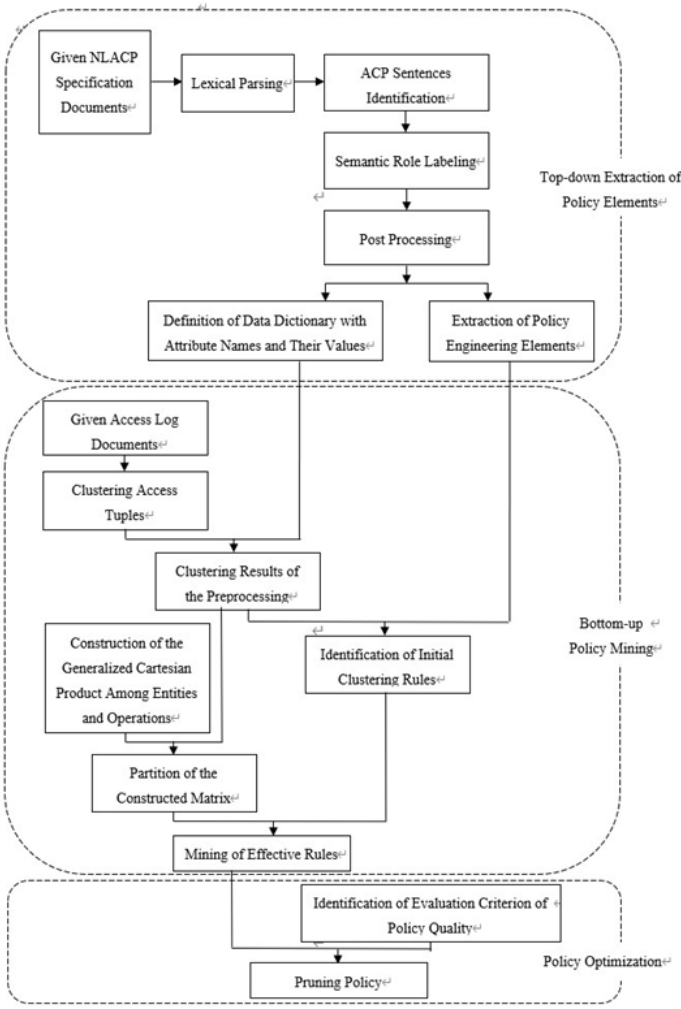


Figure 1: Flow chart of the proposed method

3.1 Top-down Extraction of Policy Elements

The policy-engineering elements are extracted using the preprocessing, access control policy sentences (ACPs) identification, semantic role labeling (SRL), and post processing such as the named-entity and argument extension. The extraction process involves the following steps:

Step 1. The whole natural language access control policy documents are read through, and the lexical analysis and correlation resolution of the sentences are conducted, in order to determine the sentence boundary and detect the words, punctuation marks and other associated items in the texts.

Specifically, to facilitate the evaluation for the text information of the requirement specifications, the toolkit CoreNLP [13] is used to determine all the sentences, so that each one is laid on a unique line. Further, there are numerous pronouns in the texts, the meanings of which are not clear. To determine whether or not different pronouns describe the same entity or event, a fast and robust algorithm for pronouns resolution [4] is utilized to replace the pronoun in the texts with specific entities, in order to determine all the different expressive patterns of the same entity.

Step 2. The ACRE method [20] is used to effectively identify and derive all the possible ACPs from the entire NLACPs. Next, to improve the accuracy of the extracted ACPs, a supervised learning algorithm such as the k-NN classifier is employed to classify the current test items according to the principle of choosing the closest items, and the Euclidean Distance is taken as the measure for evaluating the attribute value, in order to find k nearest instances to a particular instance.

Step 3. The well-known semantic role labeler (SRL), which is fast, accurate and well independent, is employed to detect the semantic arguments associated with the predicates or verbs in the ACPs, label the semantic structure of the predicate argument and then assign different semantic roles to entities that are related to a specific verb.

Specifically, for a given sentence, the main tasks of the SRL include analyzing the meanings of the sentence that contains the target verb, as well as extracting the constituents of the sentence that are filled by the semantic roles for each predicate argument. The general formalization of the SRL can be represented as $\{Arg0; Predicate; Arg1\}$, where $Arg0$ denotes the subject with respect to $Predicate$ and $Arg1$ denotes the object or resource that is access by $Arg0$. Here, the Semantic/syntax Extraction using a Neural Network Architecture (SENNA) proposed in [5] is considered as the semantic role labeler, which is a mul-

tilayer architecture of the neural network that can handle large amounts of natural language documents.

Step 4. Since the extracted ACP sentences are complicated and inaccurate, the post processing is required for the annotation parts that are generated using the SRL tool, including the named-entity recognition (NER) and argument extension. Specifically, the former is to identify the named entities and word sequences included in sentences that belong to the predefined categories, such as person name, location, institution organization, time expression, etc., in order to generate an annotated text that can highlight these names. For the latter, based on the combination of multiple subjects and objects separated from complex sentences, different policy elements that contain single subject, single object and the predicates annotated in the sentences are integrated into a policy set.

3.2 Bottom-up Policy Mining

3.2.1 Identification of Initial Clustering Rules

In this subsection, an unsupervised learning approach such as the k-modes algorithm is first utilized to train the bottom-level access tuples and detect the clustering patterns. Then, the policy-mining process is regarded as a mapping from the set of access logs to the cluster set, which can be represented as $f : AL \rightarrow C$, where AL denotes the set of access tuples, C denotes the set of clustering labels, and each label corresponds to a rule.

According to the attribute set of the existing access logs in the database, the attribute names that are stated using the natural language and the data dictionary that is defined with the attribute-value range, symbols and abbreviations that are used to represent the attribute-value pairs in the existing database are replaced with the equivalent semantic expressions in the data dictionary. For example, considering the attribute named as position in a university can take values 1,2,3. If these attribute values in the data dictionary correspond to "professor", "lecturer" and "student", then the numerical attribute values are replaced with those in the dictionary.

The cluster center indicates that its attribute values occur most frequently in the cluster. However, the attributes included in the center are not all valid. To enhance the interpretability of rule clusters, according to the original policy elements extracted in the previous phase, the concept of the effective attribute–value pair is defined first.

Definition 1. *Effective attribute–value pair: If c_i is the center of cluster i, then $S_{c_i} = \{< a, v >\}$ represents the set of all the attribute–value pairs appear in entities of c_i, where $v \in Val\_Att(c_i, a)$. For each $< a_j, v_j >$ in S_{c_i}, it is regarded as the effective attribute–value pair of rule cluster ρ for c_i, if and only if the difference between the proportion of the value v appears in the bottom-level rule cluster c_i and that of v appears in the top-level set of the policy elements is greater than threshold T. Then, $< a_j, v_j >$ is appended to the attribute-condition set of ρ. The set of all such rules is represented as the initial policy, and the clustering process is presented as shown in Algorithm 1.*

Algorithm 1 Identifying initial clustering rules

Input: Access log set AL, number of clusters k, set of policy elements $NLACP$, and threshold T
Output: Initial set of rule clusters $Init\_P$

1. The k-modes method is used to cluster AL, and the result is denoted as $C=\{c_1,c_2,\ldots,c_k\}$;
2. Identify and represent the set of all entities included in c_i as E_{c_i};
3. Identify and represent the set of all entities included in $NLACP$ as E_{NLACP};
4. Initialize $Init\_P=\varnothing$;
5. **for** each c_i in C **do**
6. $\rho_i=\varnothing$;
7. **for** each $<a_j,v_j>$ in S_{ci} **do**
8. **if** $\left(\frac{\{e_k \mid \exists e_k \in E_{c_i} : v_j \in Val\_Att(e_k, a_j)\}}{|E_{c_i}|} - \frac{\{e_{k'} \mid \exists e_{k'} \in E_{NLACP} : v_j \in Val\_Att(e_{k'}, a_j)\}}{|E_{NLACP}|}\right) > T$ **then**
9. $\rho_i=\rho_i \cup \{<a_j,v_j>\}$;
10. **end if**
11. **end for**
12. $Init\_P=Init\_P \cup \{\rho_i\}$;
13. **end for**

3.2.2 Mining of Effective Rules

To reduce mining scale, matrix AL can be separated into k different submatrices: AL_{op_1}, AL_{op_2}, $\cdots$, AL_{op_k} according to the clustering results. For any submatrix, columns correspond to the same object set, while rows correspond to different subject sets. Further, according to k sets of subjects in different submatrices, matrix UUA is separated into k different submatrices: UUA_1, UUA_2, $\cdots$, and UUA_k.

For each UUA_i in $\{UUA_1, UUA_2, \cdots, UUA_k\}$, we construct matrix $UOP^{AL_{op_i}}$ using the Cartesian product of UUA_i and OOA, and then separate it into $UOP^{AL_{op_i}}_{op=1}$ and $UOP^{AL_{op_i}}_{op=0}$, where $UOP^{AL_{op_i}}_{op=1}$ indicates that subjects are permitted to perform operation op on objects in the matrix, while $UOP^{AL_{op_i}}_{op=0}$ indicated that subjects are not permitted to perform op on objects. Thus, the process of mining effective rules is presented in Algorithm 2.

3.3 Policy Optimization

Indeed, it is possible to derive the similar rules from different rule clusters during the rule-mining stage. The Jaccard coefficient can be employed to calculate the similarity between two clustering rules ρ_1 and ρ_2 as:

$$sim(\rho_i, \rho_j) = \frac{\sum_{AC_e \in AC_E} |AC_e^{\rho_i} \cap AC_e^{\rho_j}|}{\sum_{AC_e \in AC_E} |AC_e^{\rho_i} \cup AC_e^{\rho_j}|}$$

which indicates the ratio of the size of the common attribute conditions to the size of the joint attribute conditions between $AC_e^{\rho_i}$ and $AC_e^{\rho_j}$, where $AC_e^{\rho_i}$ and $AC_e^{\rho_j}$ represent the attribute condition sets of entity e with respect to rules ρ_1, and ρ_2, respectively. ρ_1 and ρ_2 are regarded to be similar when the ratio value exceeds a given threshold.

Algorithm 2 Mining effective rules

Input: Initial set *Init_P* of rule clusters, constructed matrix set $UOP=\{UOP^{AL_{op1}}, UOP^{AL_{op2}}, \ldots, UOP^{AL_{opk}}\}$

Output: Set *Effe_P* of effective rules

1. Create and initialize rule set *Init_rules*=*Init_P*, and sort them in ascending order according to the number of attribute–value pairs;
2. Create and initialize rule set *Unin_rules*=∅;
3. Initialize *Effe_P*=∅;
4. **for** each $UOP^{AL_{opi}}$ in UOP **do**
5. Identify combinations of different attribute–value pairs present in all rows of $UOP^{A_{op'_i}}_{op=0}$, and insert them into *Unin_rules*;
6. **for** each ρ in *Effe_P* **do**
7. **for** each ρ' in *Init_rules* **do**
8. **if** (ρ is not null)∧($\rho \subseteq \rho'$) **then**
9. **continue**;
10. **else**
11. **if** $(\rho \cap \rho') \notin$ *Unin_rules* **then**
12. *Effe_P*=(*Effe_P*\{ρ}) ∪ {$\rho \cap \rho'$};
13. **else**
14. *Effe_P*=*Effe_P* ∪ {ρ'};
15. **end if**
16. **end if**
17. **end for**
18. **end for**
19. **end for**

In addition, the weighted structural complexity (WSC) is used to evaluate the policy quality. It can be represented as: $WSC(II) = WSC(P) = \sum_{\rho \in P} WSC(\rho)$, and $WSC(\rho) = WSC(AC, op) = \sum_{AC_e \cup AC_E} w_e \times WSC(AC_e)$, which is to make a comprehensive evaluation for the scale of a specific ABAC policy, where $WSC(AC_e) = |AC_e|$, $|AC_e|$ represents the number of attribute–value pairs contained in entity e, and w_i is a contribution weight that is used to adjust the complexity of the policy. It is obvious that the less the value of WSC, the more concise the policy.

According to the above criterion for evaluating the policy quality, the policy-pruning process is presented in Algorithm 3, in order to further optimize the mined rules while improving the policy conciseness.

Algorithm 3 Pruning Policy

Input: non optimized policy set *Effe_P*, threshold T_{sim}
Output: optimized policy set *Opti_P*

1. Initialize *Opti_P*=*Effe_P*;
2. **for** each ρ_i in *Opti_P* **do**
3. **for** each ρ_j in *Opti_P*\{ρ_i} **do**
4. Calculate $sim(\rho_i,\rho_j)$;
5. **if** $sim(\rho_i,\rho_j) > T_{sim}$ **then**
6. **if** $WSC(Opti\_P\backslash\{\rho_i\}) > WSC(Opti\_P\backslash\{\rho_j\})$ **then**
7. *Opti_P*=*Opti_P*\{ρ_i};
8. **else**
9. *Opti_P*=*Opti_P*\{ρ_j};
10. **end if**
11. **end if**
12. **end for**
13. **end for**

4 Experimental Analysis

To demonstrate the efficiency and effectiveness of the proposal, in this section, we conduct experiments using the real-world and synthetic datasets and compare its performances with the existing popular approaches. All the experiments are compiled and run under the Python environment.

4.1 Performance Evaluation for the Extraction of Policy Elements

First, the proposal is performed on four real-world datasets that have been used in the existing studies for the top-down policy engineering. These datasets are derived from different domains, including the health care, education, and conference management. Specifically, for the healthcare domain, Xiao *et al.* [27] have extracted policy elements from an open dataset called iTrust. Numbers of the ACPs, polices, and semantic role annotations are 418, 1070, and 1559, respectively. For the education domain, the policies were extracted from the dataset called IBM Course Registration System, and numbers of the ACPs, polices, and semantic role annotations are 169, 375, and 912, respectively. For the conference management domain, the results using the requirement documents from dataset CyberChair are 139, 386, and 696, respectively. In addition, dataset Collected, which combines 18 data resources, can be used to extract 114 ACPs, 258 policies, and 650 role annotations.

The *Precision*, *Recall*, and F_1 score can be regarded as the criteria for evaluating policy elements, which are represented as:

$$\begin{aligned} Precision &= \frac{|TP|}{|TP|+|FP|} \\ Recall &= \frac{|TP|}{|TP|+|FN|} \\ F_1 &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned}$$

where TP indicates that the true policy elements are correctly extracted, FP indicates that the false policy elements are mistakenly regarded to be correct, while FN indicates that the true elements cannot be extracted from the documents.

Table 1 presents the average values of the experimental results for the proposal, as well as those of the ACRE or SENNA method. It is observed from the viewpoint of the varying F_1 score that, the proposal performs better than the ACRE. However, it does not perform well from the viewpoint of the varying precision. This is because the ACRE method requires repetition in the sentence structure. The proposed method does not require the sentence having the repetitive structure. Each sentence is independent and is irrelevant to its structure. The proposal derives most of the policy elements as long as the semantic role set defined with predicates exists in advance.

4.2 Performance Evaluation for the Mining of Effective Rules

To derive effective rules while reducing the policy-mining scale, the synthetic datasets widely used by the ex-

Table 1: Comparison of the extracted results

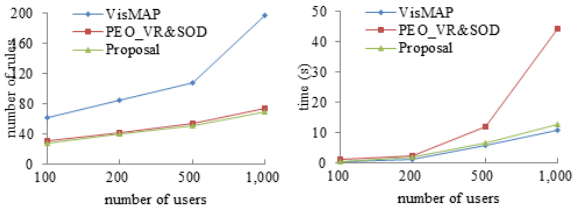
Dataset	ACRE			Proposal		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
iTrust	80%	75%	77%	73%	85%	81%
IBM Course Registration System	81%	62%	70%	55%	85%	68%
CyberChair	75%	30%	43%	46%	84%	62%
Collected	68%	18%	29%	77%	88%	83%

isting popular methods are employed to evaluate the performance of the proposal, and the parameter settings are also similar to those of the VisMAP [9] and PEO_VR&SOD [22]. Furthermore, the number of partitions is regarded as the initial number of rule clusters of the proposal, and four initial clusters are prepared before actually mining effective rules, Then, we repeatedly conduct experiments 10 times on different datasets, take the average number of rules and average execution time as output and compare its performance to the other two methods as shown in Figures 2 $\sim$ 5.

Figures 2(a) $\sim$ 5(a) show the varying trend of the number of mining rules with the increasing number of entities that contain different subjects and objects. It is observed that the number of rules increases slightly for both the PEO_VR&SOD and the proposal, while it varies more obviously for the VisMAP as the number of entities varies. Take Figure 2(a) as an example, the number of objects is fixed at 100, and the number of subjects takes 100, 200, 500, and 1000, respectively. The rule numbers increase from around 30 to 70 with 100 objects for the former two methods when the subject number varies from 100 to 1000. However, it varies obviously and increases from 62 to 197 for the third. Thus, from the viewpoint of the policy size, the proposal performs as well as the PEO_VR&SOD, and it performs better than the VisMAP.

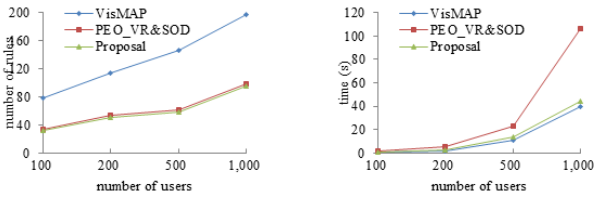
Figures 2(b) $\sim$ 5(b) show the varying trend of execution time as the number of subjects and objects varies. It is observed that the execution time tends to grow linear for both the VisMAP and proposal, while it increases significantly for the PEO_VR&SOD with the increasing number of subjects or objects. Take Figure 4(b) as an example, the number of objects is fixed at 500, and the number of subjects takes 100, 200, 500, and 1000, respectively. Execution time increases from around 5 s to 120 s for the former two methods, while it increases significantly from around 10 s to 336 s using the third method. This is because the original access records are divided into several different partitions or clusters using the VisMAP or proposal, which reduces the mining scale and spends little time for mining rules. Then, from the viewpoint of execution time, the proposal and VisMAP outperform the PEO_VR&SOD.

We also generate synthetic datasets using the particular parameters, since we find that there exist no suitable real-world datasets of large scales for the experiments. Specifically, the number of users takes 10 different value among 100 and 1000, the number of objects takes 200, 400, and 600, and the attributes of the entities are randomly selected from three small-scale datasets [11]: The



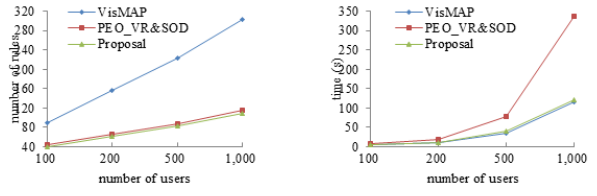
(a) (b)

Figure 2: Comparison of the mining results when $|O| = 100$. (a) Rule number, (b) Execution time.



(a) (b)

Figure 3: Comparison of the mining results when $|O| = 200$. (a) Rule number, (b) Execution time.



(a) (b)

Figure 4: Comparison of the mining results when $|O| = 500$. (a) Rule number, (b) Execution time.

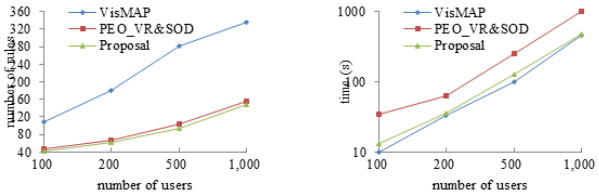


Figure 5: Comparison of the mining results when $|O| = 1000$. (a) Rule number, (b) Execution time.

University, Healthcare, and Project Management, which have been usually used for evaluating the performances of representative policy-engineering methods. To further evaluate the performance of the proposal, we consider the number of the rule clusters and the running time as measures. We repeatedly carry out the experiments and take the average value of the rule number as well as that of the execution time as outputs. The results are presented as shown in Figure 6, and in Figure 7, respectively.

It is seen from Figure 6 that the rule number tends to grow slightly as the number of users increases. From the viewpoint of the varying objects, on the other hand, the rule number also increases when the number of users is fixed. In general, the number of the extracted rules varies slightly with the small-scale entities, this is attributed to the fact that the less the user number, the fewer rules discovered in the policy extraction. Figure 7 presents the varying trend of execution time with the increasing number of users as well as that of objects. It is observed that execution time tends to grow linearly and is always below 350 s when the number of objects is less than 600. This is because the scale of the access logs increases with the increasing number of users and that of objects.

4.3 Performance Evaluation for the Policy Optimization

To evaluate the performance of the policy optimization according to Algorithm 3, the experiments are carried out on different datasets that are also used for evaluating the performances of the Xu-Stoller [28] and Cotrini [6] methods. These datasets are derived from the synthetic and the real-world policy sets. The constructed access logs come from a randomly created policy set, which includes partial datasets of University P, Healthcare P, Project Management P, University PN, Healthcare PN, and Project Management PN [11]. The authorization rules for these policies are created with randomly selected attribute and attribute value sets, and the effectiveness of policy extraction can be evaluated on access logs with different scales and continuously changing structural characteristics. To generate the input data, for each ABAC policy, an authorization tuple set is created and each access

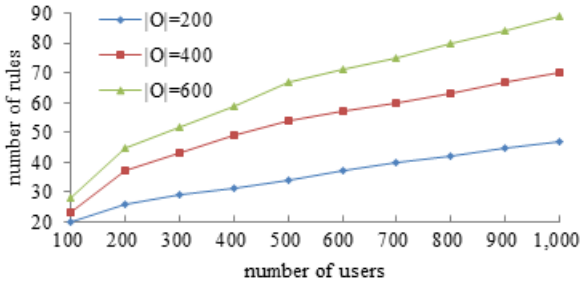


Figure 6: Evaluation of the rule number

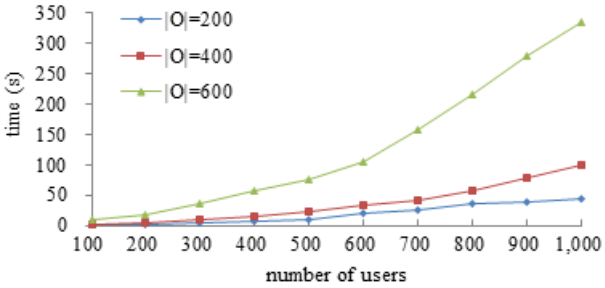


Figure 7: Evaluation of the execution time

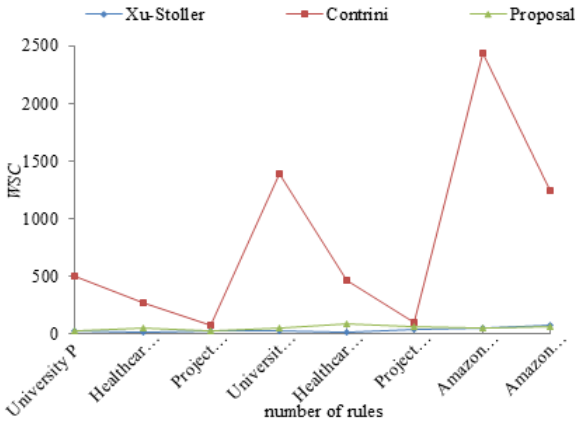


Figure 8: Comparison of the complexity

permission is evaluated with respect to its corresponding ABAC policy. The real dataset comes from publicly accessible log data provided by the Amazon Kaggle and Amazon UCI. The Amazon Kaggle records staff requests for accessing resources and whether they are authorized, as well as the attribute values of employees and the resource identifiers, which contains 12,000 users and 7,000 object resources in total. The Amazon UCI contains over 36,000 users, 27,000 permissions, and 33,000 attribute features.

According to evaluation criteria such as the running time and complexity WSC, we repeatedly conduct experiments on different policy datasets and take the best results of their performances as output. The experimental results using three different methods are presented as shown in Table 2, where "\" denotes the results are uncertain. Figure 8 presents the complexity comparison for evaluating the policy quality.

It is seen from the table that the proposed method performs better than the Xu-Stoller and Cotrini methods on half of the policy sets, especially on the University P, Project Management P, Amazon Kaggle, and Amazon UCI datasets. Specifically, for the University P, the results of the proposal are 10.6 s, and 33, respectively; for the Project Management P, the results of the proposal are 15.97 s, and 69, respectively; for the Amazon Kaggle, the results of the proposal are 211.3 s, and 49, respectively; for the Amazon UCI, the results of the proposal are 1,240.31 s, and 70, respectively. In addition, it is seen from Figure 8 that the proposal tends to grow flat and its changing trend is very close to that of the Xu-Stoller method. Thus, from the viewpoint of the complexity, both the proposal and the Xu-Stoller extract policy rules of higher quality than the Cotrini method.

Table 2: Performance evaluations with different methods

Policy set	Method	Time (s)	Complexity (WSC)
University P	Xu-Stoller	227	34
	Cotrini	126	508
	The proposal	10.6	33
Healthcare P	Xu-Stoller	32,645	16
	Cotrini	529	272
	The proposal	10.1	59
Project Management P	Xu-Stoller	-	29
	Cotrini	3,587	77
	The proposal	15.97	69
University PN	Xu-Stoller	4,230	34
	Cotrini	204	1,389
	The proposal	27	55
Healthcare PN	Xu-Stoller	45,348	17
	Cotrini	3,587	462
	The proposal	17.8	86
Project Management PN	Xu-Stoller	-	45
	Cotrini	2,848	100
	The proposal	29.83	65
Amazon Kaggle	Xu-Stoller	-	51
	Cotrini	237	2,431
	The proposal	211.3	49
Amazon UCI	Xu-Stoller	-	74
	Cotrini	1,345	1,247
	The proposal	1,240.31	70

4.4 Discussion

From the above performance evaluations and comparisons for the policy extraction and optimization, we find the proposal has the following main advantages:

1) It is time-consuming, labor-intensive, and error-prone to identify and extract the buried access control policies by manually sifting through the natural language specification documents. To address this issue in the phase of extracting policy elements, the proposal utilizes the natural language processing techniques to automatically extract the access control policy elements from the top-level policy specification documents in natural language.

2) The engineering scales are large using most conventional methods. To reduce the mining scale and complexity of the policy engineering, the proposal adopts an unsupervised leaning method to develop a policy set and derives effective rules based on the initial policy set and the constructed generalized Cartesian product among subjects, objects, as well as operations.

Table 3: Comparison of features

Feature	[10]	[22]	[9]	[28]	[6]	Proposed method
Enhancing the policy interpretability	√	√	√			
Reducing the scale of the policy engineering		√	√			√
Accuracy analysis of the policy				√	√	√
Complexity analysis of the policy				√	√	√

3) The ABAC policy should be as concise and correct as possible. However, there are numerous redundant and inaccurate rules extracted using the existing methods, and these initial derived policies are not optimized. To address this issue in the optimization phase, the proposal proposes a pruning algorithm, in order to further optimize the mined rules.

Compared to the existing research approaches, features of the proposal are presented as shown in Table 3, where a tick √ indicates that the feature is available.

Nevertheless, the security issues of the proposal such as the separation of duties constraint, various cardinality constraints on entities and attributes, as well as the conflict-detection problem, however, are not considered during either the policy extraction or optimization phase, which are the main limitations of our work.

5 Conclusions

A novel policy-engineering method that combined the top-down and bottom-up approaches was proposed in this study. We utilized the natural language processing techniques to handle the top-level policy specification documents and defined the corresponding data dictionary. We also adopted an unsupervised leaning method to develop an initial policy set, and then we presented an algorithm to mine effective rules. Last, we considered the weighted structural complexity as the evaluation criterion and proposed a pruning algorithm to eliminate redundant or similar policy rules, in order to optimize the mined rules and improve the policy conciseness. The comprehensive experiments using the real-world and synthetic datasets demonstrated that the proposed method automatically extracts the access control policy elements, reduces the engineering scale and complexity and improves the policy quality. The experimental results showed that the proposal is efficient and effective. Future work will focus on studying how to implement the proposal in practical scenarios with the IoT, blockchain, and online social networks.

Acknowledgments

This work was partially supported by the Foundation of Henan Educational Committee, under Contract No. 24A520039.

References

[1] M. Alohaly, H. Takabi, and E. Blanco, "A deep learning approach for extracting attributes of ABAC," in *23nd ACM on Symposium on Access Control Models and Technologies*, pp. 137–148, 2018.

[2] M. Alohaly, H. Takabi, and E. Blanco, "Towards an automated extraction of ABAC constraints from natural language policies," in *34th IFIP TC 11 International Conference on ICT Systems Security and Privacy Protection*, pp. 105–119, 2019.

[3] G. Batra, V. Atluri, J. Vaidya, and S. Sural, "Deploying ABAC policies using RBAC systems," *Journal of Computer Security*, vol. 27, no. 4, pp. 483–506, 2019.

[4] E. Charniak, and M. Elsner, "EM works for pronoun anaphora resolution," in *12th Conference of the European Chapter of the ACL (EACL'09)*, pp. 148–156, 2009.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, no. 12, pp. 2493–2537, 2011.

[6] C. Cotrini, T. Weghorn, and D. Basin, "Mining ABAC rules from sparse logs," in *2018 IEEE European Symposium on Se-curity and Privacy*, pp. 31–46, 2018.

[7] S. Das, B. Mitra, V. Atluri, J. Vaidya, and S. Sural, "Policy engineering in RBAC and ABAC," in *From Database to Cyber Security*, pp. 24–54, 2018.

[8] S. Das, S. Sural, J. Vaidya, and V. Atluri, "Using gini impurity to mine attribute-based access control policies with environment attributes," in *23nd ACM on Symposium on Access Control Models and Technologies*, pp. 213–215, 2018.

[9] S. Das, S. Sural, J. Vaidya, V. Atluri, and G. Rigoll, "VisMAP: Visual mining of attribute-based access control policies," in *15th International Conference on Information Systems Security, Hyderabad*, pp. 79–98, 2019.

[10] P. Iyer, and A. Masoumzadeh, "Mining positive and negative attribute-based access control policy rules," in *23nd ACM on Symposium on Access Control Models and Technologies*, pp. 161–172, 2018.

[11] L. Karimi, M. Aldairi, J. Joshi, and M. Abdelhakim, "An automatic attribute-based access control policy extraction from access logs," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2304–2317, 2021.

[12] F. Khan, H. Li, Y. Zhang, H. Abbas, and T. Yaqoob, "Efficient attribute-based encryption with repeated attributes optimiza-tion," *International Journal of Information Security*, vol. 20, pp. 431–444, 2021.

[13] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[14] E. Medvet, A. Bartoli, B. Carminati, and E. Ferrari, "Evolutionary inference of attribute-based access control policies," in *8th International Conference on Evolutionary Multi-Criterion Optimization*, pp. 351–365, 2015.

[15] D. Mocanu, F. Turkmen, and A. Liotta, "Towards ABAC policy mining from logs with deep learning," in *18th International Multiconference*, pp. 124–128, 2015.

[16] M. Narouei, H. Khanpour, and R. Nielsen, "Automatic extraction of access control policies from natural lan-guage documents," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 3, pp. 506–517, 2018.

[17] M. Narouei, H. Khanpour, H. Takabi, N. Parde, and R. Nielsen, "Towards a top-down policy engineering framework for attribute-based access control," in *22nd ACM on Symposium on Access Control Models and Technologies*, pp. 103–114, 2017.

[18] M. Narouei, and H. Takabi, "A nature-inspired framework for optimal mining of attribute-based access control policies," in *International Conference on Security and Privacy in Communication Systems*, pp. 489–506, 2019.

[19] A. Roy, S. Sural, A. K. Majumdar, J. Vaidya, and V. Atluri, "Enabling workforce optimization in constrained attribute-based access control systems," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 4, pp. 1901–1913, 2019.

[20] J. Slankas, X. Xiao, L. Williams, and T. Xie, "Relation extraction for inferring access control rules from natural language artifacts," in *30th Annual Computer Security Applications Conference*, pp. 366–375, 2014.

[21] W. Sun, "Attribute-based policy evaluation using constraints specification language and conflict detections," *Mobile Information Systems*, vol. 2022, 23 pages, 2022.

[22] W. Sun, H. Su, and H. Xie, "Policy-engineering optimization with visual representation and separation-of-duty constraints in attribute-based access control," *Future Internet*, vol. 12, no. 10, p. 164, 2020.

[23] W. Sun, X. Yuan, and H. Su, "Role-engineering optimization with user-oriented cardinality constraints in role-based access control," *International Journal of Network Security*, vol. 23, no. 5, pp. 845–855, 2021.

[24] T. Talukdar, G. Batra, J. Vaidya, V. Atluri, and S. Sural, "Efficient bottom-up mining of attribute-based access control policies," in *3rd IEEE International Conference on Collaboration and Internet Computing*, pp. 339–348, 2017.

[25] T.Y. Wu, Q. Meng, Y.C. Chen, S. Kumari, and C.M. Chen, "Toward a secure smart-home IoT access control scheme based on home registration approach," *Mathematics*, vol. 11, no. 9, 2023, 2023.

[26] Y. Xia, S. Zhai, Q. Wang, H. Hou, Z. Wu, and Q. Shen, "Automated extraction of ABAC policies from natural-language documents in healthcare systems," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM'22)*, pp. 1289-1296, Las Vegas, NV, USA, 2022.

[27] X. Xiao, A. Paradkar, S. Thummalapenta, and T. Xie, "Automated extraction of security policies from natural-language software documents," in *ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*, pp. 1–11, 2012.

[28] Z. Xu, and S. D. Stoller, "Mining attribute-based access control policies," *IEEE Transactions on Dependable and Secure Computing*, vol. 12, no. 5, pp.533–545, 2015.

Biography

Wei Sun received his B.S. and M.S. degrees from the School of Information Engineering, Zhengzhou University, China, in 2003 and 2008, respectively. He is currently an associate professor and is working in the School of Computer and Information Technology, Xinyang Normal University. His current research interests include access control and system security.

Jun Lu is currently working in the School of Computer and Information Technology, Xinyang Normal University. His current research interests include database security and mining.

Mengzhao Wang is studying in the School of Computer and Information Technology, Xinyang Normal University. His current research interests include database security and mining.

Personalized Trajectory Privacy Protection Method Based on Multiple Anonymizers Forwarding

Peng-Shou Xie, Yin-Chang Pan, Tao Feng, Ye Lu, Wan-Jun Shao, and Cun-Huan Tan
(Corresponding author: Yin-Chang Pan)

School of Computer and Communications, Lanzhou University of Technology
No. 36 Peng Jia-ping Road, Lanzhou, Gansu 730050, China
(Email: 1713974116@qq.com)

(Received May 22, 2023; Revised and Accepted Sept. 22, 2023; First Online June 22, 2024)

Abstract

In response to the limitation of a single anonymizer in the TTP structure and the problem of user personalization needs, this paper proposes a personalized trajectory privacy protection method based on multi-anonymizer forwarding. The method first allocates different privacy budgets according to the user's location sensitivity and uses differential privacy techniques to add Laplace noise, matching the privacy budget to sensitive locations. Then, multiple anonymizers are deployed between the user and the LBS server, and the noise-added trajectories are mapped to different anonymizers using a random mapping mechanism. Finally, the noisy trajectory data are forwarded to the LBS server anonymously for other users to inquire. The experimental results show that the scheme can protect users' privacy and meet personalized needs.

Keywords: Differential Privacy; Multiple Anonymizers; Personalization; Trajectory Privacy

1 Introduction

In recent years, the development of mobile communication and positioning technologies and the widespread use of Geographic Information Systems (GIS) have led to the storage of massive amounts of geographic information in GIS [5]. In location-based applications, mobile location devices such as car navigators, smartphones, tablets, and location sensors are used to collect location data from moving targets [7]. Location Service Providers (LSPs) generate huge amounts of mobile path information when performing tasks such as location reports and Location Based on Service (LBS) queries. The collected location and trajectory data are of great reference value to commercial organizations, traffic management departments, and other legitimate information search agencies and service providers. For example, traffic managers can obtain road condition information from the existing road condition information and users' travel trajectories. Commercial organizations can use users' location and trajectory data to obtain users' Points of Interest (POIs) gathering areas for accurate advertising recommendation and marketing. Meanwhile, relevant government departments can also use users' location and trajectory information to determine the spread of epidemics and other situations.

However, while people enjoy the great convenience brought by LBS but also face the risk of personal sensitive information leakage [11]. With the development of location-based applications, more and more trajectory data are collected and applied, and the trajectory data usually contain users' sensitive information, which may pose a threat to users' privacy when published directly [12]. In recent years, to address the above problems, some scholars have proposed differential privacy [2] based trajectory privacy protection schemes by adding Laplace noise to the real location. Yuan *et al.* [17] proposed a differential privacy-based regional geographic indistinguishability mechanism, which cannot guarantee the quality of LBS services because the distance between the location of the generated noise and the user's real location is unpredictable. Wu *et al.* [13] proposed a secure storage and publishing method for trajectory data satisfying differential privacy, but the data volume is large, the traversal block is not considered when constructing the storage structure, and the storage space overhead is large. Jiao *et al.* [3] proposed a trajectory differential privacy protection method based on an exponential mechanism, but the use of a more complex mathematical model increases the running time of the algorithm and reduces the running efficiency. Zhao *et al.* [20] proposed a prefix tree-based trajectory data privacy protection method, which stores the trajectory segments in the nodes of the tree and combines the minimum description length method with the Dijkstra method to select the characteristic trajectory points representing the whole trajectory, thus further re-

ducing the complexity of data processing. Cai *et al.* [1] proposed a differential privacy-based trajectory database publishing algorithm that takes the time factor into account. First, a $3D$ generalized trajectory data set was established. Then, the trajectory space is divided into several planes by timestamps, and the set of locations on each plane is further processed by clustering and generalization to reform the trajectories that can be published. Differential privacy, as a privacy-preserving method with a solid mathematical foundation, has been widely used in trajectory data publishing. However, the current methods based on differential privacy do not fully achieve personalized trajectory privacy protection.

Therefore, to address the above problems, this paper proposes a personalized trajectory privacy protection method based on multi-anonymizer forwarding in combination with differential privacy technology (TP-MAF), which can meet users' personalized needs while protecting their privacy.

2 Preliminary Knowledge

2.1 System Model

This method enhances the privacy protection of user trajectories by deploying multiple anonymizers [19] between users and servers, sending users' noisy trajectories and real trajectories to different anonymizers for forwarding so that adversaries cannot obtain users' real trajectories from anonymizers. The TP-MAF trajectory privacy protection model is shown in Figure 1.

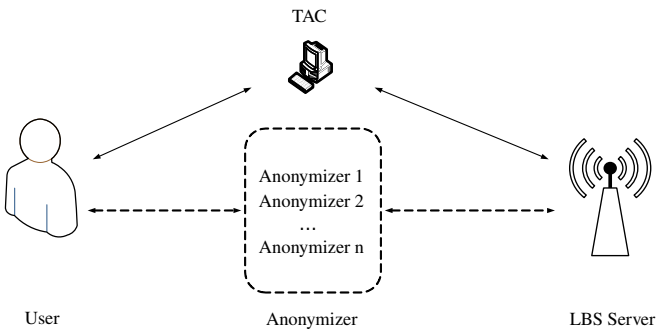


Figure 1: TP-MAF Trajectory Privacy Protection Model

The working Process is as follows:

1) Before a user initiates an inquiry request at a location, the user is verified at the Trusted Authentication Center [8] (TAC) at that location, and the pseudonym of the user and the corresponding certificate is obtained after successful verification.

2) The user sends the obtained pseudonym, the pseudonym certificate, and the noised trajectory together with the random mapping mechanism to K anonymization servers [14], which are forwarded by the anonymizer to the LBS servers for inquiry.

3) The LBS server authenticates the user's pseudonym, queries the POIs according to the inquiry content of the user, and finally returns the inquiry results to the user through each anonymizer.

4) During the search process, users only receive feedback on their search location and filter and refine the actual location of the current user to obtain accurate search results.

2.2 Related Definitions

2.2.1 Basic Concepts

Definition 1. *(Location Point) The location point is denoted by $S_i = (x_i, y_i)$, where $i = (1, 2, \cdots, n)$,x_i and y_i denote respectively the longitude and latitude of the i-th location point.*

Definition 2. *(Trajectory) Each trajectory T is a sequence of multiple position points S in time order, which can be expressed as:*

$$T = S_1 \to S_2 \to \cdots \to S_n, n \geq 0 \tag{1}$$

where n denotes the number of location points.

Definition 3. *(Trajectory Dataset) A trajectory dataset $\widetilde{T}$ is a set consisting of a series of trajectory sequences, which can be expressed as:*

$$\widetilde{T} = \{T_1, T_2, \cdots, T_m\}, m \geq 0 \tag{2}$$

where m denotes the number of trajectories.

2.2.2 Differential Privacy

Differential privacy is to add noise to the original trajectory data to make the original trajectory data distorted, to achieve the purpose of protecting the user's sensitive location.

Definition 4. *(Differential Privacy) Suppose there are inquiry algorithms A and adjacent datasets $\tilde{T}_1$,$\tilde{T}_2$ such that the inquiry result R of $\tilde{T}_1$,$\tilde{T}_2$ satisfies [10]:*

$$\frac{P_r\left[A\left(\tilde{T}_1\right) \in R\right]}{P_r\left[A\left(\tilde{T}_2\right) \in R\right]} \leq e^{\varepsilon} \tag{3}$$

Then A is said to satisfy differential privacy.

Where ε is the privacy budget, the larger the ε, the lower the privacy protection; the smaller the ε, the higher the privacy protection.

Definition 5. *(Sensitivity) For any neighboring data set $\tilde{T}_1$ and $\tilde{T}_2$, given an inquiry function G:$\tilde{T} \to V$, define the sensitivity as [9]:*

$$G = \max_{\tilde{T}_1, \tilde{T}_2} \left\| G\left(\tilde{T}_1\right) - G\left(\tilde{T}_2\right) \right\|_1 \tag{4}$$

Where $\|\cdot\|_1$ is the Manhattan distance.

Definition 6. *(Laplace Mechanism) If the randomized algorithm A satisfies ε differential privacy, its output is [18]:*

$$A\left(\tilde{T}\right) = G\left(\tilde{T}\right) + L\left(\frac{\triangle G}{\varepsilon}\right) \tag{5}$$

where L is the Laplace noise function.

Definition 7. *(Manhattan Distance) For any two points $S_i = (x_i, y_i)$, $S_j = (x_j, y_j)$ in the plane, Manhattan distance is defined as:*

$$\rho(S_i, S_j) = \begin{cases} |(x_i - y_i) - (x_j - y_j)|, x_i > x_j \& y_i > y_j \\ |(x_i - y_i) - (x_j - y_j)|, x_i < x_j \& y_i < y_j \\ |(x_i + y_i) - (x_j + y_j)|, x_i > x_j \& y_i < y_j \\ |(x_i + y_i) - (x_j + y_j)|, x_i < x_j \& y_i > y_j \end{cases} \tag{6}$$

Manhattan distance can be used to measure the shortest distance between two points.

2.2.3 Location Similarity

Suppose there exists a sensitive position $S_i = (x_i, y_i)$ on the original trajectory, which produces a release position $S_i^{'} = \left(x_i^{'}, y_i^{'}\right)$ after adding noise, then the angle change between the original position and the release position is:

$$\cos\langle S_i, S_i' \rangle = \frac{S_i \cdot S_i'}{|S_i|\,|S_i'|} = \frac{x_i y_i + x_i' y_i'}{\sqrt{(x_i^2 + y_i^2)(x_i'^2 + y_i'^2)}} \tag{7}$$

2.2.4 Description of RSA Algorithm

Step 1: Generate Key.

1) Choose two large prime numbers p and q that satisfy the need and calculate $n = p * q$, $g(n) = (p-1)*(q-1)$, where $g(n)$ is the Euler function value of n .
2) Choose an integer e that satisfies $1 < e < g(n)$, and $gcd(g(n), e) = 1$.Compute d through $d * e = 1 \bmod g(n)$.
3) Use $\{e, n\}$ as the public key and $\{d, n\}$ as the secret key. Assuming that Alice is the recipient of the secret message, only Alice knows the secret key $\{d, n\}$. All people can know the public key $\{e, n\}$.

Step 2: Encryption.

If the sender wants to send a message m to Alice that needs to be kept secret, it selects Alice's public key $\{e, n\}$, then computes $c = m^e \bmod n$, and sends the ciphertext c to the receiver Alice.

Step 3: Decryption.

The receiver Alice receives the ciphertext c, calculates $m = c^d \bmod n$ according to the private key she has, and the result m is the message the sender wants to send.

3 TP-MAF Trajectory Privacy Protection Method

In the process of user request service, the TP-MAF trajectory privacy protection method mainly consists of eight steps: privacy budget distribution, generating noise trajectory, generating user pseudonym, user inquiry request, verifying legality, server inquiry, anonymizer forwarding, and user refinement result. The working process of the TP-MAF trajectory privacy protection method is shown in Figure 2.

3.1 Privacy Budget Distribution

Different privacy budgets are assigned according to the different sensitivities of different user location points, and then noise matching their privacy budgets is added to the locations in the real trajectory, by which a noisy trajectory can be generated to meet the user's personalized needs and also to protect the user's privacy, the specific steps are as follows [4, 15].

Step 1: The user first enters his or her sensitive location $S_i = (x_i, y_i)$, the radius of the sensitive area r, and the calibration parameter $\varepsilon^{'}$.

Step 2: Calculate the Manhattan distance $\rho\left(S_i, S_i^{'}\right) = \left|x_i - x_i^{'}\right| + \left|y_i - y_i^{'}\right|$ between the inquiry location $S_i^{'} = \left(x_i^{'}, y_i^{'}\right)$ and the nearest sensitive location.

Step 3: Determine whether $\rho\left(S_i, S_i^{'}\right) > r$ holds. If it holds, calculate the privacy budget $\varepsilon_i = \frac{\rho(S_i, S_i')}{\sum_{i=1}^{n} \rho(S_i, S_i')}\varepsilon'$ for the location S_i. If it does not hold, assign the minimum privacy budget ε_{min}.

Step 4: The set of different privacy budgets obtained depending on the location points is $\varepsilon = (\varepsilon_{min}, \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$.

3.2 Generate Noise Trajectory

For ease of arithmetic, the probability density function of the plane Laplace distribution is transformed into polar coordinate form as follows:

$$P_l(r, \theta) = \frac{(\varepsilon/\Delta G)^2}{2\pi} e^{-\frac{\varepsilon}{\Delta G} r} \tag{8}$$

Where r denotes the distance between the user's real location and the published location after adding Laplace noise. The edge probability density function of distance r and angle θ is obtained from Equation (8) as:

$$P_l(r) = \int_0^{2\pi} P_l(r, \theta)\, \mathrm{d}\theta = (\varepsilon/\triangle G)^2 e^{-\frac{\varepsilon}{\triangle G} r} \tag{9}$$

$$P_l(\theta) = \int_0^{\infty} P_l(r, \theta)\, \mathrm{d}r = \frac{1}{2\pi} \tag{10}$$

To solve for the value of r, first, calculate the cumulative distribution function of r:

$$F_l(r) = \int_0^r P_l(r)\,\mathrm{d}r = \left(1 - e^{-\frac{\varepsilon}{\triangle G}r}\right)\frac{\varepsilon}{\triangle G} \tag{11}$$

By solving the inverse function of Equation (11) we obtain r:

$$r = -\frac{\triangle G}{\varepsilon}\left(W_{-1}\left(\frac{\alpha - 1}{\varepsilon}\right) + 1\right) \tag{12}$$

Where α is a random number of [0,1) subject to uniform distribution and W_{-1} is a branch of $(-\infty, -1)$ of the Lembert W function.

From Equation (10), we can see that $P_l(\theta)$ is a constant, so θ is a random number of $[0, 2\pi)$ obeying uniform distribution. Through the above analysis, r and θ are obtained, so the location of the added noise is:

$$\begin{cases} x_i' = x_i + r\cos\theta \\ y_i' = y_i + r\sin\theta \end{cases} \tag{13}$$

The real trajectory of the user is $T = S_1 \to S_2 \to \cdots \to S_n$. The Laplace noise is added by the user to the sensitive locations in the real trajectory several times to generate the corresponding $2k$ noisy trajectories $T' = S_1' \to S_2' \to \cdots \to S_n'$. The similarity between the original sensitive location and the processed released location is calculated according to Equation (7), and the trajectories with more similar locations are selected using the location similarity.

3.3 Generate User Pseudonym

When a user logs into the system for the first time, he must register with the TAC using his identity information. The user first chooses a random number r as his temporary key and sends the generated registration request message to the TAC by asymmetric encryption of the rest of the user identity ID_U together with the RSA algorithm. Then, the TAC generates a pair of public and private keys PK_U and SK_U for the user and uses the user's key r for symmetric encryption ID_U, PK_U and SK_U to generate the user's reply message E, which is then sent to the user [22]. Finally, the user decrypts the message E with the key r and obtains the PK_U and SK_U public-private key pairs.

When the user needs to apply for a certificate from TAC, he first uses his private key SK_U to sign the user's identity ID_U to get $Sig_{SK_U}(ID_U)$ and uses his identity ID_U, the digital signature $Sig_{SK_U}(ID_U)$ of the identity, and the generated random temporary key r together with TAC's public key PK_{TAC} for asymmetric encryption through RSA algorithm to generate the user's request message $E_{PK_{TAC}}$, and sends it to TAC. Then, TAC decrypts the request message with his private key SK_{TAC}. At the same time, TAC verifies the digital signature $Sig_{SK_U}(ID_U)$ with the user's public key PK_U. Only if the verification is successful, TAC generates a pseudonym for the user and distributes the pseudonym and the corresponding certificate [21]. TAC digitally signs the pseudonym PID_U with its private key SK_{TAC} to obtain the corresponding certificate $Cert_U$, $Cert_U = Sig_{SK_{CA}}(PID_U)$, and generates a pseudonym message E_{n_r} using the user's key r to symmetrically encrypt PID_U and $Cert_U$ return to the user. Finally, the user decrypts E_{n_r} with the key r to obtain the pseudonym PID_U and the pseudonym certificate $Cert_U$.

The definitions and descriptions of the symbols in the TP-MAF method are shown in Table 1.

Table 1: Definition and description of symbols in the TP-MAF method

Symbol	Description
ID_U	User identification
PK_U	User's public key
SK_U	User's private key
PK_{TAC}	TAC's public key
SK_{TAC}	TAC's private key
Key_{LBS}	The private key of the LBS server
$Sig_{SK}(ID_U)$	Digital signature of the user's identity
PID_U	User pseudonym
$Cert_U$	User pseudonym certificate
M	Inquiry content
R	Inquiry radius
T_j	Time threshold
$Hash(\cdot)$	Hash function
MUA	Messages relayed to the anonymizer by the user
MAU	Messages relayed by the anonymizer to the user
MAS	Messages relayed by the anonymizer to the LBS server
MSA	Messages relayed to the anonymizer by the LBS server

3.4 User Inquiry Request

The user has already obtained a pseudonym and the corresponding pseudonym certificate through the pseudonym generation mechanism. When an inquiry request is sent, a mapping table is first constructed to assign the pseudonym, the corresponding pseudonym certificate, and k trajectories to different anonymizers.

There are a total of K anonymizers $A_1, A_2, \cdots A_K$ in the TP-MAF trajectory privacy-preserving model, and the user inquiry uses a random mapping mechanism to assign each of the user's k trajectories to randomly selected n different anonymizers for processing, and $A \geq n$. By constructing a mapping table, using k trajectories as variables, construct a hash function and modulo it to obtain a mapping to the anonymizer A_l with the number l

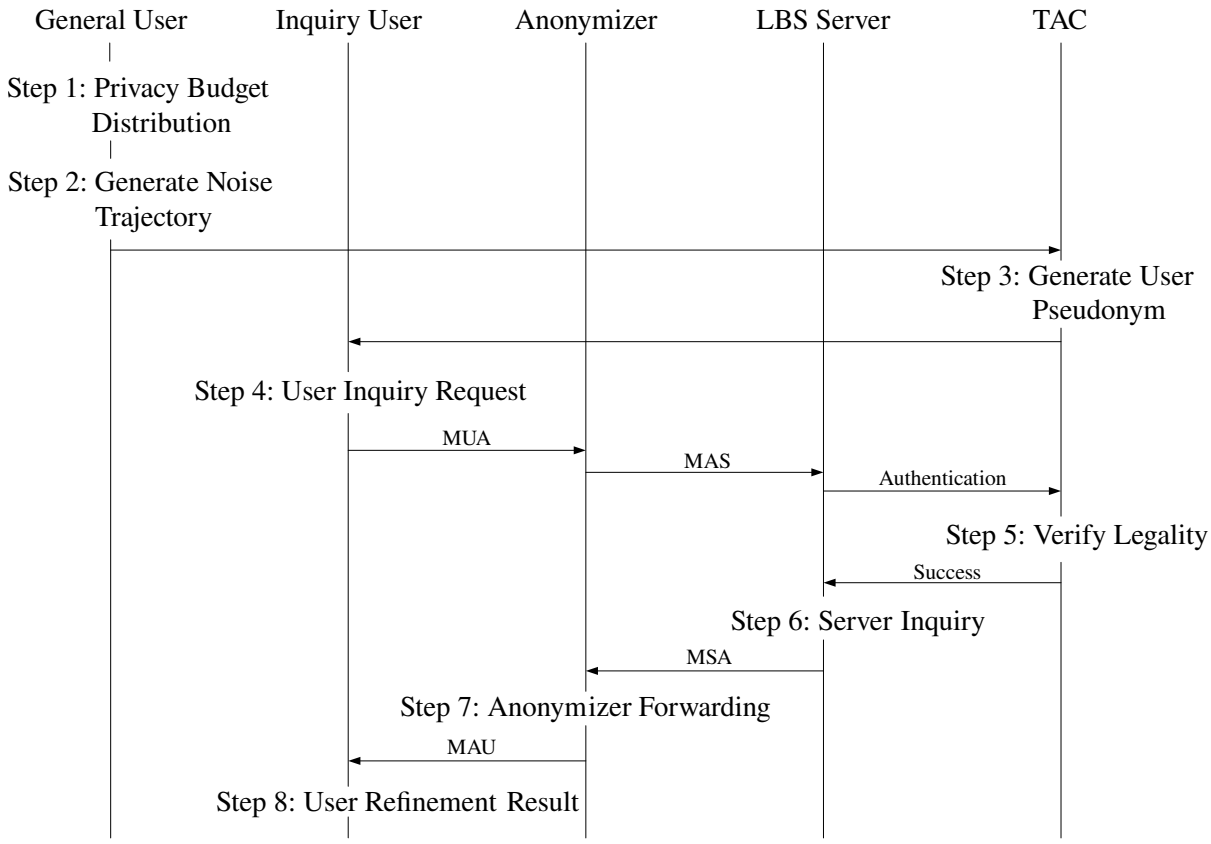


Figure 2: The working process of the TP-MAF trajectory privacy protection method

, $l = 1, 2 \cdots, K$.

$$A_l = Hash\,(k_i) \bmod K \tag{14}$$

In the above process, if there are different trajectories mapped to anonymizers with the same number, a conflict will arise. To solve this problem, this scheme uses the method of second detection and then hashing to deal with the conflicting anonymizer numbers, and then calculates by Equation (15):

$$A_l = (Hash\,(k_i) + \tau) \bmod K \tag{15}$$

In this formula, $\tau = 1$ is taken first, and if there is still a conflict in the obtained anonymizer number, the existing τ value is increased by 1 in turn until the conflict is resolved.

Finally, the user will noise-added trajectory $T^{'}$, real trajectory T, pseudonym PID_U and the corresponding certificate $Cert_U$, the location where the user initiated the inquiry (x_i, y_i), the inquiry content M, the inquiry radius R, the time threshold T_j, and the user key SK_U together to form an inquiry request message, its request message is:

$$MUA = E\{T', T, M, Cert_U, PID_U, (x_i, y_i), SK_U, T_j, R\}. \tag{16}$$

3.5 Verify Legality and Server Inquiry

After the user sends the inquiry message to different anonymizers, each anonymizer first decrypts the inquiry request message MUA using their private keys to obtain the user pseudonym PID_U and the corresponding certificate $Cert_U$, and the valid time threshold T_j. Then the TAC verifies the legitimacy of the user pseudonym PID_U and the corresponding certificate $Cert_U$ [6]. Only when the user pseudonym PID_U and the corresponding certificate $Cert_U$ are legal and within the valid time threshold T_j, the LBS server can provide inquiry service to the user, otherwise, the service stops [16].

After the user's identity is verified, the LBS server will get the message MAS relayed by the anonymizer, and the LBS server queries the POIs needed by the user in the LBS database according to the inquiry content M and the inquiry radius R to get the POIs that the user needs to inquiry, and encrypts them using the symmetric encryption algorithm and the key Key_{LBS} to get E_n respectively. Finally, the LBS server returns the encrypted inquiry result set to the corresponding anonymizer, and the message returned to the anonymizer can be expressed as:

$$MSA = \{E_n\,(POIs)\} \tag{17}$$

3.6 Anonymizer Forwarding and User Refinement Result

When these K anonymizers receive the inquiry results returned by the LBS server, they are relayed to the user respectively. The message relayed to the user can be expressed as:

$$MAU = \{E_n(POIs)\} \tag{18}$$

When the user receives the inquiry results relayed by K anonymizers, it only receives the encrypted result set of its real location and decrypts $E_n(POIs)$ using the key Key_{LBS} to obtain the exact location (x_i, y_i) of each POI, and then the user calculates the POIs included in its inquiry range to obtain the exact inquiry results.

4 Security Analysis

Theorem 1. *RSA algorithm ensures both data authenticity and data secrecy.*

Proof. Let M be the plaintext, C be the ciphertext, E be the encryption algorithm, D be the decryption algorithm, and each user is configured with a pair of keys: K_e is the public encryption key and K_d is the confidential decryption key, and the public encryption key K_e of all users is stored in the shared key store and the confidential decryption key K_d is kept properly by the users themselves. Shared Keystore is shown in Table 2.

Table 2: Shared Keystore

User	Encryption Key
A	K_{eA}
B	K_{eB}

Sender:

Step 1: The user A first decrypts the plaintext M with its decryption key K_{dA} to get the intermediate ciphertext S, $S = D(M, K_{dA})$.

Step 2: The user A finds out the public encryption key K_{eB} of user B by inquiry the shared keystore.

Step 3: The user A uses K_{eB} to encrypt S to get the ciphertext C, $C = E(S, K_{eB})$.

Step 4: The user A sends the ciphertext C to the user B.

Receiver:

Step 1: The user B receives the ciphertext C.

Step 2: The user B decrypts the ciphertext C with his decryption key K_{dB} to get the intermediate ciphertext S, $S = D(C, K_{dB})$.

Step 3: The user B gets the public encryption key K_{eA} of the user A by inquiring about the shared keystore.

Step 4: The user B encrypts with the encryption key K_{eA} disclosed by the user A to get the plaintext M, $M = E(S, K_{eA})$.

Only the user A has K_{dA}, so only the user A can decrypt M to get S, so the authenticity of the data is ensured.

Similarly, only the user B has K_{dB}, so only the user B can get the plaintext M, which ensures the secrecy of the data. □

Theorem 2. *TP-MAF trajectory privacy protection method can effectively resist anonymizer attacks.*

Proof. In the TP-MAF trajectory privacy protection method, the anonymizer tries to identify the user's real trajectory from the forwarded inquiry requests and results. Before sending an inquiry, the user first adds Laplace noise to sensitive locations on the user's trajectory several times by differential privacy techniques to form $k-1$ noisy trajectories and sends these $k-1$ noisy trajectories together with the user's real trajectory to different anonymizers for forwarding. By this message, a single anonymizer only knows the inquiry location and $d(d \leq k)$ trajectories of this user's pseudonym, but it cannot associate the user's real identity with the real trajectories. Therefore, a single anonymizer cannot know the information of the specified user. When k trajectories are sent to different anonymizers at the same time, only one anonymizer will receive the real trajectory of the user. If multiple anonymizers conspire, the probability that an attacker can identify the real trajectory of the user is only $\frac{1}{k}$, and even if the real identity of the user is identified, it is impossible to match the real identity with the real trajectory. When the LBS server sends the inquiry result to the anonymizer, it uses its private key Key_{LBS} to encrypt the inquiry result set, and in the case that the anonymizer cannot get the LBS server's private key Key_{LBS}, it cannot get the user's real trajectory through the inquiry result. □

5 Experiment Results and Analysis

This paper conducts experiments related to the impact of noisy trajectories on real trajectories generated by users' personalized privacy needs, the level of privacy protection, and data availability.

In this paper, trajectories from the Microsoft research Geolife project were used as the test dataset. This dataset captures the trajectory histories experienced by 182 users over various periods. These datasets contain a series of time-series points, each containing latitude, longitude, altitude, and other information. We take one trajectory segment every 50s, and the average length of the trajectory

segments is 10. The experimental hardware platform is 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz with 16GB of RAM. The software environment is Windows 11 JetBrains PyCharm Community Edition 2017.2.3 x64.

5.1 Effect of Personalized Differences

The impact of users' personalized needs on their real trajectories is shown in Figure 3.

A real trajectory with n location points is generated in the actual road network, and 2 noisy trajectories are determined from this real trajectory:

1) Add random noise to generate a noisy trajectory that does not take into account the user's individual needs.

2) Calculate the privacy budget of sensitive locations in the original trajectory, and then add Laplace noise matching the privacy budget to the sensitive locations to produce a trajectory that meets the user's personalization requirements. It can be seen that the method in this paper significantly deviates more from the real location at the sensitive location than the noisy trajectory without considering the user's personalized needs. The method in this paper can effectively protect the user's privacy and meet the user's personalized needs at the same time.

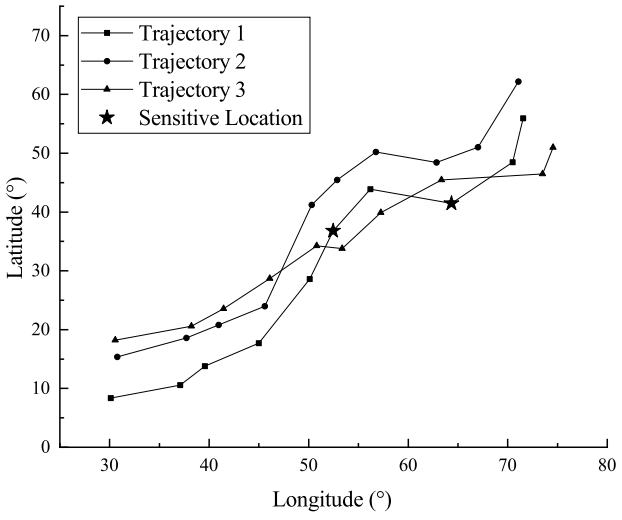


Figure 3: The impact of personalized needs on the trajectory

Where trajectory 1 represents the original trajectory of the user. Trajectory 2 represents the noisy trajectory with the user's personalized needs considered. Trajectory 3 represents the noisy trajectory without considering the user's personalized needs.

5.2 Privacy Protection Level

This paper uses the PL value to measure the level of privacy protection. Assuming that the user's sensitive location in the original trajectory T_1 is S_i, the new location generated after noise addition to the sensitive location is $S_i{}'$, the number of sensitive locations in this trajectory is n, and the number of anonymizers is k, then the PL value can be expressed as:

$$PL = \sqrt{\frac{\sum_{i=1}^{n} \left\| S_i - S_i' \right\|_1}{k}} \tag{19}$$

Where $\|S_i - S_i^{'}\|_1$ indicates the Manhattan distance between the original trajectory and the noisy trajectory corresponding to the sensitive location, PL value indicates the degree of privacy leakage, and PL value is inversely proportional to the level of privacy protection. The privacy protection of the trajectory data can be more intuitively seen by the PL value. The higher the PL value, the higher the degree of privacy leakage, the lower the level of privacy protection. On the contrary, the lower the PL value, the lower the degree of privacy leakage, and the higher the level of privacy protection.

Figure 4 shows that the method in this paper has a better level of privacy protection compared to the other two methods for the following main reasons:

1) In this paper, the method combines differential privacy techniques with multiple anonymizers. Before the trajectory is published, the original trajectory is noise-added and then published through multiple anonymizers, so that the connection between the original trajectory and the user is cut off, making it impossible for an attacker to obtain the user's trajectory data through a particular anonymizer.

2) The TP-MAF method uses Manhattan distance to calculate the distance between locations, which provides a higher level of privacy protection. Therefore, the protected locations are less likely to be identified by an attacker.

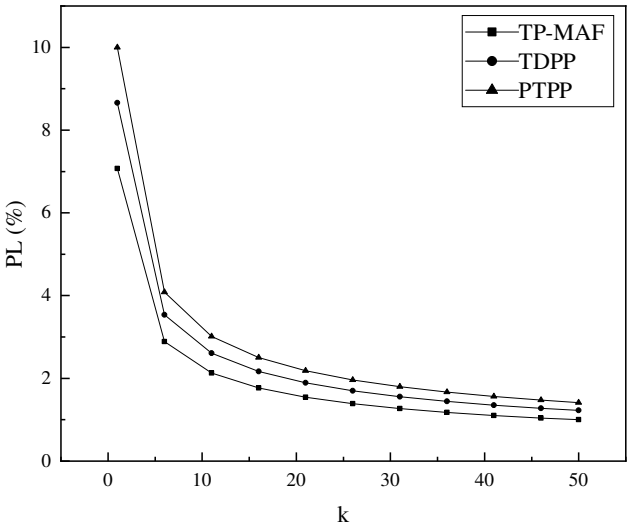


Figure 4: Privacy protection level comparison

5.3 Data Availability

This paper uses differential privacy mechanism to protect the privacy of the trajectory data, which inevitably affects the usability of the trajectory data. To test the data usability of the released trajectories, this group of experiments judges the data usability by testing the angle change between the location points of the original data set and the released data set. The smaller the angle change, the higher the similarity of the two locations, the smaller the data quality loss, and the higher the data usability. Otherwise, the lower the similarity between the two positions, the greater the data quality loss and the lower the data availability.

The purpose of this set of experiments was to test the effect of different privacy budgets on the angle change between two location points. The privacy budget values were taken as 0.1, 0.2, 0.3, 0.4, and 0.5, and each experiment was performed 10 times and the average value was taken as the final result.

Figure 5 shows that the angle change between the published dataset and the original dataset location points decreases with the increase of the privacy budget. However, the angle change of this paper's method is smaller compared to the other two methods with the same privacy budget, which indicates that the data availability of this paper's method is higher compared to the other two methods. The main reasons are as follows: the method in this paper protects only the user's sensitive location, which introduces less noise compared to the PTPP method which adds noise to the user's sensitive location and the weakly sensitive location at the same time. This paper uses a personalized differential privacy protection model, and the noise intensity added to sensitive locations is much smaller compared to TDPP, which ultimately leads to the relatively higher data availability of this paper's method.

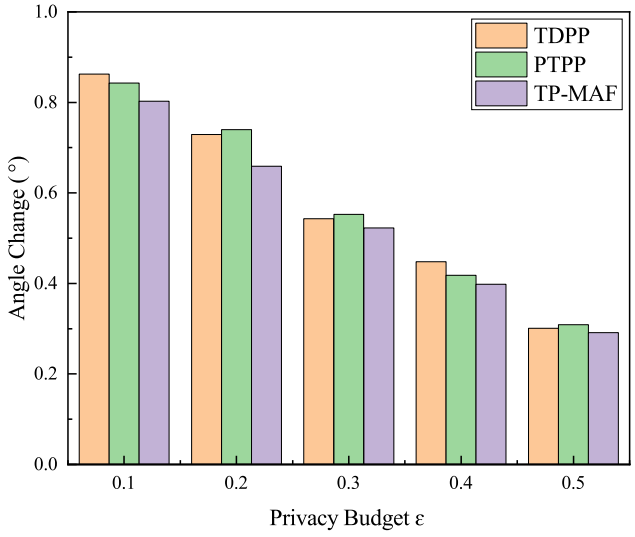


Figure 5: Data Availability Comparison

6 Conclusion

This paper proposes a privacy protection method for trajectory data. The method proposes a personalized trajectory privacy protection method based on multi-anonymizer for-warding to address the shortcomings of single anonymizer in TTP architecture and the problems such as users' personalization needs. The method first adds Laplace noise to sensitive locations using differential privacy techniques based on the user's sensitivity to geographic location. Secondly, by configuring multiple anonymizers between the user and the LBS server, the noise-added trajectories are mapped to different anonymizers using a random mapping method. Finally, the noisy trajectory data is forwarded to the LBS server by the anonymizer for other users to inquire. The experiments prove that the method proposed in this paper can protect users' privacy and meet their personalization requirements at the same time. In addition, the trajectory after noise addition does not take into account the location semantics of the user, and such phenomena as lakes and cliffs may appear in the trajectory. Therefore, the subsequent work can be carried out in terms of the location semantic analysis of the trajectories, and the unreasonable locations can be dealt with to some extent.

Acknowledgement

This study was supported by the National Natural Science Foundation of China under grants 61862040 and 62162039. The authors thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] S. J. Cai, X. Lyu, X. Li, D. H. Ban, and T. Zeng, "A trajectory released scheme for the internet of vehicles based on differential privacy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16534–16547, 2021.

[2] W. Deng, X. T. Chen, Q. H. Zhang, and G. Y. Wang, "A differential privacy-preserving algorithm based on tree model," *Journal of Chongqing University of Posts and Telecommunications*, no. 5, 2020 (in Chinese).

[3] H. C. Jiao, W. J. Liu, and Z. Wang, "Trajectory differential privacy protection method based on exponential mechanism," *Big Data Research*, pp. 141–152, 2023 (in Chinese).

[4] H. T. Li, X. Y. Ren, J. Wang, and J. F Ma, "Continuous location privacy protection mechanism based on differential privacy," *Journal on Communications*, vol. 42, no. 8, pp. 164–175, 2021 (in Chinese).

[5] J. Liu and F. L. Qin, "Protection of user data by differential privacy algorithms," *International Journal of Network Security*, vol. 22, no. 5, pp. 838–844, 2020.

[6] P. Q. Liu, S. C. Xie, Z. H. Shen, and H. Wang, "Enhancing location privacy through p2p network and caching in anonymizer.," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 5, 2022.

[7] W. F. Liu, J. J. Wu, and Z. Xi, "Privacy protection methods of location services in big data," *Open Computer Science*, vol. 12, no. 1, pp. 389–402, 2022.

[8] S. Y. Qiu, D. C. Pi, Y. X. Wang, and Y. F. Liu, "Novel trajectory privacy protection method against prediction attacks," *Expert Systems with Applications*, vol. 213, p. 118870, 2023.

[9] C. Song, D. C. Cheng, and S. P. Ni, "A privacy-preserving scheme for personalized differential privacy of k-anonymous trajectories," *Journal of Beijing University of Posts and Telecommunications*, pp. 1–6, 2023 (in Chinese).

[10] C. Song, B. Xu, and J.Y. He, "Trajectory differential privacy protection method based on exponential mechanism," *Journal of Beijing University of Posts and Telecommunications*, vol. 45, no. 1, pp. 13–18, 2022 (in Chinese).

[11] P. Wang, "Wireless positioning trajectory data privacy protection method for location-based services," in *IEEE 2nd International Conference on Power, Electronics and Computer Applications*, pp. 904–907, 2022.

[12] R. X. Wen, W. Q. Cheng, H. J. Huang, W. Miao, and C. Wang, "Privacy preserving trajectory data publishing with personalized differential privacy," in *IEEE International Conference on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking*, pp. 313–320, 2020.

[13] W. Q. Wu, Y.X. Zhao, Q. Wang, and C.F. Di, "A safe storage and release method of trajectory data satisfying differential privacy," *Journal of Computer Research and Development*, vol. 58, no. 11, pp. 2430–2443, 2021.

[14] P. S. Xie, X. M. Han, T. Feng, Y. Yan, and G. Q. Ma, "A method of constructing arc edge anonymous area based on lbs privacy protection in the internet of vehicles.," *International Journal of Network Security*, vol. 22, no. 2, pp. 275–282, 2020.

[15] P. S. Xie, X. Wang, H. X. Yang, L. X. Wang, T. Feng, and Y. Yan, "Location privacy protection algorithm based on pagerank and differential privacy in internet of vehicles," *International Journal of Network Security*, vol. 23, no. 6, pp. 1049–1057, 2021.

[16] X. Xu, M. Wen, and L. L. Wang, "A blind signature-based location privacy protection scheme for mobile social networks," *International Journal of Network Security*, vol. 23, no. 5, pp. 867–877, 2021.

[17] S. L. Yuan, D. C. Pi, and M. Xu, "Trajectory privacy protection method based on differential privacy," *Journal of Electronics*, vol. 49, no. 7, pp. 1266–1273, 2021 (in Chinese).

[18] J. Zhang, Y. Z. Li, Q. Ding, L. W. Lin, and X. C. Ye, "Successive trajectory privacy protection with semantics prediction differential privacy," *Entropy*, vol. 24, no. 9, p. 1172, 2022.

[19] S. B. Zhang, X. J. Mao, K. K. R. Choo, T. Peng, and G. J. Wang, "A trajectory privacy-preserving scheme based on a dual-k mechanism for continuous location-based services," *Information Sciences*, vol. 527, pp. 406–419, 2020.

[20] X. D. Zhao, D. C. Pi, and J. F. Chen, "Novel trajectory privacy-preserving method based on prefix tree using differential privacy," *Knowledge-Based Systems*, vol. 198, p. 105940, 2020.

[21] Z. Q. Zheng, X. S. Wu, H. Wang, K. Liu, and Z. H. Shen, "PTPM protection method for trajectory privacy in mobile social network," *Journal of Chinese Computer Systems*, vol. 42, no. 10, pp. 2153–2160, 2021 (in Chinese).

[22] H. F. Zhu, Y. L. Zhang, X. Y Wang, and L. W. Wang, "Partitioned group password-based authenticated key exchange with privacy protection," *International Journal of Network Security*, vol. 23, no. 1, pp. 116–125, 2021.

Biography

Peng-shou Xie was born in Jan.1972. He is a professor and a supervisor of master student at Lanzhou University of Technology. His major research field is Privacy Protection, Security on Internet of Vehicles, Security on Industrial Internet. E-mail: xiepsh_lut@163.com

Yin-chang Pan was born in Mar.1993. He is a master student at Lanzhou University of Technology. His major research field is network and information security.E-mail:1713974116@qq.com

Tao Feng was born in Dec.1970. He is a professor and a supervisor of Doctoral student at Lanzhou University of Technology. His major research field is modern cryptography theory, network and information security technology. E-mail: fengt@lut.cn

Ye Lu was born in May.1986. He is currently a lecturer at Lanzhou University of Technology. His research interests include cyber security and blockchain. E-mail: luye@lut.edu.cn

Wan-jun Shao was born in Jan.1998. She is a master student at Lanzhou University of Technology. Her major research field is network and information security. E-mail: 2443404684@qq.com

Cun-huan Tan was born in Jun.2000. She is a master student at Lanzhou University of Technology. Her major research field is network and information security. E-mail:1635879818@qq.com

Multi-keyword Ciphertext Sorting Search Based on Conformation Graph Convolution Model and Transformer Network in English Education

Hang Li[1], Zeyang Li[1], Xiaowei Wang[1], Muhammad Ibrar[1], and Xinjie Zhu[2]
(Corresponding author: Zeyang Li and Xiaowei Wang)
Software College, Shenyang Normal University[1]
Shenyang 110034, China
Email: lihangsoft@163.com
School of Foreign Languages, Zhengzhou University of Science and Technology[2]
Zhengzhou 450000, China
(Received Nov. 26, 2023; Revised and Accepted Feb. 7, 2024; First Online June 22, 2024)
The Special Issue on Data Fusion, Deep Learning and Optimization Algorithms for Data Privacy Protection
Special Editor: Prof. Shoulin Yin (Harbin Institute of Technology)

Abstract

With the wide application of cloud computing, the outsourcing service model for data or computing is more and more accepted by the industry. Because asymmetric searchable encryption is difficult to deal with the problem of ciphertext sorting, the existing research on multi-keyword ciphertext sorting search mainly adopts symmetric searchable encryption mechanism, the core problem of which is the data structure, construction algorithm, and search algorithm of the secure, searchable index. Therefore, a new multi-keyword ciphertext sorting search based on the conformation graph convolution model and Transformer network is constructed in this paper. First, the Transformer network uses the bottleneck features of input samples to generate the bottleneck features of pseudo-abnormal data, thereby adding abnormal data information to the training set. Then, the model constructs a multi-modal feature learning and fusion graph convolutional network to obtain contextual features of each piece of information. The security and performance analysis show that the proposed scheme is safe and feasible under the known ciphertext model. Simulation results show that the proposed scheme can realize result verification and fair payment at an acceptable cost.

Keywords: Conformation Graph Convolution; Feature Learning; Multi-Keyword Ciphertext Sorting Search; Transformer Network

1 Introduction

With the rapid development of cloud computing, more and more users migrate large amounts of data to the cloud platform to save local storage costs. At the same time, the cloud platform also provides instant services for remote storage and computing, which is convenient for users to access and use data anytime and anywhere. However, since users have lost control of the data, how to ensure data privacy security has become a key issue. In order to protect the privacy of sensitive data, the data needs to be encrypted before users upload it to the Cloud Service Provider (CSP) [8]. However, data encryption makes plaintext based keyword retrieval technology impossible to use. The proposed searchable encryption technology can not only realize the retrieval of encrypted data, but also ensure the privacy and security of data.

In searchable encryption, it is often assumed that CSP is honest and curious entities. In practice, however, in order to save computing resources while obtaining service fees, CSP may dishonestly perform search operations and send incorrect or incomplete search results to users [17, 30]. In the pay-before-use model, even if the above dishonesty occurs, the user must first pay the service fee to the CSP. In the use-before-pay-later model, there may be cases where dishonest users receive the correct results and refuse to pay the service fee. In the pay-before-use model, the value of the Data provided by the data owner to the CSP needs to be taken into account in the transaction, that is, the data user should pay the message fee for the data provided to the data owner when conducting a transaction [5]. To address these equity issues, traditional solutions often rely on a trusted third party. However, because the third party does not have the ability to directly verify the correctness and integrity of the search results, when there is a transaction dispute, the third party needs to spend a lot of time to solve the dispute, the fair payment cannot be directly guaranteed,

and the dishonest behavior generated in the transaction process will not be punished. On the other hand, when third-party organizations, CSP and users interact with each other, users' personal privacy information on third-party organizations is also at risk of being leaked.

From the above analysis, we can see that an effective method is needed to solve the fair payment problem in traditional searchable encryption schemes. With the advent of Bitcoin [20], blockchain as its underlying technology can provide support for solving this problem. Blockchain has the characteristics of decentralization and immutability, which can be well combined with cloud computing; Smart contracts on blockchains [9, 24] can be written directly into code and executed automatically, outside the control of any centralized authority. Thus, blockchains and smart contracts are suitable for performing verification operations and enabling fair payments in searchable cryptographic schemes.

In the process of combining blockchain smart contracts with searchable encryption schemes, people use blockchain smart contracts to perform verification of search results and achieve fair payment. In addition, some schemes use smart contracts to replace CSP to perform search operations. In the execution process, blockchain smart contracts need to carry out multiple transaction transactions, store indexes that take up more space and perform complex search operations, which has low scalability, high cost and large time consumption. And the cost of fees and time increases as the complexity of the operations performed by smart contracts increases. Therefore, it is necessary to consider reducing the complexity of the operations performed by smart contracts on the basis of ensuring the realization of result verification and fair payment to reduce the overhead of time and expense costs, improve efficiency, and expand the function of the scheme to be more user-friendly [16, 18, 31].

In addition, in practical applications, the schemes that only support single keyword retrieval often can not meet the needs of users. For example, in order to obtain more accurate search results, users typically enter multiple keywords when searching and want to return the first k documents that are most relevant to the entered keywords [2,32]. Therefore, it is necessary to consider designing a searchable encryption scheme with richer retrieval functions on the basis of combining blockchain technology to ensure the realization of result verification and fair payment.

In order to reduce the time and cost of realizing result verification and fair payment, our work goal in this paper is that this scheme combines the powerful and efficient retrieval ability of CSP with the advantages of blockchain smart contract to automatically execute contract contents, and realizes the sequential retrieval of ciphertext, the verification of search results, and the fair payment among data owners, CSP and data users.

2 Related Works

Searchable encryption technology is a kind of password primitive that allows users to search ciphertext data. It uses the powerful computing resources of cloud server for keyword search, and its core idea is that users have the ability to search for keywords in ciphertext domain. Mihailescu *et al.* [14] proposed the idea of searchable encryption for the first time to solve the problem of searching encrypted data on the cloud platform. Subsequently, searchable encryption under cloud storage technology became a research hotspot. In recent years, many efforts have been made to enrich the functions of searchable encryption, and schemes such as multi-keyword search [22], dynamic encryption search [27], fuzzy keyword search [4] and verifiable encryption search [11] have been proposed successively. These schemes usually assume that the cloud server will honestly perform the task, however, the cloud server is often not completely trustworthy, it may save computing resources or defrauds the service fee, after receiving the service fee to return incorrect or incomplete results; at the same time, even if the user receives the correct search results, if the user claims that the search results are incorrect, it may maliciously refuse to pay the service fee. The above situation leads to service-payment inequity, resulting in distrust between users and cloud servers. In order to solve the above problems, traditional solutions usually consider supervision and arbitration by a trusted third party.

Since its inception, blockchain has received a lot of attention from academia and industry for its ability to enable fair payments without the introduction of third-party institutions. Therefore, there are active attempts to combine blockchain with searchable encryption technology to solve the problems of traditional solutions. Niu *et al.* [15] proposed a trustworthy keyword search scheme based on cloud storage, which used bitcoin blockchain technology and hash function to achieve fair payment of search costs without a third party. The scheme established a secure index based on digital signature, which guaranteed the correctness of the retrieval results at the client side and verified the validity of the encrypted data at the server side. However, in the process of verifying the correctness of the result, the scheme needed a lot of signature verification calculation, and the user cost was high. Gao *et al.* [7] designed a fair symmetric searchable encryption scheme based on the Bitcoin blockchain, which automatically verified the search results through the blockchain to ensure the fairness of transactions between users and cloud servers. However, the scheme needed to execute six transactions each time to obtain the search results, and the verification of the search results was realized through the Bitcoin script. The transaction cycle was too long, resulting in high time cost. All the schemes in references [26,29] were searched by cloud servers, and the search results were verified based on the Bitcoin blockchain to achieve fair payment between multiple parties. However, because Bitcoin smart contracts were not Turing-complete, their

functions were limited, the transaction process was complex, the transaction cycle was long, and the efficiency was not high. Ali *et al.* [1] implemented dynamic and efficient keyword search in a distributed storage network, used smart contracts to record encrypted search logs on the Ethereum blockchain, and designed a protocol to handle disputes and issue commissions for fair search between clients and servers. The scheme was a retrieval operation performed by contracted service nodes in a distributed storage network, with searchable indexes and metadata of search results anchored to the blockchain as evidence. Arbitration nodes on the arbitrator shard in the distributed storage network checked the correctness of the search results and realized fair payment based on the Ethereum smart contract. When a data user applies for arbitration, each arbitration node needed to re-execute the search algorithm independently and determine whether the judgment request issued by the client (i.e. the stop payment request) was valid based on the re-generated search results. These individual arbitration results were then pooled into the arbitration smart contract to make a final decision. If more than 2/3 of the nodes in the arbitrator shard accepted the judgment, the time-limited payment was suspended. As we can see, this arbitration process would waste a lot of computing resources.

Li *et al.* [13] proposed a blockchain-based searchable encryption scheme that supported complex logical expression queries, which used Ethereum smart contracts to ensure the correctness of the search results and could achieve fair payment without any verification mechanism. Guo *et al.* [11] proposed a blockchain-based distributed storage system that supported fine-grained access control. The system used Ethereum smart contracts to realize keyword search function in ciphertext state, which solved the problem that the cloud server in the traditional cloud storage system should not be able to return all search results or return wrong results. Su *et al.* [21] designed an attribute-based search encryption scheme based on blockchain and supporting verification. The scheme designed search contracts and verification contracts based on Ethereum smart contracts, and realized fair payment supporting multi-keyword search without requiring additional local verification. Wang *et al.* [23] and Bi *et al.* [3] were all encrypted index and search results stored by blockchain, and search operations were performed and fair payment was achieved based on Ethereum smart contracts. However, the storage capacity of the blockchain was limited by the nodes with the smallest storage space, and complex encrypted indexes needed to be fragmented before they could be stored into blockchain transactions, and these transactions could be uploaded one by one, which would consume a lot of time. In addition, the smart contract would consume a certain amount of costs when performing operations. In the scheme, the search operation and verification operation of high complexity were performed by the smart contract, which required a large amount of calculation when executing, and would also lead to increased costs. Therefore, the above schemes have the problems of low scalability, high time cost and high cost. In addition, they only support single keyword search, and do not consider the correlation ranking of search results, which is not flexible in function and not user-friendly.

Therefore, this paper proposes a new multi-keyword ciphertext sorting search based on conformation graph convolution model and Transformer network. Using the cloud server's efficient retrieval ability and the automatic execution of Ethereum smart contracts, the cloud server stores the encrypted index tree and lookup table; The simultaneous execution of the search algorithm can effectively reduce the complexity of the smart contract execution operation, thus reducing the time cost and expense costs consumed; The verification process of the results achieved by the Ethereum smart contract not only ensures the correctness and integrity of the search results, but also completes the fair payment between the data owner, the cloud server and the data user, and can effectively reduce the cost and improve the verification efficiency. In addition, this paper uses balanced binary tree as the index, and realizes the dynamic update of multi-keyword search and the ranking of search results on the basis of ensuring the efficiency of search, which improves the flexibility and user friendliness of the scheme.

3 Data Training Based on Transformer Network

Transformer network is composed of feed-forward neural network. Its purpose is to find another feature space Z_t that is far away from the feature space Z corresponding to the input sample, and transform the bottleneck feature $z \in Z$ of the input sample into pseudo-anomaly bottleneck feature $z_t \in Z_t$. The Transformer network is defined as $f_T(\cdot) : Z \to Z_t$, then:

$$z_t = f_T(z). \tag{1}$$

From formula (1), we can get the pseudo-abnormal bottleneck feature z_t in the feature space which is far away from the bottleneck feature z of normal data. Therefore, z_t is regarded as a bottleneck feature with abnormal data. The model adds abnormal data to the training set by getting z_t.

By minimizing the reconstruction error of sample x, the model enables encoder E_1 to obtain better bottleneck characteristics, and thus obtains better reconstructed samples by decoder D_1, which is expressed as follows:

$$\min_{\theta_E, \theta_D} ||x - \hat{x}||_2^2. \tag{2}$$

Where $\hat{x}$ is the reconstructed sample. θ_E, θ_D are the parameter sets of encoder and decoder.

In order to make The Transformer network obtains a feature space Z_t that is far away from the normal data feature space Z, and generates a pseudo-abnormal bottleneck feature with abnormal data information. The

Transformer network is trained by maximizing the error between the bottleneck feature z of the input sample and the bottleneck feature z_t transformed by the Transformer network. It is expressed as follows:

$$\max_{\theta_T} ||z - z_t||_2^2. \tag{3}$$

Where θ_T is the parameter set of Transformer network.

Furthermore, in order to enable the decoder to map the bottleneck feature with abnormal data information to normal data rather than itself as much as possible, the model causes the decoder D_2 to map the transformed bottleneck feature z_t to $\hat{x}_t$ by minimizing the error between $\hat{x}$ and $\hat{x}_t$, making b$\hat{x}_t$ as similar as possible to the normal data. It is expressed as follows:

$$\min_{\theta_E,\theta_D} ||\hat{x} - \hat{x}_t||_2^2. \tag{4}$$

In order to further improve the reconstruction error of abnormal data, the model enables encoder E_3 to obtain bottleneck feature $\hat{z}_t$ from $\hat{x}_t$ by minimizing the error between z_t and $\hat{z}_t$, so that $\hat{z}_t$ is as similar as possible to bottleneck feature z_t of normal data, which is expressed as follows:

$$\min_{\theta_E,\theta_D} ||\hat{z} - \hat{z}_t||_2^2. \tag{5}$$

Comprehensively considering equations (2)-(5), the model training objective of the method in this paper is to minimize the loss function:

$$\begin{aligned}\min_{\theta_E,\theta_D,\theta_T} \frac{1}{N}(&||x - \hat{x}||_2^2 \\ &+ \alpha||x - \hat{x}||_2^2 \\ &+ \beta||\hat{z} - \hat{z}_t||_2^2 \\ &+ \gamma||z - z_t||_2^2)\end{aligned} \tag{6}$$

Where N is the number of training samples. α, β, γ are the weights of each loss function. Encoders E_1, E_2, and E_3 use the same network structure and share parameters. Decoder D_1 and D_2 use the same network structure and share parameters.

Assuming that the given test sample is normal data, the encoder maps it to the bottleneck feature of the normal data, and then the decoder maps it back to the normal data, giving the normal data a small reconstruction error. Assuming that the given test sample is abnormal data, the encoder maps it to the bottleneck feature of the abnormal data, and the decoder decodes the bottleneck feature with abnormal data information into normal data as much as possible instead of reconstructing itself, so that the abnormal data can obtain a large reconstruction error. Therefore, the transformer method can use the reconstruction error of samples as the anomaly score to classify samples.

Since the encoder used in the training phase has the same structure and shared parameters, the decoder also has the same structure and shared parameters. Therefore, the test phase only needs to use any set of trained encoders and decoders to form a new model and classify the test samples. Given a test sample x_{test}, the reconstruction sample $\hat{x}_{test}$ of x_{test} is obtained by the new model, the reconstruction error of x_{test} is calculated, and the reconstruction error is used as the anomaly score $S(x_{test})$ to classify x_{test}, which is expressed as follows:

$$S(x_{test}) = ||x_{test} - \hat{x}_{test}||_2^2. \tag{7}$$

The training process based on Transformer network is shown in **Algorithm 1**.

Algorithm 1 Transformer training

1: **Input:** training set $X = x_i{}_{i=1}^{N}$.
2: Output: encoder $f_E(\cdot)$ and decoder $f_D(\cdot)$.
3: Initialize parameter sets θ_E, θ_D, θ_T of encoder $f_E(\cdot)$, decoder $f_D(\cdot)$ and Transformer network $f_T(\cdot)$.
4: for $i = 1$ to N do
5: Through equation (2), the training sample x_i is calculated and the bottleneck feature z is obtained after encoding.
6: Equation (3) is used to calculate the bottleneck feature z_t of z after Transformer network transformation.
7: The decoded samples $\hat{x}$ and $\hat{x}_t$ of z and z_t are obtained by equation (3).
8: The bottleneck features $\hat{z}$ and $\hat{z}_t$ of $\hat{x}$ and $\hat{x}_t$ after re-encoding are calculated by equation (2).
9: The bottleneck features $\hat{z}$ and $\hat{z}_t$ of $\hat{x}$ and $\hat{x}_t$ after re-encoding are calculated by equation (3).
10: The parameter sets θ_E, θ_D, θ_T are updated by equation (7) and stochastic gradient descent.
11: End

3.1 Graph-based Feature Learning

Each discourse in the data set is taken as a graph node, and the graph $G = (v, \varepsilon)$ is constructed, where $v(|v| = N)$ represents the discourse node. $\varepsilon \subset v \times v$ is an edge between nodes.

Two nodes can be connected by different edges, representing multiple relationships of the three modal features. In this paper, the weights of nodes u_i and edges between u_j are calculated according to the following circumstances.

- Consider the feature transfer of the same mode between two nodes. Since the same modal features of two nodes are in the same semantic space, the feature transfer can be carried out regardless of whether the nodes come from the same conversation. Weight reuse Angle similarity measurement of edges between two nodes.

$$a_{ij} = 1 - \frac{arccos(sim(x_i^{mod(0)}, x_j^{mod(0)}))}{\pi}. \tag{8}$$

Where $sim(\cdot)$ is the cosine similarity function. $x_i^{mod(0)}$, $x_j^{mod(0)}$ respectively represent the initial features of some same mode of the i and j discourse, $mod \subseteq a, t, v$.

- Considering the feature transfer of different modes between two nodes, two cases can be divided according to whether the two nodes come from a conversation:

  (a) If two nodes come from different conversations, the different modal features are not passed, in which case the weight below is 0. This is because although linear transformations are carried out in the initial feature extraction process of the three modes, the features of the different modes can be considered basically aligned in the semantic space. However, different dialogue scenes and dialogue content are very different, which enlarges the gap between different modes, so this paper thinks that the feature transfer should not be carried out in this case.

  (b) If two nodes come from the same dialogue, the different modal features are also relevant due to the consistent topic and content of the dialogue, and feature transfer is required. The weight of the edge between two nodes is also measured by angular similarity:

$$a_{ij} = 1 - \frac{arccos(sim(x_i^{mod'(0)}, x_j^{mod''(0)}))}{\pi}. \tag{9}$$

  Where $x_i^{mod'(0)}$, $x_j^{mod''(0)}$ represent the initial features of different modes of discourse i and j respectively, $mod', mod'' \subseteq a, t, v$, $mod' \neq mod''$.

  The adjacency matrix is constructed according to the weight calculation method of the edges between the nodes. For a certain modal feature of a node, three kinds of adjacency matrices can be constructed to transfer and learn the feature.

  Taking the feature learning of speech mode a of a node as an example, considering the relationship between speech mode a and its own speech mode a, text mode t and image mode v, three kinds of graph adjacency matrices can be constructed, and the feature matrix $X^{a(0)}$ can be updated.

  In addition, for the feature learning of the text mode t of nodes, three kinds of graph adjacency matrices A^{tt}, A^{ta} and A^{tv} are constructed. For the feature learning of image mode v of nodes, three kinds of graph adjacency matrices A^{vv}, A^{va} and A^{vt} are constructed.

This paper takes the feature learning of node speech mode a as an example to illustrate the feature learning process of different modes. The three graph adjacency matrices A^{aa}, A^{at} and A^{av} are convolution with the initial data feature $X^{a(0)}$ of the node by multi-layer GCN, and the updated three data features $A^{aa(l)}$, $A^{at(l)}$ and $A^{av(l)}$ are obtained by using four-layer GCN for encoding. The specific process is as follows.

For the graph $G = (v, \varepsilon)$, the Laplacian matrix formula for renormalization is as follows:

$$L = \tilde{D}^{-0.5}\tilde{A}\tilde{D}^{-0.5}. \tag{10}$$

$$L = (D + I)^{-0.5}((A^{mod'mod''}) + I)(D + I)^{-0.5}. \tag{11}$$

Where D represents the degree matrix. I stands for identity matrix. L stands for the renormalized Tullapras matrix of G. $mod' = a$, $mod" \subseteq a, t, v$. The iterative representation of graph convolutional networks with different layers is:

$$\begin{aligned} X^{mod(l+1)} =& \sigma(((1-\alpha)LX^{mod(l)} + \alpha X^{mod(0)} \\ & (1-\beta^l)I + \beta^{(l)}W^{(l)})). \end{aligned} \tag{12}$$

Where $X^{mod(0)} \in R^{N \times d_0}$ is the initial feature of a certain mode of the graph node, which is $X^{a(0)}$ for multi-mode feature learning. σ is the activation function. $W^{(l) \in R^{d_{l-1} \times d_l}}$ is a learnable weight matrix. In order to solve the problem of excessive smoothing and gradient disappearance caused by the introduction of multilayer graph convolution, the initial feature $X^{mod(0)}$ is added to the high level as residual, and I is added to the weight matrix $W^{(l)}$. α and $\beta^{(l)}$ are two hyperparameters. This article sets $\beta^{(l)} = \log(\eta/l + 1)$, where η is also a hyperparameter. Using DeepGCN with l layers, it can get $X^{mod(l+1)}$.

For feature learning of data modes, $X^{mod(0)} = X^{a(0)}$, there are three kinds of adjacency matrices for feature learning, then $X^{mod(l+1)}$ corresponds to three kinds of features $X^{aa(l+1)}$, $X^{at(l+1)}$, $X^{av(l+1)}$ obtained from the feature learning of three kinds of adjacency matrices. The three features are spliced together to obtain the data modal feature $X^{a(l+1)}$ after feature learning.

$$X^{a(l+1)} = X^{aa(l+1)} \oplus X^{at(l+1)} \oplus X^{av(l+1)}. \tag{13}$$

Here $\oplus$ is concatenation operation.

The vector in row i of the above eigenmatrix is the data feature $x_i^{a(l+1)}$ corresponding to the convolution of a node u_i graph. For the feature learning of the text modes and image modes of nodes, the above multi-level DeepGCN encoding is also used to obtain three convolution features of the text modes. After the features are splicing, $X^{t(l+1)}$ is obtained. Three convolution features of the image modes are obtained, and $X^{t(l+1)}$ is obtained after the features are spliced.

$$X^{v(l+1)} = X^{va(l+1)} \oplus X^{vv(l+1)} \oplus X^{vt(l+1)}. \tag{14}$$

$$X^{t(l+1)} = X^{ta(l+1)} \oplus X^{tv(l+1)} \oplus X^{tt(l+1)}. \tag{15}$$

Finally, the three modal features are combined to obtain the total feature matrix X^{l+1} after convolutional learning.

$$X^{(l+1)} = X^{a(l+1)} \oplus X^{t(l+1)} \oplus X^{v(l+1)}. \tag{16}$$

4 Performance Analysis

The scheme in this paper uses Python to implement the searchable encryption algorithm, in which the pseudo-random function is simulated by HMAC-MD5, and the MAC function is simulated by HMAC-SHA256. Ethereum's simulation experiments are conducted on the Ethereum VirtualMachine (EVM), where an Ethereum smart contract is built using the Solidity language. The computer hardware is configured as Inter Core i5 7300HQ2. 50GHz processor, 16GB RAM, 512GB SSD, and Ubuntu18.04LTS operating system [19].

4.1 Performance Analysis of Key Algorithms

The computational overhead of GenIndex algorithm in DO index generation phase includes the construction of encrypted index tree and lookup table. The computational overhead of the Search algorithm in the CSP retrieval stage includes the retrieval of the encrypted index tree and the positioning of the lookup table. The computational overhead of GenIndex and Search is simulated below. The data set used in the experiment contains 9810 documents in 20 categories, and 5000 documents are selected as the test data in this paper.

In order to observe the relationship between the computation cost of GenIndex performed by DO and Search performed by CSP and the number of documents, the number of documents was set to 500-5000(step size 500) in experiment 1. The experimental results are shown in Figure 1. As can be seen from Figure 1(a), the GenIndex time basically increases linearly as the number of documents increases. In particular, when the number of test documents is 500 and 5000, the GenIndex time is 221.656s and 2012.545s, respectively. It should be noted that this operation can be performed in the offline state, and the time cost is acceptable.

As shown in Figure 1(b), it can be seen that the Search time basically increases logarithmically with the increase in the number of documents. Under the current number of experimental documents, the Search time is all less than 1s, and the time cost is acceptable. In particular, when the number of test documents is 500 and 5000, the Search time is 7.6ms and 50.5ms, respectively.

4.2 Computation Cost Comparison

The calculation cost of index, search and verification is compared between the proposed scheme and relevant comparison schemes, and the results are shown in Table 2. Where E_0 and E_1 represent exponential operations on two different multiplicative cyclic groups, respectively. M_M stands for modular multiplication. H stands for hash operation. F stands for pseudo-random function operation. V represents MAC function operation. L represents the operation that builds each internal node. X represents the dot product between n-dimensional vectors. E and D indicate the encryption and decryption operations, respectively. M stands for modulo operation. SIG indicates the signature algorithm, which includes two processes: signature and verification. $\times$ indicates that it does not participate in this operation. g indicates the number of records of the conditional expression. I_A represents the subscript of the gate access policy. Y indicates the number of leaf nodes in the tree access policy. j indicates the number of returned ciphertext files. a represents the number of copies into which the document identifier that satisfies the conditional expression is divided. b represents the number of predicted steps by the data user. t is the number of keywords in W. n is the size of the keyword dictionary DW. Z is the number of records in the lookup table. θ is the number of documents containing the query keyword. In the real world, θ is much smaller than the document set base m.

In the stage of index generation, the scheme of reference [28] requires one pseudo-random function operation, one hash operation and one signature algorithm operation on the keyword dictionary and the document collection containing the keywords. The reference [12] requires three pseudo-random function operations, one encryption operation, and one hashing operation on the keyword dictionary and the document collection containing the keywords. The scheme of reference [33] requires one hashing operation on the ID of the document and the set of added tags, one encryption operation on the plaintext document set, and one pseudo-random function operation on the keyword dictionary. The data owner of the scheme in reference [25] needs to perform a time-consuming modular multiplication operations and $2(a+1)g$ pseudo-random function operations. Data owners of the references [6, 10] all need to carry out time-consuming exponential and modular multiplication operations. Meanwhile, the reference [6] scheme needs to carry out two pseudo-random function operations on each keyword, and the reference [10] scheme needs to carry out one pseudo-random function operation. In this paper, the balanced binary tree is used as the index, and the retrieval efficiency is high. The construction of internal nodes requires $m-1$ times, the encryption of each internal node requires one symmetric encryption operation for each bit of its vector, and each leaf node vector is encrypted by the secure K-nearest neighbor algorithm, and the multiplication of $n \times n$ matrix and n-dimensional vector needs to be performed twice. Constructing a lookup table that matches the encrypted index tree requires Z pseudo-random function operations and Z MAC function operations.

In the search stage, reference [33] conducts $O(Z)$ times of comparison between the published list and the abstract index, and then conducts m times of hash operation search. The scheme in reference [25] requires 6 time-consuming modular multiplication operations and 26 pseudo-random function operations. The references [6,10] generate a lookup table of key-value pairs as an index, with a search efficiency of $O(Z)$. References [10, 33] all

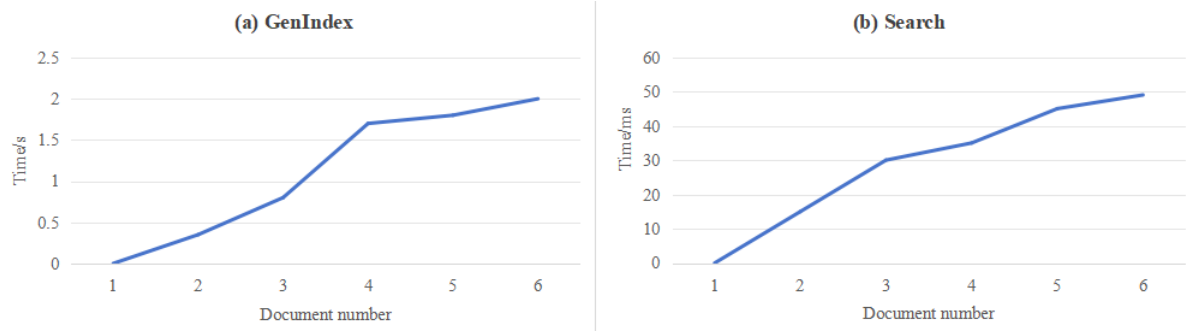


Figure 1: The relationship between the number of GenIndex and Search documents and the computational overhead

Table 1: Computing cost comparative analysis of searchable encryption schemes

Schemes	Index generation phase	Search phase	Verification stage
[28]	$n\theta(F+H+SIG)$	$m(H+SIG)$	$m(H+SIG)$
[12]	$n\theta(3F+H+E)$	$2D+4H+O(Z)$	$4H$
[33]	$2mH+mE+nF$	$mH+2O(Z)$	$(j+m)H$
[25]	$2(a+1)gF+aM_M$	$2bF+bM_M$	$\times$
[6]	$2E_0+E_1+2nF$	$O(Z)$	$\times$
[10]	$M_M+(2Y+1)E_0+E_1+nF$	$O(Z)$	$j!H$
Proposed	$(m-1)L+2nmX+n(m-1)E$ $+ZF+ZV$	$t\theta(lbm-1)D+2\theta X$ $+\theta M+O(Z)$	V

store encrypted indexes and perform search algorithms on the blockchain, resulting in high blockchain overhead, both occupying storage space and low efficiency. However, both the scheme of reference [12,28] and the scheme of this paper use CSP to search, and CSP stores ciphertext documents and encrypted indexes, which can reduce the blockchain overhead and improve the retrieval efficiency. In this scheme, the internal node is decrypted for $t\theta(lbm-1)$ times. In the worst case, the dot product operation between two n-dimensional vectors needs to be performed for 20 times and the modulo operation for θ times. The corresponding *proof* is located in the lookup table according to the *token* of DU, and $O(Z)$ comparison operations are required.

In the stage of verifying the correctness of the results, reference [28] removes the audit institution, but requires users to perform m hashing operations and m signature algorithm operations locally. As the number of documents increases, the verification time increases. The scheme in reference [12] only needs to carry out 4 hashing operations, but the process needs to involve 6 transactions on the Bitcoin script, so the verification time is long and the efficiency is low. The scheme of reference [33] requires each arbitration node on the arbitrator fragment in the distributed storage platform to perform j hashing operations on the set of added tags obtained by the client, and then re-perform the search operation to perform m hashing operations to verify the search results. The proposal in reference [10] and the proposal in this paper use smart contracts to verify search results, and the proposal in reference [10] requires $j!$ For secondary hashing operation, the scheme in this paper only needs to perform one MAC function operation to determine whether the value after MAC function operation is correct and complete the verification.

In summary, compared with other schemes with verification function, the proposed scheme has the lowest computational cost in the verification phase. This is because this scheme uses CSP to store ciphertext documents, encrypt indexes, and perform search algorithms, while smart contracts only need to perform verification and fair payment operations, which makes its computation cost low. However, this also leads to the increase of CSP storage overhead and computing overhead. However, from the analysis of computing overhead in Search phase and the experimental results of search algorithm, it can be seen that the computing overhead in search phase of CSP is acceptable.

In order to explore the relationship between the number of different documents and the retrieval time, we conducted an experiment. The results are shown in Table **??**. It can be seen that the retrieval time of reference [12] basically increases linearly with the increase of the number of documents. The retrieval time of reference [25] is slightly lower than that of reference [33]. However, the retrieval time of this proposed scheme is the fastest due to other schemes. In this scheme, the retrieval time is almost constant when the number of documents increases.

Table 2: Retrieval time comparison of different schemes/s

Scheme	1000	2000	3000	4000	5000	6000
Reference [28]	0.19	0.38	0.57	0.78	1.12	1.28
Reference [12]	0.15	0.22	0.26	0.35	0.42	0.48
Reference [33]	0.12	0.18	0.22	0.29	0.35	0.39
Reference [25]	0.11	0.16	0.21	0.26	0.31	0.36
Reference [6]	0.10	0.10	0.10	0.10	0.10	0.10
Reference [10]	0.03	0.03	0.03	0.03	0.03	0.03
Proposed	0.02	0.02	0.02	0.02	0.02	0.02

5 Conclusion

This paper proposes a new multi-keyword ciphertext sorting search based on conformation graph convolution model and Transformer network, which supports verification and fair payment, and realizes the verification of search results, the fair payment between three parties and the multi-keyword sorting retrieval of ciphertext. In order to realize the verifiability and fair payment of the search results, and reduce the time and cost, the scheme is designed by the cloud server to store the encrypted index tree and lookup table, and perform the search operation. The verification and fair payment of the search results are completed by the Ethereum smart contract, which effectively reduces the complexity of the smart contract execution operation, reduces the time and expense, and improves the verification efficiency. In addition, the scheme uses balanced binary tree as the index, which ensures the retrieval efficiency and realizes the functions of multi-keyword retrieval, ranking of search results and dynamic update, which improves the flexibility and user friendliness of the scheme. Finally, the safety and performance of the scheme are analyzed, and the simulation experiment is carried out. Performance analysis and experimental results show that the proposed scheme is feasible and practical. The results of functional comparison show that compared with the existing blockchain-based searchable encryption schemes, the proposed scheme is more comprehensive in terms of functions. In addition, the verification process of the scheme in this paper is carried out for all the search results, and there is room for further optimization. Future research on verification strategies for specific transactions to better meet user needs while reducing time and expense costs.

Acknowledgments

The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] A. Ali, M. Pasha, J. Ali, O. Fang, M. Masud, A. Jurcut, "Deep learning based homomorphic secure search-able encryption for keyword search in blockchain healthcare system: A novel approach to cryptography," *Sensors*, vol. 22, no. 2, pp. 528, 2022.

[2] M. Ali, MR. Sadeghi, X. Liu, Y. Miao, "Verifiable online/offline multi-keyword search for cloud-assisted industrial internet of things," *Journal of Information Security and Applications*, vol. 65, 2022.

[3] J. Bi, S. Yin, H. Li, L. Teng, C. Zhao, "Research on medical image encryption method based on improved Krill Herb algorithm and chaotic systems," *International Journal of Network Security*, vol. 22, no. 3, pp. 486-491, 2020.

[4] B. Deebak, F. Memon, K. Dev, S. Khowaja, *et al.*, "AI-enabled privacy-preservation phrase with multi-keyword ranked searching for sustainable edge-cloud networks in the era of industrial IoT," *Ad Hoc Networks*, vol. 125, 2022.

[5] C. K. Dehury and P. K. Sahoo, "Failure aware semi-centralized virtual network embedding in cloud computing fat-tree data center networks," *IEEE Transactions on Cloud Computing*, vol. 10, no. 2, pp. 1156-1172, 2022.

[6] T. Feng, S. Miao, C. Liu, R. Ma, "Verifiable keyword search encryption scheme that supports revocation of attributes," *Symmetry*, vol. 15, no. 4, 2023.

[7] H. Gao, H. Huang, L. Xue, F. Xiao and Q. Li, "Blockchain-enabled fine-grained searchable encryption with cloud-edge computing for electronic health records sharing," *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 18414-18425, 2023.

[8] N. Ghosh, S. K. Ghosh and S. K. Das, "SelCSP: A framework to facilitate selection of cloud service providers," *IEEE Transactions on Cloud Computing*, vol. 3, no. 1, pp. 66-79, 2015.

[9] H. Guo, X. Yu, "A survey on blockchain technology and its security," *Blockchain: research and applications*, vol. 3, no. 2, 2022.

[10] K. Guo, Y. Han, R. Wu, K. Liu, "CD-ABSE: Attribute-based searchable encryption scheme supporting cross-domain sharing on blockchain,"

Wireless Communications and Mobile Computing, vol. 2022, 2022.

[11] Y. Guo, C. Zhang, C. Wang and X. Jia, "Towards public verifiable and forward-privacy encrypted search by using blockchain," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 3, pp. 2111-2126, 2023.

[12] Y. Jiang, X. Xu and F. Xiao, "Attribute-based encryption with blockchain protection scheme for electronic health records," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 3884-3895, Dec. 2022.

[13] Y. Li, F. Zhou, D. Ji, Z. Xu, "A hierarchical searchable encryption scheme using blockchain-based indexing," *Electronics*, vol. 11, no. 22, pp. 3832, 2022.

[14] M. Mihailescu, S. Nita, "Searchable encryption," in *Pro Cryptography and Cryptanalysis with C++20. Apress, Berkeley, CA*, 2021. https://doi.org/10.1007/978-1-4842-6586-4_11

[15] S. Niu, M. Song, L. Fang, F. Yu, S. Han, C. Wang, "Keyword search over encrypted cloud data based on blockchain in smart medical applications," *Computer Communications*, vol. 192, pp. 33-47, 2022.

[16] M. S. Rahman, M. Chamikara, I. Khalil, "Blockchain-of-blockchains: An interoperable blockchain platform for ensuring IoT data integrity in smart city," *Journal of Industrial Information Integration*, vol. 30, 2022.

[17] M. Ramachandran, V. Chang, "Towards performance evaluation of cloud service providers for cloud data security," *International Journal of Information Management*, vol. 36, no. 4, pp. 618-625, 2016.

[18] S. Ramzan, A. Aqdus, V. Ravi, D. Koundal, R. Amin and M. A. Al Ghamdi, "Healthcare applications using blockchain technology: Motivations and challenges," *IEEE Transactions on Engineering Management*, vol. 70, no. 8, pp. 2874-2890, 2023.

[19] Y. S. Rao, S. Prasad, S. Bera, A. K. Das and W. Susilo, "Boolean searchable attribute-based signcryption with search results self-verifiability mechanism for data storage and retrieval in clouds," *IEEE Transactions on Services Computing*, doi: 10.1109/TSC.2023.3327816.

[20] S. Sarkodie, M. Ahmed, T. Leirvik, "Trade volume affects bitcoin energy consumption and carbon footprint," *Finance Research Letters*, vol. 48, 2022.

[21] J. Su, L. Zhang, Y. Mu, "BA-RMKABSE: Blockchain-aided ranked multi-keyword attribute-based searchable encryption with hiding policy for smart health system," *Future Generation Computer Systems*, vol. 132, pp. 299-309, 2022.

[22] Q. Tong, Y. Miao, J. Weng, X. Liu, K. -K. R. Choo and R. H. Deng, "Verifiable fuzzy multi-keyword search over encrypted data with adaptive security," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5386-5399, 2023.

[23] X. Wang, S. Yin, H. Li, L. Teng, S. Karim, "A modified homomorphic encryption method for multiple keywords retrieval," *International Journal of Network Security*, vol. 22, no. 6, pp. 905-910, 2020.

[24] X. Wang, S. Yin, M. Shafiq, A. A. Laghari, S. Karim, O. Cheikhrouhou, W. Alhakami, H. Hamam, "A new v-net convolutional neural network based on four-dimensional hyperchaotic system for medical image encryption," *Security and Communication Networks*, vol. 2022, 2022. https://doi.org/10.1155/2022/4260804

[25] N. Wu, L. Xu, L. Zhu, "A blockchain based access control scheme with hidden policy and attribute," *Future Generation Computer Systems*, vol. 141, pp. 186-196, 2023.

[26] C. Xu, P. Zhang, L. Mei, Y. Zhao, L. Xu, "Ranked searchable encryption based on differential privacy and blockchain," *Wireless Networks*, pp. 1-14, 2022.

[27] P. Xu, J. Chen, Y. Yang and J. Ning, "DuMSE: Toward practical and dynamic multiuser search over encrypted cloud data against keyword guessing attack," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1082-1095, 2023.

[28] Z. Xu, S. Zhang, H. Han, X. Dong, Z. Zheng, H. Wang, W. Tian, "Blockchain-aided searchable encryption-based two-way attribute access control research," *Security and Communication Networks*, vol. 2022, 2022.

[29] L. Yan, L. Ge, Z. Wang, G. Zhang, J. Xu, Z. Hu, "Access control scheme based on blockchain and attribute-based searchable encryption in cloud environment," *Journal of Cloud Computing*, vol. 12, no. 1, pp. 1-16, 2023.

[30] S. Yin, H. Li, L. Teng, A. Laghari, V. V. Estrela, "Attribute-based multiparty searchable encryption model for privacy protection of text data," *Multimedia Tools and Applications*, 2023. https://doi.org/10.1007/s11042-023-16818-4

[31] S. Yin, J. Liu, L. Teng, "A sequential cipher algorithm based on feedback discrete hopfield neural network and logistic chaotic sequence," *International Journal of Network Security*, vol. 22, no. 5, pp. 869-873, 2020.

[32] H. Zeng, H. Zamani, V. Vinay, "Curriculum learning for dense retrieval distillation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1979-1983, 2022.

[33] F. Zhang, Y. Zhang, G. Han, "Blockchain-based attribute-based keyword searchable encryption for health cloud system," *International Journal of Embedded Systems*, vol. 15, no. 6, pp. 493-504, 2022.

Biography

Hang Li biography. He obtained his Ph.D. degree in Information Science and Engineering from Northeastern University. Hang Li is a full professor of the software college at Shenyang Normal University. His interests are wireless networks, mobile computing, cloud computing,

social networks, network security and quantum cryptography. Prof. Li had published more than 30 international journal and international conference papers on the above research fields. Email:lihangsoft@163.com.

Zeyang Li biography. Zeyang Li is with the Software College, Shenyang Normal University. His major is computer science, information secure.

Xiaowei Wang biography. She is a full professor of the software college at Shenyang Normal University. Her interests are wireless networks, mobile computing, cloud computing, social networks, network security and quantum cryptography. Prof. Wang had published more than 10 international journal and international conference papers on the above research fields. Email:hsiaoweiw@163.com

Muhammad Ibrar biography. Muhammad Ibrar is with the Software College, Shenyang Normal University. His major is computer science, information secure.

Xinjie Zhu is with the Zhengzhou University of Science and Technology. Several papers had been puiblished related to the major. Research interest is: education data analysis, information processing.

English Data Encryption Based on U-Net Network and Attention Mechanism in Cloud Computing Environment

Ruoshuang Yin

(Corresponding author: Ruoshuang Yin)

School of Foreign Languages, Zhengzhou University of Science and Technology
Zhengzhou City 450064, China
Email: 2430175060@qq.com

(Received Nov. 29, 2023; Revised and Accepted Feb. 7, 2024; First Online June 22, 2024)

The Special Issue on Data Fusion, Deep Learning and Optimization Algorithms for Data Privacy Protection

Special Editor: Prof. Shoulin Yin (Harbin Institute of Technology)

Abstract

In the era of network information, the security of cloud storage information is related to the privacy protection of users and the security of the entire network information system. The key to the security management of cloud storage data is data encryption. Through the design of structured data encryption under the cloud storage platform, information security can be ensured. Therefore, this paper researches the English data encryption based on U-Net network and attention mechanism in cloud computing environment. By processing English unstructured big data in blocks, the input and output tensors are obtained, and the network structure based on U-Net is constructed to generate English big data sequence. The spatial attention mechanism is integrated into the jump junction of U-Net structure to extract the correlation information of spatial features. Secondly, channel attention mechanisms are incorporated simultaneously at the jump junction to effectively express the dependency of useful channels and suppress features that are not relevant to English data processing tasks. The simulation results show that this method can improve the anti-attack ability of cloud storage platform data, and the data steganography performance is good, and the data encryption and storage performance is good.

Keywords: Attention Mechanism; Cloud Computing Environment; Data Encryption; U-Net Network

1 Introduction

With the rapid development of cloud computing and Internet of Things technology, the world has entered the era of big data. Large scale is only one feature of big data [1], which also has the characteristics of variety and high value. Unstructured data is a manifestation of the diversity of big data [2, 3]. 80% of text, image, video, audio and other information is unstructured data.

With the acceleration of informatization construction in colleges and universities and the wide application of digital equipment, it has become a trend to establish digital colleges and universities [4,5] to guide the development of colleges and universities. Big data information in the English field covers large-scale unstructured data, which is not only an important basis for students' exams, but also involves students' privacy. Unstructured big data in the field of English is growing rapidly by explosion, and needs to be stored safely and accurately within a specified period, which makes data security storage face great challenges [6].

In traditional methods, cloud storage data encryption methods mainly include elliptic encryption method, ECC encryption method, and segmented NTRU public key encryption method [7]. Through the design of encryption key and code element combination, the method of encoding and key conversion is adopted to realize the encryption and decryption of data. According to the above principles, data encryption algorithms are studied in related references. Among them, a cloud storage platform encryption technology based on Elliptic Curve Cryptography (ECC) was proposed in reference [8]. The elliptic curve equation of cloud storage platform encryption was constructed in a limited domain, the encrypted data was linearly segmented coded, and the encryption and decryption key was designed to encrypt the cloud storage platform data, so as to improve the anti-attack capability of the encrypted data. This method had the problems of excessive computing overhead and poor real-time encryption performance in data encryption. Reference [9] proposed a data quantization encoding algorithm for cloud storage platform based on singular value finite domain filling en-

cryption. The public key was embedded in the limited domain of the main key distribution and the sensitive domain parameters of the key center were filled to realize big data encryption for cloud storage platform. This method required a large amount of prior data as a test set, resulting in poor anti-attack capability of data encryption. In reference [10], an adaptive linear encryption method of cloud storage platform based on multiple retransmission mechanism was proposed. MIMO hybrid pre-encoding method was used to build a chaotic encryption system to realize segmented data encryption and improve the real-time encryption. This method also had some problems, such as poor anti-interference and high sensitivity.

To solve the above problems, this paper proposes an English data encryption method based on U-Net network and attention mechanism in cloud computing environment. Firstly, the data structure is analyzed under the cloud storage platform, the chaotic sensitivity characterization and chaotic coding design of the data are carried out, the data encryption key is constructed, and the encryption algorithm is optimized based on the information entropy characteristics of plaintext attacks. Then the encryption coding protocol is designed, and finally the data encryption algorithm is designed, which shows the superiority of this method in improving the performance of data encryption.

This paper is organized as followed. In Section 2, we introduce the unstructured reorganization of big data in cloud environment. Section 3 introduces the U-net structure for data feature extraction. Experiments are conducted in Section 4. There is a conclusion in Section 5.

2 Unstructured Reorganization of Big Data in Cloud Environment

Big data is a multi-technology concept that simply refers to a collection of data whose content cannot be captured, managed, and processed with conventional software tools within a certain amount of time. IBM defines the concept of "big data" as the, Volume, Variety, Velocity and Value.

Big data will bring huge technical and business opportunities, and big data analysis mining and utilization will bring huge business value to enterprises. With the rapid increase in the scale of application data, traditional computing is facing serious challenges, large-scale data processing and industry application needs are increasing and more and more large-scale data processing application needs are urgently emerging. Traditional system is difficult to provide enough storage and computing resources for processing, cloud computing technology is the most ideal solution.

In order to realize the English data encryption design, the chaotic sensitivity feature characterization method is first used to carry out the unstructured recombination design of the big data of the cloud storage platform. The elliptic function for building cloud storage platform encryption in a limited domain is $Decrypt(sk, c')A^{-1} = T = (t_{i,j})_{i,j=1}^{m}$. The cyclic key pairing method is used for the structural feature decomposition of data encryption [11, 12], and the feature decomposition formula of data encryption key is as follows:

$$T \cdot A = \begin{bmatrix} t_{1,1} & \cdots & t_{1,m} \\ \vdots & \ddots & \vdots \\ t_{m,1} & \cdots & t_{m,m} \end{bmatrix} \begin{bmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,m} \end{bmatrix} \quad (1)$$

Random interpolation method is adopted for key expansion of un-encrypted data in the cloud storage platform, and information entropy and variance are taken as homomorphic encoding variables of encryption [13]. The unstructured reorganization of big data in the cloud storage platform is carried out to obtain the expansion key of big data in the cloud storage platform:

$$T^{(\alpha_1^{-1},\cdots,\alpha_m^{-1})^T} A^{\alpha_1,\cdots,\alpha_m} = E. \quad (2)$$

Define the length of the cloud data to be encrypted as n, subject to the standard normal distribution function. Key extension sequence is $X = x_1, x_2, \cdots, x_n$. The binomial and $S_n = x_1 + x_2 + \cdots + x_n$ of the encrypted bit sequence X are counted, and the elliptic curve and chaotic feature mapping methods [14, 15] are adopted to obtain the feature recombination output of the encrypted data as follows:

$$Decrypt(sk, c')AA^{-1} = E. \quad (3)$$

The uniform distribution value $e = h(m)$ of encode element frequency is calculated, and the binary vector quantization characteristic equation of big data encoding under cloud storage platform is obtained by using single-bit detection method as follows:

$$Decrypt(sk, c')(A^{\alpha_1,\cdots,\alpha_m})^{-1} = E. \quad (4)$$

3 U-net Structure for Data Feature Extraction

To solve the problem of different size and structure in data, a new MADoubleU-Net network is proposed in this paper. The main structure of MADoubleU-Net network includes double U-shaped network framework, multiple attention module (MAB), atrous spatial pyramid pooling (ASPP), and U-Net as the reference network, a VGG-19 encoder pre-trained on ImageNet [16] as shown in Figure 1.

The top half is the first U-shaped network with four encoder blocks and four decoder blocks, the encoder using ImageNet pre-trained VGG-19, which as a lightweight model can be easily connected to the U-Net structure. The 4 decoder blocks used 2×2 bilinear interpolation to upsample the input feature maps to restore the original feature map size. The second U-shaped network in the bottom half is basically similar to the first, also consisting of four encoder blocks and four decoder blocks, each

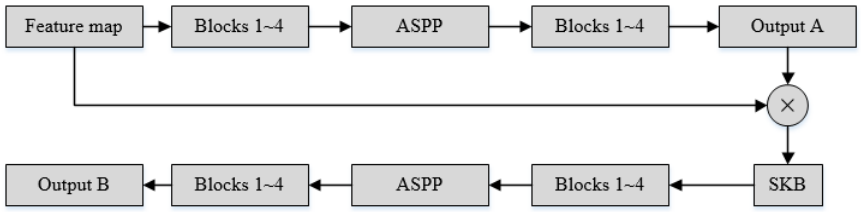


Figure 1: New network model structure

codec block performing a 3×3 convolution operation and a batch normalization (BN) operation. The Rectified Linear Unit (ReLU) activation function is used to enhance the nonlinearity, and the three operations are repeated once. In addition, the decoder performs an additional 2×2 pooling layer operation to reduce computational complexity.

The output of the first U-shaped network is formed by the Sigmoid function after a 1×1 convolution operation. The input of the second U-decoder consists of the jump connection of the first U-encoder, the output of the second U-encoder, and the jump connection through the multiple attention modules, which maintains high spatial resolution while improving the quality of the output feature map. The final Output result, Output B, synthesizes the output of the current network decoder and Output A, and uses ASPP to capture data feature information of different sizes between the encoder and decoder of two U-shaped networks. The experimental results show that this method can effectively solve the problem of different data sizes in data feature extraction.

Compared with the original U-Net network, the U-Net network in this paper has the following improvements in network structure:

1) Introduction of spatial attention module and channel attention module. In order to extract polyp features at a deeper level, MADoubleU-Net introduces a spatial attention module and a channel attention module at the jump junction of the encoder and decoder blocks with the same resolution between the up-sampling and down-sampling of two U-shaped networks. The structure uses spatial context information to obtain the correlation of spatial features, and assigns different weights to channels to improve the sensitivity of important channels and reduce the influence of unimportant background features.

2) Introduction of SKB. The input feature map of the second U-encoder is the result of SKB processing after multiplying the output Output A of the first U-encoder with the original input feature map. For convolution kernels of different sizes, the module can adaptively select appropriate receptive fields for different channels, so as to strengthen the feature characterization ability of objects of different sizes and improve the segmentation accuracy. The experiment proves that the improved multi-attention module can effectively solve the problem of similar data.

A. Multiple attention module

Based on the original DoubleU-Net network, this paper integrates multiple attention modules, including spatial attention module SAB channel attention module CAB and multi-scale selective core channel attention module (SKB). SAB can explore the correlation of spatial dimensions by connecting spatial context feature information. CAB deduces the importance of the whole channel dimension by calculating the dependency of channel dimension. SKB selects information of different receptor fields suitable for different channels for convolution cores of different sizes, so that the network can suppress the influence of unimportant background information of colon polyp images on segmentation, so that polyp feature information can play a greater role and improve segmentation accuracy.

Since local features in the data are correlated, in order to obtain the spatial context information of the data and obtain the corresponding spatial feature attention graph with spatial correlation information, this paper uses the spatial attention module SAB to avoid getting too much weight due to locally unimportant features and affecting the segmentation results. SAB mainly includes three parts: spatial feature description, correlation calculation and feature recovery, and its structure diagram is shown in Figure 2. In the figure, C is the number of channels of the feature graph, H is the height of the feature graph, and W is the width of the feature graph.

Different from the average pooling method of traditional SE module, this paper uses both maximum pooling and average pooling to describe features. The combination of the two makes the extracted spatial features more abundant and accurate. Spatial feature description S_A is calculated as follows:

$$S_A = Con(F_C^{AvgP}; F_C^{MaxP}). \tag{5}$$

In the formula, F_C^{AvgP} and F_C^{MaxP} represent the feature maps after average pooling and maximum pooling respectively, F_C^{AvgP} and $F_C^{MaxP} \in R^{1\times H\times W}$.

Correlation calculation is to obtain the final spatial correlation feature graph S_B by using Sigmoid function after 7×7 convolution operation on the connected spatial feature description S_A. The calculation formula is as

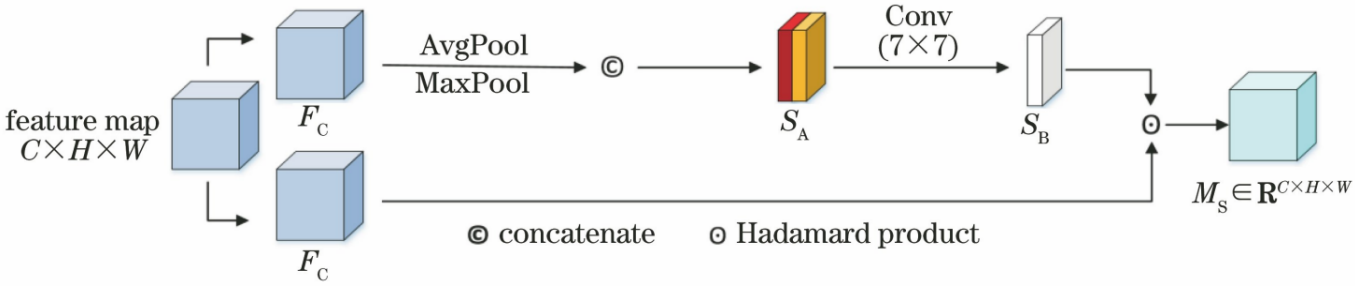


Figure 2: SAB diagram

follows:

$$S_B = \sigma[Conv(S_A)]. \tag{6}$$

Where σ represents Sigmoid activation function, $S_B \in R^{1\times H\times W}$.

Finally, multiply the elements of the original feature graph F_C and the spatial correlation feature graph S_B to obtain the final spatial attention feature graph M_S:

$$M_S = S_B \odot F_C. \tag{7}$$

Where $\odot$ stands for Hadamard product, $M_S \in R^{C\times H\times W}$.

In general, the convolution operation is to compute the fusion of all channels by default, and the importance and relevance of different channels to feature extraction are different. In this paper, the dependency between different channels is used to improve the channel semantic information in feature extraction, so as to enhance the extraction of important information and suppress the non-important information. The channel attention module mainly includes three parts: channel feature description, correlation calculation and feature recovery, and its structure is shown in Figure 3.

Similar to the spatial attention module, the channel attention module in this paper uses maximum pooling and average pooling to describe the channel dimension of the feature graph. In the following correlation calculation, the obtained two feature description graphs are fed into a fully connected network with shared weights. This fully connected network includes a full-connected layer dimension reduction operation, a ReLU activation function (to reduce computational complexity), and a dimension increase operation that reverts to the original size. The elements of the two output channel attention maps are added together and the final channel correlation matrix is obtained using the Sigmoid activation function. The calculation formula of C_A is:

$$C_A = \sigma[\sigma_{ReLU}(F_C^{AvgP}) + \sigma_{ReLU}(F_C^{MaxP})]. \tag{8}$$

Where σ_{ReLU} represents the ReLU activation function used by the fully connected layer, $C_A \in R^{C\times 1\times 1}$.

Finally, the elements of the original feature graph F_C and the channel correlation feature graph C_A are multiplied to obtain the final channel attention feature graph Mc, whose calculation formula is as follows:

$$M_C = C_B \odot F_C. \tag{9}$$

Where $\odot$ stands for Hadamard product, $M_C \in R^{C\times H\times W}$.

B. Multi-scale selection channel attention module

The channel attention mentioned above uses a single convolution kernel, which cannot realize convolution of convolution nuclei of different sizes, and has no ability to select the receptive field suitable for different channels. Since human eyes can adjust the size of the receiving domain adaptively in real life, SKB is introduced in MADoubleUNet in this paper, so that the network can adjust the size of the receptive field adaptively according to the input information of different sizes, so as to strengthen the feature extraction ability of channel dimensions. The SKB adopted in this paper includes three parts: multi-scale calculation, correlation calculation and feature selection, and its structure is shown in Figure 4.

Firstly, multi-scale calculation is carried out, using 3×3 and 5×5 convolution check respectively to calculate the input feature maps, and the feature maps M_A and M_B are obtained. The feature maps M are generated by connecting M_A and M_B. Multi-dimension convolution kernel combines multi-scale information while enlarging the receptive field, which is conducive to assigning different weights to different channels in the future. The calculation formula of the feature graph M is:

$$M = M_A + M_B. \tag{10}$$

The feature graph after multi-scale processing has the same size as the original feature graph M. Next, the feature graph M' is operated similarly to the SE module, that is, average global pooling, and then dimension reduction is carried out through a fully connected layer. After batch standardization, activation function ReLU is used to reduce the computational complexity.

$$x = F^{AvgP}(M) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} M(i,j). \tag{11}$$

$$M' = \sigma[B(\varpi x)]. \tag{12}$$

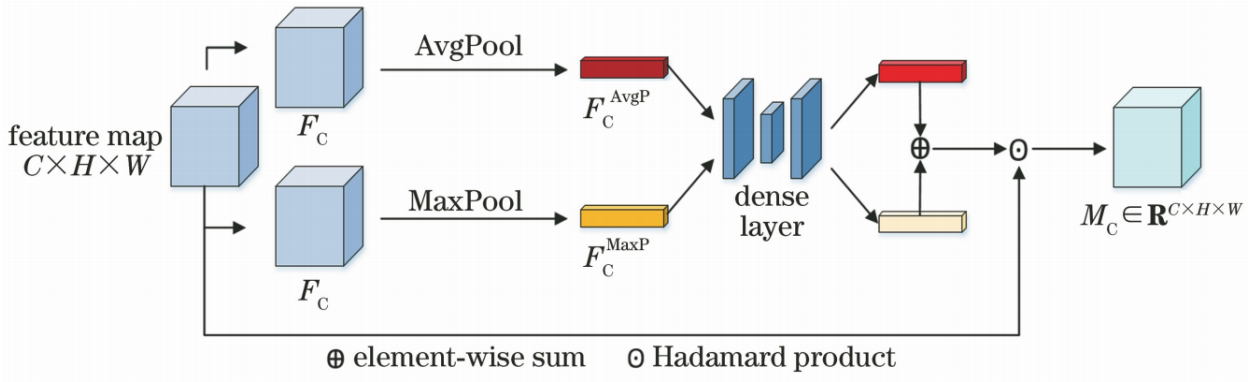


Figure 3: CAB schematic diagram

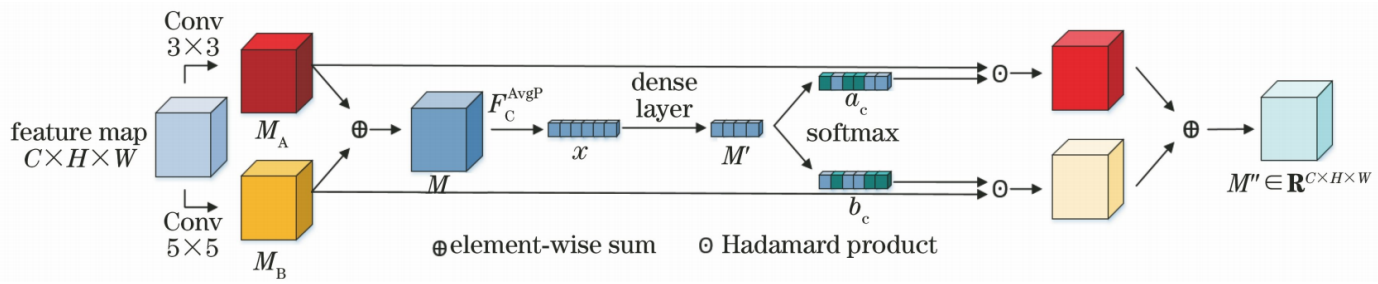


Figure 4: SKB schematic diagram

$$D = max(\frac{c}{r}, L). \tag{13}$$

Where x represents the feature map after average pooling. c is the number of channels in the feature map. D is the number of features after dimensionality reduction F^{AvgP} stands for average pooling operation. B means batch standardization. $\varpi \in R^{D\times C}$. L is set the value as 32.

Feature graph x indicates that the channel can adaptively select feature information of different scales. For the feature graph M' after dimensionality reduction, soft attention is calculated to select different spatial scales, and Softmax function is used according to channel direction. By multiplying the output attention of two channels with the feature graph after multi-scale processing, the feature graph M" with channel correlation after different scale selection can be expressed as:

$$M'' = a_c \times M_A + b_c \times M_B. \tag{14}$$

Where $M'' \in R^{C\times H\times W}$. The subscript c represents the number of rows in the matrix. a_c and b_c represent the soft attention of feature graphs of two different convolution kernels, respectively. $a_c = \frac{e^{A_c x}}{e^{A_c x}+e^{B_c x}}$. $b_c = \frac{e^{B_c x}}{e^{A_c x}+e^{B_c x}}$.

In MADoubleU-Net, SKB is placed in the input position of the second encoder, because the feature information is obtained in the calculation of the first U-structure, but some useful features are lost. Therefore, the combination of SKB and the original image effectively aggregates more information, making the channel information more rich. At the same time, the application of SKB can give higher weights to important channels and select appropriate receptive fields.

C. Unstructured big data dynamic secure storage

The distribution of F_2 in the binary domain of storage has regularity, which can be explored according to the unstructured big data scheduling model and the similarity of storage [17, 18]. And a multi-rule scheduling model is established based on the differences of users, which can be expressed as follows:

$$f_{i,j} = m_t\delta_t + m_a\delta_a + m_q\delta_q + m_s\delta_s. \tag{15}$$

In the formula, the encrypted stored data packet is δ. The storage quality is q. The data collection time is t. The loss incurred by meeting user demand is a. The consumption of stored data is s. The total number of nodes in unstructured big data is m. Each parameter must meet the following condition:

$$m_t + m_a + m_q + m_s = 1. \tag{16}$$

In the formula, s and m need to meet the following condition:

$$\sum_{s=0,1,\cdots,m} (m, s)^T = 2^{m-1}. \tag{17}$$

k represents the number of unstructured big data sources, and $k+e\varepsilon$ represents the number of nodes. Since

the number of nodes equals the number of packets stored, $k+e\varepsilon$ packets are set to $Y_i, i = 1, 2, \cdots, k+\varepsilon$, where $\varepsilon > 0$ and it is a constant. The linear mode of the encrypted source packet can be presented by the encrypted stored packet [19], so Y_i can be expressed as:

$$Y_i = g_i[X_1, X_2, \cdots, X_k]. \tag{18}$$

In the formula, the encrypted source packet is represented as $X_1, X_2, \cdots, X_k$, $i = 1, 2, \cdots, k$. The row vector is g_i. The binary field $F_2 = 0, 1$ is the interval to which the g_i value belongs. Let $g_{i,j}$ be the element of g_i, whose value is independent and can be described as:

$$P_r(g_{ij} = r) = \left\{ \begin{array}{l} (a \ln k)/k \ r = 1 \\ 1 - (a \ln k)/k \ r = 0 \end{array} \right\} \tag{19}$$

Where, the distribution of elements is expressed as $P_r(g_{ij} = r)$, $j = 1, 2, \cdots, k$. The unstructured big data node probability is $(a \ln k)/k$. The distance between nodes is r. The matrix is established according to $k + \varepsilon$ data packets, expressed as $G_{(k+\varepsilon)\times k}$, and the matrix is of order $(k + \varepsilon) \times k$, which can be described as:

$$G_{(k+\varepsilon)\times k} = [g_1, g_2, \cdots, g_{k+\varepsilon}]^T. \tag{20}$$

Through formula (18) and (19), the matrix equation can be obtained as follows:

$$labeleq21 \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{k+\varepsilon} \end{bmatrix} = G_{(k+\varepsilon)\times k} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \tag{21}$$

4 Experiment and Analysis

Multiple database child nodes for storage constitute a distributed storage unit. Each database uses flash memory to cache memory and disk, which improves data reading efficiency, reduces disk write times, minimizes data storage consumption, and greatly improves data storage efficiency. According to the relationship between the number of encrypted data packets and the number of nodes, the encrypted data packets are replaced by the linear combination of encrypted source data packets to realize the secure storage of encrypted medical unstructured big data. Encrypted storage Child nodes The controller allocates encrypted storage tasks to child nodes to balance the storage load among child nodes. The local computing controller is used to complete the data retrieval of the child nodes, and the distributed encrypted storage master node is used to implement comprehensive control over the child node controller of unstructured big data, and is responsible for the transfer of data processed by faulty or overloaded nodes, so as to realize the management and monitoring of each encrypted storage node. Each local computing controller is controlled by the parallel computing master node. The task of the parallel computing master node is to assign computing work to each computing controller and retrieve the results.

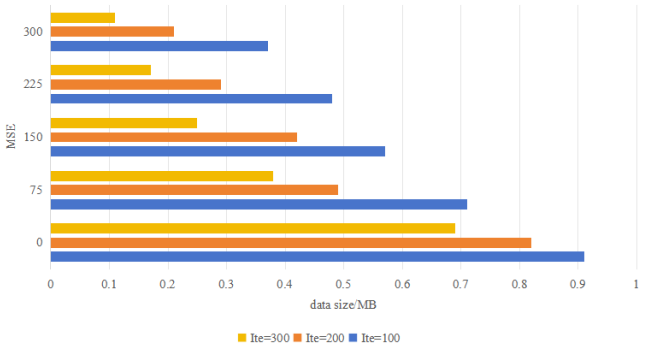


Figure 5: Training effect analysis

In order to verify the effectiveness of the above designed method of dynamic storage of unstructured data based on recurrent neural network, the following experiments are designed. The English data of a college English major is taken as the experimental object, and a variety of unstructured big data generated by the major is taken as the experimental sample and stored in encryption. Taking a 500MB data set as an example, the proposed method is used to train the big data in the data set and obtain the big data sequence. Under different iterations, the difference between the output result of the actual big data sequence and the target result is reflected by the MSE mean square error index, which verifies the convergence performance of the proposed method. The experimental results are shown in Figure 5.

According to the analysis of Figure 5, with the continuous increase of data volume, the mean square error values generated by big data series show a downward trend under different iterations, and the mean square error values are the largest after 100 times of network training. The mean square error value after 300 network training is the smallest, indicating that the big data trained by this method is trained and learned. The larger the number of iterations, the smaller the difference between the obtained actual big data sequence and the target big data, and the closer to the real data.

Security is one of the most effective indicators to verify the effect of data encryption. The value distribution of American Standard Code for Information Interchange (ASCII) is selected as the effect of the proposed method for encrypting big data, and the experimental results are shown in Figure 6.

According to the analysis of Figure 6, after encrypting English big data with this method, it can effectively scramble the arrangement of big data, cover up the effective information contained in the original big data, effectively resist external attacks, and improve data security. Experimental results show that the encryption effect of this method is good.

A data set with data size of 800MB is selected from the English big data to simulate four kinds of attacks: tampering attack, denial of service attack, forgery attack and

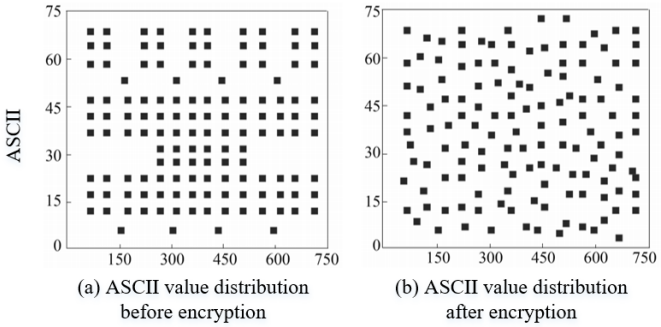


Figure 6: ASCII distribution of big data before and after encryption

injection attack. The multi-encryption storage method in reference [20], dynamic storage method in reference [21] and the resistance of big data to different external attacks after encryption are tested, and the results are shown in Table 1.

By analyzing the data in Table 1, it can be seen that the encrypted data in this paper can effectively resist four types of attacks, and the average anti-attack rate is as high as 96.99%. After encryption, the data security performance is good, and the encrypted big data can be protected from the risk of information disclosure caused by external attacks.

5 Conclusions

In order to verify the encryption and storage performance of this method for unstructured data, the English data in unstructured data of a university is taken as the research object. First, in the data series acquisition experiment, the mean square error of different iterations is compared and analyzed. The results show that with the increase of iterations, the difference between the actual big data series and the target data becomes smaller. Secondly, experiments on big data encryption and anti-attack show that the proposed method can conceal the real information in big data, and the encryption effect is significant, and the big data after encryption has a good resistance to malicious attacks, proving that it has a good anti-attack and higher storage efficiency. If the amount of data is very large, the encryption effect of this paper has some shortcomings. But in the future, we will continue to take some of the most advanced deep learning methods and apply them to real engineering.

Acknowledgments

The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] S. A. G. Ali, H. R. D. AL-Fayyadh, S. H. Mohammed and S. R. Ahmed, "A Descriptive Statistical Analysis of Overweight and Obesity Using Big Data," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey*, pp. 1-6, 2022. doi: 10.1109/HORA55278.2022.9800098.

[2] X Li, H Liu, W Wang, Y Zheng, H Lv, Z Lv, "Big data analysis of the internet of things in the digital twins of smart city based on deep learning," *Future Generation Computer Systems*, vol. 128, pp. 167-177, 2022.

[3] C. Li, Y. Chen, Y. Shang, "A review of industrial big data for decision making in intelligent manufacturing," *Engineering Science and Technology, an International Journal*, vol. 29, 2022.

[4] L. Teng, "Brief Review of Medical Image Segmentation Based on Deep Learning," *IJLAI Transactions on Science and Engineering*, vol. 1, no. 02, pp. 01-08, 2023.

[5] S. Yin, H. Li, L. Teng, A. A. Laghari, V. V. Estrela, "Attribute-based Multiparty Searchable encryption model for Privacy Protection of Text Data," *Multimedia Tools and Applications*, 2023. ttps://doi.org/10.1007/s11042-023-16818-4

[6] Z. -S. Chen, X. Zhang, R. M. Rodriguez, W. Pedrycz, L. Martinez, M. J. Skibniewski, "Expertise-Structure and Risk-Appetite-Integrated Two-Tiered Collective Opinion Generation Framework for Large-Scale Group Decision Making," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 12, pp. 5496-5510, 2022.

[7] E. Farri, P. Ayubi, "A robust digital video watermarking based on CT-SVD domain and chaotic DNA sequences for copyright protection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 10, pp. 13113-13137, 2023.

[8] S. Yin, J. Liu, L. Teng, "Improved Elliptic Curve Cryptography with Homomorphic Encryption for Medical Image Encryption," *International Journal of Network Security*, vol. 22, no. 3, pp. 419-424, 2020.

[9] L. Niu, Z. Xu, L. Zhao, D. He, J. Ji, X. Yuan, "Residual vector product quantization for approximate nearest neighbor search," *Expert Systems with Applications*, 2023.

[10] B. Padma Vijetha Dev, K. Venkata Prasad, "An Adaptive Lightweight Hybrid Encryption Scheme for Securing the Healthcare Data in Cloud-Assisted Internet of Things," *Wireless Personal Communications*, vol. 130, no. 4, pp. 2959-2980, 2023.

[11] D. Vinod, M. K. Nalini, K. Dhinakaran, D. Elantamilan and R. Gnanavel, "A Hybrid Algorithm for Secure Image based Encryption and Steganographic Technique in combination with DET and AES Algorithms," in *2022 International Conference on Advances in Computing, Communication and Applied*

Table 1: Different methods of attack resistance with attacks=200

Method	Attack activity	Attack count	Attack resistance rate/%
Proposed	Tamper attack	2	**96.67**
Proposed	Denial of service attack	20	**98.00**
Proposed	Forgery attack	3	**97.30**
Proposed	Injection attack	1	**96.00**
Reference [20]	Tamper attack	39	74.66
Reference [20]	Denial of service attack	26	82.67
Reference [20]	Forgery attack	18	88.01
Reference [20]	Injection attack	14	90.67
Reference [21]	Tamper attack	42	72.01
Reference [21]	Denial of service attack	45	70.10
Reference [21]	Forgery attack	22	85.33
Reference [21]	Injection attack	19	87.34

Informatics (ACCAI), Chennai, India, pp. 1-6, 2022. doi: 10.1109/ACCAI53970.2022.9752601.

[12] L. Teng, H. Li, S. Yin, Y. Sun, "A Modified Advanced Encryption Standard for Data Security," *International Journal of Network Security*, vol. 22, no. 1, pp. 112-117, 2020.

[13] M. Hema, S. P. Shyry, "A hybrid multimedia image encryption technique using singular value decomposition-linear sparsity regularization (SVD-LSR)," *Soft Computing*, pp. 1-11, 2023. https://doi.org/10.1007/s00500-023-07937-z

[14] S. Adhikari, S. Karforma, "A novel image encryption method for e-governance application using elliptic curve pseudo random number and chaotic random number sequence," *Multimedia Tools and Applications*, vol. 81, no. 1, pp. 759-784, 2022.

[15] D. S. Laiphrakpam, R. Thingbaijam, K. M. Singh and M. Al Awida, "Encrypting Multiple Images With an Enhanced Chaotic Map," *IEEE Access*, vol. 10, pp. 87844-87859, 2022.

[16] A. O. Topal, R. Chitic, F. Leprévost, "One evolutionary algorithm deceives humans and ten convolutional neural networks trained on ImageNet at image recognition," *Applied Soft Computing*, vol. 143, 2023.

[17] J. Wang, Y. Lu, Q. Wang, Y. Zhang, J. Shu, "Perseid: A Secondary Indexing Mechanism for LSM-based Storage Systems," *ACM Transactions on Storage*, 2023. https://doi.org/10.1145/3633285.

[18] B. Jamil, H. Ijaz, M. Shojafar, K. Munir, "IRATS: A DRL-based intelligent priority and deadline-aware online resource allocation and task scheduling algorithm in a vehicular fog network," *Ad hoc networks*, vol. 141, 2023.

[19] J. Zhu, "Research on Secure Storage of Network Data Based on Cloud Computing Technology," *International Journal of Network Security*, vol. 24, no. 1, pp. 68-74, 2022.

[20] J. Hglund, S. Raza, "BLEND: Efficient and blended IoT data storage and communication with application layer security," in *2022 IEEE International Conference on Cyber Security and Resilience (CSR), Rhodes, Greece*, pp. 253-260, 2022, doi: 10.1109/CSR54599.2022.9850290.

[21] R. U. Mustafa, M. T. Islam, C. Rothenberg and P. H. Gomes, "EFFECTOR: DASH QoE and QoS Evaluation Framework For EnCrypTed videO tRaffic," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium, Miami, FL, USA*, pp. 1-8, 2023, doi: 10.1109/NOMS56928.2023.10154448.

Biography

Ruoshuang Yin biography. Ruoshuang Yin is with the School of Foreign Languages, Zhengzhou University of Science and Technology. Interests: Data analysis, Security analysis, Foreign language teaching and translation.

Art Design Data Privacy Protection Strategy Based on Blockchain Federated Learning and Long Short-term Memory

Jing Yu, Lin Huang, and Lu Zhao
(Corresponding author: Lu Zhao)

Lu Xun Academy of Fine Arts
Shenyang 110004, China
Email: 910675024@qq.com

(Received Dec. 7, 2023; Revised and Accepted Feb. 7, 2024; First Online June 22, 2024)
The Special Issue on Data Fusion, Deep Learning and Optimization Algorithms for Data Privacy Protection
Special Editor: Prof. Shoulin Yin (Harbin Institute of Technology)

Abstract

Data visualization is a graphical method of data representation, which is often manifested as geometric structure in appearance, with the beauty of mathematical harmony and rhythm, so it is very artistic. At the same time, how to effectively and aesthetically convey information through color, graphics, composition and other means also requires the participation of design. However, data security is very important in art design, if the data is leaked, then the artistic inspiration will disappear. For most of the federation learning defense methods, there are some problems, such as reducing the usefulness of federation learning, low computing efficiency and single type of defense attack, therefore, this paper proposes a novel art design data privacy protection strategy based on blockchain federated learning and long short-term memory (LSTM). Firstly, in order to improve the computational efficiency of the algorithm, a variational autoencoder training scheme based on transfer learning is proposed to reduce the training cycle of the client. Secondly, an attribute modification scheme based on the constraint rules of attribute distribution is designed to reconstruct the client training data by using the continuous hidden variable of variational autoencoder. With the blockchain certificate issuing center as a trusted third party, initialization, key generation and distribution of the improved access control algorithm ciphertext policy attribute-based encryption (CP-ABE) are completed. The credit data is then stored on a blockchain. The data feature is extracted by LSTM. The system implemented by this model can ensure fine-grained access control and privacy protection of user rights. Function and performance evaluation show that the proposed method has important reference significance and application value in the field of digital credit privacy protection, credit deposit, access control, etc. Application system throughput and blockchain TPS can meet the needs of practical application performance.

Keywords: Art Data Privacy Protection; Blockchain; Federated Learning; LSTM

1 Introduction

With the rapid development of big data, cloud computing and other technologies, credit as a key technology in the financial field has been widely studied [1]. Big data technology provides massive and diverse data for credit bureaus, solving the problem of single data source [2]. Cloud computing provides a large capacity of computing space to solve the pain points of credit data storage and computing difficulties. However, the authenticity of the credit data provided by the data source cannot be guaranteed, and the credit data contains sensitive data in the fields of user finance and education. Therefore, when data privacy security cannot be guaranteed, data providers are reluctant to join credit rating agencies [3].

Recently, the academic research found that in multiple federation learning application scenarios, there are problems that attackers disguised as federation learning participants to spy on privacy and implant backdoors into global models. For example, in the smart medical scenario, several hospitals use medical data for federal learning cooperation to train an AI model to identify diseases, and the attacker uses non-master attribute privacy disclosure 3 and other technologies to obtain the training data of the hospital, resulting in the disclosure of personal information; In the intelligent traffic scenario, the attacker will add a trigger to the traffic sign image to participate in the federation learning training process, at this time, the model has been implanted in the backdoor vulnera-

bility [4], when the model recognizes similar traffic sign images again, it will be misclassified into other traffic sign categories.

Blockchain, as an emerging data sharing technology, has the characteristics of immutable and traceable, and uses smart contracts and consensus mechanisms to store data distributed under the premise of no trusted third party to avoid the risk of human fraud. Due to the outstanding advantages of P2P networks, anonymous transactions, and decentralized architecture in the blockchain [5, 6], the blockchain has been widely used in many privacy-protection scenarios.

At present, there are two major problems in the field of credit: data sharing is difficult and user privacy is easy to leak. At present, the combination of public and private credit investigation is adopted in China. Due to different credit investigation institutions and commercial competition among the same private institutions, a large number of effective user credit data cannot be shared, and it is easy to form an "information island". At the same time, traditional credit institutions are also accompanied by problems such as single access rights and inflexibility, which makes credit data sharing, especially restricted sharing with fine-grained access control, an urgent problem to be solved under the premise of ensuring data security.

The traditional credit system collects, processes and releases personal credit data through a centralized authority. Once the centralized server is hacked, the privacy and security of user credit information cannot be guaranteed. If blockchain is used directly to link credit data, attackers can also infer sensitive information about users by analyzing the connections between different transactions [7]. Therefore, it is also necessary to study a credit scheme to ensure user privacy security while protecting restricted sharing [8, 9].

At present, the defense measures against the above attack modes are mainly divided into three categories, namely homomorphic encryption [10], differential privacy [11] and defense against Byzantine problems [12]. Homomorphic encryption is a ciphertext computing scheme that does not need to decrypt ciphertext. It is used in federation learning to prevent internal attackers from snooping on potential private information in the model. As a cryptographic method, homomorphic encryption consumes a lot of computing resources. Differential privacy protects the privacy information of the client model by introducing noise from the federated learning local client local model. However, few of the existing federal learning differential privacy mechanisms can weigh the practicality and privacy of the algorithm. The core idea of defense against Byzantine problems is to calculate the difference between the local model uploaded by the client and the global model, and discard the local model that is too different from the global model, so the traditional backdoor attack based on data poisoning is ineffective for the federation learning defense against Byzantine problems. But recent improvements to the idea of backdoor attacks have made this defence less effective. To sum up, the above three defense measures have the problems of poor practicality, slow computing efficiency and easy to be circumvented by backdoor attacks, and it is difficult to play the expected role in many application scenarios of federated learning.

In order to solve the above problems, this paper designs an active privacy protection algorithm based on variational autoencoder (VAE) for federated learning clients, called federated learning attribute modification algorithm. The specific contributions of this paper are as follows:

1) In order to reduce the consumption of computing resources on the client side, transfer learning technology is adopted to pre-train the VAE model on the server side, and only a few rounds of adaptive adjustment are required on the client side.

2) In order to balance practicality and privacy, based on the statistical and distribution rules of image contained attributes, this paper proposes a non-main task attribute modification mechanism and adopts VAE to remove irrelevant attributes in images instead of adding noise, which ensures privacy and does not significantly reduce the accuracy of the model.

3) Experimental results show that the attribute modification scheme in this paper can effectively classify, control and modify the attributes of client training data. In federation learning, the average accuracy of global model trained by modifying single attribute data is 93.28%, and the highest accuracy of training by modifying multiple attribute data is 94.55%. The scheme of this paper can adjust the image attributes and correct the abnormal area of the image, and can also prevent the privacy leakage of non-primary attributes and the backdoor attack based on data poisoning.

The remainder of this paper is organised as follows. Section 2 presents related work on privacy protection. Section 3 presents the federated learning for this paper. Section 3 describes the scheme objective. Section 4 mainly analyses the experimental results. Finally, Section 5 summarises the work in this paper and points out future work.

2 Related Works

At present, the domestic credit investigation mode is based on public credit investigation and supplemented by private credit investigation. Credit data is collected, processed and released by several authorities (People's Bank of China, government agencies and third-party credit investigation agencies), and third-party credit investigation agencies play a complementary role [13, 14]. Credit construction in foreign developed countries has been going on for a hundred years, and the most prominent feature is that credit legislation has developed more mature.

Academically, economists believe that blockchain itself has the characteristics of privacy protection, immutable, traceable and so on, which can bring new development opportunities for the credit industry [15]. Lin *et al.* [16] pointed out that the alliance chain had the characteristics of identity authentication, public ledger, encryption algorithm, consensus mechanism, smart contract, etc., which was suitable for the collection, transmission, storage, supervision and verification of credit data. Vegesna *et al.* [17] detailed stated the issues and challenges that needed to be addressed in the application of blockchain-based distributed ledgers. Liu *et al.* [18] implemented a blockchain-based cross-platform credit data sharing model and verified the feasibility of the design model through tests. Yang *et al.* [19] proposed a user credit evaluation system based on blockchain, which realized the untamable and traceability of credit data through the technology of intelligent contract in blockchain. Szczepaniuk *et al.* [20] designed a blockchain credit model for enterprises to solve the problem of information mismatch in the credit market, but the model did not solve the problem of user privacy security. Wang *et al.* [21] considered the issue of privacy protection and proposed a decentralized blockchain credit model based on asymmetric keys, which was mainly based on the support of cited literature, without giving a detailed implementation scheme. Bai *et al.* [22] built a blockchain-based credit data sharing database, and the improved database had greatly improved the authenticity and reliability of the data.

However, the data collected so far lacks a specific blockchain-based solution for user privacy protection and credit data sharing. Therefore, it is of great significance to study how to realize both restricted sharing of credit data and privacy protection in the field of digital credit. Based on blockchain, this paper designs a multimedia data limited sharing and privacy protection model, which can complete the collection, transmission, calculation and other processing of credit data.

3 Federated Learning

Compared with traditional centralized deep learning, in order to solve the problem of data silos, federated learning trains a model independently by each client with data sets, communicates its own model through the server, and finally obtains a global model through model aggregation [23]. Assuming a total of C clients participate in the federated learning task, the server selects $K(1 \leq K \leq C)$ clients to upload the local model parameter w_t^k in the t communication round. The server uses the method shown in formula (1) to aggregate all uploaded local model parameters and send them to all clients under the global model.

$$w_{t+1} = \frac{1}{K} \sum_{k=1}^{K} w_t^k. \tag{1}$$

Where w_{t+1} represents the global model parameter of the $t+1$ round. Through the above model exchange and aggregation iteration, the global model gradually converges, and finally reaches the target accuracy or the preset communication round termination. The specific process is shown in **Algorithm 1**. FedAvg algorithm [24] has been validated by various datasets such as FederatedEMNIST, CIFAR-100, Shakespeare and StackOverflow [12].

Algorithm 1 FedAvg algorithm

Input: communication times T, local iterations E, client data set D.
Output: global model parameter w.
In server-side:
while Initializing the global model parameter w_0 **do**
  for $t = 1$ to T do
  $S \leftarrow$ Select K clients from C clients.
  $w_t^k \leftarrow$ Receive the model parameters uploaded by the selected client $k \in S$.
  $w_{t+1} \leftarrow \frac{1}{K} \sum_k w_t^k$.
end while
Client k:
while Upload the local model parameter w_t^k to the server **do**
  for $i = 1$ to E do
  for $d \in D$ do
  $w_t^k \leftarrow w_t - \eta \nabla l(w_t; d)$
end while

3.1 Variational Autoencoder (VAE)

VAE is to use autoregressive process training and sampling of latent variables to obtain structured probability distribution of training data. The latent variable z represents the intrinsic structure of the data x and satisfies some particular posterior distribution $p(x, z)$. Therefore, VAE first compresses high-dimensional images into low-dimensional latent variables, and then performs autoregression on latent variables to improve the generating effect. Formula (2) shows the optimization goal of VAE.

$$L(\theta, \varphi; x^i) = A - B. \tag{2}$$

Where, $A = E_{q\varphi(z|x^i)}[\log p_\theta(x^i|z)]$, $B = R_{KL}(q\varphi(z|x^i)||p_\theta(z))$, θ and φ are model parameters of encoder and decoder in VAE respectively. x^i is a piece of data in the data set. $p_\theta(x^i|z))$ and $p_\theta(z))$ can be calculated by Bayes formula. A is reconstruction loss. B is the KL regular term. As shown in formula (3), the addition of regular terms contributes to the potential space for VAE to have good structure and reduces overfitting on the training data.

$$KL(p(x)||q(x)) = \sum p(x) \log \frac{p(x)}{q(x)}. \tag{3}$$

3.2 Non-primary Attribute Privacy Disclosure

The principle of non-master attribute privacy disclosure is that federated learning updated model parameters will reveal additional privacy information about the client training data. For example, federation learning constructs a gender classification task of a face image, and the attacker can use this method to find out whether other attributes in the face image, such as skin color and age, appear in the current federation learning communication round. The attacker modifies the local model on the masqueraded client to build the multitask loss function l_{mt} as shown in formula (4).

$$l_{mt} = \alpha l(x, y, \theta) + (1 - \alpha) l(x, p; \theta). \tag{4}$$

Where y is the attribute (label) required by the federated learning task. p is the target attribute inferred by the attacker. $l(x, y, \theta)$ is the loss function of the federation learning main task. $l(x, p; \theta)$ is the loss function of the attacker's privacy inference task. α is the coefficient that balances the two loss functions. In this way, the attacker can use the global model to implement privacy inference without significantly reducing the accuracy of the main task.

3.3 Backdoor Attack

Backdoor attacks are a class of attacks that add adversarial triggers to training data to manipulate the output of a model, and a model trained with poisoned data will make arbitrary or targeted false predictions on other data embedded with the same trigger. Although federated learning can aggregate the model parameters of many clients to train better models, its distributed training style and non-independent and uniformly distributed data sources facilitate backdoor attacks.

First, the attacker participates in the task of federated learning as a client and converts the clean training data S_{clean} into data S_{poison} with trigger according to the poisoning data generation method [14]. Second, the local model is trained using both poison data and clean data, as shown in formula (5).

$$\begin{aligned} w = argmaxE[& \sum_{x', y' \in S_{poison}} p(G_t(x')) = y' \\ & + \sum_{x', y' \in S_{clean}} p(G_t(x)) = y]. \end{aligned} \tag{5}$$

Where w is the local model parameter. G_t is the global model client delivered by the server as the local model. x' and y' are poisoning data and labels. x and y for clean data and labels. The goal of formula (5) is to obtain the appropriate model parameter w so that the probability of poisoning data and clean data being classified as corresponding labels is maximum. Finally, the attacker uploaded the model parameter w containing the backdoor vulnerability to the server to contaminate the global model and realize the backdoor attack against federation learning.

4 Scheme Objective

Based on the attack principles of the two federative learning attack methods mentioned above, this paper uses variational autoencoder to preprocess client data, and proposes a scheme to protect data privacy and model security of federative learning client. First, on the premise of keeping the main attribute of the client data used for federated learning training, the non-main attribute that can be modified is screened and adjusted to realize the privacy protection of private data. Secondly, according to the backdoor attack trigger that may exist in the client data, the data is cleaned before the data participates in the model training, so as to avoid the damage of the poisoned data to the model security.

First, in the initial phase of federated learning, in order to save computing resources on the client side, the server uses the collected data set to pre-train the VAE model. It should be noted that the data set does not require additional label information, so relevant data can be easily collected from the Internet. Secondly, by means of transfer learning technology, the server-side VAE model is deployed to the client. The client only needs to input private data into the VAE model for far less iterations than the training rounds of the server-side VAE model, so as to realize the migration of the source domain to the target domain. Finally, the label of local data is used to separate the non-main attributes that can be modified by the attribute screening method proposed in this paper, and the protection of private data is realized by adjusting these attributes in the image, such as modifying the hair color and smile of the figure in the image.

The data is pre-processed by the above method, and the pre-processed data is used to train the federation learning model by means of inter-layer information transfer. The scheme in this paper uses the method of modifying attributes instead of adding noise to ensure the practicality of the scheme. At the same time, the adjustment of attributes can protect privacy to a certain extent, achieving a balance between practicality and privacy. And all private data only needs to be processed once, which reduces computing resource consumption compared to other protection schemes.

4.1 LSTM

Long short-term memory neural network (LSTM) [25] is based on the traditional recurrent neural network. Compared with traditional recurrent neural networks, long short-term memory neural networks have no long-term dependence defect when solving long time series problems. LSTM introduces input gate, output gate and forgetting gate to prevent gradient explosion and delay between input and feedback through the connection between the

gates, forcing a continuous error flow inside the memory unit, effectively improving the traditional recurrent neural network gradient explosion and gradient disappearance phenomenon. Therefore, LSTM can effectively control the information transmission path and greatly improve the memory ability of neural networks.

Each unit stores a memory of the input sequence, so it is also called a "memory cell". The cell has three gated units: input gate i, output gate o and forgetting gate f, and the three gated units jointly control the LSTM unit to selectively store memory and update, so as to complete the effective transmission of information. The input X_t at the current time and the output h_{t-1} at the previous time together form the input sequence of the gated unit.

The function of the forgetting door is to control the amount of memory information received from the current moment to the memory information of the previous moment, which can be expressed as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f). \tag{6}$$

The function of the input gate is to control the amount of input information received from the current moment memory information, which can be expressed as:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i). \tag{7}$$

The current time candidate input is:

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C). \tag{8}$$

The current moment memory C_t can be obtained by using the previous moment memory information C_{t-1} controlled by the forgetting gate f_t and the current moment candidate input information controlled by the input gate i_t, which is expressed as:

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t. \tag{9}$$

The amount of memory information input from the current moment to the next moment is controlled by the output gate, expressed as:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o). \tag{10}$$

$$h_t = o_t \tanh(C_t). \tag{11}$$

In equations (6)-(11), b_f, b_i, b_c and b_o are offset terms. W_f, W_i, W_c and W_o are the weight matrix. σ and $tanh$ are sigmoid and hyperbolic tangent activation functions, respectively, expressed as:

$$\sigma = sigmoid(x) = \frac{1}{1 + e^{-x}}. \tag{12}$$

LSTM neural network can be regarded as an extremely complex nonlinear function, which adjusts the internal weights through training data to approximate the input data, and then completes the task of data recognition.

4.2 Attribute Modification

In the blockchain network, based on the existing transaction ordering, transaction endorsement, transaction verification, and transaction chain functions in the fabric, ciphertext transactions that need to be stored are endorsed, sorted, linked, and verified, thus realizing distributed storage of ciphertext data. On the client, the user receives the certificate with the permission and then initiates a data request transaction. The attribute key is used to implement CP-ABE encryption and decryption of ciphertext data. In CA, when the certificate application is received, the corresponding certificate is issued or updated for the user, and CP-ABE initialization and key generation and distribution are completed at the same time to realize the user's permission setting. The user accesses the plaintext credit data through the blockchain network, which is divided into three stages: key generation, encryption and access control. The parameter description of the algorithm designed in this layer is shown in Table 1.

Table 1: Parameter description

symbol	Definition
PK_C	System public key
MK_C	System master key
λ	System safety parameter
P	Prime number field
$list_{auth}$	Permission revocation list
$prome$	User's unique prime number
S_{att}	Attributes set of user
M_C	Plaintext to be encrypted
UK_C	System user key
DO	Data owner

- Initialization. The randomization algorithm is called as the initialization algorithm of the credit system, and the initialization algorithm outputs the public key of the credit system and the main key of the system by inputting the security factor, which can be expressed as $Setup(\lambda) \rightarrow (PK_C, MK_C)$.
- Initialization of revocation organization. The access control scheme proposed in this paper can realize the function of $RevocationSetupPK_C(P, list_{auth}, list_{auth}))$, so a user permission revocation authority is set up. Through the public key of the credit system generated during system initialization, the revocation authority initialization generates a prime field, a permission revocation list, a mapping table containing user Gid and unique primes, and a mapping table containing user attributes and revocation lists.
- Private key generation. The unique private key of the user is generated dynamically by entering the public

key of the credit system, the master key of the credit system and the attribute set of the user in the CA. Select a prime number in the prime number field to assign to the user, and delete the prime number from the prime field to ensure the uniqueness of the user prime number, and finally use the prime number to update the mapping table, which can be expressed as $KeyGenPK_C, MK_C, S_{att} \rightarrow UK_C, P(user) \rightarrow map(U_{Gid}, prime)$.

- Encryption. A randomization method is invoked as the encryption algorithm, and the CP-ABE ciphertext is obtained by inputting the public key of the credit system, the plaintext message to be encrypted and the access control structure associated with the access policy, which can be represented as $Enc(PK_C, M_C, ACS_P) \rightarrow C_C$.

- Decryption. Call a deterministic method as the decryption method, which can be represented as $Dec(UK_C, C_C) \rightarrow M_C$.

- User attribute revocation. By entering the user and the corresponding unique prime number, the revocation attribute and the revocation list corresponding to the attribute into the revocation algorithm, computes a new attribute revocation $list'_{auth}$, and updates two mapping tables, which can be expressed as $Revoke(U_{att}, list_{auth}, user, prime)$, $(map(U_{Gid}, prime)$, $map(U_{att}$, $list'_{auth})$, $list'_{auth} = list \times prime$.

- User property recovery. Similar to revocation, enter the recovery algorithm for the user who needs to restore the permission, the corresponding unique prime number, the revocation attribute, and the revocation list corresponding to the attribute. Computes the new attribute revoke list $list''_{auth}$ and updates the two mapping tables, which can be expressed as $Recover(U_{att}, list_{auth}, user, prime) \rightarrow (map(U_{Gid}, prime), map(U_{att}, list''_{auth})$, $list''_{auth} = list \times prime$.

Through the above methods, you can achieve fine-grained access control over encrypted data, and DO has full control over the data. Because CP-ABE encryption does not need to know the identity of the recipient, it does not need to be encrypted multiple times when it is sent to multiple users. CP-ABE performs encryption only once after setting the access policy, and when the user has attributes that match the policy described by the encrypter, the data user can decrypt it. It solves the problem of key leakage caused by general encryption algorithm, guarantees the privacy security of credit DO, and improves the granularity of data access control.

VAE can be used to convert input data into highly structured latent variables. As shown in formula (13), all positive samples $y_{A_j} = pos$ and negative samples $y_{A_j} = neg$ are encoded and superposed into two vectors respectively. The difference between the two vectors is the attribute vector V_{A_j}, which needs to be separated.

$$V_{A_j} = \sum_{x, y_{A_j} \in D_k} encode(x, y_{A_j} = pos) - \sum_{x, y_{A_j} \in D_k} encode(x, y_{A_j} = neg). \tag{13}$$

The specific calculation process is shown in **Algorithm 1**.

Algorithm 2 Attribute modification algorithm

Input: client VAE model M, all data attributes A_{ttr}, client data set D_k.
Output: Modified the property data x_{A_j}.
Attribute separation:
if $a \in Attr$ **then**
  if $a \neq$ primary attribute and a meets the restriction **then**
    for $x, y \in D_k$ do
    $V_{y_a, y_{A_m}} += M.encode(x)$
    $V_{y_a, y_{A_m}} = V_{y_a, y_{A_m}} / num(V_{y_a, y_{A_m}})$
    return $S_{a, y_{A_m}}$
  end if
end if
Attribute control:
if $x \in D_k$ **then**
  $z \leftarrow M.encode(x)$
  for $v \in S_{a, y_{A_m}}$ do
  $\beta \leftarrow random(-1, 1)$
  $z += \beta v$
  $x_r \leftarrow M.decode(z)$
  $S_x \leftarrow x_r$
  return S_x
end if
End

After the attribute vector V_{A_j} is obtained, the scheme needs to modify the attributes of the client data to protect data privacy. At this time, the hidden variables are calculated according to formula (14) with the help of their continuity. Firstly, data x is input into VAE model to get latent variable z, and then combined with vector V_a in attribute set $S_{a, y_{A_m}}$. Finally, the calculated latent variable is decoded by VAE model to get data x_{S_A} with modified attributes. The positive and negative value of β can affect the performance of property data x_{S_A}.

$$x_{S_A} = decode(encode(x) + \sum_a \beta V_a). \tag{14}$$

5 Experimental Result

In this paper, LoadRunner, Jmeter and Tape are used to complete the test. This test uses a centos7.4 operating system, 100GB hard disk, CPU Inteli7, and 8GB memory. MySQL5.0.17Linux is installed as the database server for

data management and storage. Four application servers with the same configuration and additional installation of JDK 1.5.0_06, Apache2.2.0, and Tomcat 5.5.15 are configured. It is used to deploy and start the underlying system, and finally use the windows10 64-bit operating system with Google browser and three test tools as the client test system platform. Finally, this paper realizes a blockchain-based credit data limited sharing and privacy protection platform.

1) **LoadRunner script communication efficiency test**

The LoadRunner script is used to test the communication efficiency of the implementation platform of the above design scheme. In this paper, the efficiency is tested respectively when the number of users is 100, 500 and 1000. The test communication efficiency is shown in Table 2.

Table 2: Communication efficiency

user number	receive/tps	bandwidth/MB
100	254388	100
500	252464	100
1000	254258	100d

As can be seen from Table 2, the communication efficiency of the system platform designed in this scheme is above 252464/tps in all three cases.

2) **Jmeter block link interface test**

This article uses Jmeter to test the block link interface. Take the block ID query credit information block as an example, set the concurrent thread 1000, The system starts within 20 seconds, tests the 120s thread continuously, and sets the requested server IP address, block ID is set as a variable, and a random block ID is passed in as a parameter each time. According to the throughput test diagram and aggregate report, the block ID query block information interface of the digital credit blockchain platform designed in this paper supports no less than 1000 concurrent retrieval users. The throughput reaches 2031.1/tps, the error rate is 0, and the user concurrency performance is good.

3) **Tape the underlying blockchain test**

In this paper, Tape is used to test the throughput of the digital credit blockchain network designed in this paper. Script parameters are set in the configuration file, including node location, execution parameters, channels, etc. The execution times are set to 10000 times, and then the test script is executed three times. The results are shown in Figure 1. The throughput of the underlying blockchain available is 2518/tps, 2512/tps, 2548/tps, with an average

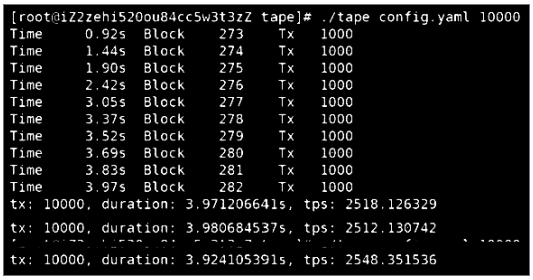


Figure 1: Tape reading test results

of 2526/tps, which can better meet the daily business needs.

According to the LoadRunner test above, the system platform supports simultaneous requests from multiple users. According to the Jmeter and Tape tests, both the block link interface and the underlying network of the block chain can achieve stable and normal access within a certain number of visits (within 1000 parallel lines). It can be seen that the system realizes the restricted sharing and privacy protection of data through the underlying blockchain technology, cryptography technology and access control algorithm, so as to ensure the reliability and security of the system.

4) **Encryption time with different schemes**

Figure 2 shows the relationship between the dimension of the attribute domain and the time required for various encryption algorithms when the data size is 500KB. According to the experimental results, all schemes' encryption times grow linearly as the attribute domain dimension rises. In the experiment, the encryption time of proposed scheme is less than that of other schemes (scheme1 (reference 26), scheme2 (reference 27)) when the attribute domain dimension is greater than 20, and the higher the attribute domain dimension, the bigger the difference between the encryption times of these two schemes.

6 Conclusions

This paper comprehensively considers the two pain points of privacy protection and restricted sharing of credit data, and proposes a credit scheme and system implementation based on blockchain. The homomorphic encryption operation is introduced into the system, and the user's credit sensitive information is stored in the cloud server for homomorphic calculation, which solves the problem of excessive cost in ciphertext operation. Combined with asymmetric encryption algorithm and access control algorithm, the fine-grained access control and privacy protection of credit data are completed. The experimental results show that the digital credit blockchain system scheme provided in this paper has good performance in communication

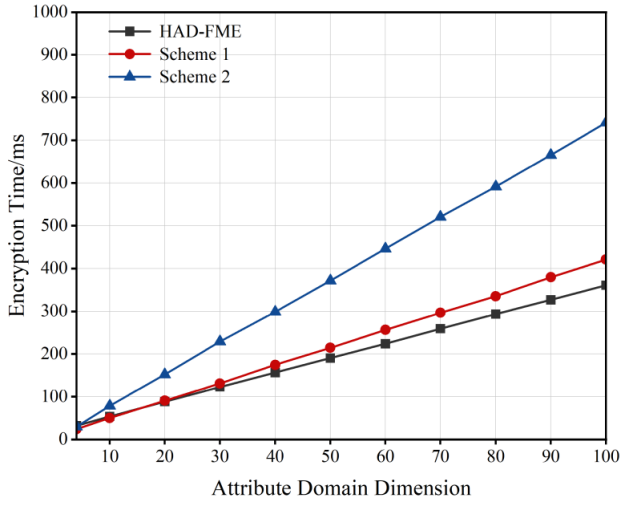


Figure 2: The relationship between encryption time and attribute domain dimension of different schemes.

efficiency, interface performance and blockchain performance. It is pointed out that the scheme and system have certain reference significance and application value in the field of actual credit. The next research direction is to combine the relevant algorithms of multi-party security computing to further improve the data security of credit institutions with multi-party participation.

Acknowledgments

The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] G. Sriram, "Edge computing vs. Cloud computing: an overview of big data challenges and opportunities for large enterprises," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 4, no. 1, pp. 1331-1337, 2022.

[2] S. Yin, H. Li, L. Teng, A. A. Laghari, V. V. Estrela, "Attribute-based Multiparty Searchable encryption model for Privacy Protection of Text Data," *Multimedia Tools and Applications*, 2023. ttps://doi.org/10.1007/s11042-023-16818-4

[3] A. Alam, "Cloud-based e-learning: scaffolding the environment for adaptive e-learning ecosystem based on cloud computing infrastructure," in *Computer Communication, Networking and IoT: Proceedings of 5th ICICC 2021, Volume 2. Singapore: Springer Nature Singapore*, pp. 1-9, 2022.

[4] Z. Zhang, G. Xiao, Y. Li, T. Lv, F. Qi, Z. Liu, Y. Wang, X. Jiang, M. Sun, "Red alarm for pre-trained models: Universal vulnerability to neuron-level backdoor attacks," *Machine Intelligence Research*, vol. 20, no. 2, pp. 180-193, 2023.

[5] T. Tan, S. Saraniemi, "Trust in blockchain-enabled exchanges: Future directions in blockchain marketing," *Journal of the Academy of marketing Science*, 2023, 51(4): 914-939.

[6] X. Wang, H. Zhu, Z. Ning, L. Guo and Y. Zhang, "Blockchain Intelligence for Internet of Vehicles: Challenges and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2325-2355, 2023.

[7] H. Han, R. Shiwakoti, R. Jarvis, C. Mordi, "Accounting and auditing with blockchain technology and artificial Intelligence: A literature review," *International Journal of Accounting Information Systems*, vol. 48: 100598, 2023.

[8] A. Pasdar, Y. Lee, Z. Dong, "Connect API with blockchain: A survey on blockchain oracle implementation," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1-39, 2023.

[9] I. Muhammad, S. Yin, H. Li, S. Karim, A. Laghari, "Comprehensive Review of Emerging Cyber security Trends and Developments," *International Journal of Electronic Security and Digital Forensics*, 2023. doi: 10.1504/IJESDF.2025.10059222.

[10] K. Munjal, R. Bhatia, "A systematic review of homomorphic encryption and its contributions in healthcare industry," *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 3759-3786, 2023.

[11] S. Yin, H. Li, L. Teng, "A Novel Proxy Re-encryption Scheme Based on Identity Property and Stateless Broadcast Encryption Under Cloud Environment," *International Journal of Network Security*, vol. 21, no. 5, pp. 797-803, 2019.

[12] Q. Xia, Z. Tao, Q. Li and S. Chen, "Byzantine Tolerant Algorithms for Federated Learning," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 6, pp. 3172-3183, 2023, doi: 10.1109/TNSE.2023.3251196.

[13] G. Yang, "Research on Financial Credit Evaluation and Early Warning System of Internet of Things Driven by Computer-Aided Technology," *Comput. Aided. Des. Appl*, vol. 19, no. S6, pp. 158-169, 2022.

[14] Y. Su, J. Li, B. Yu, Y. Zhao, "Numerical investigation on the leakage and diffusion characteristics of hydrogen-blended natural gas in a domestic kitchen," *Renewable Energy*, vol. 189, pp. 899-916, 2022.

[15] V. Stazi, M, Annesini, M. Tomei, "Anaerobic domestic wastewater treatment in a sequencing granular UASB bioreactor: Feasibility study of the temperature effect on the process performance," *Journal of Environmental Chemical Engineering*, vol. 10. no. 5, pp. 108512, 2022.

[16] S. Lin, L. Zhang, J. Li, L. Ji, Y. Sun, "A survey of application research based on blockchain smart contract," *Wireless Networks*, vol. 28, no. 2, pp. 635-690, 2022.

[17] V. Vegesna, "Using Distributed Ledger Based Blockchain Technological Advances to Address IoT Safety and Confidentiality Issues," *International*

Journal of Current Engineering and Scientific Research, vol. 9, pp. 89-98, 2022.

[18] C. Liu, T. Dong, L. Meng, "Cross-Border Credit Information Sharing Mechanism and Legal Countermeasures Based on Blockchain 3.0," *Mobile Information Systems*, vol. 2022, 2022.

[19] L. Yang, W. Zou, J. Wang, Z. Tang, "EdgeShare: A blockchain-based edge data-sharing framework for Industrial Internet of Things," *Neurocomputing*, vol. 485, pp. 219-232, 2022.

[20] H. Szczepaniuk, E. Szczepaniuk, "Cryptographic evidence-based cybersecurity for smart healthcare systems," *Information Sciences*, vol. 649: 119633, 2023.

[21] L. Wang, Y. Qi, Y. Bai, Z. Sun, D. Li, X. Li, "MuKGB-CRS: Guarantee privacy and authenticity of cross-domain recommendation via multi-feature knowledge graph integrated blockchain," *Information Sciences*, vol. 638: 118915, 2023.

[22] T. Bai, Y. Hu, J. He, H. Fan, Z. An, "HealthzkIDM: A healthcare identity system based on fabric blockchain and zero-knowledge proof," *Sensors*, vol. 22, no. 20, pp. 7716, 2022.

[23] L. Teng, Y. Qiao, M. Shafiq, G. Srivastava, A. Javed, T. Gadekallu, S. Yin, "FLPK-BiSeNet: Federated Learning Based on Priori Knowledge and Bilateral Segmentation Network for Image Edge Extraction," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1529-1542, 2023.

[24] Z. Qu, Y. Tang, G. Muhammad, P. Tiwari, "Privacy protection in intelligent vehicle networking: A novel federated learning algorithm based on information fusion," *Information Fusion*, vol. 98, 2023.

[25] J. A and S. Yin, "A New Feature Fusion Network for Student Behavior Recognition in Education," *Journal of Applied Science and Engineering*, vol. 24, no. 2, pp. 133-140, 2021.

[26] C. Wang, J. Lu, X. Li, P. Cao, Z. Zhou and Q. Wen, "A Personal Privacy Data Protection Scheme for Encryption and Revocation of High-Dimensional Attribute Domains," *IEEE Access*, vol. 11, pp. 82989-83003, 2023.

[27] G. Zhang, X. Chen, B. Feng, X. Guo, X. Hao, H. Ren, *et al.*, "BCST-APTS: Blockchain and CP-ABE empowered data supervision sharing and privacy protection scheme for secure and trusted agricultural product traceability system," *Secur. Commun. Netw.*, vol. 2022, pp. 1-11, Jan. 2022.

Biography

Jing Yu biography. Jing Yu is with Lu Xun Academy of Fine Arts. Research interests: art deign data, data analysis, big data.

Lin Huang is with Lu Xun Academy of Fine Arts. Research interests: art deign data, data analysis, big data.

Lu Zhao biography. Lu Zhao is with Lu Xun Academy of Fine Arts. Research interests: art deign data, data analysis, big data.

Intelligent Network Security with Session Initiation Protocol and Web Services

Abdallah Handoura and Daniel Bourget
(Corresponding author: Abdallah Handoura)

Computer Department, IMT-Atlantique France
Computer Department College of Telecom and Information- TVTC- KSA
Email: Abdallah.Handoura@telecom-bretagne.eu
(Received May 28, 2023; Revised and Accepted Jan. 4, 2024; First Online June 22, 2024)

Abstract

With the rapid evolution of telecommunications services, the demand in terms of security is increasingly indispensable. That necessitates the advanced method introduction of cryptography and authentication, management, and distribution of keys between the different entities on the system. The intelligent network is a basis to establish and commercialize the services by the telecommunication network. Selling services and information by a network requires a considerable increase in confidentiality and integrity. This paper discusses and proposes a method based on integrating signaling protocol SIP into Web services technologies on intelligent networks to realize a procedure of client and service authentication and data confidentiality.

Keywords: Intelligent Network Security; SIP; Web Service; WS-Security

1 Introduction

Until the recent past, telecommunications networks ensure a limited number of essential services such as telephony, telex, and data transmission. This situation has evolved by the recent advanced technologes in the communications systems and in the computer-processing. they had largely contributed to the concepts and evolution of communication architecture.

The sector of telecommunications, that replies to needs of a commercial environment spreading increasingly to all the planet, has allowed to improve the productivity and to put in relationship of communities in the whole world in almost all industrial sectors. The fact of this infrastructure of telecommunications is efficient, results from elaborate norms by organisms such that the UIT-T. The norms allow not only to maintain the efficiency of current network but also to throw the basics of the next generation network [5]. The demands of information security in industrial society have behaved of major changes in the last decades. Nevertheless, while the norms continue to reply to final user needs and industry, the interface utilization and open protocols, the multiplicity of the new actors, the even diversity of applications and platforms and the fact that the implementation are not always sufficiently tested have increased malevolent utilization risks of networks. In last years, we observed a strong increase of violations of the security in world systems that have often entailed serious economic consequences. The question is then, how to keep some open infrastructure of telecommunication without compromising information exchanged on this infrastructure.

For answer this question,we begin firstly by justifying the possibility of interconnecting intelligent network with IP network and more exactly intelligent network with the signaling protocol over IP, SIP (Session Initiation Protocol). And in the second, how to benefit with the security mechanism in SIP and Web services for securing the service and user in Intelligent network.

2 Interworking between Intelligent Network and SIP

In the intelligent network (IN) viewpoint, elements such as SCP (Service Control Point), and the request originated from an UAC SIP (User Agent Client SIP) for a call processing, these functions on a traditional switch is insignificant. Thus, it is important that the SIP entity can provide the characteristics normally provided by the traditional switch, including the operation such as a SSP (Service Switching Point) for IN. The SIP entity would also have to maintain the state of call and to release request to IN according to the service, just as a traditional switch. SIP would have to operate as a SSP. For IN some services necessitate special media (such as detection DTMF), or special for the control of call.

SIP does not operate the model of IN call directly for access to IN services, the idea, is then to exploit the machine of states of the SIP entity with the layer IN, Such that the acceptance of call and the router are executed by

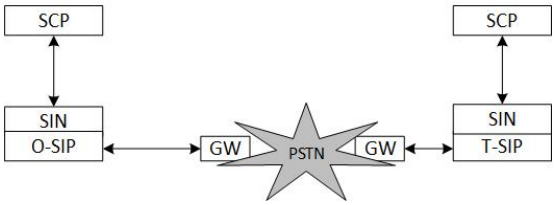


Figure 1: Interconnection SIP-IN

the native states and the services are accessed to layers IN with the model of IN call [7].

The model of service programming with SIP consists therefore of adding on SIP, an IN layer, that manages the interconnection. This model is called SIN (SIP Intelligent Network) [12] [2]. This operation necessitates the definition of a correspondence between the model of call in IN and the model of call in SIP. This correspondence is between the states machine of the SIP protocol (SIP defines the letterhead *Record_Route* that allows to order SIP to function in mode with states until liberation of the call [9]) and the states machine in IN. A call will be processed by the two machines, the sip machine processes the initiation of call and the final reply to deliverance, and the IN layer to act with the intelligent node SCP for providing services during the processing of call [2]. Figure 1 illustrates the integration SIP-IN.

Similarly to the IN states machine, two states we have, one defines originating-SIP (O-SIP) and Termination-SIP (T-SIP), that are the entities correspond, respectively, to the O_BCSM and T_BCSM of the model of call IN [2].

3 Intelligent Network Security with SIP

3.1 The Threats of Intelligent Network

The analysis of risk in the IN is depended on the different possibly of the realization in the distributed functional plan (DFP) and in the physical plan. Although enough number of interfaces are interest for the no authorization of third elements in an environment IN commercial, such that the interface between the supplier of service and the SMS, or the interface between the system of billing and the SMS.

The main security problem is in the multimedia communication area and in the IN system. Bellow some essential security problems that threats Intelligent network:

- User and terminal authentication: suppliers of a service need to know that uses their service to be able to post and possibly to correctly invoice the utilization of the service. In view of the authentication, the user and/or the terminal has first to identify with a certain identity, then to prove that the declared identity is the real identity. In order that, it is generally made call in procedures of reliable authentication (for example, password protected or numerical signature X.509). Similarly, users can wish to know that are their correspondents.

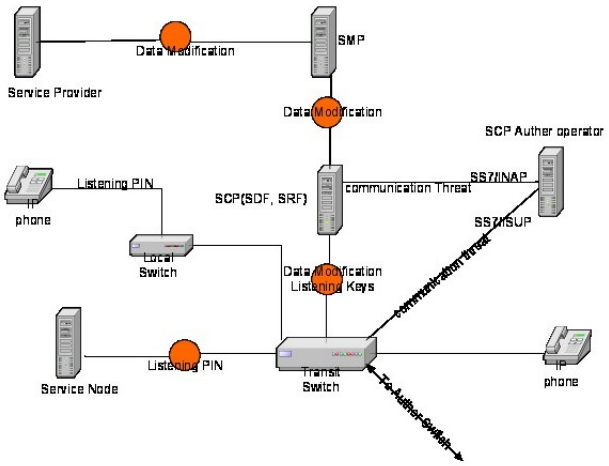


Figure 2: Threats in IN

- Server authentication: as users communicate generally between them by means of a certain infrastructure of telephony over IP with some servers (gateways, units of multi-diffusion, bridges), they wish to know if they are connected to the correct sever and/or to the supplier of correct service. This aspect concerns users fix or mobile.

The list is not complete in practice, nevertheless, we can be confronted with other problems of security that are considered as not belonging of the area of application protocol (for example problems linked to the policy of security, to the security of system management, to the placement of the security mechanism, to the security of implementation, to the operational security or to the processing of security incidents). As well as the technologies evolution on the soft plan and in the mechanisms of penetration and hacking does not cease to increase a manner not estimable. Figure 2 represents some places of threats in the IN system.

3.2 Security Mechanisms in SIP

The mechanism that provides security in SIP can be classified in two categories of protection: terminal to terminal or protection proxy -by - proxy [9]. The mechanism Terminal to terminal imply the SIP user agent and/or the visitor. The call is realized by a specific characteristic of the SIP protocol. for this mechanism we propose (SIP authentication and SIP body message crypt). The mechanism proxy-by-proxy ensures the communication between two SIP entities in covers it message of signalling. SIP does not provide specific characteristics for the protection proxy-by-proxy and use on the security problem, the standard network security protocols such as IPsec or SSL. The

mechanism proxy-by-proxy is required because intermediate elements can play an active role in the SIP processing by reading and/or writing some SIP messages.

As SIP heavily borrows HTTP for its syntax of messaging, it can also employ the model of authentication employed by http [11]. In addition, the authentication using the mechanism CHAP is proposed. The SIP body message and some headers is also suggested, but some headers must be in the transparent text. The other choice would be employ the PGP for the SIP message crypt. For the media crypt, keys of session could have been exchanged as left of the session description protocol (SDP), but that would necessitate that, the message of SIP must be crypted. In addition, [8] suggests employing a standard security protocol, such that IPsec or RTP security.

IPsec is a mechanism of network layer that can be employed to introduce the security directly on the IP layer. Habitually IPsec is employed to provide the security based on the identity of network node, and it is independently to the SIP architecture. In order that. IPsec can be employed essentially in SIP, between SIP entities with a specific and static security association.

Although mechanisms of security provide by SIP can reduce the risk of attack, there are some limitations in extended them of mechanisms that must be considered. The first limitation is with the employ of HTTP. The mechanism of integrity in HTTP does not work very well for SIP, it offers the protection only for some SIP parameters. The second, HTTP necessitates that pre-existing a security association that can be employed in SIP server where user is configured.

The problem of the agreement on the chosen security mechanism between two SIP entities (user agent and/or proxy) that want to communicate by applying a level of security sufficient is very important. As already tells, SIP has some numbers of security solution, some of them are directly defines by SIP and others are derived by low protocols (TLS, IPsec, etc.). For this reason, it is very important to define how a SIP entity can select an appropriate mechanism to communicate with a next entity of proxy [1]. When a client initiates the procedure, the SIP agent includes in the first request sent to the neighbour proxy entity the list of its mechanisms of security sustained. The other element (server side) replies with a list of its clean security mechanisms and its parameters. The client selects then the common security mechanism prefers, to make use this chosen mechanism (ex, TLS), and contacts the server by using the new mechanism of security.

Using this mechanism, the client UAC, is capable to identify himself to a server UAS, to an intermediate proxy server or to a server of registration. Therefore, the SIP authentication applies only to the communications terminal to terminal or terminal to proxy; the authentication proxy-by-proxy must use another protocols mechanisms such as IPsec or TLS.

The procedure of SIP authentication is executed when the UAS, the intermediate proxy server, or the necessary

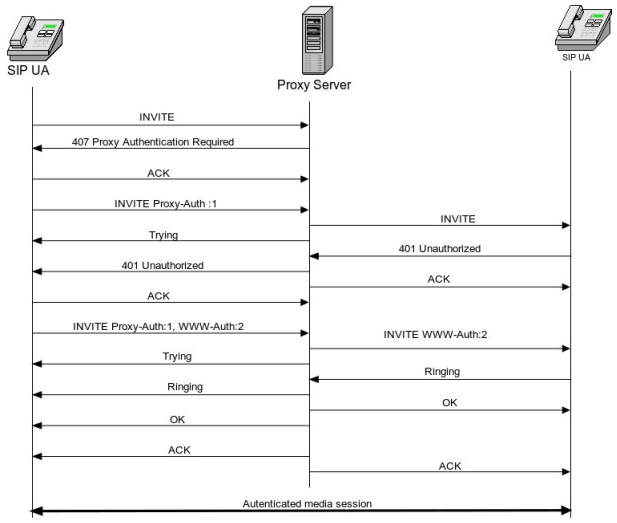


Figure 3: SIP Authentication

recording server for the call of the UAC must be authenticated before accepted call, sent the call, or accepted the recording. In the beginning, the UAS sends a request of SIP message (ex, INVITE). In the reception of this message, the UAS, proxy server, or sever of recording decides that the authentication is required and sent to the client a specific SIP message of the request of authentication. This message of error represents a challenge. In the case, where the message of error is 401 (Unauthorized) is sent by UAS and recording, if the message of error is 407 (Proxy Authentication Required) is sent by proxy server. The UAC receives the message of error, calculates the reply, and includes it in a new message of SIP request. Figure 3 shows the sequence of message for the case of request of authentication by the proxy server.

With this different securing mechanism, another mechanism is related to the key exchange, that is used by all SIP transaction between UAC. This mechanism is designate generally by ID IDENTIFIER of the session. This session key is incorporated in the SDP body message with INVITE and allows authenticating users.

Note that, the UAC send an ACK message immediately after that the message of error is received. This message closes the first transaction; then the second message INVITE opens a new transaction. The SIP message [8] allowed also the procedure based on the HTTP authentication. This authentication anticipates that username and password are sent in transparency. The SIP protocol also anticipates the employ of PGP, that could have been employed in the procedure of authentication. For the SPD, it must be referred for each IP packet, and the policy determines if the packet would have to be allowed the pipe of processing IPsec, or would have to be parted, or would have to be use in the IPsec processing. If the policy dictates that the packet must be processed by IPsec, then the point of SPD entry to a more SA (security association), that must be applied for the packet.

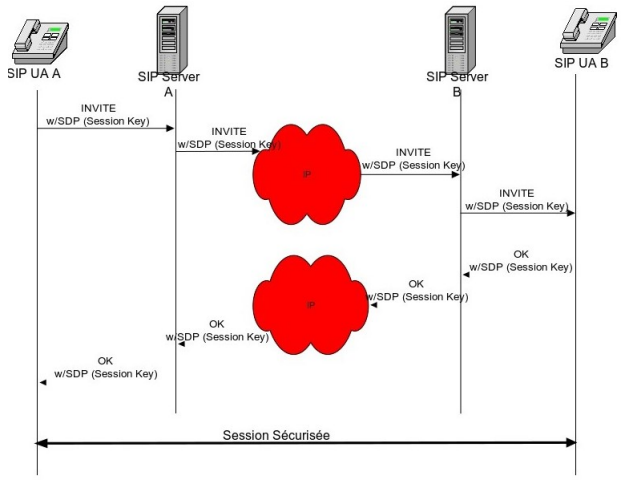


Figure 4: Security Mechanisms in Web service

4 Security Mechanisms in Web Service

The dotNet framework provides many classes allowing developing applications that consume and produce Web services. Microsoft proposes with Web Service Enhancement (WSE) [10] an improvement to the dotNet framework (See Figure 4). It allows the programmer to construct a securing Web service based on the last standards. WSE spread functionalities of security, router, and integration of enclosures in SOAP message (Simple Object Access Protocol). They allow the programmers to construct a applications based on the last specifications published by Microsoft and different industry actors (IBM, SAP,etc) such that WS-Security, WS-Policy, WS-Security policy, WS-Trust, WS-Secure Conversation and WS-Addressing. By using the WSE for Web service XML, we can:

- Secure an application all along a domain.
- Modify a transparent manner, to the level of nodes, the paths that can take SOAP message to arrive at the Web service.
- Attach a file with a SOAP message, during a communication between XML Web service, without serializing in XML.

4.1 WSE Architecture

A SOAP message is an envelope containing the body of a message and a header to describe it. The header is used to transmit information on the manner whose message must be processed. This information can be encrypted and authenticated. While the WSE (Web Service Enhancement) is contained:

- A class for implementing a new protocols (WS-Security,etc).
- A filter lodged by ASP.NET that intercepts SOAP message upstream and downstream.
- A context, that is the canal of communication between an application (Web service) and an infrastructure (filter). The class SoapContext puts in the header, the program state

The filter interprets or generates the header allowing it to take some functionalities. The WSE Filter presents functionalities of diagnostic, security, and routing. Similarly, WSE allows user to define and personalize filters (custom filters) in the input or output SOAP message. These filters make access to information contained in the SOAP message to determine how to process it. Several personalized filters can be defined in a same message. The SOAP packet exiting the client (respectively the Web service) will be intercepted by the WSE pipeline and treaty by the different filtering according to a following order: C̈ustom Output Filters; R̈outing Output Filter; S̈ecurity Output Filteränd finish by the T̈race Output Filter. This last once is activated, it allows to save in a file the track for all SOAP message over the WSE pipeline. The packet processed by the WSE, will be transmitted to the system in the form of a SOAP request (respective of a reply). When the packet arrives at a service (respectively at a client), it will again be intercepted by the WSE and processed by the same filters but in the opposite direction: the T̈race Input Filter; the S̈ecurity Input Filter; the R̈outing Input Filteränd the C̈ustom Input Filters.

4.2 WS Security

WSE uses mechanism define in WS-Security specification [6] to allow an application of Web services security and:

- Identify a user that asks a Web service (Authentication).
- Verify the role of the user and rights that are attributed it (Authorization).
- Ensure that a message has not been corrupted during the transportation (Integrity).
- Ensure that alone the owner of message can read it (Confidentiality).

The client must obtain a security token from a source (which must be trusted by the sender and the receiver). When a client sends a SOAP request, Security tokens are placed in the SOAP message. When the web server receives the request, the WSE pipeline verifies that the tokens are authenticated before sending the message to the web server without sending a request to the client to verify the integrity of the security token. The WSE provides 3 methods for secure a SOAP message:

- Security tokens to identify a user. Information allowing the implementation of the security is placed

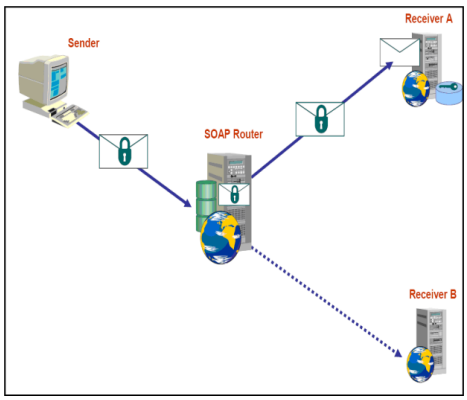


Figure 5: Encrypted Message WS

directly in SOAP header. The signature that secures the Web service by allowing a destination to verify that the message has not been modified since that it has been signed. This operation is used by a Username Token or a certificate X.509 [11].

- The digital signature that allows a recipient to verify that the SOAP message has not been modified since it was signed based on the XML Signature standard. Note that the digital signature can verify that the message has not been modified, but it does not encrypt the message; the message is transmitted in clear text.
- SOAP message encryption to guarantee the confidentiality of exchanges based on the XML Encryption standard. To encrypt a SOAP message, the sender must have the receiver's public key, which in turn must have its private key to decrypt the message. The WSE actually supports both types of symmetric and asymmetric encryption and by default the WSE encrypts the ¡Body ¿ part of a message, but you can specify which part of a SOAP message to encrypt.

5 Implementation Intelligent Network Security with SIP-Web Service

After the presentation of the different mechanisms allowing to secure the signaling protocol SIP and also the authentication of UAC, as well as methods that allow to secure data of a telecom services with WSE and WS-Security, we present in Figure 5, a logical and physical architecture for security of services, operators and users in an application of intelligent network over IP.

With the techniques already mentioned and the requirements in terms of security, we suggest that a physical structure for the implementation of a network security platform and intelligent services can be given such as that of Figure 6. The description of this architecture

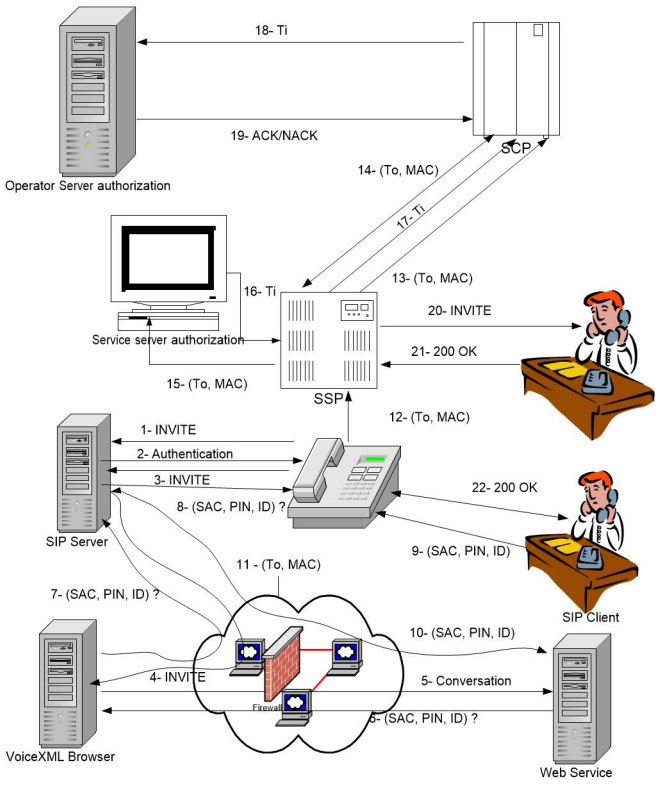


Figure 6: Implementing of Security Mechanism

is as follows: Each subscriber A has his service access code (SAC: Service Access Code), his PIN code (Personal Identification Number) and his session identifier ID [4]. A secure server calculates from these three parameters the message authentication code MAC and the ticket of the operator system associated to the subscriber A, TOA. The different requests are a secure SOAP message, with the WS-Security mechanism from the WSE. These SOAP messages are integrated in SIP request by the technique of combination SIP-SOAP [3]. The two parameters (MAC,To) will be verified through the SSP and after the SCP (user authentication). If the user is accepted, we check whether or not he has the right to a service. This verification is insured by a server of service authorization, SSA. Once, the access to the service is authorized, a ticket Ts is sent to the operator server authorization, OSA through the SSP and SCP. If the operator validates this ticket, the user will be to the service in question and a secure session will be established.

OSA: Operator server Authorization

SSA: Service Server Authorization

AKA: Authentication Key

MACA = f(ToA, AKA) : Message Authentication Code

ToA: g(SAC, IDA,) : Ticket operator

TsA: (h(ToA), MACA) : Ticket service

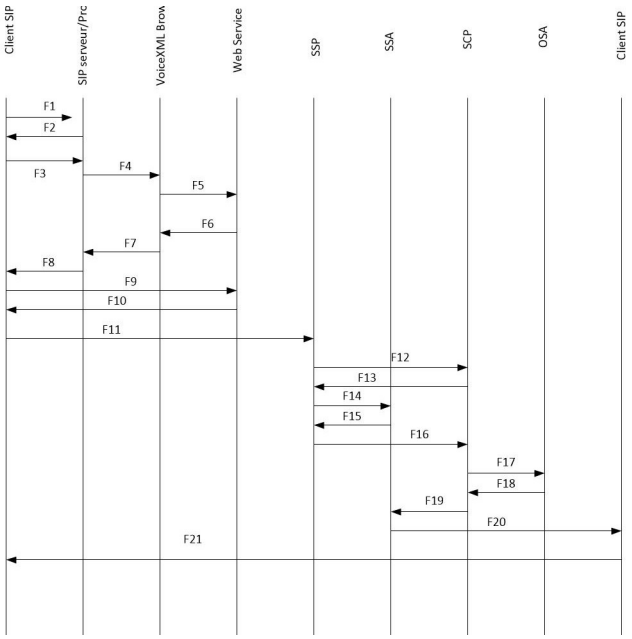


Figure 7: Diagram of Intelligent Network security

These operations:

SSA:

1) Verify if A is Authenticate
2) Calculate MACSOS for ToA
3) Generate the ticket TiA if MACSOS = MACA
4) Generate the ticket TiA
5) Send TiA to OSA

OSA:

1) Verify TiA
2) SCP authorizes or not

The diagram of the functioning of this mechanism is given in Figure 7:

F1 : INVITES SIP with authentication and ID
F2 : 407 Authentication Proxy
F3 :ACK
F4 : INVITES a session with the service
F5 :http://service1.dialog.vxml
F6 : to give your identifiers (SAC, PIN,ID)
F7 : to give your identifiers (SAC, PIN, ID)
F8 : to give your identifiers (SAC, PIN,ID)
F9 : Here is my identifiers
F10 : calculate To and MAC
F 11 : To and MAC transmitted to SSP
F12 : to SCP
F13 : validation or not
F 14 : If validated ; to send to SSA
F15 : calculate Ti
F16 : Ti to SCP
F17 : verify Ti by OSA
F18 : Validation or not of Ti
F19 : If to validate, request to SSP.
F20 : INVITE
F21 : OK.

6 Conclusions

The creation of services needs to give a clear overview to each user or to each subscriber, that its information and its orders will be processed correctly and that its requests will be processed in the network. To choose the security of user by SIP, the security of the network entity by WS-Security and the media security by SRTP, is based essentially on three reasons. First, a user has to be authenticated before its access to the different elements of network services, this is realized by SIP authentication operation. In the second, and after the authentication, the right of access to a service must to be verified, this is realized by the WS-Security and in third if a service is established, data of the service must to be also secured , this is realized by SRTP in the media case or WS-Security in the other cases.

The choose of the WS-Security is fundamental on two criteria. Firstly, WS-Security is a component of Web service. Secondly the WS-Security employs a standards algorithm of cryptography for these different operations of authentication, integrity and confidentiality such that MD-5, SHA-1, DES, AES,etc. Despite the different procedures of security employed in the different phases, one can overcome some insufficiency:

- SIP is a protocol of signalling, is not strongly conceived for such security operation.
- The techniques used by WS-Security necessitate a negotiation of a common security mechanism for an adequate transaction between two or several operators.

References

[1] J. Arkko, V. Torvinen, G. Camarillo, A. Niemi, T. Haukka, "Security mechanism agreement for the session initiation protocol (SIP)," *draft-ietfsip-sec-agree-04.txt¿, June 2002*, 2002.

[2] V. Gurbani, F. Haerens, and V. Rastogi, "Interworking SIP and intelligent network (in) applications," tech. rep., 2005.

[3] D. A.Handoura, "Combinaison de SIP et de soap via les web services," *3ème Journées scientifiques de Borj Elamri JS*, 2003.

[4] L. Herrigel.A, "Authentication and authorisation in the in," in *R3 Security Engineering AG IEEE IN '94*, 1994.

[5] ITU, *Secteur de la normalisation des télécommunications de l'UIT: Sécurité dans*

les télécommunications et les technologies de l'information. UIT, 2003.

[6] R. Kenneth.C, Madhusudhan.G, "Investigating the limits of soap performance for scientific computing," in *Proceedings of the 11 th IEEE International Symposium on High Performance Distributed Computing HPDC*.

[7] S. Pierre, *Réseaux et systèmes informatiques mobiles*. Presses internationales Polytechnique, 2003.

[8] J.Rosenberg, "SIP security," 2000. (`http://www.dynamicsoft.com/resources/pdf/SIP2000-Security.pdf`)

[9] V. Salsano.S and Papalilo.D, "SIP security issues: The SIP authentication procedure and its processing load," *IEEE Network 0890-8044/02/*, 2002.

[10] P. Sher.M, Magedanz.T, "Inter-domains security management (idsm) model for IP multimedia subsystem (ims)," in *Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06)*, 2006.

[11] M. Veltri.L, Salsano.S, "Wireless lan-3g integration: Unified mechanisms for secure authentication based on SIP," in *IEEE ICC 2006 Proceedings*, 2006.

[12] W. Wang, S. Cheng, "Accessing traditional intelligent services from SIP network," in *International Conferences on Info-Tech and Info-Net. Proceedings (Cat. No. 01EX479)*, vol. 2, pp. 772–778, IEEE, 2001.

Biography

Abdallah Handoura. An Assitant Prof of Computer science in the College of Telecoms and Information in Riyadh, Kingdom of Saudi Arabia, Since 2013. He Received his B.S from the Normal School of Sciences and Techniques of Tunis, ENSET in 1994, And his Master degree in Automatic (Artificial intelligence and speech recognition) in 1997, from the Higher School of Sciences and Techniques of Tunis, ESSTT, and his PhD in Information technologies and Communication from the Institut-Mines telecom Atlantique (ex: ENST-Bretagne) in 2009. The field of his research in PhD is Information and Network security in NGN.

Daniel Bourget. An Assitant Prof of Computer science in the IMT-Atlantique sine 1998, France. His research focuses in Web services, NGN security and web security.

A Fault Recognition Method Based on Convolutional Neural Network

Lei Chen[1], Jiaqi Shi[1], and Ting Zhang[2]
(Corresponding author: Lei Chen)
Computer School, Beijing Information Science and Technology University, Beijing 100101, China[1]
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China[2]
Email: ch913@126.com
(Received May 30, 2023; Revised and Accepted Jan. 4, 2024; First Online June 22, 2024)

Abstract

Fault recognition is an important part of seismic interpretation, but the existing methods' accuracy is not high enough. Convolutional neural networks have achieved high accuracy in handwriting recognition. Based on the similarity between fault and handwriting shape features, this paper proposes a fault recognition method based on the classical convolutional neural network. Firstly, a neural network model suitable for fault recognition is designed based on the successful LeNet5 model, which has been used for handwriting recognition. The output layer of the neural network is designed with two neurons to judge whether the seismic sample points belong to faults. Additionally, the Softmax Regression model is used to replace the original European radial basis function. This modification allows the network to output not only whether the seismic sample points belong to faults, but also the probability of belonging to or not belonging to a fault. Then, a 3D sample set and test set are established, using the most accurate manual fault recognition method, to train the neural network. Finally, the proposed method is tested on actual seismic data, and the experimental results confirm the effectiveness and progressiveness of the proposed approach.

Keywords: 3D Sample; Convolutional Neural Network; Fault Recognition; Seismic

1 Introduction

Fault recognition is an important section in seismic interpretation. The traditional method involves manually selecting the discontinuous sampling points in the seismic section and connecting them into curves. While this method can accurately identify faults, it is time-consuming and subjective. To overcome the shortcomings of the manual method, various fault recognition methods have been proposed since the early 20th century. Generally, these methods utilize discontinuities such as correlation coefficients, gradients, or variances between seismic samples from different seismic traces in seismic data to identify faults.

The classical coherence cube algorithm has devolved into its third generation, referred to as C1-C3, respectively. It is considered the earliest automatic fault identification method. Bahorich *et al.* proposed C1, where faults were extracted based on the correlation value between neighboring three seismic traces using classic normalized cross-correlation [2]. Marfurt *et al.* proposed C2, which increased the number of seismic traces used in correlative computation to those contained in a cuboid or ellipsoid [8]. Gersztenkom *et al.* proposed C3, where the eigenvalues of the covariance matrix of seismic traces within an analysis window were used for computing coherence attributes [5]. These widely-used coherence algorithms have their own advantages and disadvantages. Experiments have confirmed that C1 works faster but is sensitive to coherent noise interference. C2 provides better accuracy and noise resistance compared to C1 but requires more computational work. C3 is more efficient compared to C2 but cannot determine the exact fault location. Additionally, all three algorithms exhibit a common feature where many small faults can be easily missed.

In addition to the Coherence Cube algorithm, the Ant Tracking algorithm is famous in the field of automatic fault identification, and it has been applied in seismic data interpretation software. Based on Ant Colony Optimization (ACO), the Ant Tracking algorithm identifies faults by ants crawling on fault-like points in seismic data [9, 10, 18]. This method can detect micro faults. However, many fault-like points that should be judged as horizons are mistaken as faults. Additionally, this method has a high computational cost. Besides the Coherence Cube and Ant Tracking algorithms, other automatic fault identification methods have also been proposed. Wang *et al.* proposed a fault recognition method based on the Hough transformation [1, 16]. The discontinuities were computed with C2 and the likely fault regions were highlighted by thresholding discontinuities. Then, the Hough transformation was utilized for fault recognition. How-

ever, this method cannot work well when the number of faults in a seismic section is greater than one. Wang *et al.* also proposed a fault recognition method based on directional complex-valued coherence attribute. The Hilbert transform of a real seismic trace was used as the corresponding complex-value seismic trace. Then, the coherence value between adjacent weighted complex-value seismic traces along multiple azimuths was calculated, and the point with the minimum value was judged as a fault [14].

The mean effect still exists, and some horizontal information is left. Xiong *et al.* extracted three small 2D seismic sections in the seismic data volume in three directions: horizontal, vertical, and axial, respectively, as the three color components of a color image (RGB - Red, Green, Blue). They then attempted to identify faults using a five-layer Convolutional Neural Network (CNN) [17]. However, the obtained fault reliability was low due to the low accuracy of the training samples. Dou *et al.*introduced λ-binary cross-entropy (BCE) and λ-smooth L1 loss to train a three dimensional (3D)- CNN using a few slices from a 3D seismic volume label. This method aimed to reduce the huge workload required for image segmentation and proposed an attention module that can be used for active supervision training. This attention module was embedded in the network to suppress seismic noise [3]. However, the accuracy of the fault detection results obtained was low, and some horizons were incorrectly identified as faults. Shafiq *et al.*presented an approach for detecting faults within seismic volumes using a saliency detection framework that employed 3D-Fast Fourier Transform (FFT) local spectra and multi-dimensional plane projections [11]. This method effectively detected complex fault networks that were hardly conspicuous within the original seismic volume. However, the accuracy of the fault detection results was difficult to guarantee, as some shorter bending horizons were wrongly classified as faults, and the results were accompanied by a lot of noise. Mahadik *et al.*used multispectral coherence to characterize faults [7]. By employing spectral decomposition, spectral balance, and statistical fusion of coherence images, they were able to obtain more refined and sharper fault characteristics. However, the results were noisy, and the accuracy was still insufficient. Other methods, such as [12, 15, 19], have also been proposed.

Handwritten numerals display characteristics of significant variations in the structure of the same numeral, similar to faults in seismic data. The LeNet5 convolutional neural network has achieved high accuracy in handwriting recognition. Based on this, this study aims to identify faults in seismic data using a convolutional neural network model [4, 13]. This paper initially introduces the principle of the LeNet5 model. Then, it presents the design of the neural network model proposed for fault recognition. It also introduces the Softmax Regression algorithm and cross-entropy loss function, along with the methods used for generating training and test sample sets. Finally, the experimental results and analysis on actual seismic data are presented.

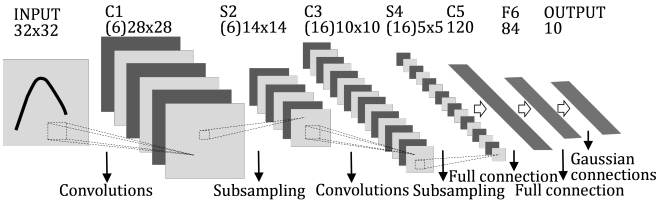


Figure 1: LeNet5 model framework and digital recognition example

2 LeNet5 Neural Network Model

The LeNet5 model was proposed by Professor Yann LeCun in 1998 [6]. It was the first convolutional neural network successfully applied to the problem of number recognition. In the MNIST (Modified National Institute of Standards and Technology database) data, it achieves an accuracy of about 99.2%. This section will provide a detailed introduction to it.

2.1 The LeNet5 Network Structure

The LeNet5 model framework is show in Figure 1. The LeNet5 model is a 7-layer deep convolutional neural network. The convolutional layer enhances the original signal features and reduces noise. The pooling layer uses the principle of image correlation to subsample the image, reducing the number of parameters and overfitting of the model, while retaining useful information. The functions of each layer are as follows:

1) Convolution layer (C1) The input of the first convolution layer is the original image, and the image size is 32×32. The size of the convolution core is 5×5, the depth is 6, the full 0 complement is not used, and the step size is 1. Since the full 0 complement is not used, the output size of this layer is 28, and the depth is 6. This convolution layer has a total of 156 parameters, which can be calculated by the formula $5 \times 5 \times 1 \times 6 + 6$, of which the last added 6 is the number of the bias term parameters. The number of parameters of the convolution layer is only related to the size and depth of the convolution kernel and the depth of the node matrix of the current layer. Since the node matrix of the next layer has 4704 i.e. $28 \times 28 \times 6$ nodes, and each node is connected to 25 i.e. 5×5 nodes of the current layer, the convolution layer has 122304 i.e. $4704 \times (25 + 1)$ total of connections.

2) Pooling layer (S2)

   The output of the first layer is a node matrix of $28 \times 28 \times 6$, which is used as the input of this layer. The size of the convolution kernel in this layer is 2×2, and the step length is 2, so the output matrix size of this layer is $14 \times 14 \times 6$.

3) Convolution layer (C3)

The size of the input matrix of this layer is $14\times14\times6$, the size of the convolution kernel used is 5×5, the depth is 16, and the step size is 1 without all zeros complement. The output size of this layer is $14-5+1$, and the depth is 16, that is, the output matrix size is $10\times10\times16$. This layer has 2416 i.e. $5\times5\times6\times16+16$ parameters and 41600 i.e. $10\times10\times16\times(5\times5+1)$ connections.

4) Pooling layer (S4)

The input matrix size of this layer is $10\times10\times16$, the convolution kernel size is 2×2, the step size is 2, and the output matrix size of this layer is $5\times5\times16$.

5) Full connection layer (C5)

The size of the input matrix of this layer is $5\times5\times16$. In the LeNet5 model paper, this layer is called the convolution layer. But because the size of the convolution core is 5×5, it is not different from the full connection layer. Here, it is directly regarded as the full connection layer. The input of this layer is a $5\times5\times16$ matrix, which is straightened to a vector of length of $5\times5\times16$, that is, a three-dimensional matrix is straightened to a one-dimensional space and expressed in the form of a vector, so that it can enter the full connection layer for training. The number of output nodes in this layer is 120, so there are 48120 i.e. $5\times5\times16\times120+120$ parameters in total.

6) Full connection layer (F6)

The number of input nodes in this layer is 120, and the number of output nodes is 84, with a total of 10164 i.e. $120\times84+84$ parameters.

7) Full connection layer (OUTPUT)

The number of input nodes in this layer is 84, and the number of output nodes is 10. There are 850 i.e. $84\times10+10$ parameters in total. This layer is the output layer, also known as the Gaussian connection layer, which is composed of Euclidean Radial Basis Function (RBF) units. Each category of output corresponds to a unit, and each unit has 84 inputs. In other words, each output RBF cell calculates the Euclidean distance between the input vector and the parameter vector. The farther the input is from the parameter vector, the larger the RBF output. An RBF output can be understood as a penalty term that measures the matching degree between the input pattern and a model associated with RBF. In terms of probability, RBF output can be understood as the negative logarithmic probability of Gaussian distribution in the sixth layer configuration space. Given an input mode, the loss function should enable the configuration of the sixth layer to be close enough to the RBF parameter vector (i.e. the expected classification of the mode).

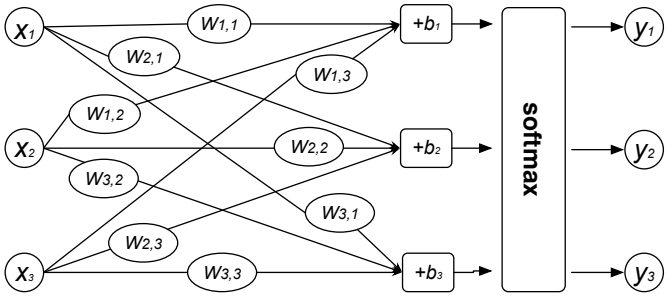


Figure 2: Softmax regression flow chart

The above is the structure of LeNet5 model. The initial features of LeNet5 are as follows:

a. Each convolution layer consists of three parts: convolution, pooling and nonlinear activation function;
b. Using convolution to extract spatial features;
c. Average pooling layer with subsampling;
d. Hyperbolic tangent (Tanh) or S-type (Sigmaid) activation function;
e. Multilayer neural network (or multi-layer perceptron, Multi Layer Perception, MLP) as the final classifier;
f. The sparse connection between layers reduces the computational complexity.

2.2 Softmax Regression Algorithm and Cross Entropy

When performing multi-classification tasks such as image recognition and text classification, the number of neurons in the output layer of the neural network is generally determined by the number of categories we want to classify. Typically, we utilize the Softmax Regression model (also known as the Softmax function) to convert the output of the neural network into probabilities. The flow chart of Softmax Regression model is shown in Figure 2.

The Softmax Regression model is used to normalize the output components corresponding to each category, so that the sum of each component is 1. This can be understood as transforming the output element values of the network output vector into the probability of classifying the input samples into each category. The calculation of Softmax Regression algorithm is shown in Formula (1):

$$\text{softmax}(x) = \text{normalize}(e^x) \tag{1}$$

The probability to be judged as class i is shown in Formula (2):

$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2}$$

After the probability value that the input sample is determined as each category is calculated, the category to which the corresponding input sample belongs can be determined according to the maximum probability value in the output vector.

Figure 3: MNIST handwriting dataset

After the output of the neural network is converted into a probability value by the Softmax Regression algorithm, the loss function value is calculated next. Before neural network training, loss function should be defined first. The smaller the loss function, the smaller the deviation between the classification result of the representative model and the real value, that is, the more accurate. At the beginning, the model is filled with all zero parameters, and there will be an initial large loss function value. The purpose of training is to continuously reduce the loss function value until a global or local optimal solution is achieved. For multi classification problems, cross entropy is usually used as the loss function. Cross entropy originated from information entropy in information theory (related to compression ratio, etc.), and then was applied to many aspects, including communication, error correction code, game theory, machine learning, etc. The definition of cross entropy is shown in Formula (3):

$$H_{\mathbf{y}'}(\mathbf{y}) = -\sum_i \mathbf{y}'_i \log(\mathbf{y}_i) \tag{3}$$

where, $\mathbf{y}$ is the predicted probability distribution and $\mathbf{y}'$ is the real probability distribution (i.e. one pot coding of Label). Usually, the value of cross entropy can be used to judge the accuracy of the model's estimation of the true probability distribution.

3 Fault Recognition Based on Convolution Neural Network

3.1 Neural Network Model Design

After using Softmax Regression model, multi-layer perceptron, Drop out, ReLU (Rectified Linear Unit) and other technologies, LeNet5 can achieve a recognition accuracy of 99% (the current latest convolutional neural network based method can reach 99.8%) for the handwriting dataset MNIST (Mixed National Institute of Standards and Technology database) collected by the National Institute of Standards and Technology. Partial screenshots of MNIST dataset are shown in Figure 3.

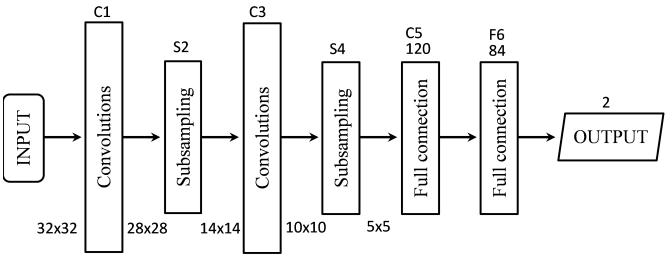


Figure 4: Fault recognition neural network model

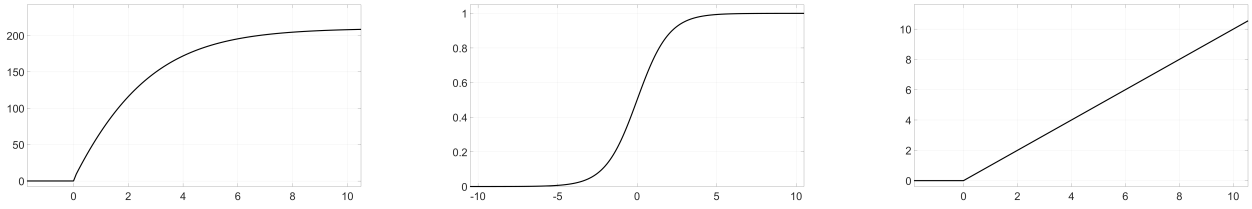
It can be seen from Figure 3 that the handwriting in MNIST is characterized by a large difference in the structure of different handwriting for the same number, which is similar to the fault in seismic data. Based on this, this study identifies faults in seismic data based on LeNet5 model, but it needs to make appropriate modifications and parameter settings to LeNet5 model to adapt to the purpose of fault identification and the needs of samples generated by seismic data. The neural network model used for fault identification in this study is shown in Figure 4.

It can be seen from Figure 4 that there are two major modifications to the fault recognition neural network model proposed in this paper compared to LeNet5. First, the number of neurons in the output layer has been changed from 10 to 2. This is because there are a total of 10 categories from 0 to 9 in the handwritten dataset. For the samples in the seismic data, there are only two categories: fault and not fault. The second modification is to change the Gaussian connection, namely the European radial basis function of the output layer, to the Softmax Region model. This study hopes that the neural network can not only output the judgment of whether the seismic sample points are faults but also output the probability of whether they are faults or not. In addition to the above modifications in the neural network structure, other aspects of fault identification are designed as follows:

1) Optimization algorithm and learning rate

   Adaptive moment estimation (Adam) optimization algorithm is adopted. Given an initial learning rate, Adam optimization algorithm can dynamically adjust the learning rate for each parameter according to the first-order moment estimation and second-order moment estimation of the gradient of the loss function for each parameter. This is an optimization algorithm to find the global optimum, and the quadratic gradient correction is introduced. Compared with the basic Stochastic Gradient Descent (SGD) algorithm, it does not lead to a large learning step because of a large gradient. The value of the parameter is relatively stable, and it is not easy to fall into local advantages. The speed is faster, which promotes the dynamic adjustment of the super parameter.

2) Activation Function and Loss Function



(a) Brain neuronal activation model (b) Sigmoid activation function (c) ReLU activation function.

Figure 5: Comparison of neuronal activation models

The ReLU activation function is used in the first and third convolution layers and the fifth and sixth fully connected layers, and the cross entropy shown in Formula (3) is used as the loss function. The reason why ReLU activation function is adopted is that it solves the problem of gradient dispersion from the front, instead of bypassing by unsupervised layer by layer training initialization weights. Compared with the traditional sigmoid function, ReLU has the advantages of unilateral inhibition, relatively wide excitation boundary, sparse activation, and can improve training speed and model accuracy. These advantages can be seen intuitively from the comparison of neuron activation models shown in Figure 5.

Corresponds to the function curve in Figure 5b, the formula definition of the sigmoid function is shown in Formula (4):

$$S(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}} \tag{4}$$

The derivative of x can be represented by itself as shown in Formula (5):

$$\begin{aligned} S'(\mathbf{x}) &= \frac{e^{-\mathbf{x}}}{(1+e^{-\mathbf{x}})^2} \\ &= S(\mathbf{x})\left(1-S(\mathbf{x})\right) \end{aligned} \tag{5}$$

Different from Sigmoid function, ReLU function generally refers to the ramp function in mathematics, as shown in Formula (6):

$$f(\mathbf{x}) = \max(0, \mathbf{x}) \tag{6}$$

In the neural network, ReLU is used as the activation function of neurons. For the input vector from the upper neural network entering the neurons, the output of neurons after ReLU function is shown in Formula (7):

$$f(\mathbf{x}) = \max\left(0, \mathbf{w}^{\intercal}\mathbf{x} + b\right) \tag{7}$$

3) Overfitting Problem

During training, Dropout technology was enabled for the sixth full connection layer. This technology is a simple but very effective method proposed by Professor Hinton's team to solve the over fitting problem. During training, part of the output node data of a certain layer of the neural network is randomly discarded. It can be understood that 50% of the points in an image are deleted randomly (that is, 50% of the points are changed into black points randomly). Even in this way, people are likely to recognize the category of the image. Of course, machines can also. This approach is essentially equivalent to creating a lot of new random samples, and preventing over fitting by increasing the sample size and reducing the number of features.

3.2 Sample Set Generation

After designing the neural network model for fault identification, optimization algorithm, loss function, activation function and other key parts, it is also necessary to obtain training sample set and test sample set from seismic data to train the neural network model and identify faults. In this paper, the training sample set and test sample set are generated from the 300 frame seismic section data containing faults in a work area in western China provided by Petro China Company Limited. Among them, 100 seismic sections whose size is 301 × 301 and with multiple uncrossed faults are used to generate training sample sets; 100 seismic sections containing a single fault and 100 seismic sections containing multiple cross faults whose size are all 301 × 101 were used to generate test sets. The specific generation procedure of sample set is as follows:

1) Training set generation

The training sample set in this paper is obtained from the seismic data volume shown in Figure 6. For this data, we have asked seismic data interpretation experts to give the fault identification results manually. For example, for the first profile, the results of manually picking up faults are shown in Figure 7.

With the seismic data shown in Figure 6 and the fault manual picking results shown in Figure 7, the training sample set can be generated. The specific method is: traverse the image obtained by manually picking up the fault (instead of the coherent volume image adopted by Xiong *et al.*[17]), if the pixel value is 0, the corresponding sample point does not belong

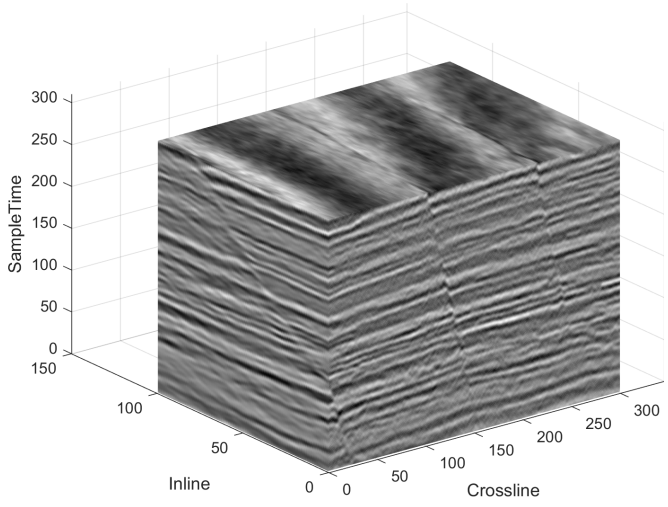


Figure 6: Seismic data volume for training set sample generation

to a fault. In the seismic data, three 2D seismic data sections with the size of $n \times n$ (affecting the final fault resolution) in the horizontal, vertical and axial directions centered on the sample point are synthesized into an RGB color image for output, which is used as a non fault sample, the label is 0; Similarly, if the pixel value is 1, a fault sample will be output, with the label of 1. The extraction of sample components is shown in Figure 8. The process of synthesizing samples from sample components is shown in Figure 9. When the number of fault samples and non fault samples reaches the required number, stop traversing the manually picked fault result image and seismic data volume, and the training set samples are created.

2) Test Set Generation

The test set used in this paper is obtained from the seismic data volume with one fault and multiple cross faults shown in Figure 10 and Figure 11.

The generation method of test set is similar to that of training set. The difference is that RGB color samples are output in the order of traversing the seismic data volume, stored in the same folder, and no longer have label information like the training set (fault and non fault samples of the training set are stored in different folders, and the folder name is the label).

4 Experiment and Analysis

To verify the effectiveness of the fault recognition method based on convolutional neural network proposed in this paper, 60000 training samples (30000 fault and 30000 non fault samples respectively) were generated with 100 seismic section data containing multiple uncrossed faults in a work area in western China provided by Petro China, with the sample size of 28×28. After the neural network model is trained, it is tested on the seismic sections (100 frames each, whose size are all 301x101) containing a single fault and multiple cross faults. The experimental tool is Tensor Flow, the second generation artificial intelligence learning system developed by Google. The initial value of learning rate during training is 0.001, the number of training is 10, the Dropout ratio is 0.5, and the batch size is 101. The experimental results are as follows:

The changes of accuracy and loss function with the batch order during training are shown in Figure 12 and Figure 13, respectively. The changes of accuracy rate and loss function with training times are shown in Figure 14 and Figure 15, respectively. After training the neural network model, the test was performed on a test set generated by seismic data containing single faults and multiple cross faults. The obtained fault recognition results are shown in Figure 16 and Figure 17 respectively.

It can be seen from the changes in accuracy during training shown in Figure 12 and Figure 14 that the convolutional neural network model designed in this paper can quickly achieve 99% accuracy during training. The value of the loss function quickly decreases to below 0.2. This shows that the neural network model can effectively extract fault-specific feature information to differentiate between faults and non-faults. This research also includes statistics on the accuracy of the test set by concealing label information, which can reach over 90%. However, Xiong *et al.*[17] pointed out in their research that the training and verification accuracy of the convolutional neural network model they used is about 73%. The above experiments on real seismic data verify the effectiveness of the fault recognition method based on convolutional neural network proposed in this paper. If we aim to detect small faults, it is necessary to train and test the neural network using smaller samples.

5 Conclusion

Based on the similarity between the irregular features of the fault and handwriting, and the success of LeNet5 in handwritten recognition, this paper proposes a fault recognition method based on a convolutional neural network. Firstly, the principle of the LeNet5 model is introduced, including the network structure, softmax regression algorithm, and cross-entropy. Then, the neural network model designed in this paper for fault identification is introduced, including the network structure, activation function, and loss function, as well as the generation method of the neural network training set and test set. Finally, the experiment and analysis on actual seismic data are provided, and the rationality of the obtained experimental results is explained through the theoretical formula of convolution and correlation operation, thereby verifying the effectiveness and feasibility of the proposed fault identification method. The future research direction is to effectively increase the hidden layers of the neural network and design more effective activation

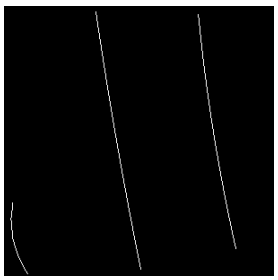


Figure 7: Manual fault picking results of training set seismic data volume

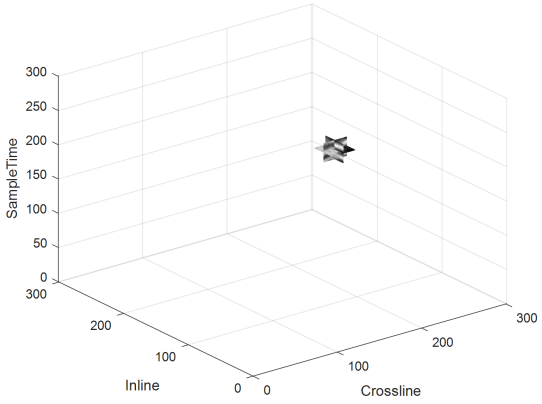


Figure 8: Sample component extraction

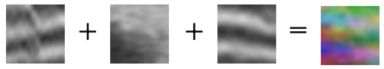


Figure 9: Synthesis of RGB color samples

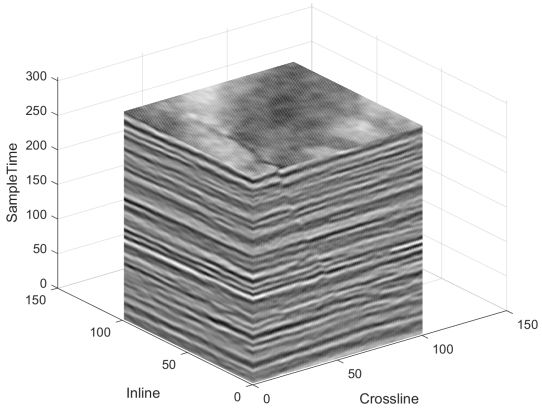


Figure 10: Seismic data volume containing a fault

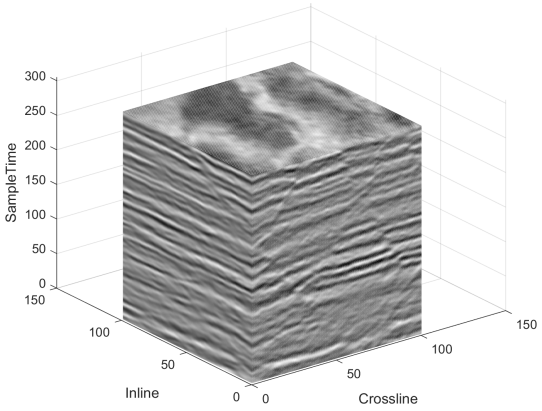


Figure 11: Seismic data volume containing multiple cross faults

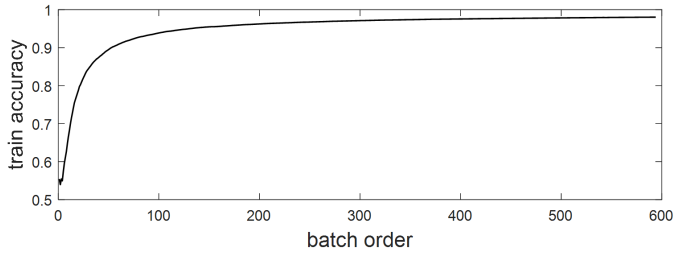


Figure 12: Relationship between train accuracy and batch order

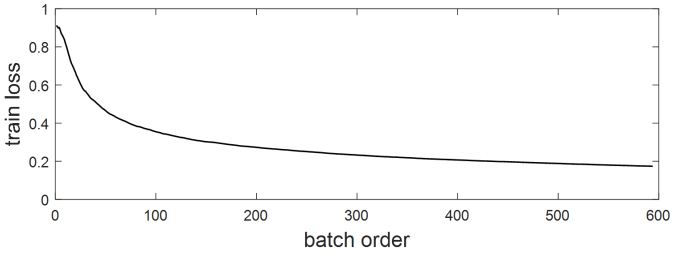


Figure 13: Relationship between train loss and batch order

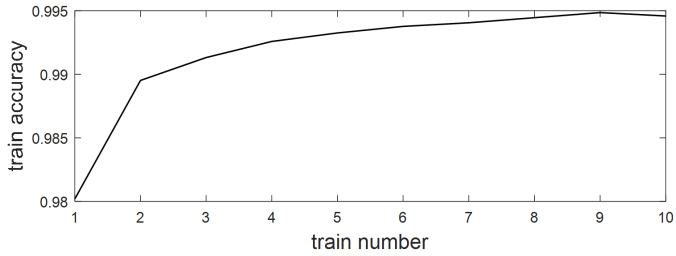


Figure 14: Relationship between train accuracy and train times

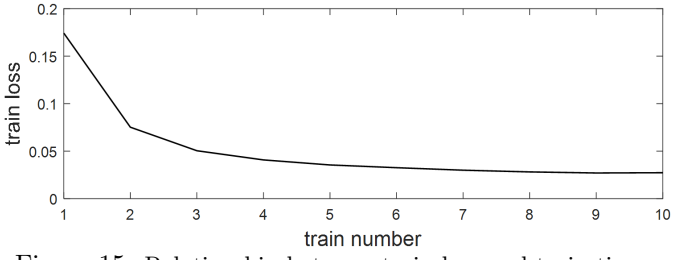


Figure 15: Relationship between train loss and train times

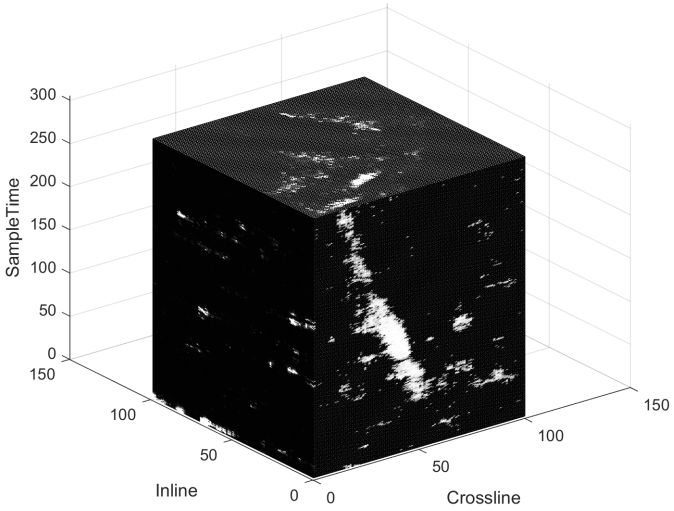


Figure 16: Fault results on seismic data containing a single fault

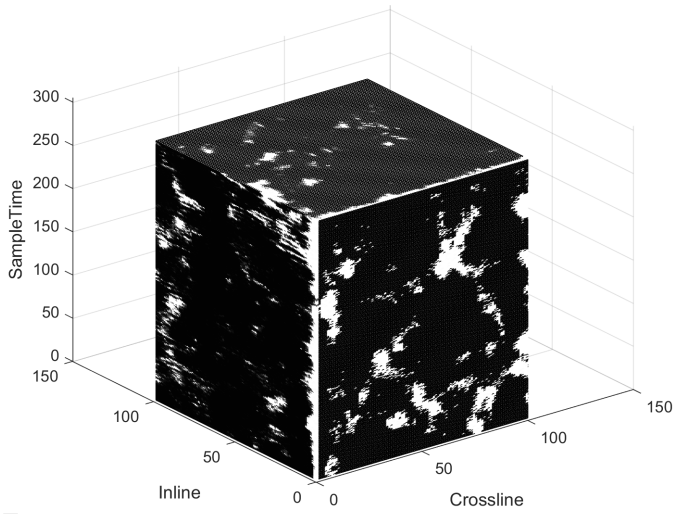


Figure 17: Fault results on seismic data containing multiple cross faults

functions to extract more advanced and effective fault features, thereby further improving the accuracy of fault recognition.

Acknowledgement

This work is supported in part by National Natural Science Foundation of China under grant 41804135, Key Laboratory of Petroleum Resources Research, Institute of Geology and Geophysics, Chinese Academy of Sciences, Open Project under grant KLOR2018-9, and Beijing Information Science and Technology University Research Fund Project under grant 2025025.

References

[1] N. M. Albinhassan and K. Marfurt, "Fault detection using hough transforms," *Seg Technical Program Expanded Abstracts*, p. 2452, 2003.

[2] M. S. Bahorich, "Stratigraphic and structural interpretation with 3-d coherence," *SEG Technical Program Expanded Abstracts*, vol. 14, no. 1, p. 1566, 1996.

[3] Y. Dou, K. Li, J. Zhu, X. Li, and Y. Xi, "Attention-based 3-d seismic fault segmentation training by a few 2-d slice labels," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.

[4] T.-T. Gao, H. Li, and S.-L. Yin, "Adaptive convolutional neural network-based information fusion for facial expression recognition," *International Journal of Electronics and Information Engineering*, vol. 13, no. 1, pp. 17–23, 2021.

[5] A. Gersztenkorn and K. J. Marfurt, "Eigenstructure-based coherence computations as an aid to 3-d structural and stratigraphic mapping," *Geophysics*, vol. 64, no. 5, p. 1468, 2012.

[6] Y. Lecun and L. Bottou, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] R. Mahadik, G. Singh, and A. Routray, "Multispectral coherence analysis for better fault visualization in seismic data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[8] K. J. Marfurt, R. L. Kirlin, S. L. Farmer, and M. S. Bahorich, "3-d seismic attributes using a semblance-based coherency algorithm," *Geophysics*, vol. 63, no. 4, p. 1150, 1998.

[9] S. I. Pedersen, T. Skov, T. Randen, and L. Sønneland, "Automatic fault extraction using artificial ants," 2002.

[10] T. Randen, "Automatic extraction of fault surfaces from three-dimensional seismic data," *SEG Technical Program Expanded Abstracts*, vol. 20, no. 1, p. 551, 2001.

[11] M. A. Shafiq, Z. Long, H. Di, G. AI Regib, and M. Deriche, "Fault detection using attention models based on visual saliency," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1508–1512.

[12] G. Singh, R. Mahadik, W. K. Mohanty, and A. Routray, "Seismic fault analysis using curvature attribute and visual saliency," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2221–2224.

[13] J.-X. Tong, H. Li, and S.-L. Yin, "Research on face recognition method based on deep neural network," *International Journal of Electronics and Information Engineering*, vol. 12, no. 4, pp. 182–188, 2020.

[14] S. Wang, S. Yuan, B. Yan, Y. He, and W. Sun, "Directional complex-valued coherence attributes for discontinuous edge detection," *Journal of Applied Geophysics*, vol. 129, pp. 1–7, 2016.

[15] Y. Wang and W. Lu, "Discontinuity enhancement based on time-variant seismic image deblurring,"

Journal of Applied Geophysics, vol. 135, pp. 155–162, 2016.

[16] Z. Wang, Z. Long, G. AlRegib, A. Asjad, and M. A. Deriche, "Automatic fault tracking across seismic volumes via tracking vectors," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5851–5855.

[17] W. Xiong, X. Ji, Y. Ma, Y. Wang, and Y. Luo, "Seismic fault detection with convolutional neural network," *Geophysics*, vol. 83, no. 5, pp. 1–28, 2018.

[18] Z. Yan, H. Gu, and C. Cai, "Automatic fault tracking based on ant colony algorithms," *Computers & Geosciences*, vol. 51, no. FEB., pp. 269–281, 2013.

[19] C. Yu, J. Zhao, and Y. Wang, "Seismic detection method for small-scale discontinuities based on dictionary learning and sparse representation," *Journal of Applied Geophysics*, vol. 137, pp. 55–62, 2017.

Biography

Chen Lei, born in 1981. He received his Ph.D degrees in School of Computer Science, Beijing University of Technology in 2017. He received his M.S. degree from Taiyuan University of Science and Technology in 2013 and received his B.S. degree from Jinan University in 2007. His research interests include the development of embedded system, pattern recognition and computer vision, data visualization, and artificial intelligence.

Jiaqi Shi, born in 1999. She received her B.S. degree from Beijing Information Science and Technology University in 2022 and she is currently a graduate student at Beijing Information Science and Technology University. Her research interests include artificial intelligence, pattern recognition and computer vision, and data visualization.

Ting Zhang, Born in 1986, associate professor. she received her Ph.D degrees in School of Computer Science, Beijing University of Technology in 2018 and stayed on as a teacher in the same year. Her main research directions include: deep learning theory, video/image processing, object detection, image recognition and image segmentation.

A Lattice-based Unidirectional Proxy Re-encryption

Lewei Wang, Mingming Jiang, Yuyan Guo, and Hui Ge
(Corresponding author: Mingming Jiang)

School of Computer Science and Technology, Huaibei Normal University
Huaibei 235000, Anhui, China
Email: jiangmm3806586@126.com

(Received June 2, 2023; Revised and Accepted Jan. 5, 2024; First Online June 22, 2024)

Abstract

Proxy re-encryption is widely used for ciphertext sharing in cloud computing environments. Proxy re-encryption enables non-trusted third parties to directly transform user Alice's ciphertext into other users' ciphertext without decrypting it, ensuring the security and privacy of data stored in third parties. However, proposed proxy re-encryption schemes are based on the number theory problem and cannot resist quantum attacks. In addition, most existing schemes are bidirectional. Therefore, in response to these issues, an identity-based unidirectional proxy re-encryption scheme on the lattice is constructed using the preimage sampling function. The proxy re-encryption key of the scheme can be generated without the interaction of two private keys, which can resist the collusion attack. The scheme is proven to be Chosen-Plaintext Attack secure under the standard model and resistant to quantum attacks.

Keywords: Identity-based Cryptography; Lattice Cryptography; LWE; Proxy Re-encryption

1 Introduction

With the development of quantum computers, lattice-based cryptographic schemes are simple to operate, resistant to quantum attacks, and have high efficiency, making them a hot research topic in the quantum era. In 2014, Kirshanova [1] proposed a single-hop lattice-based proxy re-encryption PRE scheme. The scheme was later pointed out by Fan *et al.* [2] to have subtle errors in the security proof. They constructed on its basis an agent re-encryption with multiple features over the lattice. However, the scheme has many public parameters and the key storage takes up a large amount of space. Later, lattice-based proxy re-encryption has been continuously proposed [3,4], but proxy re-encryption with special properties has been slow to develop.

In 2004, Hwang *et al.* [5] proposed an efficient scheme based on ID-based Cryptosystem. The first one-way identity-based agent re-encryption scheme was proposed by Green *et al.* [6]. After that, a series of identity-based proxy re-encryption schemes were designed [7, 8] These schemes were based on computationally number-theoretic hard problems (e.g., integer decomposition problems, discrete logarithm problems), which are insecure in the quantum computing environment. Therefore, the design of identity-based proxy re-encryption schemes resistant to quantum computing capabilities needs to be considered. Aono *et al.* [9] first proposed a one-way proxy re-encryption scheme based on the LWE problem over the lattice. They proved CPA security under their standard model and Chosen-Ciphertext Attack security under the stochastic predictor model. Singh *et al.* [10] proposed identity-based bi-directional agent re-encryption on the lattice. Later, Singh *et al.* [11] constructed a one-way agent re-encryption based on the scheme of Ano *et al.* [9]. In 2015, Jiang *et al.* [12] constructed the first lattice-only unidirectional multi-use proxy re-encryption using the original image sampling technique. In this paper, he extended the scheme to identity-based one-way multi-use agent re-encryption but did not prove its security. In 2019, Hou *et al.* [13] constructed an efficient identity-based bi-directional multi-bit proxy re-encryption algorithm in the standard model, which expanded the plaintext space and improved efficiency. Although identity-based proxy re-encryption has achieved some results, there are few identity-based one-way proxy re-encryption schemes on the grid. Most of the schemes are inefficient and under the random predictor model.

In two-way proxy re-encryption, the proxy can transform the ciphertext of both parties into each other, while one-way proxy re-encryption means that the proxy can only transform Alice's ciphertext into Bob's ciphertext. The reverse is not true. In this case, internal security needs to be considered in terms of security, i.e. collusion attacks between the proxy and the authorized person. Of course, any one-way proxy re-encryption scheme can easily be turned into a two-way proxy re-encryption scheme.

Therefore, the goal of this paper is to construct a

lattice-effective identity-based unidirectional proxy re-encryption scheme in the standard model.

2 Preliminaries

2.1 Notations

In this paper, we set up the following notations. We use $\mathbb{Z}$ and $\mathbb{R}$ to denote the integer ring and the real number ring.

2.2 Lattice

Definition 1. *Let a basis* $\mathbf{B} = \mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m \in \mathbb{Z}^{n\times m}$ *be an* $n \times m$ *matrix with linearly independent columns* $b_1, ..., b_m \in \mathbb{Z}^n$. *The m-dimensional full-rank lattice* Λ *is defined as the set of all linear combinations of integer coefficients of a vector:*

$$\Lambda = L(\mathbf{B}) = \left\{ \mathbf{Bc} = \sum_{i=1}^{m} c_i \mathbf{b}_i | c_i \in \mathbb{Z} \right\}$$

Here $\mathbf{b}_1, ..., \mathbf{b}_m$ *is a set of bases of lattice* Λ. *We respectively called m,n as the rank and dimensions of* Λ.

Definition 2. *Let* q *be a prime,* $\mathbf{A} \in \mathbb{Z}_q^{n\times m}$, $\mathbf{u} \in \mathbb{Z}_q^n$, *define:*

$\Lambda(\mathbf{A}) = \left\{ \mathbf{e} \in \mathbb{Z}^m, s.t. \exists s \in \mathbb{Z}_q^n, \mathbf{A}^T\mathbf{s} = \mathbf{e} \bmod q \right\}$
$\Lambda_q^{\perp}(\mathbf{A}) = \{ \mathbf{e} \in \mathbb{Z}^m, s.t. \mathbf{Ae} = 0 \bmod q \}$
$\Lambda_q^{u}(\mathbf{A}) = \{ \mathbf{e} \in \mathbb{Z}^m, s.t. \mathbf{Ae} = \mathbf{u} \bmod q \}$

2.3 Discrete Gaussian Distributions

For any parameter $\sigma > 0$, define the Gaussian distribution function on $\mathbf{R}^n$ centered at $\mathbf{c}$:
$\forall \mathbf{x} \in \mathbb{Z}^m$, $\rho_{\sigma,\mathbf{c}}(\mathbf{x}) = \exp\left(-\pi \|\mathbf{x} - \mathbf{c}\|^2 / \sigma^2\right)$
For any parameter $\mathbf{c} > 0, \sigma > 0$, m-dimensional lattice Λ, define the discrete Gaussian distribution function over Λ as flowing:
$\forall \mathbf{x} \in \Lambda$,
$D_{\Lambda,\sigma,\mathbf{c}}(\mathbf{x}) = \rho_{\sigma,\mathbf{c}}(\mathbf{x}) / \rho_{\sigma,\mathbf{c}}(\Lambda) = \rho_{\sigma,\mathbf{c}}(\mathbf{x}) / \sum_{\mathbf{x}\in\Lambda} \rho_{\sigma,\mathbf{c}}(\mathbf{x})$.

2.4 Important Algorithms

Lemma 1. *[14] Let* q *be odd, and* $q \geq 3$, $m = \lfloor 6n \log q \rfloor$. *There is a probability polynomial time PPT algorithm* $TrapGen(q, n)$ *that outputs a random matrice* $\mathbf{A} \in \mathbb{Z}_q^{n\times m}$ *and a group of basis* $\mathbf{T_A} \in \mathbb{Z}^{m\times m}$ *on a lattice* $\Lambda_q^{\perp}(\mathbf{A})$ *satisfying:*

$$\begin{aligned} \|\tilde{\mathbf{T}}_\mathbf{A}\| &\leq O\left(\sqrt{n \log q}\right) \\ \|\mathbf{T}\| &\leq O(n \log q). \end{aligned}$$

Lemma 2. *[15] Let* $q > 2$ *and a matrix* $\mathbf{A} \in \mathbb{Z}_q^{n\times m}$, $\mathbf{T_A}$ *is a basis of* $\Lambda_q^{\perp}(\mathbf{A})$, $\sigma \geq \left\|\widetilde{\mathbf{T}}\right\| \cdot \omega\left(\sqrt{\log m}\right)$, $\mathbf{c} \in \mathbb{Z}^m$, $\mathbf{u} \in \mathbb{Z}_q^n$. *Then do:*

1) $\Pr\left[\mathbf{x} \leftarrow D_{\Lambda_q^{\perp}(\mathbf{A}),\sigma} : \|\mathbf{x}\| > \sigma\sqrt{m}\right] \leq negl(n)$;

2) *There is a PPT algorithm* $Sample\,\Pr e(\mathbf{A}, \mathbf{T_A}, \mathbf{u}, \sigma)$ *that makes* $\mathbf{x} \in \Lambda_q^u(\mathbf{A})$ *statistically close to* $D_{\Lambda_q^u,\sigma,\mathbf{c}}$.

Lemma 3. *[16] Let* $q > 2$, $m > 2n \log q$ *and*

$$\sigma > \left\|\widetilde{\mathbf{T}}\right\| \cdot \omega\left(\sqrt{\log 2m + m_1}\right)$$

Then there exists a PPT algorithm $SampleLeft(\mathbf{A}, \mathbf{B}, \mathbf{T_A}, \mathbf{u}, \sigma)$ *that takes* $\mathbf{A} \in \mathbb{Z}_q^{n\times m}, \mathbf{B} \in \mathbb{Z}_q^{n\times m_1}$ *and a basis* $\mathbf{T_A}$ *of* $\Lambda_q^{\perp}(\mathbf{A})$ *as inputs, and returns a basis* $\mathbf{T_F}$ *of* $\Lambda_q^{\perp}(\mathbf{F})$ *where* $\mathbf{F} = (\mathbf{A} \| \mathbf{B})$.

Lemma 4. *[16] Let* $q > 2$, $m > n$ *and*

$$\sigma > \left\|\widetilde{\mathbf{T_B}}\right\| \cdot \omega\left(\sqrt{\log m}\right)$$

Then there exists a PPT algorithm $SampleRight(\mathbf{A}, \mathbf{B}, \mathbf{R}, \mathbf{T_B}, \mathbf{u}, \sigma)$ *that takes* $\mathbf{R} \in \{-1, 1\}^{m\times m}$ *and a basis* $\mathbf{T_B}$ *of* $\Lambda_q^{\perp}(\mathbf{B})$ *as inputs, and outputs a basis* $\mathbf{T_F}$ *of* $\Lambda_q^{\perp}(\mathbf{F})$ *where* $\mathbf{F} = (\mathbf{A} \| \mathbf{AR} + \mathbf{B})$.

2.5 The LWE Problem

For $n, m, q \in \mathbb{Z}$, a distribution $\chi^m \leftarrow \mathbb{Z}_q^m$. The LWE problem is to distinguish the two distributions: $(\mathbf{A}, \mathbf{b} = \mathbf{As} + \mathbf{x})$ and $(\mathbf{A}, \mathbf{b})$ where $\mathbf{A} \leftarrow \mathbb{Z}_q^{n\times m}$,$\mathbf{x} \leftarrow \chi^m$,$\mathbf{s} \leftarrow \mathbb{Z}_q^n$,$\mathbf{b} \leftarrow \mathbb{Z}_q^m$ are independently sampled.

2.6 Encoding with Full-rank Differences (FRD) Map

Lemma 5. *[16] Let* q *be a prime and* n *a positive integer. We say that a function* $H : \mathbb{Z}_q^n \to \mathbb{Z}_q^{n\times n}$ *is an encoding with full-rank differences (FRD) if:*

1) *For all distinct* $u, v \in \mathbb{Z}_q^n$, *the matrix* $H(u) - H(v) \in \mathbb{Z}_q^{n\times n}$ *is full rank;*

2) H *is computable in polynomial time (in* n *log* q*).*

3 Identity-based Unidirectional Proxy Re-encryption

3.1 Definition

An identity-based unidirectional proxy re-encryption scheme consists of six algorithms, as follows:
Setup(λ): Input the security parameter λ, then the algorithm outputs the public parameters PP and the master secret key MK.

Extract(PP, MK, ID): Input the public parameters PP, a master secret key MK, and an identity ID, this algorithm outputs a private key SK_{ID} corresponding to an identity ID.

RekeyGen$(PP, ID_i, ID_j, SK_{ID_i})$: On input the master public parameters PP, an identity ID_i, an identity ID_j and the secret key SK_{ID_i}, this algorithm outputs the re-encryption key $rk_{i\to j}$. The proxy can transform the ciphertext of the delegator ID_i into the ciphertext of the delegatee ID_j by using $rk_{i\to j}$.

Enc(PP, ID_i, μ): On input the master public parameters PP, an identity ID_i and a message μ. This algorithm will output a ciphertext CT_{ID_i}.

ReEnc$(PP, CT_{ID_i}, rk_{i\to j})$: On input the ciphertext CT_{ID_i} under identity ID_i and re-encryption key $rk_{i\to j}$, this algorithm outputs a re-encrypted ciphertext CT_{ID_j} for the identity ID_j.

Decrypt(PP, sk_{ID}, CT_{ID}): On input the ciphertext CT_{ID_i} under identity ID_i and private key sk_{ID}, this algorithm outputs either a re-encrypted ciphertext CT_{ID_j} for the identity ID_j or an error marker $\perp$ to indicate that it is an illegitimate ciphertext.

Correctness: Two conditions are needed:

$$Decrypt(PP, sk_{ID_i}, Enc(PP, ID_i, \mu)) = \mu$$

$$Decrypt(PP, sk_{ID_j}, \mathrm{Re}\, Enc(PP, CT_{ID_i}, \mathrm{Re}\, keyGen(PP, ID_i, ID_j, SK_{ID_i}))) = \mu$$

3.2 Security Model

The security model for identity-based unidirectional proxy re-encryption relies on a series of Indistinguishability under chosen-plaintext attack (IND-CPA) security games between the adversary $\mathcal{A}$ and the challenger $\mathcal{C}$, defined as follows:

Setup Phase: First, the challenger $\mathcal{C}$ inputs a security parameter λ, and then obtains the parameters param PP. Next, the $\mathcal{C}$ gives $\mathcal{A}$ the param;

Phase 1: In this phase, the $\mathcal{A}$ needs to issue a series of inquiries to the following oracles, and the $\mathcal{C}$ answers these oracles;

Public key generation oracle O_{pk}: $\mathcal{A}$ inputs an index i. The $\mathcal{C}$ uses $Extract(PP, MK, ID)$ to generate a key pair (pk_i, sk_i) and stores the pair. Finally the challenger $\mathcal{C}$ sends pk_i to $\mathcal{A}$;

Private key generation oracle O_{sk}: If user i is dishonest, when $\mathcal{A}$ inputs pk_i, the $\mathcal{C}$ will look up the table $T = (pk_i, sk_i)$and send sk_i to $\mathcal{A}$. Otherwise, outputs $\perp$;

Re-encryption key generation oracle O_{rk}: The $\mathcal{A}$ inputs (pk_i, sk_j), where (pk_i, sk_j) were generated before by $Extract(PP, MK, ID)$, then $\mathrm{Re}\, keyGen(PP, ID_i, ID_j, SK_{ID_i})$ generates $rk_{i\to j}$ to the adversary $\mathcal{A}$;

Re-encryption oracle O_{re}: If pk2 is corrupted, then $\mathcal{A}$ inputs (pk_i, sk_i, c_i) and returns a symbol $\perp$ that is not in the domains of messages and ciphertexts. Otherwise, return the re-encrypted ciphertext $c_j = \mathrm{Re}\, Enc(rk_{i\to j}, c_i)$;

Challenge Oracle O_c: The adversary $\mathcal{A}$ submits a target user id^* and two messages μ_0, μ_1. The identity id^* cannot appear in the private key query of phase 1. Next, the $\mathcal{C}$ picks a random bit $r \in \{0, 1\}$ and a random ciphertext c. If $r = 0$ it sets the challenge ciphertext to $c^* = Encrypt(pk^*, \mu_b)$. If $r = 1$ it sets the challenge ciphertext to $c^* = c$. Finally, it sends challenge ciphertext c^* to the adversary;

Phase 2: The attacker will continue to issue additional private key extraction queries, and the challenger will respond in the same way as Phase 1. However, attackers are not allowed to query at this stage.

Gauss: Finally, the adversary $\mathcal{A}$ outputs a guess $r' \in \{0, 1\}$. When $r' = r$, $\mathcal{A}$ wins the game. We generally define the advantage of adversary attacking an identity-based encryption scheme as the difference between the probability of winning the game and 1/2, i.e.

$$Adv_{\varepsilon,A}^{ind-cpa}(n) = |2\Pr[r' = r] - 1|$$

If the advantage $Adv_{\varepsilon,A}^{ind-cpa}(n)$ is negligible for all probabilistic polynomial time adversary $\mathcal{A}$, we say that this scheme is IND-CPA secure.

4 Our Scheme

4.1 Construction

The identity $id = (b_1, b_2, ..., b_l) \in \{-1, 1\}^l$ is a string and its length is bits. To reducing the size of common parameters in the scheme, we divide into l' segments $(b_1, b_2, ..., b_{l'})$, where each segment contains l/l' bits. The following is a specific description of the algorithm:

Setup(λ): Input a security parameter λ, do:

1) Run the algorithm $TrapGen(q, n)$ to produce a random matrix $\mathbf{A}_0 \in \mathbb{Z}_q^{n\times m}$ and the corresponding trapdoor basis $\mathbf{T}_{\mathbf{A}_0} \in \mathbb{Z}_q^{(m-k)\times k}$.
2) Select $l' + 1$ random matrices $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{l'}, \mathbf{B} \in \mathbb{Z}_q^{n\times m}$ randomly and a vector $\mathbf{u} \in \mathbb{Z}_q^n$.
3) Output the public parameters $PP = (\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{l'}, \mathbf{B}, \mathbf{u})$ and the master key $MK = \mathbf{T}_{\mathbf{A}_0}$.

Extract(PP, MK, ID):Input the public parameters PP, master key $MK = \mathbf{T}_{\mathbf{A}_0}$ and the user identity $id = (b_1, b_2, ..., b_{l'}) \in \{-1, 1\}^{l'}$, do:

1) Set $\mathbf{A}_{id} = \mathbf{B} + \sum_{i=1}^{l'} b_i \cdot \mathbf{A}_i \in \mathbb{Z}_q^{n\times m}$, $\mathbf{F}_{id} := (\mathbf{A}_0 \| \mathbf{A}_{id}) \in \mathbb{Z}_q^{n\times 2m}$.
2) Run the algorithm $SampleLeft(\mathbf{A}_0, \mathbf{A}_{id}, \mathbf{T}_{\mathbf{A}_0}, \mathbf{u}, \sigma)$ to produce $\mathbf{s} \in \mathbb{Z}_q^{2m}$ and $\mathbf{F}_{id} \cdot \mathbf{s} = \mathbf{u} \bmod q$.
3) Output $SK_{ID} = \mathbf{s} \in \mathbb{Z}_q^{2m}$.

RekeyGen$(PP, \mathbf{F}_i, \mathbf{F}_j, \mathbf{s}_i)$: Run the algorithm $Sample\Pr e(\mathbf{F}_i, \mathbf{s}_i, \mathbf{F}_j)$ to produce the matrix $\mathbf{R}_{i\to j} \in \mathbb{Z}^{2m\times 2m}$ and $\mathbf{F}_i \cdot \mathbf{R}_{i\to j} = \mathbf{F}_j \bmod q$. Output $rk_{i\to j} = \mathbf{R}_{i\to j}$.

Enc(PP, ID_i, m): Input $\mu \in \{0,1\}$, do:

1) Compute $\mathbf{A}_{id} = \mathbf{B} + \sum_{i=1}^{l} b_i \cdot \mathbf{A}_i \in \mathbb{Z}_q^{n\times m}$, $\mathbf{F}_{id} := (\mathbf{A}_0 \| \mathbf{A}_{id}) \in \mathbb{Z}_q^{n\times 2m}$.
2) Choose a random vector $\mathbf{t} \in \mathbb{Z}_q^n$ and l' matrices $\mathbf{R}_i \in \{-1,1\}^{m\times m}$. Compute $\mathbf{R}_{id} = \sum_{i=1}^{l'} b_i \cdot \mathbf{R}_i \in \{-l', ..., l'\}^{m\times m}$.
3) Choose an error vector $x \xleftarrow{\overline{\Psi}_\alpha} \mathbb{Z}_q$, $\mathbf{y} \xleftarrow{\overline{\Psi}_\alpha} \mathbb{Z}_q^m$. Compute $\mathbf{z} = \mathbf{R}_{id}^{\mathbf{T}} \cdot \mathbf{y} \in \mathbb{Z}_q^m$.
4) Compute $c = \mathbf{u}^T \cdot \mathbf{t} + x + \mu \cdot \lfloor q/2 \rfloor$, $\mathbf{c}_i = \mathbf{F}_{id}^T \cdot \mathbf{t} + \begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} \in \mathbb{Z}_q^{2m}$
5) Output the ciphertext $CT = (c, \mathbf{c}_i) \in \mathbb{Z}_q \times \mathbb{Z}_q^{2m}$.

ReEnc$(PP, CT_i, rk_{i\to j})$:Input the re-encryption key $rk_{i\to j} = \mathbf{R}_{i\to j}$ and the CT_i which is ciphertext of the user i, do:

1) Choose an error vector $\mathbf{x}_j \xleftarrow{\overline{\Psi}_\alpha} \mathbb{Z}_q^m$.
2) Compute $\mathbf{c}_j = \mathbf{R}_{i\to j}^T \cdot \mathbf{c}_i + \mathbf{x}_j \in \mathbb{Z}_q^{2m}$.
3) Output the ciphertext $CT' = (c, \mathbf{c}_j)$.

Decrypt(PP, sk_{ID}, CT_{ID}):

1) Compute $w = c - \mathbf{s}^T \cdot \mathbf{c}_i \in \mathbb{Z}_q$.
2) Compare the sizes of and $\lfloor q/2 \rfloor$. If $|w - \lfloor q/2 \rfloor| < \lfloor q/4 \rfloor$, output 1, otherwise output 0.

4.2 Parameters and Correctness

For ciphertext $CT = (c, \mathbf{c}_i)$, the decryption process is as follows:

$$\begin{aligned} w &= c - \mathbf{s}^T \cdot \mathbf{c}_i \\ &= \mathbf{u}^T\mathbf{t} + x + \mu\lfloor q/2\rfloor - \mathbf{s}^T \cdot \left(\mathbf{F}_{id}^T \cdot \mathbf{t} + \begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix}\right) \\ &= \mu\lfloor q/2\rfloor + \underbrace{x - \mathbf{s}^T \begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix}}_{noise} \end{aligned}$$

If the noise term is small enough, μ can be correctly restored. For re-encrypted ciphertext $CT' = (c, \mathbf{c}_j)$, we first calculate

$$\begin{aligned} \mathbf{c}_j &= \mathbf{R}_{i\to j}^T \cdot \mathbf{c}_i + \mathbf{x}_j \\ &= \mathbf{R}_{i\to j}^T \cdot \left(\mathbf{F}_{id_i}^T \cdot \mathbf{t} + \begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix}\right) + \mathbf{x}_j \\ &= \left(\mathbf{F}_{id_i} \cdot \mathbf{R}_{i\to j}\right)^T \cdot \mathbf{t} + \mathbf{R}_{i\to j}^T \begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} + \mathbf{x}_j \\ &= \mathbf{F}_{id_j}^T \cdot \mathbf{t} + \mathbf{R}_{i\to j}^T \begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} + \mathbf{x}_j \end{aligned}$$

Then, calculate

$$\begin{aligned} w &= c - \mathbf{s}_j^T \cdot \mathbf{c}_j \\ &= \mathbf{u}^T\mathbf{t} + x + \mu\lfloor q/2\rfloor - \mathbf{s}_j^T \cdot \left(\mathbf{F}_{id_j}^T \cdot \mathbf{t} + \mathbf{R}\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} + \mathbf{x}_j\right) \\ &= \left(\mathbf{u} - \mathbf{F}_{id_j}\mathbf{s}_j\right)^T \cdot \mathbf{t} + x + \mu\lfloor q/2\rfloor - \mathbf{s}_j^T \cdot \mathbf{R}\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} - \mathbf{s}_j^T\mathbf{x}_j \\ &= \mu\lfloor q/2\rfloor + \underbrace{x - \mathbf{s}_j^T \cdot \mathbf{R}\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} - \mathbf{s}_j^T\mathbf{x}_j}_{noise} \end{aligned}$$

Therefore, it can be inferred that if the parameter settings are reasonable, μ can be correctly restored after decryption as mentioned above.

The following parameters are needed to be set to ensure the correctness of the scheme:

1) The $TrapGen$ requires $m > 6n\log q$, $\|\mathbf{T}\| \le O(n\log q)$and $q = poly(n)$;
2) The algorithm $Sample$ requires $\sigma \ge \left\|\tilde{\mathbf{T}}_{\mathbf{A}}\right\| \cdot w\left(\sqrt{\log n}\right)$;
3) The algorithm $SampleLeft$ requires $q > 2$, $m > 2n\log q$ and $\sigma \ge \left\|\tilde{\mathbf{T}}_{\mathbf{A}}\right\| \cdot w\left(\sqrt{\log(m+m_1)}\right)$;
4) The algorithm $SampleRight$ requires $\sigma \ge \left\|\tilde{\mathbf{T}}_{\mathbf{B}}\right\| \cdot \sqrt{m} \cdot w\left(\sqrt{\log m}\right)$;
5) In order to meet the reduction requirements of difficult problems, we set the parameters as follows: $m = 6n^{1+\delta}, q = m^2\sqrt{n} \cdot w(\log n), \sigma = m \cdot w(\log n), \alpha = \left[m^2 \cdot w(\log n)\right]^{-1}$

Here we assume that δ satisfies $n^\delta > \lceil \log q \rceil = O(\log n)$.

4.3 Security Analysis

Theorem 1. *Under the assumption of the LWE difficulty problem, our re-encryption scheme is IND-CPA security in the standard model.*

Proof 1. *In the process of proof, we need to use a series of games to prove that the advantage of the probability polynomial time(PPT) adversarial attack is negligible. Firstly, it will be proven that the following four games are indistinguishable for the adversary.*

Game 0: *This game is the initial IND-CPA game.*

Game 1: *The difference with Game 0 is that it has changed the generation method of* $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{l'}$. *In Game 0, the challenger generates the common parameter by selecting the random matrix* $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{l'}$ *in* $\mathbb{Z}_q^{m\times n}$. *In Game 1, the challenger randomly selects* l' *matrices* $\mathbf{R}_i^* \in \mathbf{R}^{m\times m}$ *and FRD maps* $H(id^*)$ *during the initialization phase. The matrix* $\mathbf{R}_i^*$ *is composed of uniformly random polynomials with coefficients* $\{-1, 1\}$. *Then, the challenger generates matrices* $\mathbf{A}_0$ *and* $\mathbf{B}$ *according to the requirements of the initial game and calculates*

$$\mathbf{A}_i = (\mathbf{R}_i^*)^T \cdot \mathbf{A}_0 - H(id^*) \cdot \mathbf{B}, i \in [1, l']$$

Set $\mathbf{R}^* = (\mathbf{R}_1^*, \mathbf{R}_2^*, ..., \mathbf{R}_{l'}^*)$ *to generate challenge ciphertext* CT^* *during the challenge phase, where* $\mathbf{z} = \mathbf{R}_{id}^{\mathbf{T}} \cdot \mathbf{y}$. *So the distributions* $(\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_{l'}, \mathbf{z})$ *and* $(\mathbf{A}_0, (\mathbf{R}_1^*)^T \cdot \mathbf{A}_0, ..., (\mathbf{R}_{l'}^*)^T \cdot \mathbf{A}_0, \mathbf{z})$ *are statistically close. In the adversary's view, matrix* $(\mathbf{R}_i^*)^T \cdot \mathbf{A}_0$ *is statistically close to the uniform random matrix* $\mathbf{A}_i$ *and independent of vector* $\mathbf{z}$. *Therefore, Game 1 and Game 2 are indistinguishable.*

Game 2: *In Game 2, the difference from Game 1 is that it changes the generation method of* $\mathbf{A}_0$ *and* $\mathbf{B}$. *We uniformly and randomly select matrix* $\mathbf{A}_0$ *in* $\mathbb{Z}_q^{m\times n}$, *and generate matrix* $\mathbf{B}$ *by using* $TrapGen(q, n)$. *The challenger has obtained the corresponding trapdoor* $\mathbf{T_B}$. *Matrix* $\mathbf{A}_i = (\mathbf{R}_i^*)^T \cdot \mathbf{A}_0 - H(id^*) \cdot \mathbf{B}, i \in [1, l']$ *remains the same as Game 2.*
In the adversary's view, Game 2 and Game 1 are indistinguishable.

At this stage, $\mathcal{A}$ *can perform extract key queries and re-encryption key queries:*

1) *Extract key queries: The challenger* $\mathcal{C}$ *needs to generate the corresponding private key to answer the private key query that satisfies* $id = (b_1, b_2, ..., b_{l'})$ *and* $id \neq id^*$. *Let*

$$\mathbf{F}_{id} := \mathbf{A}_0 \| \mathbf{B} + \sum_{i=1}^{l'} b_i \mathbf{A}_i$$
$$= \mathbf{A}_0 \| \sum_{i=1}^{l'} b_i (\mathbf{R}_i^*)^T \mathbf{A}_0 + \left[1 - \sum_{i=1}^{l'} b_i \cdot H(id)\right] \cdot \mathbf{B}$$
$$= \mathbf{A}_0 \| (\mathbf{R}_{id})^T \cdot \mathbf{A}_0 + h_{id} \cdot \mathbf{B}$$

Where $\mathbf{R}_{id}^T = \sum_{i=1}^{l'} b_i (\mathbf{R}_i^*)^T, h_{id} = 1 - \sum_{i=1}^{l'} b_i \cdot H(id)$.

If $h_{id} = 0$, *the challenger aborts the game. Otherwise, The challenger will obtain the private key by running the right sampling algorithm, do:*
$\mathbf{s} \leftarrow SampleRight(\mathbf{A}_0, h_{id} \cdot \mathbf{B}, \mathbf{R}_{id}, \mathbf{T_B}, u, \sigma) \in \mathbb{Z}_q^{m\times n}$ *Send* SK_{id} *to the adversary* $\mathcal{A}$.

2) *Re-encryption key queries: Adversary* $\mathcal{A}$ *sends* $(\mathbf{F}_i, \mathbf{F}_j)$ *to the challenger. The challenger* $\mathcal{C}$ *calculates* $SK_i = \mathbf{s}_i$ *using the above method and then runs the sampling algorithm* $Sample\Pr e(\mathbf{F}_i, \mathbf{s}_i, \mathbf{F}_j)$ *to output the matrix* $\mathbf{R}_{i\to j} \in \mathbb{Z}^{2m\times 2m}$. *Finally,* $\mathcal{C}$ *sends the re-encrypted key* $rk_{i\to j} = \mathbf{R}_{i\to j}$ *to the adversary* $\mathcal{A}$.

Game 3: *The difference between this game and Game 2 is the method produced by the ciphertext of . In Game 2, the challenge ciphertext is generated based on the encryption algorithm, i.e.* $c = \mathbf{u}^T \cdot \mathbf{t} + x + \mu \cdot \lfloor q/2 \rfloor$,
$\mathbf{c}_i = \mathbf{F}_{id}^T \cdot \mathbf{t} + \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \in \mathbb{Z}_q^{2m}$.

In Game 3, the challenge ciphertext $(c^*, \mathbf{c}_i^*)$ *is uniformly and randomly selected from the* $\mathbb{Z}_q \times \mathbb{Z}_q^{2m}$. *It needs to be proven that Game 2 and Game 3 are computationally indivisible, and this problem can be reduced to an LWE problem.*

Reduction from LWE: *Assuming there is a PPT adversary* $\mathcal{A}$. *The advantage of distinguishing Game 2 from Game 3 is* ε. *An algorithm* $\mathcal{B}$ *can be constructed to solve the LWE decision problem. The LWE problem instance is provided by a sampling oracle* O, *where* $O_\$$ *represents the true random oracle and* O_s *represents the noise pseudo-random oracle.*

Instance: $\mathcal{B}$ *requests* O *to obtain* m *random LWE problem instances* $(\mathbf{u}_i, v_i) \in \mathbb{Z}_q^n \times \mathbb{Z}_q$.

Setup:

1) *Construct a random matrix* $\mathbf{A}_0 = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m)$.
2) *Specify the zeroth LWE sample (currently unused) as a common random n-dimensional vector* $\mathbf{u}_0 \in \mathbb{Z}_q^n$.
3) *The other common parameters, namely* $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{l'}, \mathbf{B}$, *are constructed using* id^* *and* $\mathbf{R}^*$ *in Game 2.*

Queries: $\mathcal{B}$ *answers the extract key queries and re-encryption key queries, such as Game 2.*

Challenge: *Adversary* $\mathcal{A}$ *submits the target identity* id^* *and a message* $\mu^* \in \{0, 1\}$. $\mathcal{B}$ *performs the following steps to generate the challenge ciphertext corresponding to* id^**:*

1) *Set* $\mathbf{v}^* = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \in \mathbb{Z}_q^m$;

2) Set $c^* = \mathbf{u}^T \cdot \mathbf{t} + \mu^* \lfloor q/2 \rfloor \in \mathbb{Z}_q$ *to hide the message bits* μ^**;*

3) Set $\mathbf{c}_i^* = \begin{bmatrix} \mathbf{v}^* \\ (\mathbf{R}_{id}^*)^T \cdot \mathbf{v}^* \end{bmatrix} \in \mathbb{Z}_q^{2m}$*;*

4) Choose a random bit $r \xleftarrow{R} \{0,1\}$*. If* $r = 0$*, send* $CT^* = (c^*, \mathbf{c}_i^*)$ *to* $\mathcal{A}$*. Otherwise, randomly select* $(c^*, \mathbf{c}_i^*) \in \mathbb{Z}_q \times \mathbb{Z}_q^{2m}$ *and send it to* $\mathcal{A}$*.*

When the LWE oracle is pseudorandom (i.e. $O = O_s$*), the distribution of the challenge ciphertext is the same as in Game 2. Firstly, observe* $\mathbf{F}_{id} := \left(\mathbf{A}_0 \| (\mathbf{R}_{id})^T \cdot \mathbf{A}_0 + h_{id} \cdot \mathbf{B}\right)$*. Secondly, through the definition of* O_s*, we know* $\mathbf{v}^* = \mathbf{A}_0^T \mathbf{s} + \mathbf{y}$*, where* $\mathbf{y} \in \mathbb{Z}_q^m$ *is a random noise vector distribution as* $\overline{\Psi}_\alpha$*. Therefore, the definition of in step (3) satisfies*

$$\begin{aligned} \mathbf{c}_i^* &= \begin{bmatrix} \mathbf{A}_0^T \mathbf{s} + \mathbf{y} \\ (\mathbf{R}_{id}^*)^T \cdot \mathbf{A}_0^T \mathbf{s} + (\mathbf{R}_{id}^*)^T \mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_0^T \mathbf{s} + \mathbf{y} \\ (\mathbf{A}_0 \mathbf{R}_{id}^*)^T \mathbf{s} + (\mathbf{R}_{id}^*)^T \mathbf{y} \end{bmatrix} \\ &= (\mathbf{F}_{id}^*)^T \mathbf{s} + \begin{bmatrix} \mathbf{y} \\ (\mathbf{R}_{id}^*)^T \mathbf{y} \end{bmatrix} \end{aligned}$$

Obviously, the quantity on the right side of the equation is the $\mathbf{c}_1$ *part of the challenge ciphertext in Game 2. Note that* $v_0 = \mathbf{u}^T \mathbf{s} + x$ *is the* $\mathbf{c}_0$ *in the ciphertext. When* $O = O_\$$*,* v_0 *is uniform on* $\mathbb{Z}_q$*, and* $\mathbf{v}^*$ *is uniform on* $\mathbb{Z}_q^m$*. Therefore, the* $\mathbf{c}_i^*$ *defined in Step (3) is uniformly independent in the* $\mathbb{Z}_q^{2m}$*. Therefore, the challenge ciphertext defined in Game 3 is also uniform in* $\mathbb{Z}_q \times \mathbb{Z}_q^{2m}$*.*

Guess: *After being allowed to perform other queries,*$\mathcal{A}$ *guessed whether he was interacting with Game 2 or Game 3. Simulator* $\mathcal{B}$ *outputs the guess of* $\mathcal{A}$ *as the answer to the LWE challenge they are attempting to solve.*

We have already discussed that when $O = O_s$*, it appears to the* $\mathcal{A}$ *that it is the same as Game 2. When* $O = O_\$$*, it appears to the opponent that it is the same as Game 3. Therefore, the advantages of* $\mathcal{B}$ *in solving the LWE problem are the same as the advantages of* $\mathcal{A}$ *in distinguishing Game 2 and Game 3.*

5 Comparison

This scheme has been compared with relevant scheme literature in terms of private key size, directionality and security model. The comparison results are shown in Table 1.

Table 1: Comparison to related works

Cryptosystem	Private key size	Directionality	Security model
[11]	$O(mn \log q)$	Bidirectional	Random oracle model
[12]	$O(2m \log q)$	Unidirectional	Random oracle model
[13]	$O(mn \log q)$	Bidirectional	Standard model
Our scheme	$O(2m \log q)$	Unidirectional	Standard model

6 Conclusions

In this article, a lattice-based one-way proxy re-encryption scheme has been proposed, and its security has been proven under the LWE problem. The difference between the scheme in this paper and others is that the proxy re-encryption scheme in this article is unidirectional. In the standard model, it can also be proven to be CPA secure.

Acknowledgments

We acknowledge the financial support the Nature Science Foundation of Anhui Higher Education Institutions No. 2022AH050388, No. 2022AH050374.

References

[1] E. Kirshanova, "Proxy re-encryption from lattices," in *Public-Key Cryptography - PKC 2014 - 17th International Conference on Practice and Theory in Public-Key Cryptography, Buenos Aires, Argentina, March 26-28, 2014. Proceedings*, ser. Lecture Notes in Computer Science, H. Krawczyk, Ed., vol. 8383. Springer, 2014, pp. 77–94. [Online]. Available: https://doi.org/10.1007/978-3-642-54631-0_5

[2] X. Fan and F. Liu, "Various proxy re-encryption schemes from lattices," *IACR Cryptol. ePrint Arch.*, p. 278, 2016. [Online]. Available: http://eprint.iacr.org/2016/278

[3] X. Liang, J. Weng, A. Yang, L. Yao, Z. Jiang, and Z. Wu, "Attribute-based conditional proxy re-encryption in the standard model under LWE," in *Computer Security - ESORICS 2021 - 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4-8, 2021, Proceedings, Part II*, ser. Lecture Notes in Computer Science, E. Bertino, H. Shulman, and M. Waidner, Eds., vol. 12973. Springer, 2021, pp. 147–168. [Online]. Available: https://doi.org/10.1007/978-3-030-88428-4_8

[4] L. Wu, X. Yang, M. Zhang, and X. A. Wang, "IB-VPRE: adaptively secure identity-based proxy re-encryption scheme from LWE with re-encryption verifiability," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 1, pp. 469–482, 2022. [Online]. Available: https://doi.org/10.1007/s12652-021-02911-9

[5] M. Hwang, J. Lo, and S. Lin, "An efficient user identification scheme based on id-based cryptosystem," *Comput. Stand. Interfaces*, vol. 26, no. 6, pp. 565–569, 2004. [Online]. Available: https://doi.org/10.1016/j.csi.2004.03.006

[6] M. Green and G. Ateniese, "Identity-based proxy re-encryption," in *Applied Cryptography and Network Security, 5th International Conference, ACNS 2007, Zhuhai, China, June 5-8, 2007, Proceedings*, ser. Lecture Notes in Computer Science, J. Katz and M. Yung, Eds., vol. 4521. Springer, 2007, pp. 288–306. [Online]. Available: https://doi.org/10.1007/978-3-540-72738-5_19

[7] G. Kan, C. Jin, H. Zhu, Y. Xu, and N. Liu, "An identity-based proxy re-encryption for data deduplication in cloud," *J. Syst. Archit.*, vol. 121, p. 102332, 2021. [Online]. Available: https://doi.org/10.1016/j.sysarc.2021.102332

[8] S. Maiti and S. Misra, "P2B: privacy preserving identity-based broadcast proxy re-encryption," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5610–5617, 2020. [Online]. Available: https://doi.org/10.1109/TVT.2020.2982422

[9] Y. Aono, X. Boyen, L. T. Phong, and L. Wang, "Key-private proxy re-encryption under LWE," in *Progress in Cryptology - INDOCRYPT 2013 - 14th International Conference on Cryptology in India, Mumbai, India, December 7-10, 2013. Proceedings*, ser. Lecture Notes in Computer Science, G. Paul and S. Vaudenay, Eds., vol. 8250. Springer, 2013, pp. 1–18. [Online]. Available: https://doi.org/10.1007/978-3-319-03515-4_1

[10] K. Singh, C. P. Rangan, and A. K. Banerjee, "Lattice based identity based proxy re-encryption scheme," *J. Internet Serv. Inf. Secur.*, vol. 3, no. 3/4, pp. 38–51, 2013. [Online]. Available: https://doi.org/10.22667/JISIS.2013.11.31.038

[11] ——, "Cryptanalysis of unidirectional proxy re-encryption scheme," in *Information and Communication Technology - Second IFIP TC5/8 International Conference, ICT-EurAsia 2014, Bali, Indonesia, April 14-17, 2014. Proceedings*, ser. Lecture Notes in Computer Science, Linawati, M. S. Mahendra, E. J. Neuhold, A. M. Tjoa, and I. You, Eds., vol. 8407. Springer, 2014, pp. 564–575. [Online]. Available: https://doi.org/10.1007/978-3-642-55032-4_58

[12] M. Jiang, Y. Hu, B. Wang, F. H. Wang, and Q. Lai, "Lattice-based multi-use unidirectional proxy re-encryption," *Secur. Commun. Networks*, vol. 8, no. 18, pp. 3796–3803, 2015. [Online]. Available: https://doi.org/10.1002/sec.1300

[13] J. Hou, M. Jiang, Y. Guo, and W. Song, "Efficient identity-based multi-bit proxy re-encryption over lattice in the standard model," *J. Inf. Secur. Appl.*, vol. 47, pp. 329–334, 2019. [Online]. Available: https://doi.org/10.1016/j.jisa.2019.05.015

[14] X. Wang, A. Hu, and H. Fang, "Feasibility analysis of lattice-based proxy re-encryption," in *Proceedings of the 2017 International Conference on Cryptography, Security and Privacy, ICCSP 2017, Wuhan, China, March 17 - 19, 2017*. ACM, 2017, pp. 12–16. [Online]. Available: https://doi.org/10.1145/3058060.3058080

[15] C. Gentry, C. Peikert, and V. Vaikuntanathan, "Trapdoors for hard lattices and new cryptographic constructions," in *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, C. Dwork, Ed. ACM, 2008, pp. 197–206. [Online]. Available: https://doi.org/10.1145/1374376.1374407

[16] S. Agrawal, D. Boneh, and X. Boyen, "Efficient lattice (H)IBE in the standard model," in *Advances in Cryptology - EUROCRYPT 2010, 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Monaco / French Riviera, May 30 - June 3, 2010. Proceedings*, ser. Lecture Notes in Computer Science, H. Gilbert, Ed., vol. 6110. Springer, 2010, pp. 553–572. [Online]. Available: https://doi.org/10.1007/978-3-642-13190-5_28

Biography

Lewei Wang was born in 1998. She received the B.S degree in the School of Software Engineering from Huaiyin Normal University, Huaian, China, in 2020. She is currently pursuing the M.Sc degree in software engineering with Huaibei Normal University, Huaibei. Her current research interests include identity-based encryption, lattice-based cryptography.(12111080775@chnu.edu.cn)

Mingming Jiang was born in 1984. He received the PhD degree in cryptography from Xidian University of China. Currently, He is an associate professor in the school of computer science and technology, Huaibei Normal University, China. His research interests include cryptography and information security.(jiangmm3806586@126.com)

Analysis of One Multifactor Authenticated Key Agreement Scheme for Industrial IoT

Zhengjun Cao[1], Jiahua Zhu[1], and Lihua Liu[2]
(Corresponding author: Zhengjun Cao)
Department of Mathematics, Shanghai University[1]
Shangda Road 99, Shanghai 200444, China
Department of Mathematics, Shanghai Maritime University, China[2]
Email: caozhj@shu.edu.cn
(Received June 9, 2023; Revised and Accepted Jan. 5, 2024; First Online June 22, 2024)

Abstract

We show that the key agreement scheme [IEEE ITJ, 2021, 3801–3811] is flawed. (1) It is insecure against internal attack because any unauthorized sensing device (not revoked) can retrieve the final session key. (2) It could be insecure against external attack because an adversary can recover the secret divisor of two parameters. We want to stress that a key agreement scheme integrated with secret-sharing could be vulnerable to internal attack.

Keywords: External Attack; Internal Attack; Key Agreement; Secret Sharing

1 Introduction

Industrial IoT (IIoT) is a system of devices, sensors, and applications that work together to collect, monitor, and analyze data from industrial operations. It enables organizations to get actionable data in order to increase efficiencies, reduce costs, and improve safety and security. It has a wide range of applications, including medical treatment, house, agriculture, and surveillance.

In 2018, Das *et al.* [4] used biometrics to design privacy-preserving user authentication scheme for cloud-based IIoT. Li *et al.* [9] presented a robust ECC-based provable secure authentication protocol with privacy preserving for IIoT. Hussain and Chaudhry [6] remarked that Das *et al.*'s key agreement scheme is insecure. Gope and Sikdar [5] proposed a privacy-aware authenticated key agreement scheme for secure smart grid communication. Kumar *et al.* [8, 13, 22] designed lightweight authentication and key agreement schemes for smart metering in smart energy networks and internet of drones.

Wazid *et al.* [23] also proposed a secure remote user authenticated key establishment protocol for amart home environment. In 2020, Zhang *et al.* [25] have presented a key agreement scheme for internet of drones. Liu and Cao [11] have shown its flaws. Pan, Yang, and Hwang [16] presented an enhanced secure smart card-based password authentication scheme. Ali *et al.* [2] designed a lightweight privacy-aware IoT-based metering scheme for smart industrial ecosystems.

In 2022, Pirayesh *et al.* [17] proposed an authentication and key agreement scheme for smart home networks. Verma and Bhardwaj [20] presented a secure lightweight anonymous elliptic curve cryptography-based authentication and key agreement scheme for fog assisted-IoT enabled networks. Malik, Gandhi and Narwal [14] discussed the problem for anonymous mutually authenticated key agreement scheme for wireless sensor networks. Wazid and *et al.* [24] designed a blockchain-enabled user authentication and key agreement scheme for crowdsourcing system. Abbasinezhad-Mood *et al.* [1] also presented an efficient provably-secure dynamic ID-based authenticated key agreement scheme with enhanced security provision. But it has shown the insecurity against password-guessing attack [3]. Lu and Hwang [12] presented a cryptographic key generation scheme without a trusted third party for access control in multilevel wireless sensor networks. In 2023, Lin and Hsu [10] designed a chaotic maps-based privacy-preserved three-factor authentication scheme for telemedicine systems. Hwang *et al.* [7, 15, 19] proposed some authentication and session key agreement schemes for different scenarios.

Recently, Vinoth *et al.* [21] have presented a key agreement scheme for industrial Internet of Things. The scheme makes use of password, biometrics, and smart card to identify the user, and utilizes the secret-sharing technology to construct a session key among the user and authorized sensing devices. In this note, we show that the Vinoth *et al.*'s scheme is insecure against either internal attack or external attack.

2 Review of the Scheme

In the Vinoth *et al.*'s key agreement scheme, there are many entities: a user, the Gateway Node (GWN), n sensing devices. Its security goals include entity authentica-

Table 1: The Vinoth *et al.*'s key agreement scheme

User U_i	Gateway Node (GWN)	Sensing Device(SD_j)
$Gen(\cdot), Rep(\cdot)$ are generation and reproduction algorithms of fuzzy extractor, respectively, and $h(\cdot)$ is a hash function.	For the dealer P_0 and n devices $P_1, \cdots, P_n$, compute $x_i = \varphi(P_i), i = 0, \cdots, n$. Pick n-dimensional $Vector_1, Vector_2$, and a secret value S, s.t., $S = Vector_1 \cdot x_0, S^2 = Vector_2 \cdot x_0$. Pick $\mathrm{ID}_{\mathrm{SD}_j}$, compute $s_j = Vector_1 \cdot x_j$, $f_j = Vector_2 \cdot x_j$. Pick pairwise coprime positive integers $k_1, \cdots, k_n$. Compute $\mathrm{Mul}_j = \prod_{t=1}^{n} k_t/k_j$, Nonce_j, s.t., $\mathrm{Mul}_j \times \mathrm{Nonce}_j \equiv 1 \bmod k_j$. Set $\gamma = \sum_{j=1}^{n} \mathrm{Mul}_j \times \mathrm{Nonce}_j$. $\xRightarrow[\text{[secure channel]}]{\mathrm{ID}_{\mathrm{SD}_j}, s_j, f_j, k_j}$	
Choose ID_i, PW_i, imprint biometrics B_i. Compute $(BK_i, \tau_i) = Gen(B_i)$. Pick a nonce a, compute $TPW_i = h(\mathrm{ID}_i \| PW_i \| BK_i) \oplus a$. $\xRightarrow{\mathrm{ID}_i, TPW_i}$	Generate the key $\mathrm{KEY}_{\mathrm{GWN}}$. Set $KEY_{GWN-U_i} = h(\mathrm{ID}_i \| \mathrm{KEY}_{\mathrm{GWN}})$, $A_i = KEY_{GWN-U_i} \oplus TPW_i$ $C_i = \mathrm{ID}_{\mathrm{GWN}} \oplus TPW_i$. Pick a 128-bit temporary identity TID_i. $\xLeftarrow{\{TID_i, A_i, C_i\}}$	
Compute $RPW_i = h(\mathrm{ID}_i \| PW_i \| BK_i)$, $A'_i = A_i \oplus TPW_i \oplus RPW_i$, $C'_i = C_i \oplus TPW_i \oplus h(\mathrm{ID}_i \| BK_i)$, $D_i = a \oplus h(\mathrm{ID}_i \| BK_i)$, $V_i = h(RPW_i \| A_i \| a \| h(\mathrm{ID}_i \| BK_i)) \bmod \omega$, where ω is a medium integer to thwart online guessing attack. Store $\{TID_i, A'_i, C'_i, D_i, V_i, \tau_i, \omega\}$.		
Pick a nonce r_i and timestamp TS_1, compute $BK_i = Rep(B_i, \tau_i)$, $RPW_i = h(ID_i \| PW_i \| BK_i)$, $ID_{GWN} = C'_i \oplus h(ID_i \| BK_i)$, $M_1 = A_i \oplus RPW_i \oplus r_i$, $M_2 = h(TID_i \| M_1 \| ID_{GWN} \| r_i \| TS_1)$. $\xrightarrow[\text{[open channel]}]{TID_i, M_1, M_2, TS_1}$	Check $\lvert TS_1 - TS'_1 \rvert \le \triangle TS$. Use TID_i to look up ID_i, KEY_{GWN-U_i}, and compute $r_i = M_1 \oplus KEY_{GWN-U_i}$. Check $M_2 = h(TID_i \| M_1 \| ID_{GWN} \| r_i \| TS_1)$. If so, pick r_{GWN} and TS_2 to compute $M_4 = r_{GWN} \times \gamma, M_5 = Enc_{r_{GWN}}(ID_i, ID_{GWN}, r_i, r_{GWN} \oplus KEY_{GWN-U_i})$, $M_6 = h(ID_i \| ID_{GWN} \| r_i \| M_4 \| KEY_{GWN-U_i} \| TS_2)$. $\xrightarrow{M_4, M_5, M_6, TS_2}$	Check $\lvert TS_2 - TS'_2 \rvert \le \triangle TS$. Compute $r_{GWN} = M_4 \bmod k_j$, $Dec_{r_{GWN}}(M_5) = (ID_i, ID_{GWN}, r_i, r_{GWN} \oplus KEY_{GWN-U_i})$, check $M_6 = h(ID_i \| ID_{GWN} \| r_i \| M_4 \| r_{GWN} \oplus KEY_{GWN-U_i} \oplus r_{GWN} \| TS_2)$. If so, generate TS_3, compute $M_8 = Enc_{r_{GWN}}(s_j, f_j, ID_{SD_j})$ $\xleftarrow{M_8, TS_3}$
	Check $\lvert TS_3 - TS'_3 \rvert \le \triangle TS$. Compute $Dec_{r_{GWN}}(M_8) = (s_j, f_j, ID_{SD_j})$, $\theta_1 = \sum_{t=1}^{l} \lambda_t s_t, \theta_2 = \sum_{t=1}^{l} \lambda_t f_t$. Check $\theta_2 = \theta_1^2$. Set $S = \theta_1$. $M_9 = h(S \| r_{GWN}), M_{10} = M_9 \times \gamma$, $M_{11} = h(M_9 \| M_{10})$. Generate TID_i^{new}, TS_4. $\xrightarrow{M_{10}, M_{11}}$	$\xleftarrow{\{M_8^{(SD_i)}, TS_3^{(SD_i)}\}_{SD_i \text{is in the authorized set}}}$ Compute $M_9 = M_{10} \bmod k_j$. Check $M_{11} = h(M_9 \| M_{10})$. Compute $SK = h(ID_i \| ID_{GWN} \| r_{GWN} \| r_i \| M_9 \| KEY_{GWN-U_i})$, $M_{16} = h(SK \| ID_{GWN} \| ID_i)$
Check $\lvert TS_4 - TS'_4 \rvert \le \triangle TS$. $Dec_{KEY_{GWN-U_i}}(M_{12}) = (r_{GWN}, r_i, M_9)$. Check $M_{14} = h(M_{12} \| M_9 \| r_i)$. Compute $SK = h(ID_i \| ID_{GWN} \| r_{GWN} \| r_i \| M_9 \| KEY_{KEY-U_i})$ Check $M_{16} = h(SK \| ID_{GWN} \| ID_i)$. Set $TID_i^{new} = h(ID_i \| KEY_{GWN-U_i} \| TS_4) \oplus M_{13}$ Update TID_i with TID_i^{new}.	Compute $M_{12} = Enc_{KEY_{GWN-U_i}}(r_{GWN}, r_i, M_9)$, $M_{13} = h(ID_i \| KEY_{GWN-U_i} \| TS_4) \oplus TID_i^{new}$, $M_{14} = h(M_{12} \| M_9 \| r_i)$. $\xleftarrow{M_{12}, M_{13}, M_{14}, TS_4}$ $\xleftarrow{M_{16}}$	

tion, data confidentiality, and user anonymity. To make it easier to follow the below discussion, we now depict the scheme as follows (see Table 1, or Figure 2, [21]).

3 Insecure Against Internal Attack

By the registration (see §V.B, [21]), we know, GWN registers the devices using secret-sharing technology and Chinese remainder theorem. GWN picks a unique identity ID_{SD_j} for each device SD_j, and pairwise coprime positive integers $k_1, \cdots, k_n$, where $j = 1, 2, \cdots, n$. GWN computes $\text{Mul} = \prod_{j=1}^{n} k_j, \text{Mul}_j = \text{Mul}/k_j$ and Nonce_j, s.t., $\text{Mul}_j \times \text{Nonce}_j \equiv 1 \bmod k_j$. Set

$$\gamma = \sum_{j=1}^{n} \text{Mul}_j \times \text{Nonce}_j \tag{1}$$

Note that γ *is set for the whole group of* n *devices, not for any authorized set of* $l\,(< n)$ *devices.* We find the secret γ and shares $k_j, j = 1, \cdots, n$, are not harmonically invoked. Concretely, GWN invokes γ to hide the nonce r_{GWN} as

$$M_4 = r_{GWN} \times \gamma, \tag{2}$$

and the device SD_j invokes k_j to recover the nonce

$$r_{GWN} \equiv M_4 \bmod k_j \tag{3}$$

Clearly, a corrupted device SD_s (not revoked), even unauthorized for the current session, can also recover the same nonce by computing

$$r_{GWN} \equiv M_4 \bmod k_s, \tag{3'}$$

because M_4 is transported via an open channel (see the blue-colored parts, Table 1).

Using r_{GWN}, the corrupted device can compute

$$Dec_{r_{GWN}}(M_5) = (ID_i, ID_{GWN}, r_i, r_{GWN} \oplus KEY_{GWN-U_i})$$

where M_5 is also publicly accessible, and $Dec(\cdot)$ is a symmetric key decrypting algorithm. By the recovered nonce r_{GWN} and the component $r_{GWN} \oplus KEY_{GWN-U_i}$, it is easy to recover KEY_{GWN-U_i}. Now, all components

$$ID_i, ID_{GWN}, r_{GWN}, r_i, KEY_{GWN-U_i}, M_9 = M_{10} \bmod k_s$$

can be obtained by the adversary for computing the final session key

$$SK = h\left(ID_i \| ID_{GWN} \| r_{GWN} \| r_i \| M_9 \| KEY_{GWN-U_i}\right) \tag{4}$$

We want to stress that in a secret sharing scheme [18], an owner of a share is not assumed to directly use it for transporting data. The below simple relation

$$M_4 = r_{GWN} \times \gamma \quad \Longrightarrow \quad r_{GWN} \equiv M_4 \bmod k_j$$

is insufficient to securely transfer the nonce r_{GWN}.

4 Insecure Against External Attack

We find the scheme is also insecure against external attack, because the secret divisor γ could be retrieved.

In fact, the calculations of

$$M_4 = r_{GWN} \times \gamma, \quad M_{10} = M_9 \times \gamma$$

are actually computed over the ring $\mathbb{Z}_k$, where

$$k = [k_1, k_2, \cdots, k_n]$$

is the lowest common multiple. Since they are pairwise coprime, $k = k_1 \times \cdots \times k_n$. In view of that the residue r_{GWN} modulo k_j is used as the key for $Dec(\cdot)$, the bit-length of k_j is greater than 256. In general,

$$\text{BitLength}(r_{GWN}) = \text{BitLength}(h(\cdot)) = 256,$$

and $\text{BitLength}(k) \geq 256n$, such as the popular SHA-256, and AES-256. By the equations

$$\gamma = \sum_{j=1}^{n} \text{Mul}_j \times \text{Nonce}_j \bmod k, \;\; M_9 = h(S \| r_{GWN}),$$

it is very likely that $r_{GWN} \times \gamma < k, M_9 \times \gamma < k$. So,

$$M_4 = r_{GWN} \times \gamma, \;\; M_{10} = M_9 \times \gamma \tag{5}$$

are two common equalities. An external adversary can recover the common divisor γ from M_4 and M_{10}, both are transported via open channels. Thus, r_{GWN}, M_9 can also be exposed. Now, the adversary can compute $Dec_{r_{GWN}}(M_5)$ to obtain

$$ID_i, \; ID_{GWN}, \; r_i, \; r_{GWN} \oplus KEY_{GWN-U_i},$$

which means that all components for the final hashing (see Eq.(4)) can be successfully retrieved. In result, the adversary can retrieve the session key by computing $h\left(ID_i \| ID_{GWN} \| r_{GWN} \| r_i \| M_9 \| KEY_{GWN-U_i}\right)$.

5 Conclusion

We show that the Vinoth *et al.*'s key agreement scheme is flawed. It is worth noting that a key agreement scheme being integrated with secret-sharing technology could be vulnerable to internal attack. One should carefully design such a scheme and balance its security goals. The findings in this note could be helpful for the future work on designing such schemes.

Acknowledgment

We thank the National Natural Science Foundation of China (61411146001). We are grateful to the reviewers for their valuable suggestions.

References

[1] D. Abbasinezhad-Mood and *et al.*, "Efficient provably-secure dynamic id-based authenticated key agreement scheme with enhanced security provision," *IEEE Trans. Dependable Secur. Comput.*, vol. 19, no. 2, pp. 1227–1238, 2022.

[2] W. Ali and *et al.*, "A lightweight privacy-aware IoT-based metering scheme for smart industrial ecosystems," *IEEE Trans. Ind. Informatics*, vol. 17, no. 9, pp. 6134–6143, 2021.

[3] Z. J. Cao, "A note on 'efficient provably-secure dynamic ID-based authenticated key agreement scheme with enhanced security provision'," *IEEE Trans. Dependable Secur. Comput.*, doi: 10.1109/TDSC.2023.3302300.

[4] A. K. Das and *et al.*, "Biometrics-based privacy-preserving user authentication scheme for cloud-based industrial internet of things deployment," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4900–4913, 2018.

[5] P. Gope and B. Sikdar, "Privacy-aware authenticated key agreement scheme for secure smart grid communication," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3953–3962, 2019.

[6] S. Hussain and S. A. Chaudhry, "Comments on "biometrics-based privacy-preserving user authentication scheme for cloud-based industrial internet of things deployment"," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10936–10940, 2019.

[7] M. S. Hwang, H. W. Li, and C. Y. Yang, "An improved of enhancements of a user authentication scheme," *Int. J. Netw. Secur.*, vol. 25, no. 3, pp. 508–514, 2023.

[8] P. Kumar and *et al.*, "Lightweight authentication and key agreement for smart metering in smart energy networks," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4349–4359, 2019.

[9] X. Li and *et al.*, "A robust ECC-based provable secure authentication protocol with privacy preserving for industrial internet of things," *IEEE Trans. Ind. Informatics*, vol. 14, no. 8, pp. 3599–3609, 2018.

[10] T. W. Lin and C. L. Hsu, "Chaotic maps-based privacy-preserved three-factor authentication scheme for telemedicine systems," *Int. J. Netw. Secur.*, vol. 25, no. 2, pp. 194–200, 2023.

[11] L. H. Liu and J. Cao, "Analysis of one lightweight authentication and key agreement scheme for internet of drones," *Int. J. Electr. Inf. Engineering*, vol. 13, no. 4, pp. 142–148, 2021.

[12] Y. C. Lu and M. S. Hwang, "A cryptographic key generation scheme without a trusted third party for access control in multilevel wireless sensor networks," *Int. J. Netw. Secur.*, vol. 24, no. 5, pp. 959–964, 2022.

[13] I. Makhdoom and *et al.*, "Anatomy of threats to the internet of things," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 2, pp. 1636–1675, 2019.

[14] M. Malik, K. Gandhi, and B. Narwal, "AMAKA: anonymous mutually authenticated key agreement scheme for wireless sensor networks," *Int. J. Inf. Secur. Priv.*, vol. 16, no. 1, pp. 1–31, 2022.

[15] M. Nikooghadam and H. Amintoosi, "Secure communication in cloud IoT through design of a lightweight authentication and session key agreement scheme," *Int. J. Commun. Syst.*, vol. 36, no. 1, 2023.

[16] H. Pan, H. Yang, and M. Hwang, "An enhanced secure smart card-based password authentication scheme," *Int. J. Netw. Secur.*, vol. 22, no. 2, pp. 358–363, 2020.

[17] J. Pirayesh and *et al.*, "A PLS-HECC-based device authentication and key agreement scheme for smart home networks," *Comput. Networks*, vol. 216, p. 109077, 2022.

[18] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[19] S. Uppuluri and G. Lakshmeeswari, "Secure user authentication and key agreement scheme for iot device access control based smart home communications," *Wirel. Networks*, vol. 29, no. 3, pp. 1333–1354, 2023.

[20] U. Verma and D. Bhardwaj, "A secure lightweight anonymous elliptic curve cryptography-based authentication and key agreement scheme for fog assisted-Internet of Things enabled networks," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 23, 2022.

[21] R. Vinoth and *et al.*, "Secure multifactor authenticated key agreement scheme for industrial iot," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3801–3811, 2021.

[22] M. Wazid and *et al.*, "Design and analysis of secure lightweight remote user authentication and key agreement scheme in internet of drones deployment," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3572–3584, 2019.

[23] M. Wazid and *et al.*, "Secure remote user authenticated key establishment protocol for smart home environment," *IEEE Trans. Dependable Secur. Comput.*, vol. 17, no. 2, pp. 391–406, 2020.

[24] M. Wazid and *et al.*, "BUAKA-CS: blockchain-enabled user authentication and key agreement scheme for crowdsourcing system," *J. Syst. Archit.*, vol. 123, p. 102370, 2022.

[25] Y. Zhang and *et al.*, "A lightweight authentication and key agreement scheme for internet of drones," *Computer Communications*, vol. 154, pp. 455–464, 2020.

Biography

Zhengjun Cao, associate professor with Department of Mathematics, Shanghai University, received his PhD degree in applied mathematics from Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He had served as a post-doctor in Computer Sciences Department, Université Libre de Bruxelles. His research interests include cryptography, discrete logarithms and quantum computation.

Jiahua Zhu is currently pursuing his master degree from

Department of Mathematics, Shanghai University. His research interests include information theory and applied mathematics.

Lihua Liu, associate professor with Department of Mathematics, Shanghai Maritime University, received her PhD degree in applied mathematics from Shanghai Jiao Tong University. Her research interests include combinatorics and cryptography.

A Group Repair Codes with Low Recovery-overhead in Distributed Storage System

Wenjie Deng, Cong Li, Tieyuan Hong, and Dan Tang
(Corresponding author: Dan Tang)

School of Software Engineering, Chengdu University of Information Technology
Sichuan Province Engineering Technology Research Center of Support Software of Informatization Application
Chengdu 610225, China
Email: tangdan@foxmail.com

(Received June 12, 2023; Revised and Accepted Jan. 5, 2024; First Online June 22, 2024)

Abstract

In distributed storage systems, erasure codes are widely used to ensure data reliability and provide higher storage efficiency than replication techniques. However, the cost of using erasure codes is the increased network bandwidth required during the repair process, which can lead to performance bottlenecks in the system. This paper proposes a CGRC (Crisscross Group Repair Codes), which constructs an encoding array to generate multiple types of parity blocks in a cross-interleaved manner and adjusts the distribution of data blocks within the parity block groups. This approach aims to reduce the bandwidth overhead for recovery while maintaining fault tolerance performance. Additionally, the paper presents an efficient decoding algorithm that achieves optimal repair efficiency for CGRC. Simulation results indicate that CGRC, compared with other block codes, the recovery efficiency of single-node can be improved by 48.56% at most, 77.5% compared with RS, and the recovery efficiency of multi-nodes can be improved by 16.9% at most compared with other block code. Moreover, CGRC ensures higher fault tolerance performance and increases reliability under the same storage efficiency.

Keywords: Bandwidth Overhead; Crisscross;Erasure Codes; Group Repair Codes

1 Introduction

With the development of information technology, the explosion of data has become a significant challenge. To meet the storage demands of such vast amounts of data, distributed storage systems such as HDFS (Hadoop Distributed File System) [1], and GFS (Google File System) [2] have emerged. However, these distributed storage systems still face numerous challenges, particularly in large-scale data centers and distributed storage environments. As data volume continues to increase, traditional backup and replication methods require significant additional redundant space, greatly impacting storage efficiency. To overcome this drawback, erasure codes with MDS (Maximum Distance Separable) properties [3], such as RS (Reed-Solomon) codes [4], have been introduced as an effective storage technique.

Erasure codes [5] are encoding methods that import redundant information into data blocks to achieve redundant storage and fault tolerance. They divide the original data into multiple encoding blocks and generate redundant blocks by performing calculations on these blocks. By utilizing the redundant blocks during decoding, the original data can be recovered even if some data blocks are lost.

Paper [6] provides an LRC (Local Reconstruction Codes) code constructed based on RS code. However, network bandwidth in storage systems is a critical bottleneck [7], especially in large-scale data centers and distributed storage environments. With the continuous increase in data volume, the high repair bandwidth cost and computational complexity of traditional MDS codes like RS codes have a significant impact on system bandwidth. Although subsequent improvements like RS codes based on Cauchy matrices [8] have optimized their computational performance, the node recovery overhead remains substantial, thereby limiting the efficiency and speed of data transmission.

Papers [9–11] have proposed array codes that use a simple XOR operation instead of complex finite field encoding, reducing computational complexity. Moreover, most array codes have MDS properties when considering nodes as the minimum unit of data loss. They also have obvious drawbacks. Their special array encoding method makes them strictly dependent on the number of nodes, lacks flexible scalability and still incurs high repair bandwidth costs. LDPC (Low-Density Parity Check) [12] is

based on XOR and graph structures, offering efficient repair performance. However, their structure is not fixed due to the use of probabilistic distributions for constructing Tanner graphs, making them more commonly used in channel coding. Paper [13] provides a construction method for group codes.

Research shows that most node failures in distributed storage systems are caused by single-node failures [14]. Huang proposed a novel code, LRC [15], for WAS (Windows Azure Storage), aiming to provide lower repair bandwidth and efficient data repair capabilities, by encoding within stripe groups. LRC only requires repairs within the group, thereby reducing the bandwidth overhead and repair time. Pyramid [16] provides a dimension-based approach, offering more flexible encoding constructions while improving multi-node repair performance to some extent. However, this comes at the cost of higher storage overhead. EXPyramid [17], based on Pyramid, adopts two-dimensional array encoding, performing Pyramid encoding in both horizontal and vertical directions. It possesses superior fault-tolerance capability, but still incurs significant storage overhead. Meng proposed DLRC [18], where each local parity block involves different data blocks, resulting in more flexible encoding. SHEC [19] employs a stacked encoding scheme, effectively reducing the repair overhead of individual nodes, but lacks global parity blocks, leading to lower multi-node repair performance. Paper [20] introduced a three-layer coding structure, adding additional redundant blocks to parity blocks, thereby reducing repair costs when parity blocks are lost.

To address these issues, this paper proposes a three-layer coding scheme called CGRC (Crisscross Group Repair Codes), utilizing cross-group encoding of local parity blocks. CGRC effectively divides local parity blocks into two categories: low repair overhead and low storage overhear.

The following is the general structure of this paper, the second part introduces the basics including MDS properties, and LRC codes. The third part describes the design of CGRC, which includes the basic concepts of CGRC, the coding and decoding process, and reliability analysis. The fourth part presents an experimental comparison with other coding schemes in several aspects. The fifth part summarizes the article and analyzes the advantages and shortcomings of this scheme.

2 Preliminary Knowledge

2.1 MDS Property

Let the code length be n and the length of the information bits be k, where $d_{\min}$ represents the minimum distance. For all erasure codes, the minimum distance must satisfy the Singleton bound shown in Equation (1):

$$d_{\min} \leq n - k + 1 \tag{1}$$

A larger $d_{\min}$ typically indicates stronger fault-tolerance capabilities of the code. When $d_{\min}$ reaches the upper limit of the Singleton bound, it is referred to as an MDS (Maximum Distance Separable) code, which means the code has achieved theoretically optimal fault-tolerance performance. Furthermore, when the data bits and parity bits of an MDS code are independent of each other after encoding, it is referred to as a systematic MDS code.

2.2 LRC Code

(k, l, r)LRC [15] is a special optimization of encoding that focuses on single-node repair. It achieves the goal of reducing repair overhead by dividing all data blocks into several equal-length subsets and generating local parity blocks for each subset. The generation of parity blocks in LRC is based on systematic MDS encoding. Currently, there are many systematic MDS encodings available, but both LRC and the proposed CGRC in this paper use RS codes for generating parity blocks. Figure 1 illustrates the encoding process of RS(7,5).

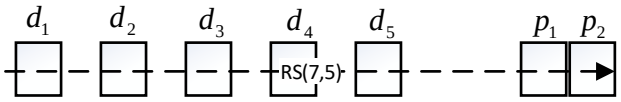


Figure 1: RS(7,5) Example. (k = 7 data and parity blocks, m = 5 data blocks)

The calculation of its parity blocks is usually performed in $GF(2^w)$ finite field, where w can be adjusted according to the encoding scope. The parity blocks encoding formula is shown in Equation (2), g_i is an element in GF(2^w).

$$p_i = \sum_{j=1}^{k} g_i^{j-1} \times d_j, i \in [1, 2, \ldots, n-k] \tag{2}$$

Based on the aforementioned RS code encoding process, Figure 2 illustrates an example of (6,2,2) LRC.

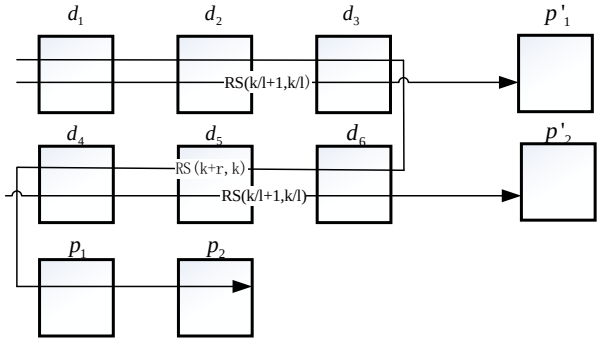


Figure 2: (6,2,2)LRC Example. (k = 6 data blocks and l = 2 local parity blocks, r = 2 global parity blocks)

The core idea of LRC is to divide k data blocks into l groups, so that when a single error occurs within a

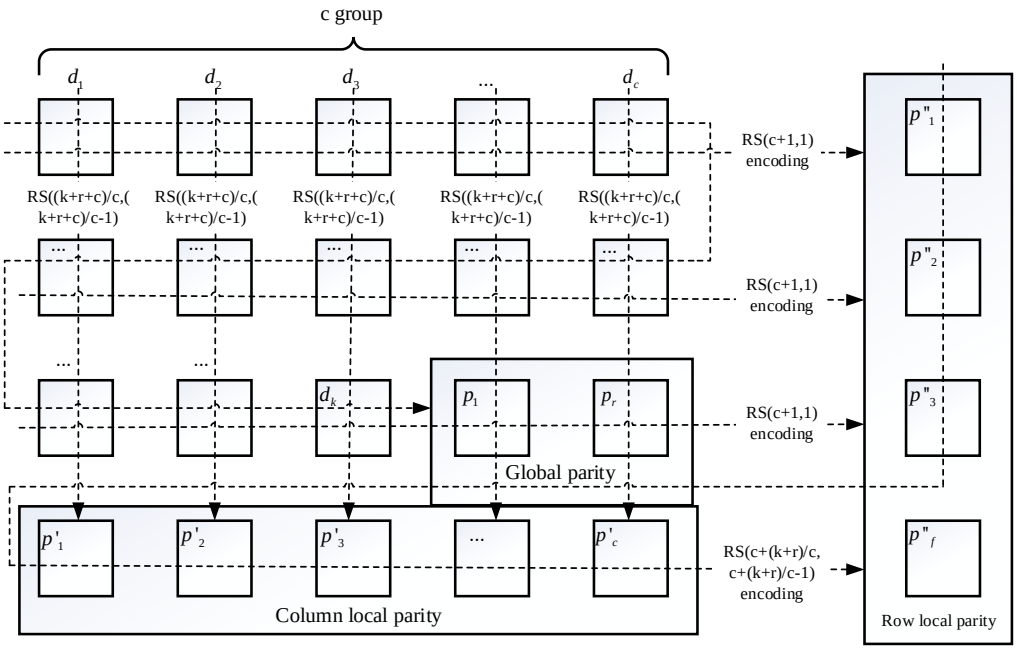


Figure 3: CGRC encoding structure. (k data block and c column local parity block, r global parity block)

group, it can be quickly recovered using the data within the group. In Figure 2, the data blocks are named as $d_1, d_2, d_3, d_4, d_5, d_6$. p'_1 is calculated using the data from group d_1, d_2, d_3. and p'_2 is calculated using the data from group d_4, d_5, d_6. The encoding formula is given as Equation (3) and Equation (4):

$$p_i = \sum_{j=1}^{k} g_i^{j-1} \times d_j, i \in [1, 2, \ldots, r] \quad (3)$$

$$p'_i = \sum_{j=(i-1)\times k/l+1}^{i\times k/l} g_i^{(j-1)\%(k\backslash l)} \times d_j, i \in [1, 2, \ldots, l] \quad (4)$$

3 Design of CGRC

This chapter mainly discusses the parameter settings, encoding and decoding methods of CGRC. An example is used to illustrate the encoding layout and various performance aspects of CGRC.

CGRC inherits the excellent single-node repair performance of LRC codes and further optimises the encoding of local parity blocks. By cross-generating local parity blocks, the local parity blocks can be further divided into two types: column local parity blocks with low repair cost and row local parity blocks with low storage overhead local parity blocks. This approach further improves the single-node repair performance of the encoding while taking into account storage efficiency.

In addition, CGRC incorporates global parity blocks into the calculation of local parity blocks. This means that if a global parity block is lost, there is no need to recalculate all data blocks as in traditional codes like LRC, reducing the repair overhead.

The specific encoding structure of CGRC is shown in Figure 3.

3.1 Basic Concept of CGRC

This section explains the concepts and symbols used in CGRC:

- Parity Block: A block generated after applying erasure coding to the data blocks.
- Column Local Parity Block: Local a parity block generated by column-wise encoding in the encoding array.
- Row Local Parity Block: Local parity block generated by row-wise encoding in the encoding array.
- Stripe: The minimum unit of erasure coding, typically consisting of data blocks and parity blocks.
- Repair Cost: The total amount of data that needs to be transmitted in the event of data loss in erasure coding.
- Fault Tolerance: The maximum number of node failures that the erasure coding scheme can tolerate.
- Storage Efficiency: The ratio of the space occupied by data blocks within a stripe after encoding to the total space.

The following is meaning of symbols used in (k, c, r)CGRC as shown in Table 1.

Table 1: Symbol Meaning

Symbols	meaning
k	Number of data blocks per stripe
c	Number of column local parity blocks per stripe
r	Number of global parity blocks per stripe
A	Parity matrix during encoding
D	Data matrix during encoding
$a_{(i,j)}$	Element at i-th row and j-th column in A
$d_i, d_{i,j}$	Data block at position i within current encoding stripe; Data block at the i-th row and j-th column in the encoding array
p_i, p'_i, p''_i	Global parity block at position i; Column local parity block at position i; Row local parity block at position i

3.2 Encoding of CGRC

In (k, c, r)CGRC, the encoding process begins by encoding k data blocks within a stripe, resulting in r global parity blocks. The encoding of the local parity blocks is performed using a cross-coding approach. Therefore, the data blocks and global parity blocks are arranged in an array with c columns. Then, the two types of local parity blocks are cross-generated in a sequential manner. After completing the encoding process, there will be r global parity blocks, c column local parity blocks, and $\lceil (k+r+c)/c \rceil$ row local parity blocks. The parameter settings of CGRC must satisfy the requirements of Equation (5):

$$k + r + c \leq c^2 \tag{5}$$

The encoding process of CGRC primarily involves the encoding of three types of local parity blocks. It starts with the encoding of global parity blocks P_r. The matrix A_{Global} is defined as an $r \times k$ matrix. A_{Global} is constructed according to Equation (6). In all the encoding of parity blocks, both $a_{(i,j)}$ and g_i is an element over GF(2^w) and constructed as shown in Equation (7). Additionally, any two g_i values are distinct, and the subsequent g_i construction follows the same pattern as.

$$A_{Global} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & \ddots & \vdots \\ a_{r,1} & \cdots & a_{r,k} \end{bmatrix} \tag{6}$$

$$a_{i,j} = g_i^{j-1} \, (1 \leq i \leq r, 1 \leq j \leq k) \tag{7}$$

The corresponding global check block A_{Global} can be calculated through P_r, and the specific calculation process is shown in Equation (8):

$$\begin{aligned} P_r &= A_{Global} \times D \\ &= \begin{bmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & \ddots & \vdots \\ a_{r,1} & \cdots & a_{r,k} \end{bmatrix} \times \begin{bmatrix} d_1 \\ \vdots \\ d_k \end{bmatrix} \\ &= \begin{bmatrix} p_1 & \cdots & p_r \end{bmatrix}^T \end{aligned} \tag{8}$$

Then generate the column local check block P'_r, define the matrix A_{CP} as a $c \times \lceil (k+r+c)/c \rceil$ matrix, and construct it as Equation (9). The data matrix D_{CP} is a $\lceil (k+r+c)/c \rceil \times c$ matrix constructed as Equation (10) $\sim$ (11):

$$A_{CP} = \begin{bmatrix} a'_{1,1} & \cdots & a'_{1,\lceil (k+r)/c \rceil} \\ \vdots & \ddots & \vdots \\ a'_{c,1} & \cdots & a'_{c,\lceil (k+r)/c \rceil} \end{bmatrix} \tag{9}$$

$$D_{CP} = \begin{bmatrix} d_1 & d_2 & \cdots & d_c \\ d_{c+1} & d_{c+2} & \cdots & d_{2\times c} \\ \vdots & \vdots & \ddots & \vdots \\ d_{\lceil (k+r)/c \rceil} & \begin{matrix} a_{1,1} \times d_1 + \cdots \\ + a_{1,k} \times d_k \end{matrix} & \cdots & \begin{matrix} a_{r,1} \times d_1 + \cdots \\ + a_{r,k} \times d_k \end{matrix} \end{bmatrix} \tag{10}$$

$$a'_{i,j} = g_i^{j-1} \, (1 \leq i \leq r, 1 \leq j \leq k,) \tag{11}$$

By performing calculation A_{CP}, a matrix with c column-level local parity blocks P'_r can be obtained. Taking the diagonal elements of this matrix will yield c column-level local parity blocks. The calculation process is shown in Equation (12):

$$\begin{aligned} P'_r &= A_{CP} \times D_{CP} \\ &= \begin{bmatrix} a'_{1,1} & \cdots & a'_{1,\lceil (k+r)/c \rceil} \\ \vdots & \ddots & \vdots \\ a'_{c,1} & \cdots & a'_{c,\lceil (k+r)/c \rceil} \end{bmatrix} \times \\ &\quad \begin{bmatrix} d_1 & \cdots & d_c \\ \vdots & \ddots & \vdots \\ d_{\lceil (k+r)/c \rceil} & \cdots & a_{r,1} \times d_1 + \cdots + a_{r,k} \times d_k \end{bmatrix} \\ &= \begin{bmatrix} p'_1 & \cdots & \cdots & \vdots \\ \vdots & p'_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & p'_c \end{bmatrix} \end{aligned} \tag{12}$$

Finally, generate the row local check block P''_r. The definition matrix A_{RP} is a matrix with $\lceil (k+r)/c \rceil$ rows and c columns, while the data matrix D_{RP} is a matrix with c rows and $\lceil (k+r)/c \rceil$ columns, as shown in Equations (13) $\sim$ (15):

$$A_{RP} = \begin{bmatrix} a''_{1,1} & \cdots & a''_{1,c} \\ \vdots & \ddots & \vdots \\ a''_{\lceil \lceil (k+r)/c \rceil \rceil,1} & \cdots & a''_{\lceil (k+r)/c \rceil,c} \end{bmatrix} \tag{13}$$

$$D_{RP} = \begin{bmatrix} d_1 & \cdots & d_{1+k+r-c} \\ d_2 & \cdots & a_{1,1} \times d_1 + \cdots + a_{1,k} \times d_k \\ \vdots & \ddots & \vdots \\ d_c & \cdots & a_{r,1} \times d_1 + \cdots + a_{r,k} \times d_k \end{bmatrix} \quad (14)$$

$$a''_{i,j} = g_i^{j-1} \, (1 \le i \le r, 1 \le j \le k,) \quad (15)$$

By calculating A_{RP}, $\lceil (k+r)/c \rceil$ column local check blocks can be obtained. For row parity block P''_f, it can be calculated using Equation (16):

$$\begin{aligned} p''' &= G_{PGlobal} \times D_{PGlobal} \\ &= \begin{bmatrix} a'''_{1,1} & a'''_{1,2} & \cdots & a'''_{1,c+m} \\ \vdots & \vdots & \ddots & \vdots \\ a'''_{s,1} & a'''_{s,2} & \cdots & a'''_{s,c+m} \end{bmatrix} \times \\ &\begin{bmatrix} a'_{1,1} \times d_{1,1} + \cdots + a'_{1,m} \times d_{m,1} \\ \vdots \\ a'_{c,1} d_{1,c} + \cdots + a'_{c,m} (a_{r,1} d_{1,1} + \cdots + a_{r,k} d_{m,k\%c}) \\ a''_{1,1} \times d_{1,1} + \cdots a''_{1,c} \times d_{1,c} \\ \vdots \\ a''_{m,1} \times d_{m,1} + \cdots + a''_{m,c} \times \\ (a_{1,1} \times d_{1,1} + \cdots + a_{1,k} \times d_{m,k\%c}) \end{bmatrix} \end{aligned} \quad (16)$$

3.3 Decoding of CGRC

The decoding method of CGRC is mainly related to the number of lost blocks and the parity blocks involved in decoding. We categorize errors into two types based on the number of lost blocks: single-node loss and multi-node loss with a number of lost nodes greater than 1. Each class has three decoding methods: Intra-column decoding, Intra-row decoding, and check block decoding, which correspond to the three different ways in which check blocks participate in decoding.

3.3.1 Single-node Failure Decoding

For single-node failures, if the failed node is at row n and column j in the original encoding array, and does not belong to row-level local parity block, the decoding process prioritizes the corresponding column-level local parity block for decoding. This decoding method ensures better decoding performance and is known as Intra-column decoding. Additionally, all calculations in the repair methods are rounded up to the nearest integer. The specific steps for repair are as follows:

1) Determine the column local check block p'_j corresponding to the lost node.

2) Obtain the encoding equation of p'_j, as shown in Equation (17):

$$p'_j = \sum_{i=1}^{k+r/c} a'_{j,i} \times D_{CP_{i,j}} \quad (17)$$

3) By decoding the encoding equation of p'_j and the data blocks within the group using Equation (18), the missing node $d_{(n,j)} \notin P''_r$ is obtained

$$d_{(n,j)} = \left(p'_j - \sum_{i=1, i \ne n}^{k+r/c} a'_{j,i} \times D_{CP_{i,j}} \right) / a'_{n,j} \quad (18)$$

The specific intra-column decoding process is shown in Figure 4.

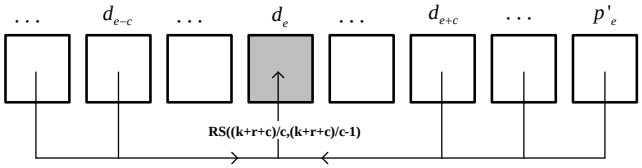


Figure 4: Intra-column Decoding

If the error node $d_{n,c+1} \in P''_r, d_{n,c+1} \ne p''_f$, is a row local checksum block and not p''_f. Decoding is performed through the encoding equation of $d_{n,c+1}$, which is called inline decoding. The specific steps are as follows.

1) Determine the row check block p''_n corresponding to the lost node in the original encoding array.

2) Determine the original encoding equation of an according to Equation (19), and calculate to obtain p''_n, which is the missing block $d_{n,c+1}$.

$$p''_n = \sum_{i=1}^{c} a''_{n,i} \times D_{RP_{i,n}} \quad (19)$$

The specific Intra-row decoding process is shown in Figure 5.

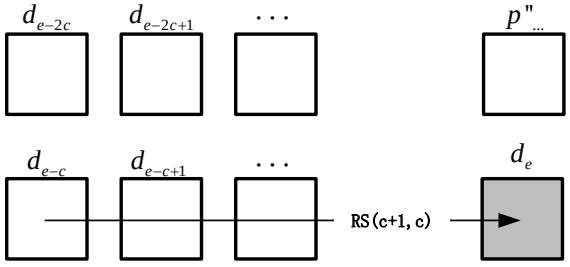


Figure 5: Intra-row Decoding

If the error node is p''_f, the original encoding equation can be obtained through Equation (20), and the encoding can be performed again.

$$p''_f = \sum_{i=1}^{c+(k+r)/c} a'''_{n,i} \times D_{f_i} \quad (20)$$

Table 2 shows the single-node recovery read overhead for four encoding schemes: CGRC (k, c, r), LRC (k, l, r), DLRC (k, m, n, l), TLRC (k, m, s, x, r).

Table 2: Reconstruction read cost for single-node failure

Code	Reconstruction read cost
CGRC(k,c,r)	$\frac{(k+r)/c\times(k+r+c+1)+k+r+c}{((k+r+c)/c)+k+r+c}$
LRC(k,l,r)	$\frac{k\times(k+l)}{l\times(k+l+r)}+\frac{k\times r}{(k+l+r)}$
DLRC(k,m,n,l)	$\frac{n\times(k+m)}{k+m+l}+\frac{k\times l}{k+m+l}$
TLRC(k,m,s,x,r)	$\frac{(k+m)^2}{(k+m+s\cdot x+r)/s}+\frac{(s\cdot x+r)\times s\cdot x}{k+m+s\cdot x+r}$

3.3.2 Multi-node Failure Decoding

Theorem 1. *Setting the number of lost nodes to x can fix this error when there are x check blocks associated with the lost data block.*

Proof. If there are x check blocks, this means that x check equations can be generated with unknown block data. When encoding in section 2.2, the g_i in the generation matrix of each check block is pairwise and distinct, and each set of equations is generated according to the RS code encoding process. Thus in a sufficiently large GF(2^w), each calibration equation within the band is linearly independent. By combining x calibration equations into a calibration equation system, a linear equation system $Ax = b$ can be obtained. Since each calibration equation is linearly independent, it can be concluded that the coefficient matrix A is of full rank, and the rank of $Ax = b$ satisfies the following conditions: $r(A) = r(A, b) = x$.

When the system of linear equations can have a unique solution, then the error can be corrected and verified. □

The basic idea of Algorithm 1 is to first repair the column with the highest repair performance to repair the lost nodes that can be directly repaired, and then use row repair and global repair in order to repair other repairable nodes. Then obtain each encoding equation associated with the remaining missing blocks, and repeat the loop traversal. Assuming that the decoding condition $r(A) = r(A, b) = x$ can be satisfied, query whether each decoding equation belongs to the column check equation group. If it does not belong to the column check equation group, delete the equation. Then the decoding equation set has the best repair cost.

Taking (11,4,1) CGRC as an example, the encoding and decoding process, as well as the execution process of Algorithm 1, will be explained below. The layout of CGRC (11,4,1) is shown in Figure 6.

The encoding algorithm in the first section of this chapter can obtain the calculation equation of the check block,

Algorithm 1 Multi-node Decoding Algorithm

Input: $fail\_loc$: record the location of the current error nodes. *decode_eqa*: decoding equations. *encode_row*: row check block encoding equations

Output: Decoding result

```
Begin
for i in range len(fail_loc) do
    if SingleNodeDecode(fail_loc[i]) == true then
        fail_loc.remove(fail_loc[i])
    end if
end for
for i in range len(fail_loc) do
    if result=SearchEquation(fail_loc(i))!=null then
        if result is not in decode_eqa then
            decode_eqa.add(result)
        end if
    end if
end for
if len(decode_eqa ≥)len(fail_loc) then
    while len(decode_eqa ≥)len(fail_loc) do
        if decode_eqa(i) is in encode_row then
            decode_eqa.remove(decode_eqa(i))
        end if
    end while
    MultiNodeDecode(decode_eqa)
    return True
end if
return False
End
```

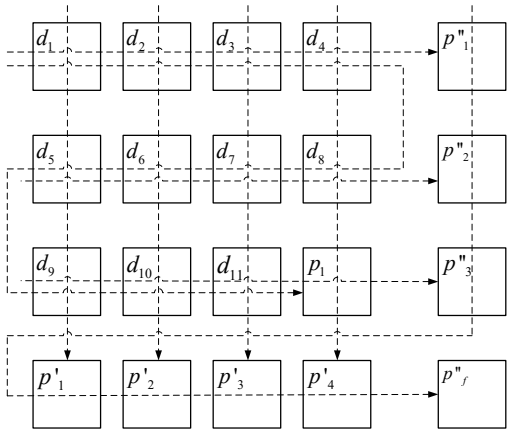


Figure 6: (11,4,1)CGRC Example. (k = 11 data block and c = 4 column local parity block, r = 1 global parity block)

as shown in the following equation:

$$\begin{aligned} p_1 &= d_1 + a_1 d_2 + a_1^2 d_3 + a_1^3 d_4 + a_1^4 d_5 + a_1^5 d_6 + a_1^6 d_7 \\ &\quad + a_1^7 d_8 + a_1^8 d_9 + a_1^9 d_{10} + a_1^{10} d_{11} \\ p_1' &= d_1 + a_2 d_5 + a_2^2 d_9 \\ p_2' &= d_2 + a_3 d_6 + a_3^2 d_{10} \\ p_3' &= d_3 + a_4 d_7 + a_4^2 d_{11} \end{aligned}$$

$$
\begin{aligned}
p'_4 &= d_4 + a_5 d_8 + a_1 a_5^2 d_2 + a_1^2 a_5^2 d_3 + a_1^3 a_5^2 d_4 \\
&\quad + a_1^4 a_5^2 d_5 + a_1^5 a_5^2 d_6 + a_1^6 a_5^2 d_7 + a_1^7 a_5^2 d_8 \\
&\quad + a_1^8 a_5^2 d_9 + a_1^9 a_5^2 d_{10} + a_1^{10} a_5^2 d_{11} \\
p''_1 &= d_1 + a_6 d_2 + a_6^2 d_3 + a_6^3 d_4 \\
p''_2 &= d_5 + a_7 d_6 + a_7^2 d_7 + a_7^3 d_8 \\
p''_3 &= d_9 + a_8 d_{10} + a_8^2 d_{11} + a_1^2 a_8^3 d_3 + a_1^3 a_8^3 d_4 \\
&\quad + a_1^4 a_8^3 d_5 + a_1^5 a_8^3 d_6 + a_1^6 a_8^3 d_7 + a_1^7 a_8^3 d_8 \\
&\quad + a_1^8 a_8^3 d_9 + a_1^9 a_8^3 d_{10} + a_1^{10} a_8^3 d_{11} \\
p''_f &= d_1 + a_2 d_5 + a_2^2 d_9 + a_9 \left(d_2 + a_3 d_6 + a_3^2 d_{10}\right) \\
&\quad + a_9^2 \left(d_3 + a_4 d_7 + a_4^2 d_{11}\right) + a_9^3 p'_4 \\
&\quad + a_9^4 \left(d_1 + a_6 d_2 + a_6^2 d_3 + a_6^3 d_4\right) \\
&\quad + a_9^5 \left(d_5 + a_7 d_6 + a_7^2 d_7 + a_7^3 d_8\right) + a_9^6 p''_3
\end{aligned}
$$

From the encoding equation, it can be seen that for each part of the verification block, its computational complexity and required the data blocks are different. The column local parity block has the best computational performance, requiring only three data blocks for computation, ensuring the reliability of a few nodes while also possessing excellent repair performance. For row check blocks, which have better fault tolerance performance than column check blocks, only three additional check blocks are required to provide fault tolerance for all data blocks. Therefore, it generally provides additional support when column check blocks cannot recover from errors. Global check blocks and p''_f provide the most basic fault tolerance capabilities for all data blocks and all local check blocks, respectively.

For a single-node error, if the error node $d_e \notin P''_r$ is present, for example, if the error node is d_1, priority should be given to using the In-column repair. By using the repair equation of the local check block p'_1 in the column, only the data of the three nodes of d_5, d_9, p'_1 can be read to recover the lost node d_1. This type of error repair has the lowest cost and accounts for 80% of the total single-node error probability.

If there is an error node $d_e \in P''_r$, for example, if the error node here is p''_1, priority should be given to repairing in the column. By using the repair equation of the row local check block p''_1, the lost node p''_1 can be recovered by simply reading the data of the d_1, d_2, d_3, d_4 node. This type of error repair has a high cost, but only accounts for 15% of the total error.

If the error node is p''_f, it must to be decoded by concatenating all row and column check blocks. It is the most expensive single-node error to fix, but due to its small number, which representing only accounts for 5% of the total probability, its high cost can be ignored.

For multi-node errors, follow the process of Algorithm 1 to repair them. If the error node is set to $d_1, d_2, d_3, d_5, d_6, d_7$, the algorithm begins to search for the best performing Intra-column check block p'_1, p'_2, p'_3 for each node. It finds the corresponding Intra-row check block p''_1, p''_2 for each node, and finally use the global check block p''_1. A total of six linearly independent coding equations can be obtained, and after judging by Theorem 1, the multi node error can be fixed.

3.3.3 Reliability Model

Markov models are commonly used to estimate the reliability of systems, so we take (8,3,1) CGRC as an example and establish a Markov model as shown in Figure 7. In industry, MTTF (Mean Time To Failure) is commonly used to analyze the reliability of products, while MTTDL (Mean Time To Data Loss) is often used in storage systems to analyze the reliability of data. In this article, it is used to analyze the specific reliability of CGRC. In Figure 7, λ and μ are used to represent the failure rate and conversion rate of a node. The higher λ of a node, the higher the probability of the node will fail. The higher the sum of all nodes λ, the greater the probability of the distributed storage system failing. The larger the μ value of a node, the higher the repair rate, which represents the current state transitioning to the next state, the higher the stability of the system, and the safer the data.

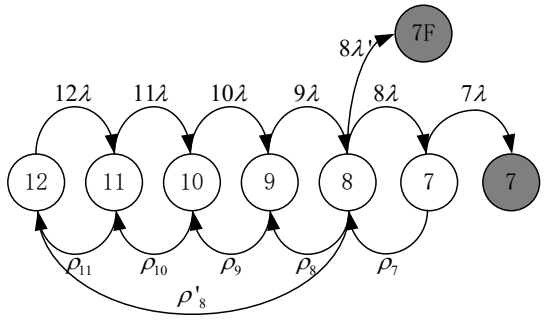


Figure 7: Markov Model for (8,3,1) CGRC

From the Markov model of the (6,3,0) CGRC, it can be seen that the encoding scheme generates a total of 6 check blocks, including 3 row check blocks and 3 column check blocks. According to Chapter 1, the maximum fault tolerance of this coding theory is 6. The transition in the positive direction of the model represents the process of system nodes entering the next state due to errors. According to the model, it can be seen that if there are fewer faulty nodes in the system, the probability of node failure is the highest. In this figure, the maximum error probability occurs in state 12, which is 12λ. From this, it can be seen that the probability of a few nodes in the system making mistakes, especially a single-node making mistakes, is much higher than the probability of multiple nodes making mistakes. Therefore CGRC optimized for a few nodes, especially single-node errors, can have better repair performance.

State 12 represents that the system is in a healthy state. When one node fails, it enters state 11 with a probability of 12λ. Conversion in a positive direction in sequence, indicating an increase in the number of node errors. When the system enters state 8, there are two transition results, which means that when a node error

occurs in this state, the system will enter state 8 and state 8F respectively, where state 7 is recoverable and state 7F is irreparable, with probabilities of $8\lambda, 8\lambda'$, respectively. Here, the probability of an unsolvable error is $8\lambda' = 89.3\%$.

The white circle in the model indicates that the current system state is healthy or repairable, while the gray circle indicates the current system has entered an irreparable state. Converrsely, μ represents the repair rate, where ρ_{11} represents the repair probability in the case of a single-node error. Equation (21) is the calculation formula for MTTDL, where assuming MTTF is 4 years, let the system have a total of N nodes, γ total bandwidth, and the storage space of each node is M.

$$MTTDL = \frac{\rho^m}{\prod_{i=0}^{m} (n-1)\,\lambda^{m+1}} \tag{21}$$

For (6,3,0) CGRC, 9 out of 12 nodes can be repaired through 2 nodes, 1 node can recoverde by 5 nodes, and 2 nodes can be restored through 3 nodes. Therefore, the average cost of repairing a single disk $\rho_7, \rho_9, \cdots, \rho_{11}$ is $t = (9 \times 2 + 1 \times 5 + 2 \times 3)/12 = 2.41$. ρ_8' represents an error in the system that cannot be repaired by a single node, meaning that it cannot be restored to state 9 through a single-node repair. Four nodes are required to decode a quaternion encoding equation system, and after decoding is completed, it returns to state 12, which is the healthy state. So the average repair cost of ρ_8' is $t_4 = 6$. We set the parameters number of nodes $M = 16TB$, network bandwidth $\gamma = 1Gbps$, number of nodes $N = 400$, and a hard disk fault-free time of 4 years. The final MTTDL is shown in Table 3.

Table 3: The reliability of RS,LRC,CGRC

Code	MTTDL
RS(12,6)	7.4×10^{11}
LRC(6.3.3)	4.7×10^{12}
CGRC(6,3,0)	6.8×10^{12}

It can be seen that CGRC has better reliability compared to RS and LRC. And can ensure system stability, while also providing higher reliability to the system for higher fault tolerance performance.

4 Experiment and Discussion

This experiment mainly deploys CGRC and other erasure codes on a Ceph-based distributed erasure code testing platform and compares their performance. In order to obtain the most realistic experimental results.

4.1 Experimental Environment

Ceph is a large and reliable distributed storage system. This erasure code testing experimental platform is built based on the Ceph Pacific (16.2.13) version, and its main components include Monitor and OSD. OSD is a data node responsible for storing and managing data. It is divided into the Primary OSD main data node and the OSD ordinary data node. We extend the erasure code OSD based on its existing framework to add different erasure code schemes. The Monitor is responsible for managing data nodes and interacting with the clients. The specific structure diagram is shown in Figure 8.

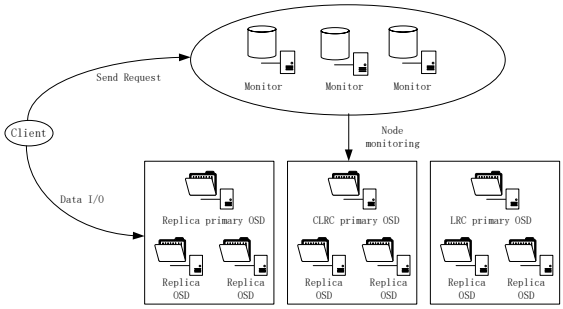


Figure 8: Architecture Diagram of Erasure Code Experimental Platform

Ceph's OSD often uses the three-replica technology for Data redundancy. This experimental platform adds CGRC Primary OSD and other contrast erasure codes OSD in the native Ceph, so as to obtain accurate experimental data more conveniently and quickly.

4.2 Fault Tolerance

Fault tolerance refers to the number of errors that can be corrected by erasure codes, which is the maximum number of lost nodes that can be corrected in a distributed storage system. The fault-tolerant ability of erasure codes is a very important parameter in storage systems. Figure 9 mainly shows the multi-node fault tolerance capabilities of three different encoding schemes: CGRC (13,7,1), CGRC (13,5,2), and CGRC (13,5,3). Its horizontal axis represents the current number of lost nodes, and its vertical axis represents the repair probability under the current number of lost nodes.

For CGRC with three different parameters, their respective additional storage space is 11,11,13. Comparing CGRC (13,7,1) and CGRC (13,5,2), it can be seen that under the same cost, different parameters and encoding layouts, the repair efficiency will also vary. Although both encoding schemes use 11 additional storage spaces, the encoding array is different. The latter has a coding layout that is closer to a square, resulting in better repair efficiency. Therefore, CGRC can provide more flexible encoding methods to meet more differentiation needs. CGRC (13,5,3) achieves the most perfect repair efficiency by adding a small number of storage nodes. However,

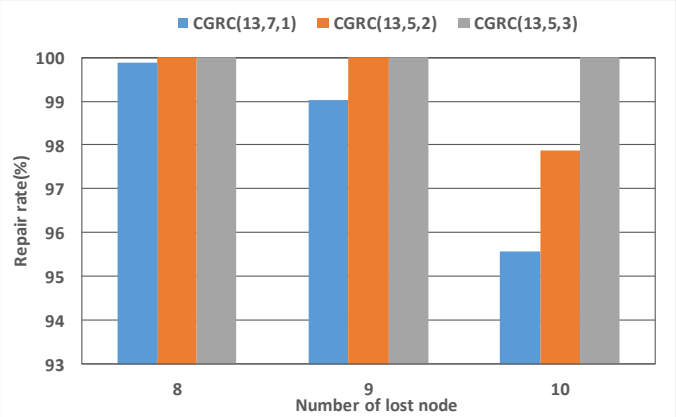


Figure 9: Comparison of CGRC multi-node repair rates under different parameters

overall, the multi-node fault tolerance capabilities of different CGRCs are excellent, achieving a repair rate of over 95%.

4.3 Single-node Recovery Performance

According to the reliability analysis in section 2.3.3, the probability of node failures in the system is highest when there are more surviving nodes, so a small number of node errors are the most common errors faced in storage systems. The repairability of individual nodes can greatly determine the stability and reliability of distributed storage systems and is also one of the most critical performance indicators of erasure code schemes. This section compares the single-node repair costs of RS (24,12), Pyramid (24,12), LRC (12,6,6,6), DLRC (12,6,3,6), as well as CGRC (12,7,2) and CGRC (12,5,2). The specific results are shown in Figure 10, where the horizontal axis represents different encoding schemes, and the vertical axis represents the average cost of repairing a single node under the current scheme.

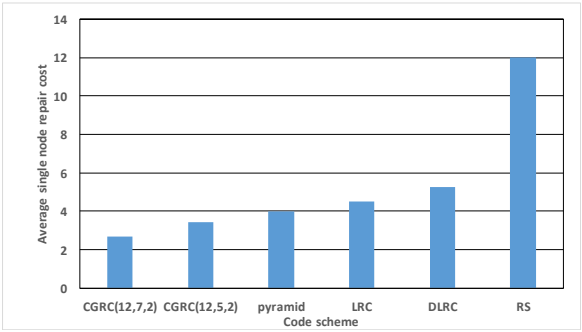


Figure 10: Single-node repair overhead

Figure 10 shows that for the same storage cost, the two different parameter encoding schemes of CGRC provide the best average single-node repair cost. Among them, CGRC (12,7,2) requires only an additional 2.7 auxiliary nodes per single-node repair on average, and CGRC with different coding layouts only requires 3.458 auxiliary nodes, which has certain advantages over other coding schemes.

In the real testing environment, this section used six different encoding schemes, namely RS (24,12), Pyramid (24,12), LRC (12,6,6), DLRC (12,6,3,6), and CGRC (12,7,2), CGRC (12,5,2), to encode six files of different sizes, corresponding to the horizontal axis in Figure 11. Finally, the single-node repair cost was calculated for each of the six different encoding schemes, and the specific experimental results are shown in Figure 11.

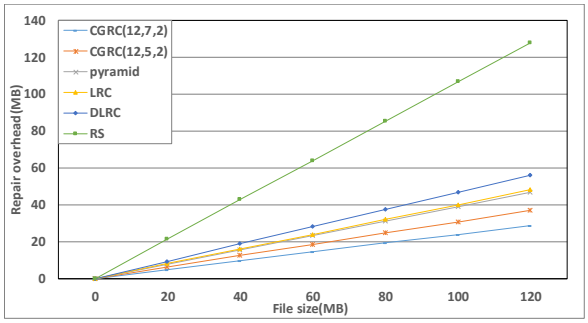


Figure 11: Single-node repair overhead under different file sizes

According to Figure 11, it can be seen that the two encoding schemes using CGRC have lower single-node repair costs for files of different sizes, and have better single-node repair performance compared to LRC, DLRC, and Pyramid. Among them, the repair cost of CGRC (12,7,2) decreased by 39.23% compared to Pyramid, 40% compared to LRC, 48.56% compared to DLRC, and 77.5% compared to RS.

4.4 Multi-node Recovery Performance

In distributed storage systems, although the probability of multi-node errors is much lower than the probability of a single-node error, it is still one of the factors contributing to system instability. CGRC uses Algorithm 1 to achieve recovery from multi-node errors, which still has a good ability in multi-node recovery efficiency and can ensure system reliability. This section deployed and experimented with Pyramid (24,12), LRC (12,6,6), DLRC (12,6,3,6), and CGRC (12,7,2). The specific results are shown in Figure 12, where the horizontal axis represents the number of faulty nodes and the vertical axis represents the average recovery cost under the current number of faulty nodes.

According to Figure 12, it can be seen that CGRC has excellent multi-node fault tolerance ability. When the number of lost nodes is small, the repair cost is better compared to other solutions. For example, when three nodes are lost, the multi-node repair cost of CGRC is

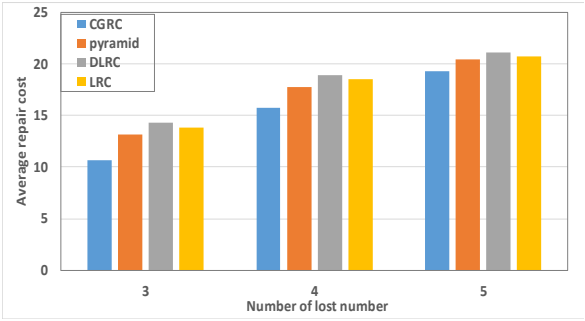


Figure 12: Multi-node repair overhead

reduced by 18.9%, 25.1%, and 22.4% compared to Pyramid, DLRC, and LRC, respectively. When four nodes are lost, the multi-node repair cost of CGRC is reduced by 11.7%, 16.9%, and 15.1% compared to Pyramid, DLRC, and LRC, respectively. Even when 5 nodes are lost, the average improvement compared to three other encoding schemes is 6.83%.

In the real testing environment, this section used four different encoding schemes, namely Pyramid (24,12), LRC (12,6,6), DLRC (12,6,3,6), and CGRC (12,7,2), to encode four files of different sizes. The file sizes correspond to the horizontal axis in Figure 13. Finally, the 4-node repair cost was calculated for each of the four different encoding schemes, and the specific experimental results are shown in Figure 13.

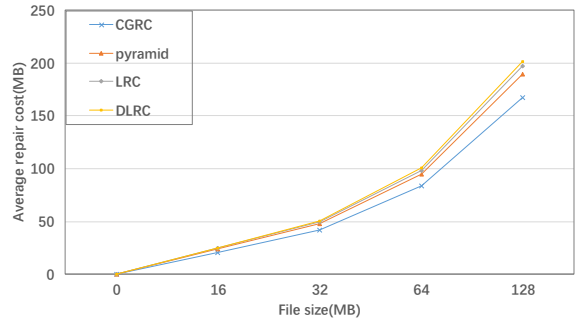


Figure 13: Multi-node repair overhead under different file size

According to Figure 13, it can be seen that in real distributed environments, CGRC has excellent multi-node repair costs for different file sizes. At the same time, as the size of the encoded file, the advantage of repair cost becomes larger.

4.5 Storage Efficiency

Storage efficiency is one of the key indicators in erasure code schemes, which determines the proportion of the real data storage space in the system to the total usage space. The larger the proportion, the higher the storage efficiency. Figure 14 shows the storage efficiency diagrams of five different encoding schemes: CGRC (14,5,1), Pyramid (12,10), LRC (16,8,2), DLRC (14,2,4,8), and RS (14,10). The horizontal axis represents different encoding schemes, and the vertical axis represents the storage efficiency corresponding to the scheme.

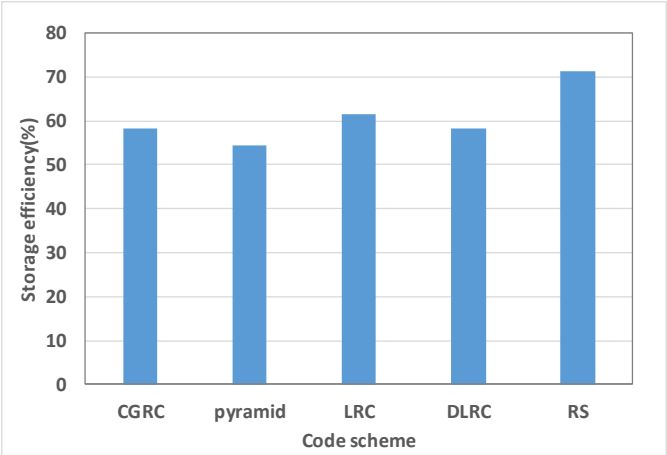


Figure 14: Storage efficiency

According to Figure 14, CGRC sacrifices a little storage efficiency compared to RS and LRC, with a storage efficiency difference of about 3.2% compared to LRC. However, CGRC provides better node repair performance and lower node repair costs. So this small additional cost can be ignored. Compared to Pyramid and DLRC, CGRC has similar fault tolerance and better single-node repair performance, while also improving storage efficiency by 3.79% compared to Pyramid. So overall, the storage efficiency of CGRC is within an acceptable range.

5 Conclusions

With the increase of information, the risk of data loss in distributed storage systems is increasing, so more efficient fault-tolerant solutions are highly needed. At the same time, according to research [13], more than 90% of data loss in storage systems is single-node loss.

In this paper, a systematic comparison between CGRC and multiple erasure code schemes is conducted, and the advantageous points of CGRC mainly focus on the single-node repair cost, and the repair cost remains excellent when a small number of nodes are lost. In terms of single-node repair overhead, CGRC(12,7,2) reduces 39.23% compared to pyramid, 40% compared to LRC, 48.56% compared to DLRC, and 77.5% compared to RS. So CGRC is optimized for a small amount of node loss, and can better ensure the reliability of data in distributed storage systems. CGRC can adjust the encoding parameters to change the layout during encoding, allowing CGRC to have different fault tolerance and repair costs under the same storage overhead, which can better adapt to different storage requirements in distributed storage systems.

CGRC still has shortcomings. When the number of lost nodes is large, there are certain limitations, recovery still requires large I/O overhead, secondly, how to balance the coding parameters to achieve the best and most balanced state of fault tolerance and repair overhead in the distributed storage system, further research is needed.

6 Acknowledgments

This work was supported by the Science and Technology Support Project of Sichuan Province (2022YFG0037) and (2022YFG0033).

References

[1] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Ieee, 2010, pp. 1–10.

[2] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *Proceedings of the nineteenth ACM symposium on Operating systems principles*, 2003, pp. 29–43.

[3] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*. Elsevier, 1977, vol. 16.

[4] H. C. T Dan and W. Geng, "Decoding method of reed-solomon erasure codes(in chinese)," *Computer Research and Development*, vol. 3, no. 59, pp. 582–596, 2022.

[5] H. Weatherspoon and J. D. Kubiatowicz, "Erasure coding vs. replication: A quantitative comparison," in *Peer-to-Peer Systems: First InternationalWorkshop, IPTPS 2002 Cambridge, MA, USA, March 7–8, 2002 Revised Papers 1*. Springer, 2002, pp. 328–337.

[6] C. Rajput and M. Bhaintwal, "Optimal rs-like lrc codes of arbitrary length," *Applicable Algebra in Engineering, Communication and Computing*, vol. 31, no. 5, pp. 1–19, 2020.

[7] Z. Zhang, A. Deshpande, X. Ma, E. Thereska, and D. Narayanan, "Does erasure coding have a role to play in my data center?" 2010.

[8] R. M. Roth and A. Lempel, "On mds codes via cauchy matrices," *IEEE transactions on information theory*, vol. 35, no. 6, pp. 1314–1319, 1989.

[9] M. Blaum, J. Brady, J. Bruck, and J. Menon, "Evenodd: An efficient scheme for tolerating double disk failures in raid architectures," *IEEE Transactions on computers*, vol. 44, no. 2, pp. 192–202, 1995.

[10] L. Xu and J. Bruck, "X-code: Mds array codes with optimal encoding," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 272–276, 1999.

[11] P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong, and S. Sankar, "Row-diagonal parity for double disk failure correction," in *Proceedings of the 3rd USENIX Conference on File and Storage Technologies*. San Francisco, CA, 2004, pp. 1–14.

[12] X. Xin, V. Bane, L. Shu, L. Juane, and A. G. Khaled, "Quasi-cyclic ldpc codes with parity-check matrices of column weight two or more for correcting phased bursts of erasures," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 2812–2823, 2021.

[13] N. Ghadbane and D. Mihoubi, "A construction of some group codes," *International Journal of Electronics and Information Engineering*, vol. 4, no. 1, pp. 12–21, 2016.

[14] L. Xianghong and S. Jiwu, "Summary of research for erasure code in storage system," *Journal of computer research and development*, vol. 49, no. 1, pp. 1–11, 2012.

[15] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin, "Erasure coding in windows azure storage," in *Presented as part of the 2012 {USENIX} Annual Technical Conference ({USENIX}{ATC} 12)*, 2012, pp. 15–26.

[16] C. Huang, M. Chen, and J. Li, "Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems," *ACM Transactions on Storage (TOS)*, vol. 9, no. 1, pp. 1–28, 2013.

[17] S. Zhou and Y. Wang, "Expyramid: A flexible high error tolerance and low repair cost coding scheme based on array structure(in chinese)," *Computer Research and Development*, vol. 1, 2011.

[18] Y. Meng, L. Zhang, D. Xu, Z. Guan, and L. Ren, "A dynamic erasure code based on block code." in *EWSN*, 2019, pp. 379–383.

[19] T. Miyamae, T. Nakao, and K. Shiozawa, "Erasure code with shingled local parity groups for efficient recovery from multiple disk failures," in *10th Workshop on Hot Topics in System Dependability (HotDep 14)*, 2014.

[20] Z. Wang, Z. Xie, and D. Tang, "An erasure code with low recovery-overhead based on a particular three-hierarchical redundancy structure," *International Journal of Network Security*, vol. 24, no. 5, pp. 965–974, 2022.

Biography

Wenjie Deng master candidate. He is currently an Master student in Chengdu University of Information Technology, Chengdu, China. His research interests include coding theory and information security.

Cong Li master candidate. He is currently an Master student in Chengdu University of Information Technology, Chengdu, China. His research interests include coding theory and information security.

Tieyuan Hong master candidate. He is currently an Master student in Chengdu University of Information Technology, Chengdu, China. His research interests include coding theory and information security.

Dan Tang received his Ph.D. degree from Graduate University of Chinese Academy of Sciences (CAS), Beijing,

China in 2010. He is currently an associate professor with Chengdu University of Information Technology, Chengdu, China. His research interests include coding theory and secret sharing scheme.

A Novel Malicious Code Propagation Model Based on Dual Defense and Honeypot Feedback

Chenxi Li, Jianguo Ren, and Fengjiao Li
(Corresponding author: Jianguo Ren)
School of Computer Science and Technology, Jiangsu Normal University
Xuzhou Jiangsu 221116, China
Email: jsnucs1119@163.com
(Received July 5, 2023; Revised and Accepted Jan. 5, 2024; First Online June 22, 2024)

Abstract

A SUDRS model based on double defense is built in this paper, considering the hiding of malicious code during transmission, the dynamic nature of the network, and the coexistence of various defenses in the real network. Honeypot technology is used to create the H-SUDRS model based on the SUDRS model. Furthermore, some related dynamics properties are studied, analyzing the model's basic reproduction number and demonstrating its stability at the disease-free equilibrium point. The SUDRS model is more realistic than the original SEIRS model, according to numerical simulation and experiments, and the double defense in the SUDRS model has a containing impact on the spread of malicious code in the network. Through the honeypot's direct feedback and cloud feedback, the H-SUDRS model enhances the network's active defensive capability and lessens the damage of new malicious code on the network.

Keywords: Double Defense; Honeypot; Malicious Code; Network Security; Propagation Model

1 Introduction

Security hazards persist even when network-related technology makes life easier. Network security issues have always been present along with the evolution of the network. As more individuals become aware of the value of network security following the enactment of the "Network Security Law of the People's Republic of China" in 2017, network security concerns have increased [1]. Malicious code that is still operating on a device across the network without the owner's permission is referred to as network malicious code. In the broadest sense, malicious code includes spam, harmful pop-ups, computer viruses, and network worms, all of which will negatively affect user experience, especially network worms [2]. While worms are a type of computer virus as well, they are more hazardous when used in a network. Computer viruses resemble biological viruses in many ways. They can replicate themselves and spread to new hosts, which is incredibly dangerous. Studying the dissemination of malicious code is therefore crucial for preserving computer security.

The study of network malicious code propagation today can be split into micro and macro components. The micro aspect is based on the characteristics of the network malicious code itself to develop anti-malicious code software [3, 4]. The macro aspect is to learn from the relevant models of biological infectious diseases, establish a corresponding malicious code propagation model [5, 6], and study the law of the evolution of the number of malicious code over time. In order to successfully prevent the transmission of malicious code on the network, it is necessary to foresee the evolution trend of malicious code. The dynamic propagation model is a workable method for examining network malicious code from a macro perspective. The first consideration is the device and the user together as a node, and the transformation relationship between the states is an edge. The appropriate propagation model is then built. The magnitude of malicious code diffusion is examined, and the development trend of malicious code is projected, using numerical simulation experiments and dynamic equation analysis. Finally, network security is preserved and network malicious code is prevented and controlled at the macro level [7]. It is essential to correspond to the actual network environment when creating the malicious code propagation model in order to better understand the law of malicious code propagation at the macro level. The effect of defense mechanisms on the spread of malicious code must therefore be carefully considered in the existing network.

The two primary categories of malware defense technology today are static defense and dynamic defense. The static defense relies on the virus library, which makes it somewhat out-of-date and ineffective at identifying newly created malicious code. Sandbox technology can be used to implement dynamic defense, which is a behavior-based dynamic detection mechanism. Even though dynamic detection technology is capable of detecting brand-new

malicious code, it will only do so when that malicious code exhibits harmful behavior. Additionally, as malicious code produces a number of sandbox detection technologies when it antagonizes security software, malicious code can use these sandbox detection technologies to determine if it is running in the sandbox, thereby avoiding its own behavior exposure [8]. The dual defense, which combines static and dynamic defense, is still too passive in the attack and defense of networks. A proactive defense technology, termed honeypot technology, is introduced to change the passive state of defense technology in network confrontation. The honeypot technology may trap and analyze attack behavior, infer the attack aim and motivation, and then feed back the relevant preventive and control information by deceiving attackers into using it unlawfully. These prevention and control measures are intended to improve the computer's security protection capabilities through technology and management, protecting the hardware from the destruction of new dangerous programs [9].

In-depth research on the propagation of network malicious code has been done recently by researchers both domestically and internationally [10, 11]. The dynamic evolution of the network virus propagation model was examined in the literature [12]. There could be a new node accessed or an old node offline at any time, and the number of network nodes is always changing due to differences in global time, region, and user preferences. Dynamic networks are more in accordance with reality than the static networks used in earlier studies. Studying the law of malicious code spread in dynamic networks is therefore more useful. A worm propagation model in a cloud security environment was proposed by Zhang W et al. [13], with an emphasis on the influence of cloud security deployment and data gathering capabilities on the worm propagation model. The related ideas of cloud security have gained popularity alongside the development of cloud computing technology. More and more devices are entering the cloud security environment by installing cloud security software, and some popular operating systems themselves feature cloud security protection for devices, allowing for safe network usage. By updating the cloud virus library, the cloud server may more rapidly and effectively transmit prevention information against new malicious code to the device that installs the cloud protection software. Therefore, the study of network malicious code propagation models in a cloud security environment is more in line with the real network environment. With an emphasis on the latency of computer viruses in the propagation process, Peng M et al. [14] established an SEIR direct immunization-based computer virus propagation model. The study's consideration of the virus's latency is a complement to the earlier SIR model [15, 16], however, the latency taken into account in the literature [14] still has certain drawbacks. In actuality, the device is infectious as soon as it contracts the network virus. Despite not having been discovered to be infected, or in the latent state described in the article [14], the device can distribute the relevant malicious code throughout the network. Therefore, the incubation period without infectivity is not consistent with the actual scenario when a virus spreads over a network. The defense methods mentioned above are static defense. It is of practical significance to analyze the influence of behavior-based dynamic detection technology on the propagation law of malicious code in the network. However, there hasn't been much research done on the impact of behavior-based dynamic detection technology on the macro-scale spread of malicious code in the network. The depth of the research on how dynamic protection mechanisms affect the spread of malicious code within a network is insufficient.

Based on an analysis of the above research's limitations and the SEIRS model [17, 18], this paper takes into account several scenarios in which malicious code might appear during the propagation process in the actual network environment, including the network's dynamics, the ability of infected nodes in the incubation period to spread infection, and the fact that the current defense measures in the actual network environment is not only static defense. The SEIRS model is enhanced, and a more precise model for the spread of malicious code the Susceptible -Unaware-Discoverable-Recovered-Susceptible (SUDRS) model is suggested considering the influence of these crucial aspects on the real network. After considering the crucial elements in the real network, the SUDRS model focuses on the influence of the dual defense of static defense and dynamic defense on the spread of malicious code in the network.

This paper also introduces honeypot technology into the SUDRS model to construct a new malicious code propagation model H-SUDRS. The H-SUDRS model is built using the SUDRS model as a foundation and then considers the influence of the honeypot feedback mechanism on the dissemination of malicious code in the network. It focuses on the analysis of the feedback mechanism in honeypot technology to curb the spread of malicious code in the network.

2 Model Construction

We may create a new malicious code propagation model, the SUDRS model, by upgrading the SEIRS model. Then, on the basis of the SUDRS model, honeypot technology is used to build the H-SUDRS model. The H-SUDRS model not only adds a new chamber based on the SUDRS model, but it also expands the immune pathway of nodes, altering the original immune way.

2.1 Construction of SUDRS Model

By using the SEIRS model as a foundation, the SUDRS model is enhanced. In a cloud-secure defensive setting, the SUDRS model also includes behavior-based dynamic detection defense. Node security is maintained through the dual defense of static defense and dynamic de-

fense. Cloud security software begins by gathering a lot of suspicious data from the client, identifies new risks either automatically or manually, and then promptly feeds that information back to the cloud server. Finally, it keeps the node's virus library updated to enable the node to fend off attacks from malicious code. Cloud security software can update new threat solutions to nodes earlier and faster than conventional anti-malicious code techniques. This solution, however, continues to rely on the viral library's static defense. Even though behavior-based dynamic detection technology can't entirely compensate for static defense's drawbacks, it can give nodes the capacity to identify undiscovered malicious code and improve their dynamic defensive capabilities.

The four states that each node in the SUDRS model can be in are as follows: (1) Susceptible state (S), where S stands for the node of the susceptible state and is defined as the node that lacks immunity and malicious code infection. (2) Unaware state, the unaware node is designated by the symbol U. Unaware nodes are those that have been compromised by malicious code but have gone undetected by cloud security software on users or devices. (3) Discoverable (D), where D denotes the node of the discovery state and "discovery node" refers to the node discovering malicious code by a person or antiviral software. (4) Recovered state (R), Recovered state can also be understood as the immune state, R represents the node of the immune state, and the node with the immune ability is called immune node.

N is used to represent the overall number of nodes in the network to reflect the relationship between the number of nodes in each state. One way to describe the total number of nodes at time t is as follows: $N(\mathrm{t}) = S(\mathrm{t}) + U(\mathrm{t}) + D(\mathrm{t}) + R(\mathrm{t})$.

The network's nodes will go into an Unaware state when they are infected with malicious code. The nodes in this condition are infected nodes that neither people nor cloud security software has found. They are not aware that they are sick, which is the meaning of the term "unaware state." The Unaware node (U node), however, has the capacity to spread malicious code even when neither people nor cloud security software has found it.

When U node is discovered, there will be multiple development directions, and the cause of multiple development directions is multi-layer defense. SUDRS model is a malicious code propagation model based on dual defense, and its key mechanism is multi-layer defense that combines static defense and dynamic defense. The construction of multi-level defense algorithm is helpful to analyze the complexity of node development. The concrete implementation process of multi-level defense algorithm is shown in Algorithm 1.

By analyzing multi-level defense algorithms, we can better understand the joint action of static defense and dynamic defense, which is helpful to further study the relationship between nodes in state transition.

Figure 1 depicts the state transition connection and the four states of the SUDRS model. Figure 1 shows four

Algorithm 1 State transition algorithm in multi-layer defense

Require: All of S
Ensure: Nodes after state transition

```
Begin
Malicious code detection
while Malicious code exist And Successful detection do
    Eliminate malicious code, upload information to the server.
    R ⇐ S, return R
end while
while Malicious code exist And detection failure do
    U ⇐ S
    if Found Malicious code then
        F ⇐ U
        R ⇐ F(Static defense Or external intervention solves malicious code, return R)
        Or S ⇐ F(Dynamic defense deals with malicious code, return S)
        Or Keep F, return F
    end if
    return U
end while
End
```

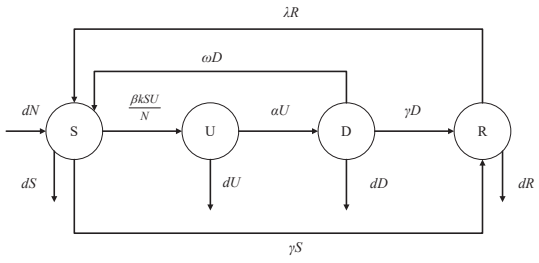


Figure 1: SUDRS model state transition diagram

circular boxes, each of which corresponds to a different condition. In Figure 1, the directed line segments indicate the node state transition's direction, and the symbols next to the directed line segments indicate its parameters. Table 1 displays the relevant variables and parameters.

In some previous studies [20, 21], the contact behavior between susceptible nodes and infected nodes (this paper's U node is the infected node) is expressed as a single mathematical symbol and directly defined as the infection rate after analysis, abstraction, and generalization. This expression is not specific enough. To reflect the contact behavior more specifically between susceptible nodes and infected nodes, it is assumed that the probability of each contact infection between susceptible nodes and infected nodes is β_1, and the number of contacts between an infected node and other nodes per unit of time is n. According to the research of literature [22], the number of contacts per unit time n is directly proportional to the network node degree k_i of i nodes. Assuming that the

Table 1: The symbols involved in the proposed model are explained

Sign	Meaning of each sign
S(t)	The number of susceptible nodes at time t
U(t)	The number of unaware nodes at time t
D(t)	The number of discoverable nodes at time t
R(t)	The number of recovered nodes at time t
H_A(t)	The number of susceptible honeypot nodes at time t
H_B(t)	The number of infected honeypot nodes at time t
N	Total number of nodes
M	Total number of honeypots deployed
d	The rate of node joining and leaving
β	Infection coefficient of susceptible nodes
k	Degree of node
α	Probability of malicious code being discovered
γ	Immune rate of nodes
λ	Immune failure rate of nodes
ω	Successful isolation rate of dynamic defense
β_0	Infection coefficient of susceptible nodes after the introduction of honeypot
α_0	Probability of malicious code being discovered after the introduction of honeypot
γ_0	Immune rate of nodes after the introduction of honeypot
λ_0	Immune failure rate of nodes after the introduction of honeypot
ξ_1	Feedback rate of infected honeypot nodes to susceptible nodes
ξ_2	Feedback rate of infected honeypot nodes to discovery nodes
β_H	Infection coefficient of susceptible honeypot node
δ	Rate of honeypot deployment and revocation

degree k_i of each node on the network is approximately equal to k, that is, $k_i \approx k$, it can be obtained:

$$n = \beta_2 k \tag{1}$$

Therefore, the number of infections per unit time of a single infected node is $\beta_1\beta_2 kS/N$, and the number of infections per unit time of all infected nodes is $\beta_1\beta_2 kSU/N$. Let $\beta = \beta_1\beta_2$ be the infection coefficient of the virus, then the number of infections can be simplified as $\beta kSU/N$.

According to the state transition relationship in Figure 1, the differential dynamic equation of the SUDRS model can be written as:

$$\begin{cases} \dfrac{dS}{dt} = dN + \omega D + \lambda R - \dfrac{\beta kSU}{N} - \gamma S - dS \\ \dfrac{dU}{dt} = \dfrac{\beta kSU}{N} - \alpha U - dU \\ \dfrac{dD}{dt} = \alpha U - \omega D - \gamma D - dD \\ \dfrac{dR}{dt} = \gamma D + \gamma S - \lambda R - dR \end{cases} \tag{2}$$

In addition to considering the dynamic nature of the real network and the ability of unaware nodes to spread malicious code, the SUDRS model focuses on the analysis of the containment effect of the dual defense system combining static defense and dynamic defense on the spread of malicious code in the network.

2.2 Construction of H-SUDRS Model

Honeypot technology is an active defense technology in network defense. To make network defense no longer in a passive position in network attack and defense confrontation, honeypot technology is very important. At present, honeypot technology has been widely used [23], and there is more and more research on honeypot technology [24, 25]. The honeypot induces the attacker to attack itself through the particularity of its configuration and records the attack behavior, studies and finds the attacker's attack purpose and attack means from the attacker's attack behavior record, helps the network defense system to collect malicious code data information and master the attack information, analyze the collected data, to enhance the security protection ability of the actual system through technical and management means, realize the continuous improvement of the system defense means, and improve the security of the network [26].

This research builds the H-SUDRS model on the foundation of the SUDRS model in order to incorporate honeypot technology into the SUDRS model and further investigate the limiting impact of honeypot active defense technology on the dissemination of malicious code in networks. Figure 2 depicts the state transition relationship of the H-SUDRS model. Table 1 provides a description of the parameters and variables used in Figure 2.

The honeypot is introduced into the H-SUDRS model as a state variable. H_A is a susceptible honeypot node and H_B is an infected honeypot node. In order to reflect the contact behavior more specifically between the susceptible honeypot node and the infected node, it is assumed that the susceptible honeypot node and the infected node contact each time and the probability of infection is β_3. Considering that the honeypot node captures the threat by constructing a deception environment, it is more likely to be infected by the infected node after contact, so β_3¿β_1. The number of honeypots that a single infected node can infect per unit time is $\beta_2\beta_3 kH_A/(N+M)$, and the number of honeypots that all infected nodes can infect per unit time is $\beta_2\beta_3 kH_AU/(N+M)$, denoted by β_H=$\beta_2\beta_3$

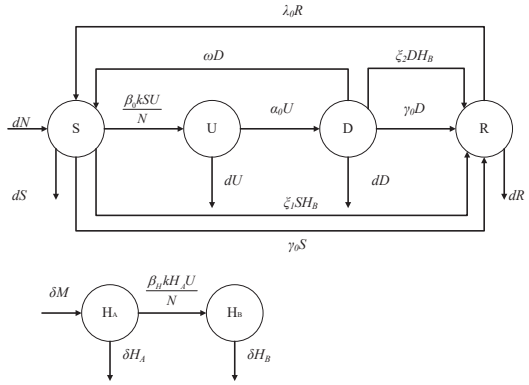


Figure 2: H-SUDRS model state transition diagram

and β_H¿β, the number of infections can be simplified to $\beta_H kH_A U/(N+M)$. Considering that the number of honeypots M is very small compared to the number of total nodes in the network N, the number of infections can also be expressed as $\beta_H kH_A U/N$.

In addition to trapping attacks and analyzing and solving security threats, the key to the honeypot technology also includes feedback on the means and methods that will solve the threat. When the honeypot obtains the attacker 's attack behavior by trapping, analyzes the malicious behavior record, and obtains the defense means against malicious code, the improved defense means can be fed back in two ways, one is direct contact feedback, and the other is cloud feedback. The ideal situation is that after the honeypot node is infected, it can analyze more perfect defense methods and generate corresponding prevention information based on malicious behavior records and related data.

The direct contact feedback is that the honeypot node directly feeds back the prevention and control information to other nodes in the surrounding contact after obtaining the prevention and control information against malicious code. It increases the way for susceptible nodes and discovery nodes to obtain immunity, so that susceptible nodes can obtain immunity through honeypot feedback. The feedback rate of susceptible nodes to obtain immunity is ξ_1. The immune pathway of unaware nodes is not expanded, only the immune pathway of discovery nodes, because nodes that are unaware, they are infected would be instantly transformed into discovery nodes after the discovery of malicious code. The immune node is discovered through direct contact feedback with the aid of the honeypot node that has preventive and control information, and the immune node's feedback rate is discovered to be ξ_2. The meaning of each node is not only the device, but also the user of the device. The device and the user are a whole. Therefore, considering the subjective factors of human beings, people who are under threat are more likely to accept the defensive help provided by others than those who are not threatened, so $\xi_2 > \xi_1$.

The cloud feedback is that the honeypot node immediately feedbacks the prevention and control information to the cloud server of the cloud security software after obtaining the prevention and control information against malicious code, updates the cloud virus library in time, and then helps the node with cloud security software protection. Update the virus library, and jointly maintain the security of the node through honeypot feedback and cloud security software. The cloud feedback mechanism of the honeypot has an impact on the defense of the entire system, and has an impact on other parameters other than the infection coefficient, such as the immune failure rate of the node, the infection coefficient of the susceptible node, the discovery rate of the unaware but infectious node, and the immune rate of the node. Following is the link between these parameters and the pertinent parameters prior to the use of honeypot technology:

$$\lambda_0 = p\lambda \tag{3}$$

$$\alpha_0 = \frac{\alpha}{p} \tag{4}$$

$$\beta_0 = p\beta \tag{5}$$

$$\gamma_0 = \frac{\gamma}{p} \tag{6}$$

In the above formula, p represents the adjustable constant controlled by the deployed honeypot, $0 < p \leq 1$. The more comprehensive and effective the deployment of the honeypot is, the lower the p value is. When $p = 1$, it means that there is no deployment of the honeypot or the deployment location of the honeypot is invalid.

The differential dynamic equation of the H-SUDRS model can be constructed as follows by the state transition relationship in Figure 2:

$$\begin{cases} \frac{dS}{dt} = dN + \omega D + \lambda_0 R - \frac{\beta_0 kSU}{N} - \gamma_0 S - dS - \xi_1 SH_B \\ \frac{dU}{dt} = \frac{\beta_0 kSU}{N} - \alpha_0 U - dU \\ \frac{dD}{dt} = \alpha_0 U - \omega D - \gamma_0 D - \xi_2 DH_B - dD \\ \frac{dR}{dt} = \xi_1 SH_B + \xi_2 DH_B + \gamma_0 D + \gamma_0 S - \lambda_0 R - dR \\ \frac{dH_A}{dt} = \delta M - \frac{\beta_H kH_A U}{N} - \delta H_A \\ \frac{dH_B}{dt} = \frac{\beta_H kH_A U}{N} - \delta H_B \end{cases} \tag{7}$$

3 Model Analysis

In the model construction of Chapter 2, the H-SUDRS model and its dynamic equation have been obtained. Then the basic reproduction number of the H-SUDRS model can be obtained by the dynamic equation and the stability analysis can be carried out.

3.1 Basic Reproduction Number

System (7) can be simplified as:

$$\begin{cases} \frac{dS}{dt} = dN + \omega D + \lambda_0(N - S - U - D) \\ \quad - \frac{\beta_0 kSU}{N} - \gamma_0 S - dS - \xi_1 S H_B \\ \frac{dU}{dt} = \frac{\beta_0 kSU}{N} - \alpha_0 U - dU \\ \frac{dD}{dt} = \alpha_0 U - \omega D - \gamma_0 D - \xi_2 D H_B - dD \\ \frac{dH_A}{dt} = \frac{\beta_H k(M - H_B)U}{N} - \delta H_B \end{cases} \tag{8}$$

N and M represent the total number of nodes and honeypots in the network, respectively, where N=S+U+D+R, M=M_A+M_B. In a dynamical system, the set of all possible points is a feasible region. In the dynamical system (8), the feasible region Ω is: $\Omega = \{S, U, D, M_B \in \Re_+^4 | S \geq 0, U \geq 0, D \geq 0, 0 \leq S + U + D \leq N, 0 \leq M_B \leq M\}$.

The above set is the positive invariant set of system (8).

By calculating the following equations, the disease-free equilibrium point of the H-SUDRS model can be obtained. The equilibrium point is determined by the following equations:

$$\begin{cases} \frac{dS}{dt} = 0 \\ \frac{dU}{dt} = 0 \\ \frac{dD}{dt} = 0 \\ \frac{dH_A}{dt} = 0 \end{cases} \tag{9}$$

Obviously, the disease-free equilibrium of the H-SUDRS model can be obtained: $E_0 = (S_0, U_0, D_0, M_{B0}) = (\frac{(d+\lambda_0)N}{\lambda_0+\gamma_0+d}, 0, 0, 0)$

Let $x = (U, S, D, H_B)$, the dynamical system (8) of the H-SUDRS model can be written as: $\frac{dx}{dt} = F(x) - V(x)$.

Among them,

$$F(x) = \begin{pmatrix} \frac{\beta_0 kSU}{N} \\ 0 \\ 0 \end{pmatrix}, V(x) =$$

$$\begin{pmatrix} \alpha_0 U + dU \\ \frac{\beta_0 kSU}{N} + \gamma_0 S + dS + \xi_1 S H_B - dN - \omega D - \lambda_0(N - S - U - D) \\ \omega D + \gamma_0 D + dD + \xi_2 D H_B - \alpha_0 U \\ \delta H_B - \frac{\beta_H k(M - H_B)U}{N} \end{pmatrix}$$

Then the Jacobian matrices of F(x) and V(x) are obtained respectively:

$$\phi = \begin{pmatrix} \frac{\beta_0 kS}{N} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\varphi = \begin{pmatrix} \alpha_0 + d & 0 & 0 & 0 \\ \frac{\beta_0 kS}{N} + \lambda_0 & \gamma_0 + \lambda_0 + d & -\omega + \lambda_0 & \xi_1 S \\ -\alpha_0 & 0 & \omega + \gamma_0 + d & 0 \\ \frac{\beta_H k(M - M_B)}{N} & 0 & 0 & 0 \end{pmatrix}$$

The basic reproduction number can be known from the next generation matrix method $R_0 = \rho(\phi\varphi^{-1})$. The basic reproduction number R_0 of system (8) is obtained by calculation, as follows:

$$R_0 = \frac{\beta_0 k(d + \lambda_0)}{(\alpha_0 + d)(\lambda_0 + \gamma_0 + d)} = \frac{p\beta k(d + p\lambda)}{(\alpha_0 + d)(p\lambda + \frac{\gamma}{p} + d)} \tag{10}$$

By analogy, the basic reproduction number of SUDRS model dynamical system (2) can also be obtained as follows:

$$R_0 = \frac{\beta k(d + \lambda)}{(\alpha + d)(\lambda + \gamma + d)} \tag{11}$$

3.2 Local Stability of Disease-free Equilibrium

Theorem 1. *When $R_0 < 1$, the disease-free equilibrium E0 of the dynamical system (8) is locally asymptotically stable in the feasible region Ω.*

Proof. Known $E_0 = (S_0, U_0, D_0, M_{B0}) = (\frac{(d+\lambda_0)N}{\lambda_0+\gamma_0+d}, 0, 0, 0)$, then the Jacobian matrix of system (8) at the disease-free equilibrium point E_0 is: $J(E_0) = \begin{pmatrix} -\lambda_0 - \gamma_0 - d & -\lambda_0 - \frac{\beta_0 kS}{N} & \omega - \lambda_0 & -\xi_1 S \\ 0 & \frac{\beta_0 kS}{N} - \alpha_0 - d & 0 & 0 \\ 0 & \alpha_0 & -\omega - \gamma_0 - d & 0 \\ 0 & 0 & 0 & -\delta \end{pmatrix}$

The corresponding eigenvalue of $J(E_0)$ is:

$$\begin{cases} \lambda_1 = -(\lambda_0 + \gamma_0 + d) \\ \lambda_2 = \frac{\beta_0 kS}{N} - (\alpha_0 + d) \\ \lambda_3 = -(\omega + \gamma_0 + d) \\ \lambda_4 = -\delta \end{cases} \tag{12}$$

According to the stability theory, when $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are negative, the system (8) is locally asymptotically stable at the disease-free equilibrium point.

Assuming that each parameter in the model is positive, so when $R_0 < 1$, we can get from equation (10):

$$\beta_0 k(d + \lambda_0) - (\alpha_0 + d)(\lambda_0 + \gamma_0 + d) < 0$$

From the disease-free equilibrium point E_0 and formula (12), we can get: $\lambda_2 = \frac{\beta_0 kS}{N} - (\alpha_0 + d) = \frac{\beta_0 k(\lambda_0+d)}{\lambda_0+\gamma_0+d} - (\alpha_0 + d) = \frac{\beta_0 k(d+\lambda_0) - (\alpha_0+d)(\lambda_0+\gamma_0+d)}{\lambda_0+\gamma_0+d}$

Then when $R_0 < 1$, we can get $\lambda_2 < 0$, so when $R_0 < 1$, system (8) is locally asymptotically stable at the disease-free equilibrium point E_0. □

3.3 Global Stability of Disease-free Equilibrium

Theorem 2. *When $R_0 \leq 1$, the disease-free equilibrium E_0 of the dynamical system (8) is globally asymptotically stable on the feasible region.*

Proof. The Lyapunov function is constructed as follows:

$$\begin{aligned} V_1 &= U \\ V_1' &= U' = \frac{\beta_0 kSU}{N} - \alpha_0 U - dU \\ &= (\frac{\beta_0 kS}{N} - \alpha_0 - d)U \\ &= (\frac{(\alpha_0+d)\beta_0 kS}{(\alpha_0+d)N} - (\alpha_0+d))U \\ &\leq (\frac{(\alpha_0+d)\beta_0 k(\lambda_0+d)N}{(\alpha_0+d)N(\lambda_0+\gamma_0+d)} - (\alpha_0+d))U \\ &= ((\alpha_0+d)R_0 - (\alpha_0+d))U \\ &= (\alpha_0+d)(R_0-1)U \end{aligned}$$

When $R_0 \leq 1$, it can be obtained $V'_1 < 0$. Therefore, according to the Lasalle invariance principle [27], when $R_0 \leq 1$, the disease-free equilibrium E_0 of system (8) is globally asymptotically stable on the feasible region. □

4 Numerical Simulation and Analysis

This part uses the fourth-order Runge-Kutta (RK4) method to numerically simulate the system's dynamic behavior and subsequently analyze the model system. The implementation effect of the dynamic system of the SUDRS model, meaning the joint defensive system integrating static defense and dynamic defense, can be evaluated by assessing the SUDRS model at the disease-free equilibrium point and conducting relevant numerical simulation and simulation experiment. The H-SUDRS model introduced honeypot technology was then analyzed based on the SUDRS model to study the containment effect of the number of honeypot nodes on the network malicious code, and the influence of direct feedback of honeypot and cloud feedback on the spread of malicious code in the network was simulated. In order to model the propagation law of malicious code in large-scale networks, the number of beginning nodes is set to 100000 by default in most numerical simulation and simulation experiments in this section.

4.1 SUDRS Model

The SUDRS model is put through numerical simulation studies, such as comparing it to the original SEIRS model and examining how important static defense and dynamic defense factors affect the spread of malicious code throughout the network.

4.1.1 Comparative Simulation of Dynamic Behavior of Different Models

The first and second numerical simulation experiments are to compare the SUDRS model proposed in this paper with the original SEIRS model.

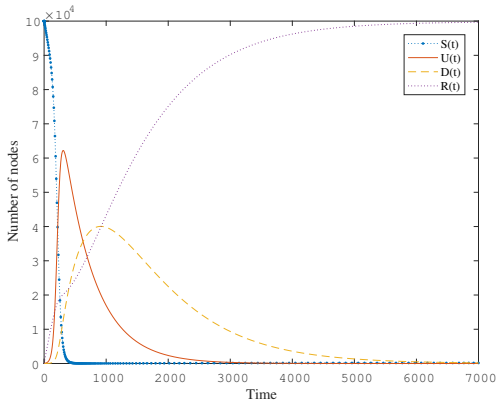


Figure 3: Dynamic behavior of SUDRS model

The first numerical simulation experiment is to study the change in the number of nodes in each state of the SUDRS model with time. As shown in Figure 3, the abscissa of Figure 3 is the unit time, and the ordinate is the number of nodes. In the first experiment, the number of nodes in each state is divided. Among them, the number of initial nodes of susceptible nodes, unaware nodes, discoverable nodes, and recovered nodes is set to 99950,50,0 and 0 respectively. In the process of state transition, the model parameters are $d = 0.0000006, \omega = 0.000003, \alpha = 0.002, k = 5, \lambda = 0.000001, \beta = 0.008, \gamma = 0.001$. According to Formula (11), the basic reproduction number of SUDRS model $R_0 = 0.0319 < 1$ can be obtained. It can be seen from Figure 3 that when the new malicious code appears, most of the nodes lack the necessary prevention and control measures, and the infected nodes infect many susceptible nodes through direct contact in the early stage of the system, so the number of infected nodes has been greatly increased in a short time, but then began to gradually decline and approach zero. Therefore, in the long run, the entire network is in an immune state.

The second numerical simulation experiment is to study the change in the number of nodes in each state of the original SEIRS model with time. To construct the SEIRS model on the premise that it is more conducive to comparing with the SUDRS model in this paper, the method of constructing the dynamic SEIRS model in reference [12] is used, but the immune method of system reloading and latent nodes killing in time in reference [12] is not considered. The differential dynamic equation of the constructed SEIRS model is as follows:

$$\begin{cases} \frac{dS}{dt} = dN - \frac{\beta kSI}{N} - \gamma S - dS + \lambda R \\ \frac{dE}{dt} = \frac{\beta kSI}{N} - \alpha E - dE \\ \frac{dI}{dt} = \alpha E - \gamma I - dI \\ \frac{dR}{dt} = \gamma S + \gamma I - \lambda R - dR \end{cases} \tag{13}$$

In the second numerical simulation experiment, the

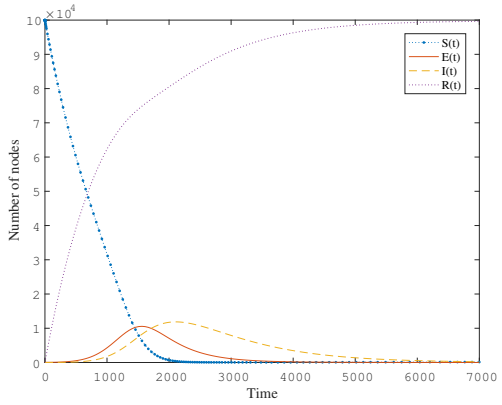


Figure 4: Dynamic behavior of each state of SEIRS model

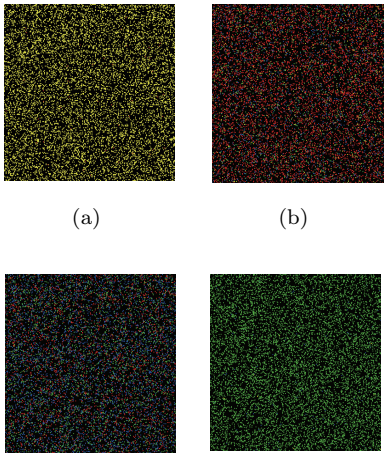


Figure 5: The development process of malicious code in SUDRS model

initial number of nodes and the values of each parameter of system (13) are the same as those in experiment 1. As shown in Figure 4, the abscissa of Figure 4 is unit time, and the ordinate is the number of nodes.

From the results of simulation experiments 1 and 2, there are three problems that do not conform to the actual situation when simulating the trend of malicious code spreading in the network in the simulation system (13).

First, the peak value of the latent node in Figure 4 is less than the peak value of the infected node, and in the normal case of the SEIRS model, the infection will experience the latent state, and the peak value of the latent node should not be less than the infected node. Considering that not all latent nodes will be converted into infected nodes, therefore, the peak value of the latent node in the system (13) is greater than the peak value of the infected node, which is not in line with the actual situation. For example, in Figure 3, the number of peaks of unaware nodes is far greater than the number of peaks found.

Second, when the new malicious code appears, the number of infected nodes infected by malicious code should reach a peak in the early stage of the system because of the lack of necessary prevention and control measures for the new threat, and then gradually decrease and eventually disappear because of the emergence of prevention and control measures. Obviously, Figure 3 is more in line with this actual situation. The peak of infected nodes in Figure 4 appears in the middle of the system, which is not in line with the actual situation.

Thirdly, in the simulation experiment of the system (13), the number of immune nodes has been growing rapidly. When the time reaches 400, the growth rate of immune nodes in the SEIRS model is about 2.93 times that of the SUDRS model. This is obviously not the development trend of immune nodes in line with the actual situation. The actual situation should be that the immune nodes rise rapidly in the early stage of the system, which is because the fact that nodes that are equipped to fight new types of malicious code take timely preventive measures. Then, the growth rate of immune nodes slows down, as shown in Figure 3. The growth rate of the number of immune nodes at time 400 is about 56.27% lower than that at time 200, which is since the nodes that can resist the new malicious code are immune, and most of the remaining nodes cannot enter the immune state due to the lack of prevention measures.

Finally, due to the popularization of prevention and control measures, most nodes also enter the immune state, and the growth rate of immune nodes increases gradually. As shown in Figure 3, the growth rate of immune nodes at the time 1000 increases by about 74.75% compared with that at the time 400. Therefore, SUDRS model can more accurately reflect the development trend of malicious code in real network, and put forward more accurate prevention and control measures against new malicious code.

In order to reflect the spread of malicious code more intuitively in the network, this paper uses NetLogo tools to simulate the dynamic propagation process of malicious code under the SUDRS model, such as Figure 5(a), Figure 5(b), Figure 5(c) and Figure 5(d). In the 100 * 100 interface, the number of initial nodes is 10000, and 9995 susceptible nodes and 5 unaware nodes are set initially. The number of initial nodes of discovery nodes and immune nodes is 0. The yellow, red, blue, and green nodes represent susceptible nodes, unaware nodes, discovery nodes, and immune nodes respectively. Figure 5 shows the four stages of malicious code in the process of network transmission, which is sorted in chronological order. Among them, Figure 5(a) is the initial state, that is, when the unit time is 0, Figure 5(b) is the peak time of unaware but infectious unaware nodes, that is, when the unit time is about 200, Figure 5(c) is when the number of immune nodes in the network exceeds the number of infectious unaware nodes, and Figure 5(d) is when the number of immune nodes in the network is close to the peak.

4.1.2 The Impact of Dynamic Defense on the Spread of Malicious Code in the Network

In the SUDRS model, the key parameter of dynamic defense is ω, which means the successful isolation rate of dynamic defense, that is, the probability that malicious code behavior is discovered and isolated under the protection of dynamic defense. The following experiment 3 studies the influence of different dynamic defense success isolation rates on system (2) through numerical simulation experiments.

In the third numerical simulation experiment, the number of initial nodes and other parameters except ω in each state is consistent with the first experiment. The values of the successful isolation rate of dynamic defense are 0.000003, 0.00001, 0.0001, and 0.0003, respectively. The simulation results are shown in Figure 6. The abscissa of Figure 6 is unit time, and the ordinate is the number of infected nodes. In Figure 6, by changing the successful isolation rate of dynamic defense, the peak value that the infected nodes can reach is not reduced, but when the number of infected nodes decreases, the reduction rate slows down with the increase of the isolation rate. It can be seen that increasing the successful isolation rate of dynamic defense without considering the actual situation will not only not accelerate the reduction of infected nodes, but also slow down the reduction of infected nodes. The important factor causing this situation is that the defense method of dynamic defense will not make the nodes immune, but make the nodes in the discovery state convert into susceptible nodes, and the susceptible nodes will still face the risk of being infected. Therefore, in the early stage of the system, when there are no effective means of prevention and control, the implementation effect of dynamic defense can be increased to reduce the infection rate of susceptible nodes. In the later stage of the system, when the prevention and control measures have been effectively popularized, the implementation effect of dynamic defense can be reduced, and the static defense that has been updated the virus library can be mainly used to increase the immunization rate of nodes.

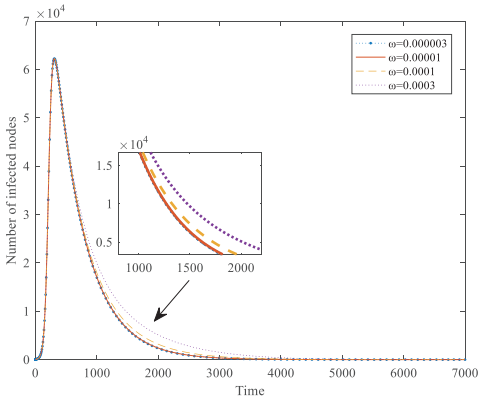


Figure 6: Changes of infected nodes with time under different isolation rates

4.1.3 The Impact of Static Defense on the Spread of Malicious Code in the Network

In the SUDRS model, the key parameter of static defense is γ, which means the node immunization rate under static defense, that is, under the protection of static defense, the node updates the virus library in time to obtain the probability of immunization. The next experiment 4 studies the influence of different immunization rates on the system (2) through numerical simulation experiments.

In the fourth numerical simulation experiment, the number of initial nodes and other parameters except the immune rate γ of each state is consistent with the first experiment, and the values of the node immune rate are 0.001, 0.0015, 0.0017, and 0.002, respectively. The simulation results are shown in Figure 7. The abscissa of Figure 7 is unit time, and the ordinate is the number of infected nodes. In Figure 7, it can be seen from the trend of the curve that as the immunization rate increases, the number of infected nodes at the same time shows a downward trend, and the peak value of the number of infected nodes in the initial stage of the system is getting smaller and smaller. Therefore, by updating the virus library in a timely, comprehensive, and effective manner, the static defense capability of the system can be improved, and the outbreak of malicious code in the early stage of the system can be reduced. At the same time, it can also reduce the number of nodes infected in the system at various time periods.

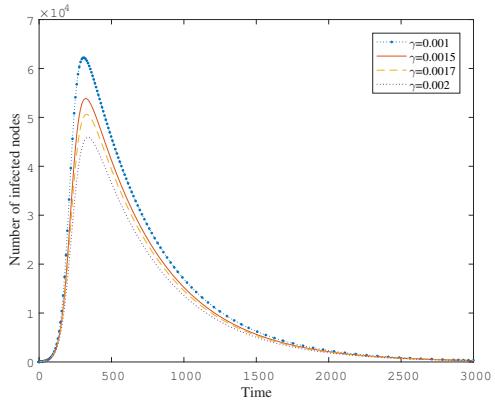


Figure 7: Changes of infection nodes with time under different immunization rates

4.2 H-SUDRS Model

To study the influence of the feedback function of the honeypot on the spread of malicious code in the network, this section focuses on the analysis of the H-SUDRS model that introduces the honeypot technology based on the SUDRS model. Firstly, the influence of different numbers of unaware but infectious nodes on the system (7) in the ini-

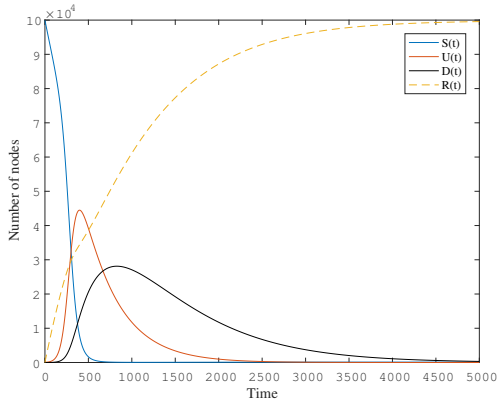


Figure 8: Dynamic behavior of H-SUDRS model

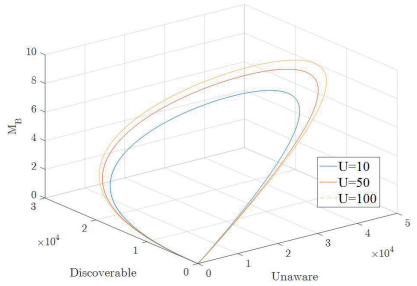


Figure 9: Phase diagram (U, D, M_B) of H-SUDRS model

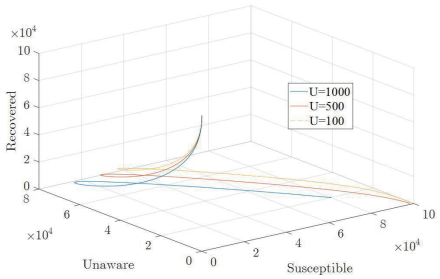


Figure 10: Phase diagram (S, U, R) of H-SUDRS model

tial state is studied. Secondly, the effect of the number of honeypot nodes on the containment of malicious code in the network is studied. Finally, the direct feedback and cloud feedback of the honeypot are simulated respectively.

4.2.1 Dynamic Behavior of H-SUDRS Model

The fifth numerical simulation experiment is to simulate the dynamic system (7) of the H-SUDRS model over time, that is, to study its dynamic behavior. The results are shown in Figure 8.

In this simulation, the number of initial nodes and some parameters of each state are consistent with experiment one, and some parameters are increased and changed compared with experiment one. For example, the adjustable constant p of deploying honeypot control is 0.8. From the formulas (3), (4), (5), and (6), some parameters changed in experiment five are: $\lambda_0 = 0.0000008, \beta_0 = 0.0064, \gamma_0 = 0.0008, \alpha_0 = 0.0016$. The values of the new parameters and variables are: $\xi_1 = 0.00005, \xi_2 = 0.0001, \delta = 0.25, \beta_H = 0.01, M = 100$. Currently, according to Formula (10), the basic reproduction number of the H-SUDRS model is $R_0 = 0.0349$ ¡ 1. According to Theorems 1 and 2, malicious code will eventually disappear from the network.

By comparing Figure 3 and Figure 8, it can be found that after the introduction of honeypot technology, the peak value of the number of unaware nodes with infectious ability in the system is reduced. The peak value of unaware but infectious nodes in the H-SUDRS model is reduced by 28.49 % compared with the SUDRS model and the time to reach the peak is delayed. The number of immune nodes at the same time is also increasing. Therefore, honeypot technology plays an effective role in curbing the spread of malicious code in the network.

4.2.2 Different Number of Initial Nodesl

The sixth and seventh simulation experiments are to study the influence of different initial node numbers on the dynamic behavior of the system (7). The results are shown in Figure 9 and Figure 10.

The values of the parameters in Experiment 6 are consistent with those in Experiment 5. In the initial node values, the values of the susceptible nodes are 99990,99950, and 99900, respectively, and the corresponding values of the unaware nodes are 10,50, and 100, respectively. The values of each parameter in Experiment 7 are consistent with those in Experiment 5. In terms of the initial node value, to see the difference of the initial node more intuitively from the graph, this experiment increases the magnitude of the change in the number of initial nodes, so that the initial position is significantly different. The values of the susceptible nodes are 99000, 99500 and 99900 respectively, and the corresponding values of the unaware nodes are 1000,500 and 100 respectively. It can be seen from Figure 9 and Figure 10 that no matter where the initial state starts, the network will eventually tend to the same stable state, that is, although the number of nodes in each state is different at the beginning, the number of nodes in each state is consistent when the network is stable.

4.2.3 Different Initial Number of Honeypots

The eighth numerical simulation experiment is to study the influence of the change in the initial number of honeypots on the feedback function of honeypots, and

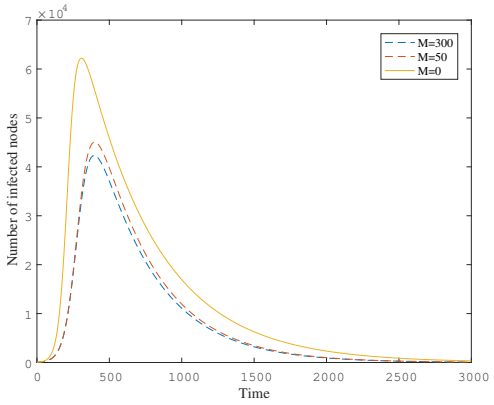


Figure 11: Changes of infected nodes with time under different initial number of honeypots

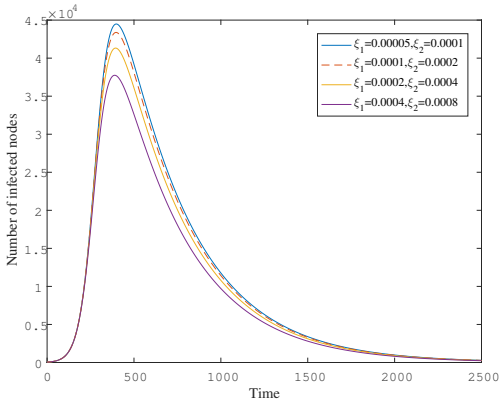


Figure 12: Changes of infected nodes with time under different direct feedback rates

the results are shown in Figure 11.

In the eighth simulation experiment, the initial number of nodes and the values of each parameter except the honeypot are consistent with those in the fifth experiment. The number of initial honeypot nodes is 300,50 and 0 respectively. It can be seen from Figure 11 that when the honeypot is not deployed in the system, the H-SUDRS model becomes the SUDRS model. The number of infected nodes at the same time when the honeypot is not deployed (M = 0) exceeds the number of infected nodes when the honeypot is deployed, and the peak value of infected nodes when M = 0 is far greater than the peak value of infected nodes when the honeypot is deployed. From Figure 11, it can also be found that when only increasing the number of initial honeypot deployments without considering whether they are in the right position, the peak value that the infected node can reach can still be reduced. Therefore, increasing the number of deployed honeypots can improve network security.

4.2.4 Direct Feedback of Honeypot

The ninth numerical simulation experiment is to study the impact of the direct feedback function of the honeypot on the spread of malicious code in the network. The results are shown in Figure 12.

The direct feedback of the honeypot is that the honeypot analyzes its malicious behavior and records it by trapping malicious code attacks. After analyzing the defense means against malicious code, the prevention and control information against new malicious code is fed back to the surrounding nodes in direct contact. Generally, the security protection ability of surrounding nodes can be enhanced by means of technology and management. The nodes of the direct contact feedback of the honeypot are mainly divided into two types, one is the susceptible node, which has not really understood the harm of the new malicious code; the other is to discoverable nodes, such nodes have been personally aware of the harm caused by new malicious code. Considering the complexity of the node, a node means more than just a device, the node contains the device and the user of the device, because people will instinctively make defense behavior when they feel the threat, so when the honeypot feeds the prevention and control information to the discovery node, compared to the susceptible node, the discovery node is more likely to accept the feedback from the honeypot node and improve its own defense capability. Therefore, in the system (7), ξ_1 is always less than ξ_2. The number of initial nodes and the values of parameters except ξ in Experiment 9 are consistent with those in Experiment 5. The values of ξ_1 are 0.00005, 0.0001, 0.0002 and 0.0004, respectively, and the corresponding values of ξ_2 are 0.0001, 0.0002, 0.0004 and 0.0008, respectively. It can be seen from Figure 12 that with the increase of the direct feedback rate ξ_1 and ξ_2 of the honeypot, the number of infected nodes at the same time is decreasing. Therefore, by increasing the direct feedback rate of the honeypot, the spread of malicious code in the network can be curbed.

4.2.5 Cloud Feedback on Honeypots

The tenth numerical simulation experiment is to study the impact of the cloud feedback function of the honeypot on the spread of malicious code in the network. The results are shown in Figure 13.

The cloud feedback of the honeypot, first, the honeypot obtains the attacker's attack behavior by trapping, analyzes the malicious behavior record, and then feeds back the relevant prevention and control information against the new malicious code to the cloud server to help the cloud server discover the new malicious code earlier and faster and propose more comprehensive prevention and control measures for the new malicious code. Finally, the cloud server helps the node update the virus library of the cloud security software for the first time. The key to the cloud feedback of the honeypot is to adjust the deployment strategy of the honeypot. When the honeypot is deployed in a suitable position, it can better

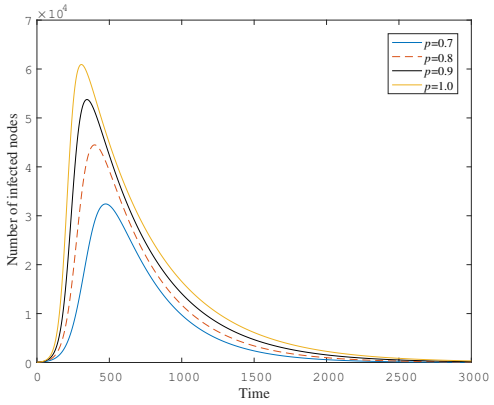


Figure 13: Changes of infected nodes over time under different honeypot deployment efficiency

trap malicious code, and then analyze its malicious behavior and propose corresponding prevention measures. Constantly adjusting the appropriate honeypot deployment strategy, can help the immune node to reduce the immune failure rate of the node, increase the discovery rate of the unaware but infectious node, reduce the probability of contact between the susceptible node and the infected node, and the probability of infection and increase the probability of the node obtaining the immune ability, that is, the all-round enhancement of the new malicious code containment effect. The more comprehensive and effective the honeypot deployment, the lower the p-value. When $p = 1$, it means that no honeypot is deployed or the deployment location of the honeypot is invalid. According to Figure 13, with the optimization of the honeypot deployment strategy, honeypot deployment is more and more comprehensive and effective, and the cloud feedback effect of the honeypot will be better and better, which can greatly reduce the number of infected nodes at the same time and also greatly reduce the peak value that infected nodes can reach. Therefore, combined with the results of Experiment 8, when conducting network defense, we can increase the number of deployed honeypots or optimize the deployment strategy of honeypots, so as to improve the security of the network, enhance the active defense capability of the network, and reduce the loss caused by new malicious code in the early stage of the system.

5 Conclusions

This paper explores several scenarios in which malicious code will occur during the propagation process of the actual network environment to develop a malicious code propagation model that is better in line with the real network environment. These conditions include nodes in the network being dynamic, nodes having the ability to infect when latent, and nodes having a variety of protection strategies when fighting malicious behavior. A malicious code propagation model that is more by the real network environment is built on top of the original SEIRS model, and a SUDRS model is produced. The SUDRS model is used to study a dual defense system that combines static and dynamic defense. The results of numerical simulation reveal that when confronted with new malicious code, the dual defense can be carried out in stages. The defense is carried out in a combination of dynamic defense and static protection in the early stage of the system, with the dynamic detection technology based on behavior and the un-updated virus library of cloud security software. In the middle and late stages of the system, with the updated virus library of cloud security software, the defense is primarily based on static defense, which can fully utilize the advantages of static defense and dynamic defense, improve the immune rate of nodes to reduce the damage caused by malicious code. To increase the network's active defensive ability against new malicious code, honeypot technology is implemented based on the SUDRS model, and the H-SUDRS model is formed. The H-SUDRS model is used to investigate the impact of the honeypot's feedback mechanism on the spread of malicious code in the network. Dynamic analysis is used to determine the basic reproduction number R_0 of the newly suggested H-SUDRS model, and its stability at the disease-free equilibrium point is demonstrated. The numerical simulation and simulation experiments show that the spread of malicious code in the network can be contained not only by enhancing the direct feedback of honeypots, but also by increasing the number of honeypots deployed or optimizing the deployment strategy of honeypots to enhance the active defense capability of the network, and then further contain the spread of malicious code in the network through the active defense. In the future, the feedback function of the honeypot will be combined with the individual feedback function of other nodes to reduce the error information in the feedback process and build a more comprehensive information feedback system.

Acknowledgments

This study is supported by the Natural Science Foundation of Jiangsu Province under Grant No. BK20201462, and the Natural Science Foundation of Xuzhou City under Grant No. KC21018, and Postgraduate Research and Practice Innovation Program of Jiangsu Normal University under Grant No. 2022XKT1532.

The author would like to thank everyone who helped with the topic selection, conception, writing, or revision of the paper. The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] W. Wu, and C. Y. Yang, "An Overview on Network Security Situation Awareness in Internet", *Interna-*

tional Journal of Network Security, vol. 24, no. 3, pp. 450-456, 2022.

[2] S. J. Achar, C. Baishya, and M. K. A. Kaabar, "Dynamics of the worm transmission in wireless sensor network in the framework of fractional derivatives", *Mathematical Methods in the Applied Sciences*, vol. 45, no. 8, pp. 4278-4294, 2022.

[3] M. Kumar, "Scalable malware detection system using distributed deep learning", *Cybernetics and Systems*, pp. 1-29, 2022.

[4] A. Darem, J. Abawajy, and A. Makkar, et al, "Visualization and deep-learning-based malware variant detection using OpCode-level features", *Future Generation Computer Systems*, vol. 125, pp. 314-323, 2021.

[5] Z. Yu, H. Gao, and D. Wang, et al, "Sei2rs malware propagation model considering two infection rates in cyber–physical systems", *Physica A: Statistical Mechanics and its Applications*, vol. 597, pp. 127207, 2022.

[6] D. T. Le, K. Q. Dang, and Q. L. T. Nguyen, et al, "A behavior-based malware spreading model for vehicle-to-vehicle communications in VANET networks", *Electronics*, vol. 10, no. 19, pp. 2403, 2021.

[7] J. R. C. Piqueira, M. A. M. Cabrera, and C. M. Batistela, "Malware propagation in clustered computer networks", *Physica A: Statistical Mechanics and its Applications*, vol. 573, pp. 125958, 2021.

[8] C. X. Dai, G. Liu, and C. C. Han, et al, "Research on Technologies of Windows Malware Detecting Based on Abnormal Behavior in KVM Environment", *Journal of Information Security Research*, vol. 6, no. 06, pp. 514-522, 2020.

[9] J. W. Zhuge, Y. Tang, and X. H. Han, et al, "Honeypot Technology Research and Application", *Journal of Software*, vol. 24, no. 04, pp. 825-842, 2013.

[10] W. Kang, and S. Wang, "Virus Propagation Behavior Simulation Based on Node Movement Model of Wireless Multi-hop Network", *International Journal of Network Security*, vol. 21, no. 3, pp. 471-476, 2019.

[11] F. Yang, and Z. Zhang, "Hopf bifurcation analysis of SEIR-KS computer virus spreading model with two-delay", *Results in Physics*, vol. 24, pp. 104090, 2021.

[12] J. L. Wang, Z. R. Luo, and C. F. Guo, et al, "Dynamically Evolving SEIRS Network Virus Propagation Model and Control", *Computer Simulation*, vol. 39, no. 08, pp. 383-388, 2022.

[13] W. Zhang, R. C. Wang, and P. Li, "Worm propagation modeling in cloud security", *Journal on Communications*, vol. 31, no. 04, pp. 17-24, 2012.

[14] M. Peng, C. D. Li, and X. He, "SEIR Computer Virus Propagation Model Based on Direct Immunization", *Journal of Chongqing Normal University (Natural Science)*, 2013, 30(01): 77-80.

[15] S. Kumari, and R. K. Upadhyay, "Exploring the dynamics of a malware propagation model and its control strategy", *Wireless Personal Communications*, vol. 121, no. 3, pp. 1945-1978, 2021.

[16] A. M. del Rey, R. C. Vara, and S. R. González, "A computational propagation model for malware based on the SIR classic model", *Neurocomputing*, vol. 484, pp. 161-171, 2022.

[17] G. Liu, J. Chen, and Z. Liang, et al, "Dynamical analysis and optimal control for a SEIR model based on virus mutation in WSNs", *Mathematics*, vol. 9, no. 9, pp. 929, 2021.

[18] Y. W. Chang, B. Yang, and Y. L. Gao, et al, "Modeling and Analysis of WeChat Official Account Information Dissemination Based on SEIR", *Computer Science*, vol. 49, no. 04, pp. 56-66, 2022.

[19] X. Zhao, D. M. Chen, and X. X. Yan, et al, "Review of Sandbox Technology", *Journal of Zhongyuan University of Technology*, vol. 25, no. 04, pp. 1-5, 2014.

[20] B. Xie, and M. Liu, "Dynamics stability and optimal control of virus propagation based on the e-mail network", *IEEE Access*,vol. 9, pp. 32449-32456, 2021.

[21] Q. Zhu, X. Luo, and Y. Liu, "Modeling and Analysis of the Spread of Malware with the Influence of User Awareness", *Complexity*, vol. 2021, pp. 1-9, 2021.

[22] Z. H. Guan, Y. J. Qi, and X. W. Jiang, et al, "Virus propagation dynamic model and stability on complex networks", *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, vol. 39, no. 01, pp. 114-117, 2011.

[23] S. Sharma, and A. Kaul, "A survey on Intrusion Detection Systems and Honeypot based proactive security mechanisms in VANETs and VANET Cloud", *Vehicular communications*, vol. 12, pp. 138-164, 2018.

[24] Y. Sun, Z. Tian, and M. Li, et al, "Honeypot identification in softwarized industrial cyber–physical systems", *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5542-5551, 2020.

[25] J. G. Ren, C. Zhang, and Q. Hao, "A theoretical method to evaluate honeynet potency", *Future Generation Computer Systems*, vol. 116, pp. 76-85, 2021.

[26] L. Y. Shi, Y. Li, and M. F. Ma, "Latest Research Progress of Honeypot Technology", *Journal of Electronics & Information Technology*, vol. 41, no. 02, pp. 498-508, 2019.

[27] J. P. Lasalle, The stability of dynamical systems. *Society for Industrial and Applied Mathematics*, 1976.

Biography

Chenxi Li, born in 1999, postgraduate. His main research interests include computer virus propagation model and network security.

Jianguo Ren, born in 1978, Ph.D, associate professor . His main research interests include cyberspace security and network attack and defense confrontation.

Fengjiao Li, born in 1999, postgraduate. Her main research interests include network security.

Study on Nega-Hadamard Transform and Nega-crosscorrelation of Vectorial Boolean Functions

Jingjing Zhang[1], Zepeng Zhuo[1], and Guolong Chen[2]
(Corresponding author: Zepeng Zhuo)
School of Mathematical Sciences, Huaibei Normal University[1]
Huaibei, Anhui 235000, China
Email: zzp781021@sohu.com
School of Computer Engineering, Bengbu College[2]
Bengbu, Anhui 233030, China
(Received July 4, 2023; Revised and Accepted Jan. 5, 2024; First Online June 22, 2024)

Abstract

Cryptography security is closely related to network security, so to ensure network security, it is particularly important to study the properties of cryptographic functions. In this paper, we define a vectorial negabent function. Several nega-crosscorrelation properties of vectorial Boolean functions are investigated, and some properties of nega-crosscorrelation and nega-Hadamard transform are presented. The relationship among nega-crosscorrelation of arbitrary four vectorial Boolean functions is presented. Finally, vectorial Boolean function decomposition in codimension subspace and the conditions of a class of functions are vectorial negabent are studied.

Keywords: Nega-Crosscorrelation Function; Nega-Hadamard Transform; Vectorial Boolean Function; Vectorial Negabent Function;

1 Introduction

Cryptography is a supporting technology of network security. The whole network security is based on the basis of cryptography. It can be said that there is no network security without cryptography. As an important cryptographic function in the cryptographic system, the Boolean function has a great research value. Boolean functions play a central role in symmetric key cryptosystems. So the study of Boolean function is of great significance in the field of cryptography.

There are many kinds of tools to study Boolean functions, one of which is the Walsh-Hadamard transform. The Walsh-Hadamard transform is not only used in cryptography but also in coding theory [1–3].

By means of the spectral value of the Walsh-Hadamard transform, a special class of functions, the Bent function, can be defined. The spectral values of these functions are flat and have the greatest distance from the set of affine functions, which means that Bent functions have the highest nonlinearity.

They were introduced by Rothaus [4] and already studied first by Dillon [5] and next by many researchers for more than three decades ago.

The Walsh-Hadamard transform is just a special case of a unitary transformation, and unitary transformations have other forms, the nega-Hadamard transform being another. Riera and Parker [6] generalizes the concept of bent by requiring that Boolean functions have a flat spectrum with respect to one or more transformations.

The transforms they applied are n-fold tensor products of the identity $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, the Walsh-Hadamard matrix $H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, and the nega-Hadamard matrix $N = \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}$, where $i^2 = -1$. The nega-Hadamard transform of Boolean functions was first proposed by Parker [7]. As in the case of the Walsh-Hadamard transform, a Boolean function is called negabent if the spectrum under the nega-Hadamard transform is flat. For an even number of variables, a function is bent-negabent if it is both bent and negabent. Moreover, there are many good results for the research on this area [8–10].

In addition to the two transformations, the tools for studying Boolean functions include the correlation functions. The correlation functions are divided into autocorrelation and cross-correlation. While there are various connections between the Walsh-Hadamard transform and correlation function of the Boolean functions. *Stănică* and Gangopadhyay [11] gives another characterization of negabent function by analyzing the properties of nega-

Hadaqmard transform of Boolean function.

Applying the cross-correlation functions as a fundamental tool, Sarkar and Maitra [12] studied cryptographic properties of Boolean functions, and got the Cross-correlation Theorem. Zhuo [13] given the relationship among cross-correlation of arbitrary four Boolean functions. *Stănică* and Gangopadhyay [11] also considered the decomposition of negabent functions with respect to codimension one subspaces.

If a class of Boolean functions has relatively ideal cryptographic properties, then it is also a very interesting research topic to consider the vectors of this class of functions, namely vectorial Boolean functions(e.g.vectorial bent functions comes from Boolean bent functions [14]).

So similar to the Boolean function, the relationship between the nega-Hadamard transform of the vectorial Boolean function and the nega-correlation function is also worth research. In this article, we introduce the notion of vectorial negabent functions.

The paper is organized as follows. In Section 2, the basic concepts and notations are presented. In Section 3, we first given the characterization of vectorial negabent function; secondly, we proved the relationship between nega-crosscorrelation and the nega-Hadamard transform, and proved the relationship among nega-crosscorrelation of arbitrary four vectorial Boolean functions and its consequences.

Finally, we investigate the decomposition of vectorial Boolean function in space and the condition that a class of vectorial Boolean function is vectorial negabent function. Section 4 gives the conclusions of this paper.

2 Preliminaries

In this section, we present the basic definitions and some of the preparation work required for this paper. Let the Galois field $\mathbb{GF}(2^n)$ be denoted by $\mathbb{F}_{2^n}$ and its corresponding vector space $\mathbb{GF}(2^n)$ by $\mathbb{F}_2^n$. Let $\mathbb{F}_2$ denote the finite field with two elements.

The set of integers, real numbers, and complex numbers are denoted by $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$, respectively. To avoid confusion, we denoted addition in $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ by "+", and the addition in $\mathbb{F}_2$ by "$\oplus$". If $z = a + bi \in \mathbb{C}$, then $|z| = \sqrt{a^2 + b^2}$ denotes the absolute value of z, and $\bar{z} = a - bi$ denotes the complex conjugate of z, where $i^2 = -1$, $a, b \in \mathbb{R}$.

If $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two elements of $\mathbb{F}_2^n$, we define the scalar (or inner) product $x \cdot y$ and, the intersection $x * y$ by

$$x \cdot y = x_1y_1 \oplus x_2y_2 \oplus \cdots \oplus x_ny_n$$
$$x * y = (x_1y_1, x_2y_2, ..., x_ny_n).$$

The Hamming weight $wt(x)$ of an element $x = (x_1, x_2, ..., x_n) \in \mathbb{F}_2^n$ is the number of components equal to 1. There is a well-known conclusion about Hamming weight

$$wt(x \oplus y) = wt(x) + wt(y) - 2wt(x * y).$$

The set of all Boolean function in n-variable, $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is denoted by $\mathcal{B}_n$. We say that a Boolean function is balanced if its truth table contains an equal number of $1'$s and $0'$s, that is, if its Hamming weight equals $wt(f) = 2^{n-1}$. Any Boolean function, $f \in \mathcal{B}_n$, is generally represented by its algebraic normal form(ANF)

$$f(x_1, x_2, ..., x_n) = \bigoplus_{u \in \mathbb{F}_2^n} \lambda_u (\prod_{i=1}^{n} x_i^{u_i}),$$

where $\lambda_u \in \mathbb{F}_2$ and $u = (u_1, u_2, ..., u_n) \in \mathbb{F}_2^n$. The algebraic degree of f, denoted by $deg(f)$, is the maximal value of $wt(u)$ such that $\lambda_u \neq 0$. A Boolean function is affine if there exists no term of degree strictly greater than 1 in the ANF and the set of all affine functions is denoted by A_n. An affine function with constant term equal to zero is called a liner function.

The cryptographic properties of a Boolean function are most easily reflected through its Walsh-Hadamard transform. The Walsh-Hadamard transform of $f \in \mathcal{B}_n$ at any point $u \in \mathbb{F}_2^n$ is defined by

$$\mathcal{W}_f(u) = 2^{-\frac{n}{2}} \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) \oplus u \cdot x}.$$

Similarly, the Fourier transform of $f \in \mathcal{B}_n$ at any point $u \in \mathbb{F}_2^n$ is defined by $\widehat{f}(u) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) \oplus u \cdot x}$. A function $f \in \mathbb{F}_2^n$ is a bent function if $|\mathcal{W}_f(u)| = 1$ for all $u \in \mathbb{F}_2^n$. Another useful tool is the correlation function. The cross-correlation between $f, g \in \mathcal{B}_n$ at any point $u \in \mathbb{F}_2^n$ is defined by

$$C_{f,g}(u) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) \oplus g(x \oplus u)}.$$

The auto-correlation of f is denoted by $C_f(u)$ above $f = g$. It is known [3] that a function $f \in \mathcal{B}_n$ is bent if and only if $C_f(u) = 0$ for all $u \neq 0$. In addition to Walsh-Hadamard transform, the Nega-Hadamard transform is also an important tool to study the properties of Boolean functions. The Nega-Hadamard transform of $f \in \mathcal{B}_n$ at any point $u \in \mathbb{F}_2^n$ is defined by

$$\mathcal{N}_f(u) = 2^{-\frac{n}{2}} \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) \oplus u \cdot x} i^{wt(x)}.$$

A function $f \in \mathcal{B}_n$ is a negabent function if $|\mathcal{N}_f(u)| = 1$ for all $u \in \mathbb{F}_2^n$. If f is both bent and negabent, we say that f is bent-negabent. The sum

$$\mathcal{C}_{f,g}(u) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) \oplus g(x \oplus u)} (-1)^{x \cdot u}$$

is the nega-crosscorrelation of f and g at z. Takeing $f = g$ above, we get the nega-autocorrelation $\mathcal{C}_f(u)$ of $f \in \mathcal{B}_n$ at $u \in \mathbb{F}_2^n$.

For $m \geq 2$, a mapping $F : \mathbb{F}_2^n \to \mathbb{F}_2^m$ is called vectorial (Boolean) function. The set of all vectorial (Boolean)

function is denoted by $\mathcal{VB}_n^m$. The Walsh-Hadamard transform of $F \in \mathcal{VB}_n^m$ at any point $(u,v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$ is defined by $\mathcal{W}_F(u,v) = 2^{-\frac{n}{2}} \sum_{x\in\mathbb{F}_2^n} (-1)^{v\cdot F(x)\oplus u\cdot x}$. The definition of vectorial bent function is given in [15]. One can show that a vectorial function F is bent if for all $u \in \mathbb{F}_2^n$ and all $v \in \mathbb{F}_2^m$ with $v \neq 0$ the walsh transform satisfies $|\mathcal{W}_F(u,v)| = 1$. The sum

$$C_{F,G}(u,v) = \sum_{x\in\mathbb{F}_2^n} (-1)^{v\cdot(F(x)\oplus G(x\oplus u))}$$

is the cross-correlation of F and G at (u,v). Taking $F = G$ above, we get the auto-correlation $C_F(u,v)$ of $F \in \mathcal{VB}_n^m$ at $(u,v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$.

3 Main Conclusions

3.1 The Characterization Of Vectorial Negabent Functions

Let F is vectorial Boolean function, then $v \cdot F$ is Boolean function. The nega-Hadamard transform of $v \cdot F$ is $\mathcal{N}_{v\cdot F}(u) = 2^{-\frac{n}{2}} \sum_{x\in\mathbb{F}_2^n} (-1)^{v\cdot F(x)\oplus u\cdot x} i^{wt(x)}$, if $|\mathcal{N}_{v\cdot F}(u)| = 1$, then $v\cdot F$ is negabent. Next, we present the definition of vectorial negabent functions is the following.

Definition 1. *The nega-Hadamard transform of $F \in \mathcal{VB}_n^m$ is*

$$\mathcal{N}_F(u,v) = 2^{-\frac{n}{2}} \sum_{x\in\mathbb{F}_2^n} (-1)^{v\cdot F(x)\oplus u\cdot x} i^{wt(x)}.$$

If F is a vectorial negabent function, then $|\mathcal{N}_F(u,v)| = 1$ for all $(u,v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$.

We know that the inverse of the Walsh-Hadamard transform $\mathcal{W}_f(u)$ of $f \in \mathcal{B}_n$ is

$$(-1)^{f(x)} = 2^{-\frac{n}{2}} \sum_{u\in\mathbb{F}_2^n} \mathcal{W}_f(u)(-1)^{u\cdot x},$$

for all $x \in \mathbb{F}_2^n$. The inverse of the nega-Hadamard transform of the Boolean function is given in [16]. Next, we give the inverse of the nega-Hadamard transform of $v \cdot F$ in Lemma 1.

Lemma 1. *Suppose $F \in \mathcal{VB}_n^m$. Then*

$$(-1)^{v\cdot F(x)} = 2^{-\frac{n}{2}} i^{-wt(x)} \sum_{u\in\mathbb{F}_2^n} \mathcal{N}_F(u,v)(-1)^{u\cdot x},$$

for all $x \in \mathbb{F}_2^n$.

On this basis, the properties of nega-Hadamard transform of vectorial Boolean functions are given in Theorem 1.

Theorem 1. *Let F, G, H be vectorial Boolean functions. The following statements are ture.*

1) *For any affine function $l_{a,c}(x) = a \cdot x \oplus c$, c is constant, and $v \cdot F \in \mathcal{B}_n$, for all $v \in \mathbb{F}_2^m$, $\mathcal{N}_{v\cdot F\oplus l_{a,c}}(u) = (-1)^c \mathcal{N}_F(a \oplus u, v)$.*

2) *If $H(x) = F(x) \oplus G(x)$ on $\mathbb{F}_2^n$, then for $u \in \mathbb{F}_2^n$, $v \in \mathbb{F}_2^m$,*

$$\begin{aligned}\mathcal{N}_H(u,v) &= 2^{-\frac{n}{2}} \sum_{w\in\mathbb{F}_2^n} \mathcal{N}_F(w,v)\mathcal{W}_G(u\oplus w, v)\\ &= 2^{-\frac{n}{2}} \sum_{w\in\mathbb{F}_2^n} \mathcal{W}_F(w,v)\mathcal{N}_G(u\oplus w, v).\end{aligned}$$

3) *If $H(x,y) = F(x) \oplus G(y)$, $x, y \in \mathbb{F}_2^n$, then $\mathcal{N}_H(u,v,w) = \mathcal{N}_F(u,w)\mathcal{N}_G(v,w)$.*

4) *If $F(x) \in \mathcal{VB}_n^m$, $G(y) \in \mathcal{VB}_k^m$, and $H(x,y) = F(x) * G(y)$, then*

$$\begin{aligned}2^{\frac{k}{2}}\mathcal{N}_H(u,v,w) &= \mathcal{N}_F(u,w)A_{G1}(v) + \omega^n i^{-wt(u)} A_{G0}(v)\\ &+ 2^{-\frac{n}{2}} \sum_{(x,y)\in\mathbb{F}_2^{n+k}, G(y)\neq 0,1} (-1)^{w\cdot(F(x)*G(y))\oplus u\cdot x\oplus v\cdot y} i^{wt(x,y)},\end{aligned}$$

where

$$A_{G0}(v) = \sum_{y\in\mathbb{F}_2^k, G(y)=0} (-1)^{y\cdot v} i^{wt(y)},$$

$$A_{G1}(v) = \sum_{y\in\mathbb{F}_2^k, G(y)=1} (-1)^{y\cdot v} i^{wt(y)}.$$

and $\omega = \frac{1+i}{\sqrt{2}}$ is an eighth primitive root of 1.

Proof. The (1) is direct from the definition of nega-Hadamard transform. We show the first identity of (2) (the second follows by symmetry).

$$\begin{aligned}&\sum_{w\in\mathbb{F}_2^n} \mathcal{N}_F(w,v)\mathcal{W}_G(u\oplus w, v)\\ =\ & 2^{-n} \sum_{w\in\mathbb{F}_2^n}\sum_{x\in\mathbb{F}_2^n} (-1)^{v\cdot F(x)\oplus w\cdot x} i^{wt(x)}\\ &\times \sum_{y\in\mathbb{F}_2^n} (-1)^{v\cdot G(y)\oplus(w\oplus u)\cdot y}\\ =\ & 2^{-n} \sum_{x,y\in\mathbb{F}_2^n} (-1)^{v\cdot(F(x)\oplus G(y))\oplus u\cdot y} i^{wt(x)}\\ &\times \sum_{w\in\mathbb{F}_2^n} (-1)^{w\cdot(x\oplus y)}\\ \overset{x=y}{=}\ & \sum_{x\in\mathbb{F}_2^n} (-1)^{v\cdot(F(x)\oplus G(x))\oplus u\cdot x} i^{wt(x)}\\ =\ & 2^{\frac{n}{2}} \mathcal{N}_{F\oplus G}(u,v).\end{aligned}$$

To show item(c), if $H(x,y) = F(x) \oplus G(y)$, $x, y \in \mathbb{F}_2^n$,

then

$$\begin{aligned}\mathcal{N}_{F\oplus G}(u,v,w) &= 2^{-n}\sum_{x,y\in\mathbb{F}_2^n}(-1)^{w\cdot(F(x)\oplus G(y))\oplus u\cdot x\oplus v\cdot y}\\ &\quad\times i^{wt(x,y)}\\ &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}(-1)^{w\cdot F(x)\oplus u\cdot x}i^{wt(x)}\\ &\quad\times\sum_{y\in\mathbb{F}_2^n}(-1)^{w\cdot G(y)\oplus v\cdot y}i^{wt(y)}\\ &= \mathcal{N}_F(u,w)\mathcal{N}_G(v,w).\end{aligned}$$

To show item(d), we write,

$$\begin{aligned}2^{\frac{k}{2}}\mathcal{N}_H(u,v,w) &= 2^{-\frac{n}{2}}\sum_{(x,y)\in\mathbb{F}_2^{n+k}}(-1)^{w\cdot(F(x)*G(y))\oplus u\cdot x\oplus v\cdot y}\\ &\quad\times i^{wt(x)+wt(y)}\\ &= 2^{-\frac{n}{2}}\sum_{y\in\mathbb{F}_2^k,G(y)=1}(-1)^{y\cdot v}i^{wt(y)}\\ &\quad\times\sum_{x\in\mathbb{F}_2^n}(-1)^{w\cdot F(x)\oplus u\cdot x}i^{wt(x)}\\ &\quad+2^{-\frac{n}{2}}\sum_{y\in\mathbb{F}_2^k,G(y)=0}(-1)^{y\cdot v}i^{wt(y)}\\ &\quad\times\sum_{x\in\mathbb{F}_2^n}(-1)^{u\cdot x}i^{wt(x)}\\ &\quad+2^{-\frac{n}{2}}\sum_{(x,y)\in\mathbb{F}_2^{n+k},G(y)\neq 0,1}(-1)^{w\cdot(F(x)*G(y))}\\ &\quad\times(-1)^{u\cdot x\oplus v\cdot y}i^{wt(x,y)}\\ &= \mathcal{N}_F(u,w)\sum_{y\in\mathbb{F}_2^k,G(y)=1}(-1)^{y\cdot v}i^{wt(y)}\\ &\quad+\omega^n i^{-wt(u)}\sum_{y\in\mathbb{F}_2^k,G(y)=0}(-1)^{y\cdot v}i^{wt(y)}\\ &\quad+2^{-\frac{n}{2}}\sum_{(x,y)\in\mathbb{F}_2^{n+k},G(y)\neq 0,1}(-1)^{w\cdot(F(x)*G(y))}\\ &\quad\times(-1)^{u\cdot x\oplus v\cdot y}i^{wt(x,y)},\end{aligned}$$

from which we obtain the desired identity. □

Next, we study nega-crosscorrelation function of vectorial Boolean functions and give a definition. The sum

$$\mathcal{C}_{F,G}(u,v)=\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus G(x\oplus u))}(-1)^{x\cdot u}$$

is the nega-crosscorrelation of F and G at $(u,v)\in\mathbb{F}_2^n\times\mathbb{F}_2^m$. Taking $F=G$ above, we get the nega-autocorrelation $\mathcal{C}_F(u,v)$ of $F\in\mathcal{VB}_n^m$ at $(u,v)\in\mathbb{F}_2^n\times\mathbb{F}_2^m$.

Based on the study of correlation functions in [17] and the study of the nega-Hadamard transform in [18,19], the conclusion in Theorem 2 can be obtained.

Theorem 2. *If F,G are vectorial functions, then the nega-crosscorrelation equals*

$$\mathcal{C}_{F,G}(z,v)=i^{wt(z)}\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F(u,v)\overline{\mathcal{N}_G(u,v)}(-1)^{u\cdot z}.$$

Proof. We start with the sum at the right hand side.

$$\begin{aligned}&i^{wt(z)}\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F(u,v)\overline{\mathcal{N}_G(u,v)}(-1)^{u\cdot z}\\ &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}\sum_{y\in\mathbb{F}_2^n}(-1)^{v\cdot F(x)\oplus v\cdot G(y)}i^{wt(x)-wt(y)+wt(z)}\\ &\quad\times\sum_{u\in\mathbb{F}_2^n}(-1)^{u\cdot(x\oplus y\oplus z)}\\ &= \sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus G(x\oplus z))}(-1)^{x\cdot z}\\ &= \mathcal{C}_{F,G}(z,v).\end{aligned}$$

We complete the proof. □

Next, we give the inverse of the nega-crosscorrelation of F and G in Theorem 2.

Theorem 3. *If F,G are vectorial functions, then*

$$\begin{aligned}&\mathcal{N}_F(u,v)\overline{\mathcal{N}_G(u,v)}\\ =&2^{-n}\sum_{z\in\mathbb{F}_2^n}\mathcal{C}_{F,G}(z,v)(-1)^{u\cdot z}i^{-wt(z)}.\end{aligned}\tag{1}$$

Proof. We start with the sum at the left hand side.

$$\begin{aligned}\mathcal{N}_F(u,v)\overline{\mathcal{N}_G(u,v)} &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot F(x)\oplus u\cdot x}i^{wt(x)}\\ &\quad\times\overline{\sum_{y\in\mathbb{F}_2^n}(-1)^{v\cdot G(y)\oplus u\cdot y}i^{wt(y)}}\\ &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot F(x)\oplus G(x\oplus z)}\\ &\quad\times\sum_{z\in\mathbb{F}_2^n}(-1)^{u\cdot z}i^{wt(x)-wt(x\oplus z)}\\ &= 2^{-n}\sum_{z\in\mathbb{F}_2^n}\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot F(x)\oplus G(x\oplus z)}\\ &\quad\times(-1)^{u\cdot z}(-1)^{x\cdot z}i^{-wt(z)}\\ &= 2^{-n}\sum_{z\in\mathbb{F}_2^n}\mathcal{C}_{F,G}(z,v)(-1)^{u\cdot z}i^{-wt(z)}.\end{aligned}$$

We complete the proof. □

If we take $F=G$ in the previous Theorem 3, then we obtain

$$|\mathcal{N}_F(u,v)|^2=2^{-n}\sum_{z\in\mathbb{F}_2^n}\mathcal{C}_F(z,v)(-1)^{u\cdot z}i^{-wt(z)}.\tag{2}$$

Based on Theorem 2, we then investigate a property of nega-Hadamard transform of vectorial Boolean functions.

Corollary 1. *(The nega-Parseval's Identity of vectorial Boolean function) We have*

$$\sum_{u\in\mathbb{F}_2^n}|\mathcal{N}_F(u,v)|^2=2^n.$$

Proof. If we take $F = G$ in the previous Theorem 2, then we obtain

$$\begin{aligned}&\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus F(x\oplus z))}(-1)^{x\cdot z}\\ =&i^{wt(z)}\sum_{u\in\mathbb{F}_2^n}|\mathcal{N}_F(u,v)|^2(-1)^{u\cdot z}.\end{aligned}\tag{3}$$

If $z = 0$, we obtain a proof of this fact for the particular case of nega-Hadamard transforms. □

Based on the above research, Lemma 2 provides another characterization of vectorial negabent functions.

Lemma 2. *A vectorial function $F \in \mathcal{VB}_n$ is vectorial negabent if and only if $\mathcal{C}_F(z, v) = 0$, for all $z \in (\mathbb{F}_2^n)^*$.*

Proof. If F is a vectorial negabent function, then $|\mathcal{N}_F(u, v)| = 1$ for all $u \in \mathbb{F}_2^n$. For all $z \neq 0$ then by Equation (3), we obtain $\mathcal{C}_F(z, v) = 0$. The converse also follows from Equation (3). □

3.2 Links Between The Nega-crosscorrelation And The Nega-Hadamard Transform

In this subsection, we research the nega-crosscorrelation by the nega-Hadamard transform of the vectorial function. The following Theorem 4 shows that the relations between the nega-crosscorrelation and the nega-Hadamard transform.

Theorem 4. *Let $F, G \in \mathcal{VB}_n^m$. Then for any $u, a \in \mathbb{F}_2^n$ and $v \in \mathbb{F}_2^m$,*

$$\begin{aligned}\mathcal{N}_F(u,v)\mathcal{N}_G(u,v) = 2^{-n}&\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_{F,G}(a,v)\\ &\times(-1)^{u\cdot a\oplus wt(x)}i^{wt(a)}.\end{aligned}$$

Proof. According to the definition, for any $u \in \mathbb{F}_2^n$, $v \in \mathbb{F}_2^m$,

$$\begin{aligned}\mathcal{N}_F(u,v)\mathcal{N}_G(u,v) &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot F(x)\oplus u\cdot x}i^{wt(x)}\\ &\quad\times\sum_{y\in\mathbb{F}_2^n}(-1)^{v\cdot G(y)\oplus u\cdot y}i^{wt(y)}\\ &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}\sum_{a\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus G(x\oplus a))\oplus u\cdot a}\\ &\quad\times i^{wt(x)+wt(x\oplus a)}\\ &= 2^{-n}\sum_{x\in\mathbb{F}_2^n}\sum_{a\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus G(x\oplus a))\oplus x\cdot a}\\ &\quad\times(-1)^{u\cdot a}i^{2wt(x)+wt(a)}\\ &= 2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_{F,G}(a,v)\\ &\quad\times(-1)^{u\cdot a\oplus wt(x)}i^{wt(a)}.\end{aligned}$$

We complete the proof. □

If $F(x) = G(x)$, we have

$$\mathcal{N}_F^2(u,v) = 2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_F(a,v)(-1)^{u\cdot a\oplus wt(x)}i^{wt(a)}.\tag{4}$$

Theorem 5. *Let $F, G \in \mathcal{VB}_n^m$. Then for any $u, a \in \mathbb{F}_2^n$ and $v \in \mathbb{F}_2^m$,*

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F^2(u,v)\mathcal{N}_G^2(u\oplus e,v)\\ =&2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_F(a,v)\mathcal{C}_G(a,v)(-1)^{a\cdot e\oplus wt(a)+wt(x,y)}\end{aligned}\tag{5}$$

Proof. According to Equation (4), we note that the left-hand side in Equation (5) can be rewritten, for any $e \in \mathbb{F}_2^n$,

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F^2(u,v)\mathcal{N}_G^2(u\oplus e,v)\\ &= 2^{-2n}\sum_{u\in\mathbb{F}_2^n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_F(a,v)(-1)^{u\cdot a\oplus wt(x)}i^{wt(a)}\\ &\quad\times\sum_{b\in\mathbb{F}_2^n}\mathcal{C}_G(b,v)(-1)^{(u\oplus e)\cdot b\oplus wt(y)}i^{wt(b)}\\ &= 2^{-2n}\sum_{u\in\mathbb{F}_2^n}\sum_{a\in\mathbb{F}_2^n}\sum_{b\in\mathbb{F}_2^n}\mathcal{C}_F(a,v)\mathcal{C}_G(b,v)\\ &\quad\times(-1)^{u\cdot a\oplus(u\oplus e)\cdot b\oplus wt(x,y)}\\ &= 2^{-2n}\sum_{a\in\mathbb{F}_2^n}\sum_{b\in\mathbb{F}_2^n}\mathcal{C}_F(a,v)\mathcal{C}_G(b,v)\\ &\quad\times(-1)^{b\cdot e\oplus wt(x,y)}\sum_{u\in\mathbb{F}_2^n}(-1)^{u\cdot(a\oplus b)}i^{wt(a,b)},\end{aligned}$$

where

$$\sum_{u\in\mathbb{F}_2^n}(-1)^{u\cdot(a\oplus b)} = \begin{cases}2^n, & if\, a = b\\ 0, & if\, a \neq b.\end{cases}$$

Then we have for all $a \in \mathbb{F}_2^n$,

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F^2(u,v)\mathcal{N}_G^2(u\oplus e,v)\\ =&2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_F(a,v)\mathcal{C}_G(a,v)(-1)^{e\cdot a\oplus wt(a)\oplus wt(x,y)}.\end{aligned}$$

We complete the proof. □

Next, we have two Corollaries.

Corollary 2. *If $F = G$ in Equation (5), we have*

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F^2(u,v)\mathcal{N}_F^2(u\oplus e,v)\\ =&2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_F^2(a,v)(-1)^{e\cdot a\oplus wt(a)\oplus wt(x,y)}.\end{aligned}$$

Moreover, if $e = 0$, then

$$\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F^4(u,v) = 2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_F^2(a,v)(-1)^{wt(a)\oplus wt(x,y)}$$

Corollary 3. *Let $F, G \in \mathcal{VB}_n^m$, then*

$$\sum_{u\in\mathbb{F}_2^n}\mathcal{N}_F^2(u,v)\mathcal{N}_G^2(u,v) = 2^{-n}\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_{F,G}^2(a,v)(-1)^{wt(a)\oplus wt(x,y)}$$

3.3 The Relationship Among Nega-crosscorrelation Of Four Vectorial Boolean Functions

Let F, G, H, K are vectorial Boolean functions. The partial results of this subsection can be derived by Theorem 6, namely, as some special cases of Theorem 6, which involves the four nega-crosscorrelation functions $\mathcal{C}_{F,G}(u,v), \mathcal{C}_{H,K}(u\oplus e, v), \mathcal{C}_{F,H}(a,v)$,and $\mathcal{C}_{G,K}(a\oplus e, v)$.

Theorem 6. *Let $F, G, H, K \in \mathcal{VB}_n^m$, then for any $e, r, a \in \mathbb{F}_2^n$*

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{H,K}(u\oplus e,v)(-1)^{e\cdot u}\\ =&\sum_{a\in\mathbb{F}_2^n} \mathcal{C}_{F,H}(a,v)\mathcal{C}_{G,K}(a\oplus e,v)(-1)^{e\cdot a}.\end{aligned} \tag{6}$$

Proof. Using the definition of the nega-crosscorrelation, for any $e, r, a \in \mathbb{F}_2^n$

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{H,K}(u\oplus e,v)(-1)^{e\cdot u}\\ =&\sum_{u\in\mathbb{F}_2^n}\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus G(x\oplus u))\oplus u\cdot x}\\ &\times\sum_{y\in\mathbb{F}_2^n}(-1)^{v\cdot(H(y)\oplus K(y\oplus u\oplus e))\oplus(u\oplus e)\cdot y}(-1)^{e\cdot u}\\ =&\sum_{x\in\mathbb{F}_2^n}\sum_{y\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus H(y))}\\ &\times\sum_{u\in\mathbb{F}_2^n}(-1)^{v\cdot(G(x\oplus u)\oplus K(y\oplus u\oplus e))\oplus(u\oplus e)\cdot y\oplus u\cdot x}(-1)^{e\cdot u}\\ =&\sum_{x\in\mathbb{F}_2^n}\sum_{y\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus H(y))}\\ &\times\sum_{r\in\mathbb{F}_2^n}(-1)^{v\cdot(G(r)\oplus K(r\oplus x\oplus y\oplus e))\oplus(x\oplus r)\cdot(x\oplus y)\oplus e\cdot(y\oplus x\oplus r)}\\ =&\sum_{x\in\mathbb{F}_2^n}\sum_{a\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus H(x\oplus a))\oplus a\cdot x}\\ &\times\sum_{r\in\mathbb{F}_2^n}(-1)^{v\cdot(G(r)\oplus K(r\oplus a\oplus e))\oplus r\cdot a\oplus e\cdot r\oplus a\cdot e}\\ =&\sum_{a\in\mathbb{F}_2^n}\mathcal{C}_{F,H}(a,v)\mathcal{C}_{G,K}(a\oplus e,v)(-1)^{e\cdot a}\end{aligned}$$

This proves the equality in the theorem. □

In Equation (6), if $F = H$, $G = K$, then

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{F,G}(u\oplus e,v)(-1)^{e\cdot u}\\ =&\sum_{a\in\mathbb{F}_2^n} \mathcal{C}_F(a,v)\mathcal{C}_G(a\oplus e,v)(-1)^{e\cdot a}.\end{aligned} \tag{7}$$

In particular, if $e = 0$, then we have the following corollary.

Corollary 4. *Let $F, G \in \mathcal{VB}_n^m$, then*

$$\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}^2(u,v) = \sum_{a\in\mathbb{F}_2^n} \mathcal{C}_F(a,v)\mathcal{C}_G(a,v). \tag{8}$$

Thus Equation (8) gives the relationship between $\mathcal{C}_{F,G}(u,v)$ and $\mathcal{C}_F(a,v)$, $\mathcal{C}_G(a,v)$. In Equation (6), if $G = K$, then

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{H,G}(u\oplus e,v)(-1)^{e\cdot u}\\ =&\sum_{a\in\mathbb{F}_2^n} \mathcal{C}_{F,H}(a,v)\mathcal{C}_G(a\oplus e,v)(-1)^{e\cdot a}.\end{aligned} \tag{9}$$

Moreover, when G is a vectorial negabent function, we have the following corollary.

Corollary 5. *Let $F, G, H \in \mathcal{VB}_n^m$, and G is vectorial negabent function. Then*

1)

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{H,G}(u\oplus e,v)(-1)^{e\cdot u}\\ =&\begin{cases}2^n\mathcal{C}_{F,H}(e,v)(-1)^{wt(e)}, & if a = e\\ 0, & if a \neq e.\end{cases}\end{aligned}$$

2) $\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}^2(u,v) = 2^{2n}$.

3) If $e \neq 0, v \neq 0$ and F is a vectorial negabent function, then

$$\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{F,G}(u\oplus e,v) = 0.$$

Proof. According to Equation (9), since G is vectorial negabent, then

$$\mathcal{C}_G(a\oplus e,v) = \begin{cases}2^n, & if a = e\\ 0, & if a \neq e.\end{cases}$$

Therefore, (1) holds.

In (1), if $a = e, F = H$, we get

$$\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{F,G}(u\oplus e,v)(-1)^{e\cdot u} = 2^n\mathcal{C}_F(e,v)(-1)^{wt(e)},$$

when $e = 0$, (2) holds.

When $e \neq 0$ and F is vectorial negabent, then

$$\begin{aligned}&\sum_{u\in\mathbb{F}_2^n} \mathcal{C}_{F,G}(u,v)\mathcal{C}_{F,G}(u\oplus e,v)(-1)^{e\cdot u}\\ =&2^n\mathcal{C}_F(e,v)(-1)^{wt(e)}\\ =&2^n\sum_{x\in\mathbb{F}_2^n}(-1)^{v\cdot(F(x)\oplus F(x\oplus e))\oplus e\cdot x}(-1)^{wt(e)}\\ =&0.\end{aligned}$$

(3) holds. We complete the proof. □

3.4 Decomposition Of Vectorial Boolean Functions And The Condition That A Class Of Functions Are Vectorial Negabent Functions

In this subsection, we mainly study some results of decomposition of vectorial Boolean function. Suppose $1 \leq r \leq n$. Then, a vectorial function can be decomposed as a vectorial function from $\mathbb{F}_2^r \times \mathbb{F}_2^{n-r}$ to $\mathbb{F}_2^m$. If $v \in \mathbb{F}_2^r$, the vectorial function $F_v \in \mathcal{VB}_{n-r}^m$ is defined as $F_v(x) = F(v, x)$, for all $x \in \mathbb{F}_2^{n-r}$.

Theorem 7. *Let $F \in \mathcal{VB}_n^m$ be expressed as $F : \mathbb{F}_2^r \times \mathbb{F}_2^{n-r} \to \mathbb{F}_2^m$. Then*

$$\mathcal{C}_F(u, w, a) = \sum_{v \in \mathbb{F}_2^r} \mathcal{C}_{F_v, F_{v\oplus u}}(w, a)(-1)^{v\cdot u}.$$

Proof. By definition,

$$\begin{aligned}\mathcal{C}_F(u, w, a) &= \sum_{v\in\mathbb{F}_2^r} \sum_{z\in\mathbb{F}_2^{n-r}} (-1)^{a\cdot(F(v,z)\oplus F(v\oplus u, z\oplus w))} \\ &\quad \times (-1)^{v\cdot u\oplus z\cdot w} \\ &= \sum_{v\in\mathbb{F}_2^r} (-1)^{v\cdot u} \sum_{z\in\mathbb{F}_2^{n-r}} (-1)^{a\cdot(F_v(z)\oplus F_{v\oplus u}(z\oplus w))} \\ &\quad \times (-1)^{z\cdot w} \\ &= \sum_{v\in\mathbb{F}_2^r} \mathcal{C}_{F_v, F_{v\oplus u}}(w, a)(-1)^{v\cdot u}.\end{aligned}$$

□

If the nega-autocorrelation functions of the vectorial Boolean functions F and G satisfy $\mathcal{C}_F(u, w) + \mathcal{C}_G(u, w) = 0$, for all nonzero $u \in \mathbb{F}_2^n$, then F and G are said to be complementary nega-autocorrelation. The following Lemma 3 establishes a connection between the nega-autocorrelations of F, G and their nega-Hadamard transformations.

Lemma 3. *Two vectorial functions $F, G \in \mathcal{VB}_n^m$ have complementary nega-autocorrelations if and only if $|\mathcal{N}_F(u, v)|^2 + |\mathcal{N}_G(u, v)|^2 = 2$, for all $(u, v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$.*

Proof. Let F, G be two vectorial functions with complementary nega-autocorrelations. Then

$$\begin{aligned}|\mathcal{N}_F(u, v)|^2 + |\mathcal{N}_G(u, v)|^2 &= 2^{-n} \sum_{z\in\mathbb{F}_2^n} (\mathcal{C}_F(z, v) + \mathcal{C}_G(z, v)) \\ &\quad \times i^{-wt(z)}(-1)^{u\cdot z} \\ &= 2^{-n} \cdot 2^{n+1} = 2.\end{aligned}$$

Conversely, suppose $|\mathcal{N}_F(u, v)|^2 + |\mathcal{N}_G(u, v)|^2 = 2$, for all $(u, v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$. Then

$$\begin{aligned}&\mathcal{C}_F(z, v) + \mathcal{C}_G(z, v) \\ =&i^{wt(z)} \sum_{u\in\mathbb{F}_2^n} (|\mathcal{N}_F(u, v)|^2 + |\mathcal{N}_G(u, v)|^2)(-1)^{u\cdot z} \\ =&2i^{wt(z)} \sum_{u\in\mathbb{F}_2^n} (-1)^{u\cdot z} \\ =&2^{n+1} i^{wt(z)} \delta_0(z),\end{aligned}$$

where

$$\delta_0(z) = \begin{cases} 0, & if\, z \neq 0 \\ 1, & if\, z = 0. \end{cases}$$

Thus, the functions F and G have complementary nega-autocorrelations. □

Theorem 8. *Suppose $F \in \mathcal{VB}_{n+2}^m$ is expressed as*

$$\begin{aligned}F(x, x_{n+1}, x_{n+2}) &= F_1(x) + x_{n+1}(F_1(x) + F_2(x)) \\ &\quad + x_{n+2}(F_1(x) + F_3(x)) \\ &\quad + x_{n+1}x_{n+2}(F_2(x) + F_3(x))\end{aligned}$$

for all $(x, x_{n+1}, x_{n+2}) \in \mathbb{F}_2^n \times \mathbb{F}_2 \times \mathbb{F}_2$, where $F_1, F_2, F_3 \in \mathcal{VB}_n^m$. If F_1 is vectorial negabent function, F_2 and F_3 have complementary nega-autocorrelations and $\mathcal{C}_{F_1,F_2} = \mathcal{C}_{F_1,F_3}$, for all $(u, v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$ with $wt(u) \equiv 1(\mathrm{mod}2)$, then F is vectrial negabent function.

Proof. Suppose F is a vectorial negabent fuction. Then $\mathcal{C}_F(u, a, b, v) = 0$ for all nonzero $(u, a, b) \in \mathbb{F}_2^n \times \mathbb{F}_2 \times \mathbb{F}_2$. If F_1 is vectorial negabent function, $\mathcal{C}_{F_2}(u, v) + \mathcal{C}_{F_3}(u, v) = 0$ and $\mathcal{C}_{F_1,F_2}(u, v) = \mathcal{C}_{F_1,F_3}(u, v)$ for all $(u, v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$ with $wt(u) \equiv 1(\mathrm{mod}2)$.

We obtain

$$\mathcal{C}_F(u, 0, 0, v) = 2\mathcal{C}_{F_1}(u, v) + \mathcal{C}_{F_2}(u, v) + \mathcal{C}_{F_3}(u, v) = 0,$$

$$\mathcal{C}_F(u, 1, 1, v) = 2\mathcal{C}_{F_1}(u, v) - (1 + (-1)^{wt(u)})\mathcal{C}_{F_2,F_3}(u, v) = 0,$$

for all $u \in (\mathbb{F}_2^n)^*$ and

$$\begin{aligned}\mathcal{C}_F(u, 1, 0, v) &= (1 - (-1)^{wt(u)})(\mathcal{C}_{F_1,F_2}(u, v) - \mathcal{C}_{F_1,F_3}(u, v)) \\ &= 0, \\ \mathcal{C}_F(u, 0, 1, v) &= (1 - (-1)^{wt(u)})(\mathcal{C}_{F_1,F_3}(u, v) - \mathcal{C}_{F_1,F_2}(u, v)) \\ &= 0.\end{aligned}$$

We complete the proof. □

4 Conclusions

In this paper, we have investigated the nega-Hadamard transform of vectorial Boolean functions in detail. First, we study the properties of the nega-Hadamard transform and the definition of vectorial negabent functions. Next, we also investigate some properties of the nega-Hadamard transform and nega-crosscorrelation, and generalize certain known properties of the correlation function. Thirdly, the relationship among nega-crosscorrelations of four vectorial Boolean functions has been shown, and some important equations have been obtained. Finally, we concentrate on decompositions of vectorial Boolean functions and the condition that a class of functions are vectorial negabent functions are given. The above research on the excellent properties of cryptographic functions not only has a profound impact on other problems of cryptography, but also makes a great contribution to network security.

Acknowledgments

This study was supported by the Huaibei Normal University graduate student innovation fund project(CX2023045).The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] C. Carlet, "Boolean functions for cryptography and error correcting codes," in Boolean Models and Methods in Mathematics, Computer Science, and Engineering, Cambridge, the United Kingdom: Cambridge University Press, pp. 257-397, 2010.

[2] C. Carlet, "Vectorial Boolean functions for cryptography," in Boolean Models and Methods in Mathematics, Computer Science, and Engineering, Cambridge, the United Kingdom: Cambridge University Press, pp. 398-469, 2010.

[3] T. W. Cusick and P. Stănică, "Cryptographic Boolean functions and Applications," New York: Academic, 2009.

[4] O. S. Rothaus , "On "bent" functions," Journal of Combinatorial Theory, Series A, vol. 20, no. 3, pp. 300-305, May 1976.

[5] J. F. Dillon, "Elementary Hadamard difference sets," Ph.D. dissertation: University of Maryland, 1974.

[6] C. Riera and M. G. Parker, "Generalized bent criteria for Boolean functions," IEEE Transactions on Inforamation Theory, vol. 52, no. 9, pp. 4142-4159, Feb. 2006.

[7] M. G. Parker, "The constant properties of Goley-Devis-Jedwab sequences," International symposium on information theory, Sorrento, Italy, 2000. http://www.ii.uib.no/matthew/mattweb.html.

[8] X. Chen , L. J. Qu , S. J. Fu,et al."The number of affine equivalent classes and extended affine equivalent classes of vectorial Boolean functions," Discrete Applied Mathematics, vol. 289, no. 20, pp. 477-491, Jan. 2021.

[9] N. Jiang, M. Zhao, Z. Y. Yang, et al. "Characterization and Properties of Bent-Negabent Functions," Chinese Journal of Electronics, vol. 31, no. 4, pp.786-792, Jul. 2022.

[10] Z. Y. Yang, P. H. Ke, Z. X. Chen. "New Secondary Constructions of Generalized Bent Functions", Chinese Journal of Electronics, vol. 30, no. 6, pp.1022-1029, Nov. 2021.

[11] P. Stănică, S. Gangopadhyay, A. Chaturvedi, et al., "Investigations on Bent and Negabent Functions via the Nega-Hadamard Transform," IEEE Transactions on Information Theory, vol.58, no. 6, pp. 4064-4072, Jun. 2012.

[12] P. Sarkar and S. Maitra, "Cross-correlation analysis of cryptographically useful Boolean functions and S-boxes," Theory Computer Systems, vol. 35, pp. 39-57, Jan. 2002.

[13] Z. P. Zhuo, "On Cross-Correlation Properties of Boolean Functions," Internation Journal of Computer Mathematics, vol. 88, no. 10, pp. 2035-2041, Nov. 2011. (SCI:000291460400004)

[14] A. Pott, "Almost perfect and planar functions," Designs, Codes Cryptography, vol. 78, no. 1, pp. 141-195, Jan. 2016.

[15] A. A. Polujan and A. Pott, "On Design-Theoretic Aspects of Boolean and Vectorial Bent Function," IEEE Transactions on Inforamation Theory, vol. 67, no. 2, pp. 1027-1037, Feb. 2021.

[16] M. G. Parker and A. Pott, "On Boolean functions which are bent and negabent," in Proceeding of the 2007 International Conference on Sequences, Subsequences, and Consequences, pp. 9-23, Los Angeles, California, the United States, May 2007.

[17] P. Sarkar and S. Maitra, "Cross correlation analysis of cryptographically useful Boolean functions and S-Boxes," Theory of Computing Systems, vol. 35, pp. 39-57, Feb. 2002.

[18] L. E. Danielsen, T. A. Gulliver, and M. G. Parker, "Aperiodic propagation criteria for Boolean functions," Information and Compution, vol. 204, no. 5, pp.741-770, May 2006.

[19] C. Riera and M. G. Parker, "One and two-variable interlace polynomials: A spectral interpretation," in Proceeding of the International Conference Coding Cryptography, vol. LNCS-3969, pp. 397-411, 2006.

Biography

Jingjing Zhang received the B.S. degrees in 2020 from the School of Information , Huaibei Normal University. She is currently a master course student at the School of Mathematical Sciences, Huaibei Normal University. Her research interests include cryptography and information theory.

Zepeng Zhuo received the M.S. degree from Huaibei Normal University in 2007, and the Ph.D. degree from Xidian University in 2012. Since 2002, he has been with the School of Mathematical Science, Huaibei Normal University, where he is now a professor. His research interests include cryptography and information theory.

Guolong Chen received the M.S. degree from Beijing Normal University in 1993, and the Ph.D. degree from Beijing Normal University in 1998. Since 2020, he has been with the School of Computer Engineering, Bengbu College, where he is now a professor. His research interests include mathematical logic and its applications.

Distributed Parallel Algorithm for Finite Element Multi-Computer System Considering Network Security Performance Evaluation

Yi Li

(Corresponding author: Yi Li)

Department of Information Technology, Henan Judicial Police Vocational College
Zhengzhou, Henan 450046, China
Email: jiuersedi@outlook.com

(Received Dec. 20, 2023; Revised and Accepted May 9, 2024; First Online June 22, 2024)

The Special Issue on Cybersecurity and Privacy in the Industrial Internet of Things (IIoT)

Guest Editor: Prof. Zhengyi Chai (Tiangong University, China)

Abstract

In the complex network topology of traditional finite element multi-microcomputer systems, the data transmission process can cause a series of errors. In order to improve the operation effect of finite element multi-computer systems, this paper proposes a distributed parallel algorithm for finite element multi-computer systems considering network security performance evaluation. Taking advantage of the complementarity of DE (Differential Evolution) and EP (Evolutionary Programming), this paper introduces EP under the framework of DE. It constructs a hybrid algorithm named DEEP (Differential Evolution Evolutionary Programming), which is dominated by DE and supplemented by EP. Moreover, this paper adopts the partition clustering technique to obtain a more accurate partition scheme. To verify the correctness of the distributed parallel algorithm, the distributed parallel algorithm proposed in this paper is studied through experiments. This article studies the use of parallel computing technology to accelerate the computational speed of DE for solving reactive power optimization problems and implements it on a microcomputer cluster platform. The required population size is reduced by improving the algorithm itself, thereby further accelerating computation or using smaller clusters to reduce costs. The distributed parallel algorithm proposed in this paper has been verified to have good stability at various frequencies through square wave experiments; the experimental analysis shows that the distributed parallel algorithm proposed in this paper is effective, and it verifies that the distributed parallel algorithm proposed in this paper has a certain promotion effect on the security improvement of the finite element multi-computer system.

Keywords: Finite Element; Multi-computer System; Network Security; Parallel Algorithm

1 Introduction

The microcomputer monitoring system is an important component of the modernization of railway equipment. It integrates the latest modern technologies, such as sensors, fieldbus, computer network communication, databases, and software engineering, to monitor and record the main operating status of signal equipment, providing scientific basis for the electrical department to grasp the quality of equipment application and fault analysis. At the same time, the system also has the function of data logic judgment. When the working condition of the signal equipment deviates from the predetermined limit or abnormal occurs, it will alarm in a timely manner to avoid affecting the safety and punctual operation of the train due to equipment failure or illegal operation

The microcomputer monitoring system enables signal equipment to have self diagnosis function, thereby greatly improving the safety of the signal system. It can reflect the operating status of signal equipment 24/7 during operation, detect potential faults, eliminate hidden dangers, and use computer technology to logically judge, which is conducive to capturing instantaneous and intermittent faults. Through playback and reproduction, it is conducive to analyzing faults and distinguishing responsibilities. The microcomputer monitoring system can grasp the working status and trend of signal equipment, which is the technical basis for promoting signal equipment status maintenance and providing scientific basis for maintenance decision-making.

The transmission system functions and is technically safe according to the same process according to European standard EN50129. However, the use of untrusted transport systems limits the course of this functional approach. Therefore, the safety-related transmission system should be characterized by a functional specification including

an overall error model, and the safety integrity requirements specification should be formulated on the basis of a functional analysis of this error model. It is proposed in the functional integrity requirements that the design organization should provide six protective measures: check transmission identifier error, check data type error, check data value error, check the error of expired data or data not received within a predetermined time, check the data loss after a predefined delay, and ensure the independence of the security transmission function and the level of use of the untrusted transmission system. Meanwhile, six requirements should be fulfilled in the security integrity requirements: security protection should be applied to the generation of transmitted data; security response should be generated in the case of mis-operation, which is consistent with the security requirements of the receiver, the receiver applies an error checking mechanism and should be consistent with the receiver's security requirements; the use of the second clause in an untrusted transmission system should be functionally independent; the residual data error rate for each information exchange between sender and receiver within a safety-relevant transmission system should be less than a predetermined value. This ratio should be adapted to the safety integrity level (SIL) of each receiver.

Authenticity, integrity and accurate time of data should be ensured when communicating with each other between safety-related devices. In order to maintain the security required for communication between safety-related devices, the following requirements should be met: if the source of the transmission system is not uniquely specified, authenticity should be provided by adding a source indicator to user data; integrity should be provided by the user The security code is added to the data to provide, the security process can not only rely on the generated transmission code, but must be checked by the overall circuit as part of the untrusted transmission system; the timing of user data should be provided by adding time information to the user data., the time delay can be independent of the application [1–3]; it is necessary to set the order in which the safety process should check the information: the safety program of the safety-related device is functionally independent of the program used for the untrusted transmission system, if the two programs use the For the same coding structure, the parameters should also be different [4]; if the transmission quality falls below the pre-established transmission requirement specification level, an appropriate security response should be triggered. When safety and non-safety devices communicate with each other, safety-related and safety-independent information should have different structures, which is accomplished by applying secure coding to safety-related information [5]. This safety code shall protect the system to the required SIL from safety non-safety information to safety relevant information. The safety procedures of safety-related equipment should be functionally independent from the procedures used by untrusted transmission systems and safety-independent equipment [6].

In a complex network topology, the data transmission process will cause a series of errors, such as: packet loss, duplication, insertion, delay, timing errors, user data corruption and addressing errors [7]. From the perspective of the communication receiver, the following two communication errors may lead to dangerous situations:

1) data errors, such as: sending (receiving) address error, data type error, data value error;

2) timing errors, such as communication delay If it is too long, the sequence of sending and receiving data frames is inconsistent.

The goal of secure communication is to establish a protection mechanism to avoid the above errors during the transmission of security-related data, or when the above errors occur, the system can detect these errors in time and take necessary security measures [8].

In the existing computerized automatic inter-station blocking system, the real-time information is not checked in the secure communication layer, and only relies on the TCP/IP (Transmission Control Protocol/Internet Protocol) protocol of the untrusted transmission system to ensure that TCP is a connection-oriented end-to-end Reliable protocol and guarantees the order in which packets are delivered [9]. The sequence is guaranteed by the response sequence number, which tells the receiver the next packet the sender expects [10]. If the confirmation response is not received within the specified time, the packet needs to be resent. The reliable mechanism of TCP allows devices to handle lost, deleted and misread packets, and the pause mechanism allows devices to monitor lost packets and request retransmission [11].

The computer monitoring system is an important part of the modernization of railway equipment [12]. It integrates the latest modern technologies, such as sensors, fieldbus, computer network communication, database and software engineering, etc., to monitor and record the main operating status of the signal equipment, and provide the electric affairs department to master the use quality of the equipment and provide fault analysis. Scientific basis [13, 14]. At the same time, the system also has the function of data logic judgment. When the working condition of the signal equipment deviates from the predetermined limit or abnormality occurs, it will alarm in time to avoid equipment failure or illegal operation affecting the safe and punctual operation of the train [15]. The computer monitoring system enables the signal equipment to have a self-diagnosis function, thereby greatly improving the safety of the signal system [16]. When the signal equipment is running, it can reflect the operation status of the equipment all day long, and can find potential faults and eliminate hidden troubles. Moreover, it uses computer technology to make logical judgments, which is conducive to capturing instantaneous faults and intermittent faults, and reproducing them through playback. It is beneficial to analyze faults and distinguish responsibilities [17]. The types of detection objects can be roughly divided

into analog quantities and switch quantities. The analog quantities include: power screen voltage, track circuit voltage, switch operating current, cable insulation resistance and power supply-to-ground leakage current, etc. Switch quantities include: key relay status, Console button and identification lamp status, fuse status, filament status and turnout indicating gap status, etc. [18].

In order to improve the operation effect of the finite element multi-computer system, this paper proposes a distributed parallel algorithm for the finite element multi-computer system considering the performance evaluation of network security, and combines the experimental analysis to verify the algorithm effect to improve the operation reliability of the microcomputer system.

The innovation of this article lies in the study of using parallel computing technology to accelerate the calculation speed of DE for solving reactive power optimization problems, and implementing it on a microcomputer cluster platform. By improving the algorithm itself, the required population size is reduced, thereby further accelerating computation or using smaller clusters to reduce costs.

The contribution of this article is that the proposed distributed parallel algorithm has a certain promoting effect on the security improvement of finite element multi microcomputer systems, and also provides new references for related research.

2 Distributed Parallel Algorithms

2.1 Introduction of Differential Evolution Algorithms

The reason DE is named "differential evolution" is that the mutation operation of DE is not driven by a given random distribution function as in EP or ES, but by genetic differences between individuals randomly sampled from contemporary populations. For example, the form called DE/rand/1 shown in Equation (1):

$$\mathrm{DE}/rand/1 : u_{i,j}[k] + F(u_{r2,j}[k] - u_{r3,j}[k] \tag{1}$$

Among them, $r_1 \neq r_2 \neq r_3 \neq i \in \{1, 2, L, N\}$, that is, r1, r2 and r3 are integers randomly selected from the set $\{1, 2, \ldots, N\}$ that are not equal to each other and not equal to i; F is the proportional coefficient, which can be called the variation control coefficient.

According to the nomenclature of DE, $(u_{r1}[k])$ means to use the randomly selected individual as the basis of variation to form the increment of variation by the difference of a pair of randomly selected individuals.

We assume that there are only two control variables, then Equation (1) can be graphically illustrated by Figure 1 in a two-dimensional plane, and the ellipse in the figure represents the contour of the objective function.

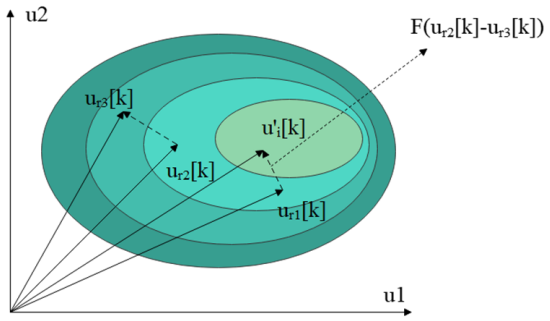


Figure 1: Schematic diagram of the variation of DE/rand/1

It is also possible to use the i-th individual $u_i[k]$ or the best contemporary individual $u_{best}[k]$ as the basis of variation instead of randomly selected individuals, and use the differences of more than one pair of individuals to form the variation increment, such as DE/current/1, DE/best/1, DE/rand/2, DE/current/2 and DE/best/2 as shown in Equations (2) - (6):

$$\mathrm{DE/current/1}: u_{i,j}[k] = u_{i,j}[k] + F(u_{r1,j}[k] - u_{r2,j}[k]) \tag{2}$$

$$\mathrm{DE/best/1}: u_{i,j}[k] = u_{best,j}[k] + F(u_{r1,j}[k] - u_{-2,j}[k]) \tag{3}$$

$$\mathrm{DE/rand/2}: u_{i,j}[k] = u_{i,j}[k] + F_1(u_{r2,j}[k] - u_{r3,j}[k]) + F_2(u_{r4,j}[k] - u_{rj}[k]) \tag{4}$$

$$\mathrm{DE/current/2}: u_{i,j}[k] = u_{i,j}[k] + F_1(u_{r1,j}[k] - u_{r2,j}[k]) + F_2(u_{r3,j}[k] - u_{r4,j}[k]) \tag{5}$$

$$\mathrm{DE/best/2}: u_{i,j}[k] = u_{best,j}[k] + F_1(u_{r1,j}[k] - u_{r2,j}[k]) + F_2(u_{r3,j}[k] - u_{r,j}[k]) \tag{6}$$

To speed up the search, DE also uses the hybridization operation. Discrete crossover was adopted in the early stage, mainly in two forms: "Binary Crossover" and "Exponential Crossover". Taking as an example, after introducing binary or exponential hybridization, two complete DE forms of DE/rand/1/bin and DE/rand/1/exp can be formed, as shown in Equations (7) and (8), respectively:

$$\mathrm{DE/rand/1/bin\ is}: u_{i,j}[k] = \begin{cases} u_{r1,j}[k] + F(u_{r2,j}[k] - u_{r3,j}[k], \\ if U_j(0,1) \leq \mathrm{CR}U_j = l_i \\ u_{i,j}[k], \quad otherwise \end{cases} \tag{7}$$

$$\mathrm{DE/rand/1/exp\ is}: \begin{cases} \mathrm{flag} = 1 \ \ if\ U_j(0,1) \leq \mathrm{CR}U_j = l_i \\ u_{r1,j}[k] = \begin{cases} u_{r1,j}[k] + F(u_{r2,j}[k] \\ -u_{r3,j}[k] if \mathrm{flag} = 1 \\ u_{i,j}[k], otherwise \end{cases} \end{cases} \tag{8}$$

Among them, $U_j(0,1)$ is a random number in the (0,1) interval, which is generated once for each control variable of individual i; $CR \in [0,1]$ is the given hybrid control coefficient; $l_i \in \{1,2,L,M\}$, which is generated only once for individual i.

Regarding the use of binary hybridization or exponential hybridization, the article points out that there is no significant difference in the effectiveness of the two hybridization forms in most cases, but the ability of binary hybridization to search the corners of hypercubes is stronger. Figure 2 graphically depicts a binary cross, in which a total of 8 control variables are assumed.

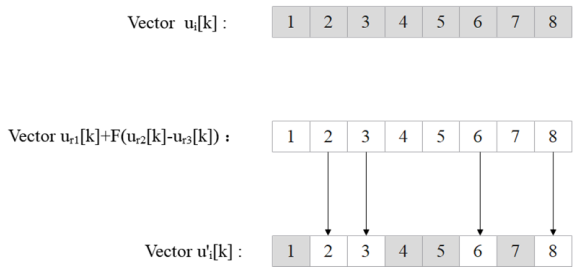


Figure 2: Schematic diagram of binary hybridization

The selection operation usually used by DE is "Binary Tournament Selection", which can also be more vividly called "one-to-one father-son competition", that is, comparing each child individual $u\mathbf{Mei}[k]$ with its parent individual $u_i[k]$, the one with higher fitness wins and enters the next generation, which becomes $u_i[k+1]$, as shown in Equation (9):

$$u_i[k+1] = \begin{cases} u_{i,i}[k], & \text{if} f'_i(k) \geq f_i(k) \\ u_i[k], & \text{if} f_i(k) \geq f'_i(k) \end{cases} \tag{9}$$

In Equation (9), the greater than or equal sign is used instead of the greater than sign, which helps to cope with the situation that the shape of the search space contains a platform part, so that the search process can "climb" over the platform.

2.2 Optimization Mechanism and Parameter Setting Analysis of Differential Evolution Algorithm

The advantages of DE's mutation operation are analyzed in detail, and it is considered that although it is very simple, it meets three conditions for becoming an excellent mutation operation:

Condition 1: The generated variation increments must conform to a random distribution centered at 0.

Condition 2: The magnitude of variation of each control variable should be dynamically changed to suit the shape of the search space. This condition actually contains the law that the author summarized for EP in the previous chapter, and it is more demanding.

Condition 3: The variation of each control variable should be correlated with each other to ensure rotational Invariance.

For Condition 1, if the center is not 0, the subgroup will have a fixed cumulative drift relative to the parent group, thus weakening the global search ability of the algorithm. The Gaussian distribution, the Cauchy distribution, or the more general Levy distribution used by EP and ES mentioned in the previous chapter all satisfy this condition. In Equation (1), since the selection of r1 and r2 is random, the probability of $U_{r1,j}[k] - u_{r2,j}[k]$ and $U_{r2,j}[k] - u_{r1,j}[k]$ appearing is the same, so the $F(U_{r1,j}[k] - u_{r2,j}[k])$ term also conforms to the random distribution centered on 0.

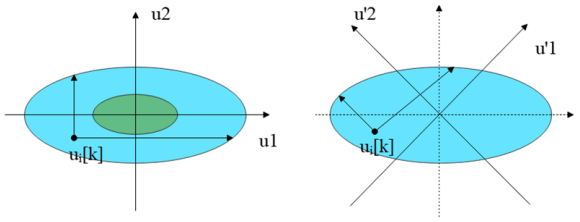


Figure 3: Illustration of Condition 2 and Condition 3

For Conditions 2 and 3, it can be understood with reference to Figure 3, which assumes that there are only two control variables, and the shape of the contour of the objective function is an ellipse. The left picture shows the case where the directions of the major and minor axes of the ellipse are the same as the coordinate axes. From this, it can be seen that at the same time, the effective variation range of different control variables is different, and with the progress of the optimization process, the effective variation range of each control variable is generally smaller and smaller, so as to adapt to the situation that the contour of the objective function changes from a large ellipse to a small ellipse.

Figure 4 shows the distribution of a population with four individuals (left panel) and its corresponding distribution of difference vectors (right panel, scale factor F=1). It can be seen that the distribution of the difference vector is a symmetrical distribution centered at 0, and its shape changes dynamically with the change of the population distribution. With the assistance of the selection operation, the shape of the search space can be gradually adapted, and only a scalar parameter F is needed to adjust the variation amplitude in each direction. Moreover, since the coordinate rotation will not change the relative position between individuals and the relative position of the entire population in the search space, it will not affect the performance of DE.

However, the aforementioned discrete hybridization would destroy the rotational independence. Figure 5

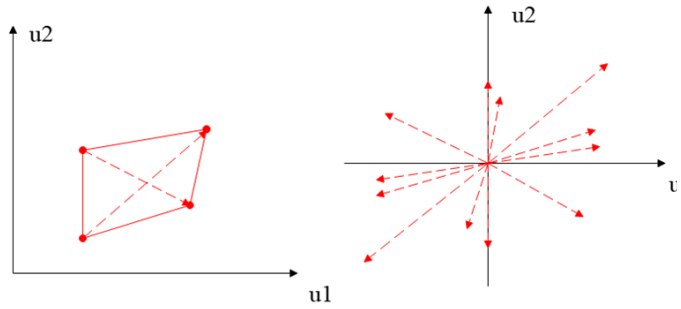


Figure 4: Schematic diagram of difference vector distribution

shows the hybridization of two individuals u and u in two dimensions. The position of the discrete hybridization result ($uXuu'$) will change with the rotation of the coordinate axis. However, if continuous linear hybridization is used like some ES methods, that is, the result of hybridization is on the connecting line of u and u, it will not be affected by the rotation of the coordinate axis.

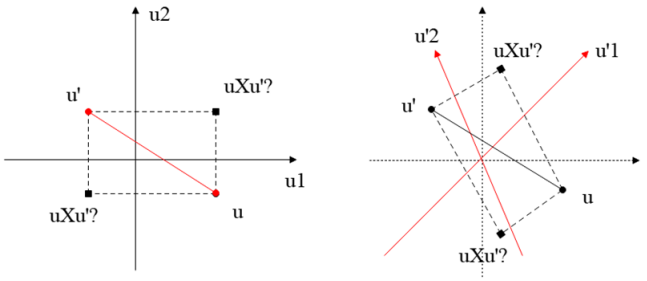


Figure 5: Schematic diagram of discrete hybridization and linear hybridization

Therefore, DE is more inclined to use linear hybridization, and the hybridization and mutation are directly combined together to form a concise and complete DE form, such as DE/current-to-rand/1 and DE shown in Equations (10) and (11). /current-to-best/1, where DE/current-to-rand/1 is the preferred form of DE recommended by K.V.Price.

DE/*current - to - rand*/1:

$$u'_{i,j}[k] = u_{i,j}[k] + K(u_{r1,j} - u_{i,j}[k]) + F(u_{r2,j}[k] - u_{r3,j}[k]) \tag{10}$$

DE/*current - to - best*/1:

$$u'_{i,j}[k] = u_{i,j}[k] + K(u_{best,j} - u_{i,j}[k]) + F(u_{r1,j}[k] - u_{r2,j}[k]) \tag{11}$$

Among them, K is called the hybridization control coefficient; u_{best} is the best individual that has appeared until the current generation, which may be called "the best individual in history".

The optimization mechanism of DE/current-to-best/1 can be illustrated by Figure 6: The function of the

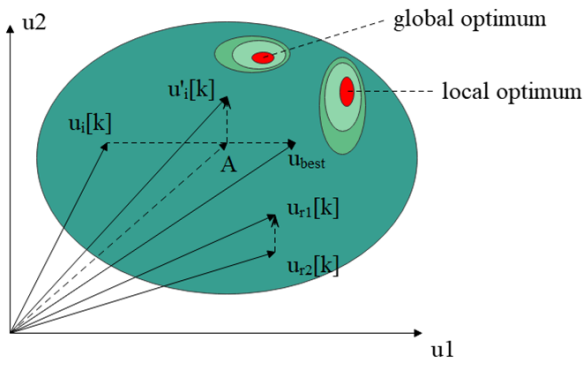


Figure 6: Schematic diagram of the mechanism of DE/current-to-best/1

$K(u_{best,j} - u_{i,j}[k])$ term is to "accelerate" each individual u_{best} (a search point) in the contemporary group towards u_{best} to a point A halfway (the case of $0 \leq K \leq 1$). The function of the $F(u_{r2,j}[k] - u_{r3,j}[k])$ term is to impose a perturbation on the search point at A, and the direction and intensity of the perturbation are determined by the difference between F and the two individuals u_{r1} and u_{r2} randomly selected from the current population, and finally a new individual u_i is obtained.

The flow chart of reactive power optimization based on DE/current-to-best/1 is shown in Figure 7.

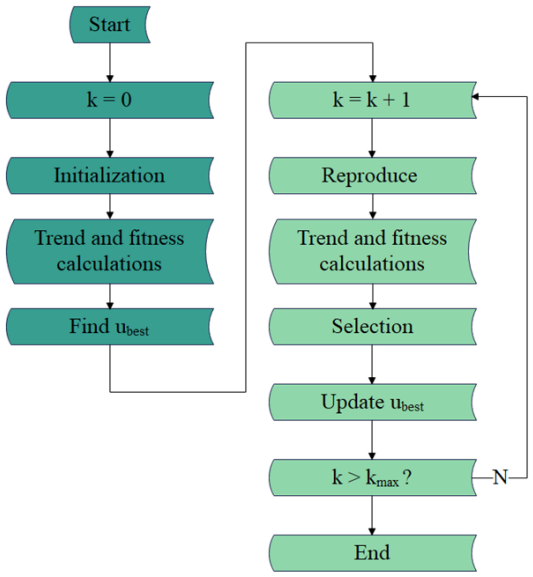


Figure 7: Flowchart of reactive power optimization based on DE/current-to-best/1

In the 0-th generation, the initialization is performed first to generate an initial group containing N individuals, then the power flow and fitness are calculated for each individual, and the best individual is found, which is the initial u_{best} .

In the k-th generation ($1 = k = k_{\max}$), firstly, the existing group (parent group) is reproduced according to Equation (11) to generate a sub-group, then the trend and fitness are calculated for each individual in the subgroup. Then, the individuals with the same numbers in the subgroup and the parent group conduct "one-to-one father-son competition", and the winner enters the next generation of the parent group, and the best contemporary individual $u_{best}[k]$ is found. If $u_{best}[k]$ is better than the historical best individual u_{best}, it is replaced by u_{best}.

2.3 Complementarity of Differential Evolution and Evolutionary Programming Algorithms

Regarding the reason why DE is prone to fall into premature convergence, it is believed that because DE adopts a relatively "greedy" selection operation of "one-to-one father-son competition", which leads to a rapid decline in population diversity and premature convergence. It adopts the selection operation of "one-to-one father-son competition", or adopts the relatively less "greedy" q-stakes selection like EP, or adopts the "greedier" deterministic selection operation like ES, and the performance shown by DE are not much different. It can be seen that the selection operation of "one-to-one father-son competition" is not the main reason.

The main reason is that there is no truly random mutation operation like in EP in the reproduction of DE, and the article also holds this view. For the convenience of explanation, the reproduction formulas of DE and EP are rewritten as follows:

DE/current-to-best/1 is:

$$u'_{i,j}[k] = u_{i,j}[k] + K(u_{best,j} - u_{i,j}[k]) + F(u_{r1,j}[k] - u_{r2,j}[k]) \quad (12)$$

$$EP : u_{i,j}[k] = u_{i,j}[k] + \rho_{i,j}[k]\dot{N}_{i,j}(0,1)[k] \quad (13)$$

Due to the mutation operation of DE, the last term in Equation (12), it consists of the difference between two individuals randomly selected from the current population. Therefore, the variation direction of each individual is limited. Because of the population size, there are (N-1) (N-2) mutation directions in total. The limited variation direction may affect the global search ability of DE in solving complex problems, and the population will be rapidly homogenized and fall into premature maturity. The expansion of the population size can not only expand the distribution of sampling, but also increase the variation direction of each individual, which is obviously beneficial to preventing precocious puberty, but it will increase the calculation time.

2.4 Design of a Hybrid Algorithm of Differential Evolution and Evolutionary Programming

For reactive power optimization, DE is a relatively excellent new evolutionary algorithm that deserves further research and application. However, it was also found that DE requires a relatively large population size to avoid premature convergence. This will result in a long calculation time, which cannot meet the needs of online reactive power optimization. By adopting parallel DE and appropriately sized clusters, online reactive power optimization of the power system can be effectively achieved; However, it is still necessary to find ways to reduce the required group size for DE to further accelerate computation or use smaller clusters to reduce costs.

A simple and effective hybrid solution of DE and EP was found and named DEEP. The scheme adopts the mechanism of main group plus auxiliary group. It refers to the group originally used in the evolution of DE as the main group, and the group size is N. Before reproduction in each generation, based on the first L (1=L=N) individuals in the main population, an auxiliary population of size L is generated according to an EP-type random mutation operation. The situation at the k-th generation is shown in Figure 8.

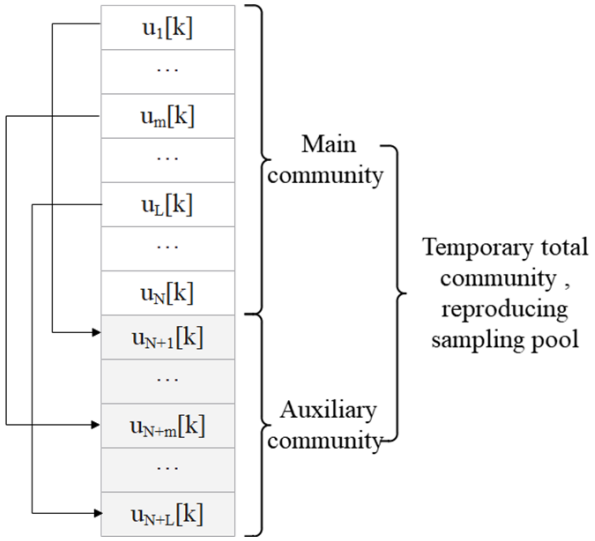


Figure 8: Schematic diagram of the generation of helper populations in DEPP (k-th generation)

After trial and error, it is found that the simple mutation operation shown in Equation (14) works well:

$$u_{N+m,j}[k] = u_{m,j} + U_{m,j}(0,1)\frac{f_{best} - f_m[k]}{f_{best} - f_{worst}}(u_j^{max} - u_j^{min}) \quad (14)$$

Among them, 1=m=L, 1=j=M, and M is the number of control variables. $U_{m,j}(0,1)$ generates a uniformly distributed random number located in the (0,1) interval each time; $f_m[k]$ is the fitness value of the m-th individual $u_m[k]$; f_{best} is the fitness value of the best individual u_{best} in history. f_{worst} is the fitness value of the worst individual u_{worst} in history; $(f_{best} - f_m[k])/(f_{best} - f_{worst})$ term provides an adaptive adjustment mechanism to the variation range of u_m, $[k]$, and its value gradually changes from large to small with the progress of the evolution process.

In each generation, the worse individuals have a larger variation range, and the better individuals have a smaller variation range. The $(u_j^{max} - u_j^{min})$ term is used to limit the variation range of the j-th control variable of each individual, and the value after guiding the variation should fall within $[u_j^{min}, u_j^{max}]$ as much as possible.

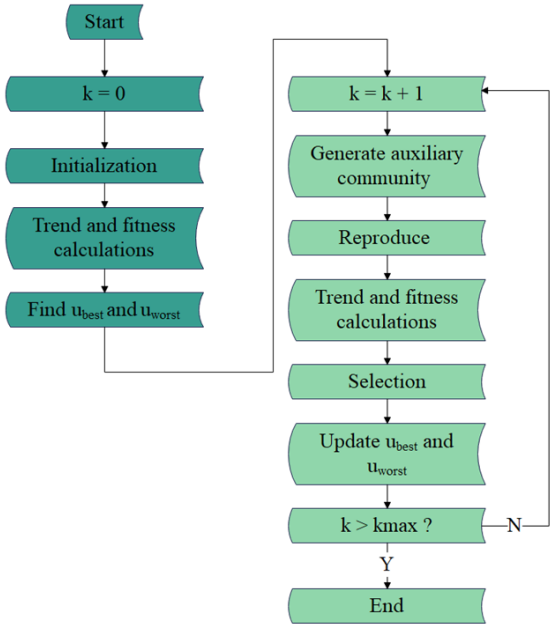


Figure 9: Flowchart of reactive power optimization based on DEEP

Since the hybrid scheme is very simple, the calculation process of DEEP differs little from that of DE. Figure 9 is a flow chart of DEEP-based reactive power optimization.

In the 0-th generation, initialization is performed first to generate an initial population containing N individuals. Then, the trend and fitness are calculated for each individual. Then, the best individual u_{best} and the worst individual u_{worst} are found, that is, the initial historical best individual and the historical worst individual. The initial population is used as the main parent population in generation 1.

In the k-th generation $(1 = k = k_{\max}())$, based on the first L individuals in the main parent group, an auxiliary parent group of size L is generated according to formula (14). Then, according to the above description, a subgroup of size N is generated based on the temporary general parent group of size N+L synthesized by the main and auxiliary parent groups. Then, the power flow and fitness are calculated for each individual in the subgroup. Then, the individuals with the same number in the subgroup and the main parent group conduct "one-to-one father-son competition", and the winner enters the main parent group of the next generation. The contemporary best individual $u_{best}[k]$ and the contemporary worst individual $u_{worst}[k]$ are also found from the winners, if $u_{best}[k]$ is better than the historical best individual u_{best}, it is replaced by u_{best}. If a new parameter appears in the $u_{worst}[k]$ ratio history DEEP, that is, the size of the auxiliary population L, it needs to be determined through experiments. More generally, the determination of L may be replaced by the determination of the ratio of L to N, L/N.

DE and EP can be described as follows: DE is better at mining existing genetic information in the population, but not at introducing new genetic information, that is, exploring new fields in the search space, Therefore, the convergence speed is fast but it is easy to fall into premature puberty; EP, on the other hand, constantly introduces new genetic information without fully utilizing it, making it less likely to fall into early puberty but with slow convergence. It can be clearly seen that DE and EP have good complementarity and can be combined to construct a hybrid algorithm that leverages strengths and avoids weaknesses

2.5 Grid Partitioning Based on Reactive Voltage Sensitivity and Clustering Technology

In order to realize grid partitioning, the electrical distance between nodes is firstly defined based on the reactive power and voltage sensitivity between nodes. The definition is shown in Equation (15):

$$d_{ij} = -\log(a_{ij}a_{ji} \tag{15}$$

$$a_{ij} = \frac{\partial V_i}{\partial Q_i} / \frac{\partial V_j}{\partial Q_j} \tag{16}$$

Among them, dij represents the electrical between nodes i and j; Vi and Vj are the voltages at nodes i and j, respectively; Qj is the reactive power injection at node j.

Commonly used interval distances are defined as maximum distance, minimum distance, average distance and Ward distance, as shown in Equations (17) to (20), respectively.

The maximum distance is:

$$d_{max}(C_i, C_j) = max_{b\in C_i, b\in C_j} d(b, b') \tag{17}$$

The minimum distance is:

$$d_{min}(C_i, C_j) = min_{b\in C_i, b\in C_j} d(b, b') \tag{18}$$

The average distance is:

$$d_{avr}(C_i, C_j) = \frac{1}{k_i, k_j} \sum_{b\in C_i} \sum_{b\in C_j} d(b, b') \tag{19}$$

Ward distance is:

$$d_{ward}(C_i, C_j) = \frac{k_i k_j}{k_i + k_j} d(o_i, o_j) \tag{20}$$

Among them, C_i and C_j represent the partitions numbered i and j; $d()$ represents the distance; b represents the node; k_i and k_j represent the number of nodes in partitions C_i and C_j ; o_i and o_j represent the "central

nodes" of partitions and . The central node of a partition is defined as the node with the smallest sum of squared distances from all other nodes in the same region.

The clustering validity index (CVI) is an index used to evaluate the quality of the clustering scheme (here, the partition scheme), and there are many definitions of CVI. Among them, the four definitions of Global Silhouette Index(GSI), Davis-Bouldin Index(DBI), Dunn's Index(DI) and C Index(CI) have clear and relatively simple physical meanings, which are introduced as follows.

2.5.1 Global Silhouette Index (GSI)

$$GSI = \frac{1}{L}\sum_{j=1}^{L} S(C_j)$$

$$S(C_j) = \frac{1}{K_j}\sum_{i=1}^{k_j} S(b_i)$$

$$S(b_i) = \frac{y(b_i)-x(b_i)}{\max\{x(b_i),y(b_i)\}} \quad (21)$$

$$x(b_i) = \frac{1}{KJ}\sum_{b_i,b\in c_j} d(b_i,b_l)$$

$$y(b_i) = min_{j\neq j\in\{1,2,L,L_j\}}(\frac{1}{k_j}\sum_{b_i\in C_j}^{b_i\in C_j} d(b_i,b_j))$$

Among them, $x_(b_i)$ represents the average distance between the i-th node b_i in the partition C_j and other nodes in the same region. $S(b_i)$ is called the "Silhouette Width" of , and is used to represent the "Confidence Index" that b_i belongs to $c_j\dot{S}(b_i)$. is close to 1, indicating that b_i has been assigned to the correct partition. $S(b_i)$ is close to 0, indicating that b_i can also be classified into adjacent partitions; When $S(b_i)$ is close to -1, it means that b_i is classified into the wrong partition. L is the number of partitions, and the GSI indicator is the average Silhouette width of the entire partition scheme. The larger the GSI, the better the partitioning scheme.

2.5.2 Davis-Bouldin Index (DBI)

$$DBI = \frac{1}{L}\sum_{i=1}^{L} min_{j\neq i\in\{1,2,L,L\}} D_{i,j}$$

$$D_{ij} = \frac{y(C_i,C_j)}{x(C_i)+x(C_j)} \quad (22)$$

$$x(C_i) = \frac{1}{k_i}\sum_{b_l\in C_i} d(b_l,o_i)$$

Among them, $x(C_i)$ represents the average distance between all nodes in partition C_i and the central node of the partition; $y(C_i,C_j)$ represents the distance between the central nodes of partitions C_i and C_j. The larger the DBI, the more compact the nodes in the same partition, and the more scattered the centers of different partitions, that is, the better the clustering results.

2.5.3 Dunni Index (DI)

Among them, D_{min} represents the minimum distance between two nodes belonging to different partitions, and D_{max} represents the maximum distance between two nodes belonging to the same partition. The larger the DI, the better the clustering result.

$$DI = D_{min}/D_{max}$$

$$D_{min} = min_{b_l\in C_c,b_{12}\in C_j,i\neq j\in\{1,2,L,L_j\}} d(b_{l1},b_{l2} \quad (23)$$

$$D_{max} = max_{b_l,h_2\in C_i,i\in\{1,2,\perp,L\}} d(b_{l1},b_{l2})$$

2.5.4 C Index (CI)

Among them, S represents the sum of the distances between all nodes in the same area of the L partitions; there are N_d such distances in total, and S_{min} and S_{max} respectively represent the sum of the N_d minimum and maximum distances among the distances between all N_b nodes in the entire power grid $N_b * (N_b - 1)/2$total. Likewise, the larger the CI, the better the clustering result.

$$\min F = \sum_{j=1}^{L}\sum_{b_i\in C_j} d^2(b_i,o_j) \quad (24)$$

Among them, the meaning of each symbol is the same as the previous one, which is reiterated as follows: L is the number of partitions, b_i is the i-th node in partition c_j, o_j, is the central node of partition C_j, and $d(b_j,o_j)$ represents the electrical distance between b_i and o_j.

3 Distributed Parallel Algorithm for Finite Element Multi-computer System Considering Network Security Performance Evaluation

The function of the microcomputer measurement and control protection is to collect the power information of the line and monitor the running status of the system in real time. When a fault is detected, it immediately performs a protection operation to trip the circuit breaker, or quickly uploads the fault information to the upper computer to achieve fault location, and executes the control commands sent by the upper computer. The system block diagram of the microcomputer measurement and control protection device designed in this paper is shown in Figure 10. In order to verify the correctness of the distributed parallel algorithm, input a 100V, 50Hz square wave to analyze the distributed parallel algorithm, and use the DA module to output the detection results of the fundamental wave instantaneous value. At the same time, the FFT function of the oscilloscope is used to detect the effective value of the fundamental wave of the input signal and compare it with the calculation result of the distributed parallel algorithm, as shown in Figure 11.

Figure 11(a) shows the waveform of the instantaneous value of the fundamental wave calculated by the input square wave and the distributed parallel algorithm. The

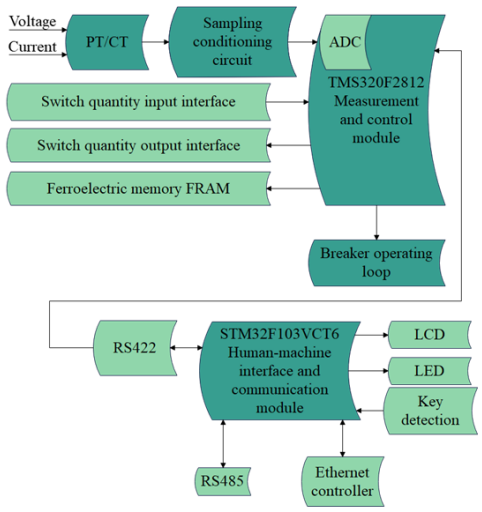


Figure 10: System structure of microcomputer protection device

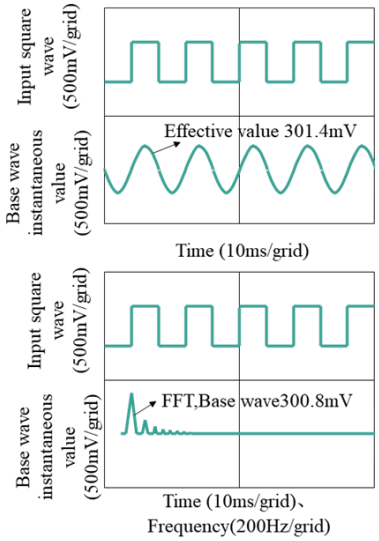


Figure 11: Experimental results of distributed parallel algorithm for detecting square waves

oscilloscope shows that the effective value of the fundamental wave is 301.4mV (attenuated by 300 times). (b) is the result of square wave analysis using the oscilloscope FFT function, and the effective value of the fundamental wave is 300.8V. The two are approximately equal within the allowable error range, indicating that the distributed parallel algorithm correctly detects the fundamental wave of the input signal.

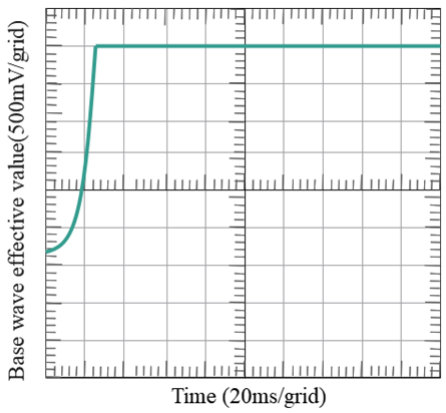
Sinusoidal signals with a voltage of 100V and frequencies of 50Hz, 49.5Hz, and 50.5Hz are input, and the DA module is used to output the fundamental RMS calculation result of the distributed parallel algorithm. The experimental waveform is shown in Figure 12. When using the further method, the system poles are outside the unit circle. Obviously, the RMS calculation result of the distributed parallel algorithm gradually diverges from 100V at this time, and finally is limited because it exceeds the range of DA representation. During simulation, the limit is limited because it exceeds the fixed-point representation range of Q15, and the limit value here is smaller than that during simulation. The divergence mode is different at different frequencies, but the overall increase is gradually increased, and the distributed parallel algorithm is unstable.

Table 1: The security improvement of the distributed algorithm for the finite element multi-computer system

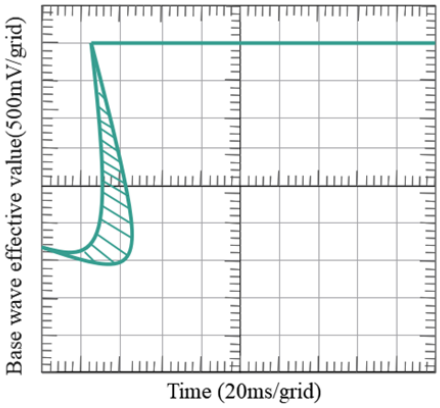
No.	Safety	No	Safety	No	Safety
1	91.117	11	89.725	21	86.174
2	87.075	12	88.076	22	89.049
3	87.538	13	90.386	23	85.512
4	84.366	14	90.339	24	91.662
5	88.596	15	89.399	25	89.533
6	91.908	16	85.001	26	88.239
7	88.716	17	88.991	27	87.546
8	87.759	18	85.247	28	89.188
9	86.625	19	91.337	29	87.604
10	91.994	20	86.513	30	90.129

Through the above research, the effectiveness of the distributed parallel algorithm proposed in this paper is verified. On this basis, the effect of the distributed algorithm in the security improvement of the finite element multi-computer system is carried out, and 30 sets of data are tested, and the results shown in Table 1 below are obtained. To further verify the effectiveness of the model proposed in this paper, the same experimental method was used for simulation. The model was compared with the method proposed in Reference [10], and their algorithm performance and network security were compared. The comparison results are shown in Table 2 and Table 3 below

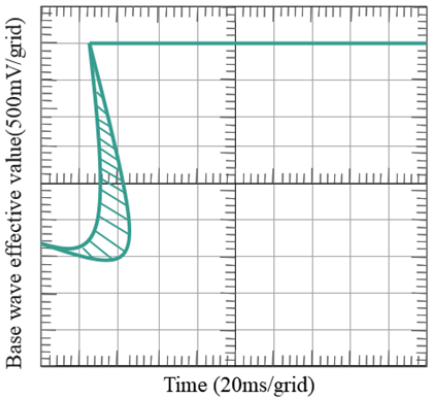
It can be seen from the above research that the distributed parallel algorithm proposed in this paper has a certain promotion effect on the security improvement of



(a) 20ms/grid



(b) 49.6Hz



(c) 50Hz

Figure 12: The experimental results of the distributed parallel algorithm to calculate the fundamental wave RMS at different frequencies

Table 2: Comparison of Algorithm Performance

	The Proposed Method	**Reference [10]**
1	84.72	79.41
2	85.67	79.27
3	83.94	79.04
4	85.61	81.63
5	83.85	82.20
6	84.47	79.24
7	81.70	78.39
8	88.99	78.38
9	83.54	82.19
10	81.98	75.07
11	86.18	75.56
12	84.24	82.99
13	82.23	81.18
14	82.88	81.10
15	83.85	81.59

Table 3: Network security comparison

	The Proposed Method	**Reference [10]**
1	93.93	81.68
2	93.45	80.96
3	91.49	78.00
4	87.56	76.85
5	93.04	80.57
6	90.16	82.07
7	87.24	75.86
8	91.08	77.80
9	92.32	78.74
10	92.09	75.58
11	90.50	79.60
12	93.58	78.60
13	87.41	81.00
14	91.36	78.17
15	91.63	82.58

the finite element multi-computer system.

4 Conclusions

In order to test the safety communication program of the computerized automatic inter-station blocking system, a computerized automatic inter-station blocking system safety communication simulation subsystem is designed to analyze the inter-station communication messages and simulate the inter-station communication fault setting. The microcomputer monitoring system can grasp the working status and changing trend of the signal equipment, which is the technical basis for the implementation of the status repair of the signal equipment, and provides a scientific basis for the maintenance decision. The computer monitoring system uses computers and information acquisition machines to detect various signal devices in real time. This paper proposes a distributed parallel algorithm for finite element multi-computer system considering network security performance evaluation, and combines the experimental analysis to verify the algorithm effect. The experimental analysis shows that the distributed parallel algorithm proposed in this paper is effective, and it verifies that the distributed parallel algorithm proposed in this paper has a certain promotion effect on the security improvement of the finite element multi-computer system.

The subsequent research work is to combine DE with interior point method, complement each other's strengths and weaknesses, and construct a universal and effective hybrid algorithm suitable for solving reactive power optimization problems.

References

[1] A. Al-Halafi and Shihada B. Uhd, "video transmission over bidirectional underwater wireless optical communication," *IEEE Photonics Journal*, vol. 10, no. 2, pp. 1–14, 2018.

[2] Nourah Almrezeq, Mamoona Humayun, Madallah Alruwaili, Saad Alanazi, and NZ Jhanjhi, "Cyber security attacks and challenges in saudi arabia during covid-19," *International Journal of Computer Science & Network Security*, vol. 23, no. 10, pp. 179–187, 2023.

[3] Neda Azizi and Omid Haass. "Cybersecurity issues and challenges,". in *Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications*, pp. 21–48. IGI Global, 2023.

[4] J. Barowski, M. Zimmermanns, and I. Rolfes, "Millimeter-wave characterization of dielectric materials using calibrated fmcw transceivers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 8, pp. 3683–3689, 2018.

[5] B. Behroozpour, "Sandborn p a m, wu m c, et al. lidar system architectures and circuits," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 135–142, 2017.

[6] Q. Y. Cheng, X. L. Zhao, Y. X. Weng, et al., "Fully sustainable, nanoparticle-free, fluorine-free, and robust superhydrophobic cotton fabric fabricated via an eco-friendly method for efficient oil/water separation," *ACS Sustainable Chemistry & Engineering*, vol. 7, no. 18, pp. 15696–15705, 2019.

[7] Y. Jiang, S. Karpf, and B. Jalali, "Time-stretch lidar as a spectrally scanned time-of-flight ranging camera," *Nature photonics*, vol. 14, no. 1, pp. 14–18, 2020.

[8] H. C. Kumawat and A. B. Raj, "Extraction of doppler signature of micro-to-macro rotations/motions using continuous wave radar-assisted measurement system," *IET Science, Measurement & Technology*, vol. 14, no. 7, pp. 772–785, 2020.

[9] N. Maring, P. Farrera, K. Kutluer, et al., "Photonic quantum state transfer between a cold atomic gas and a crystal," *Nature*, vol. 551, no. 7681, pp. 485–488, 2017.

[10] Z. Meng, J. Li, C. Yin, et al., "Dual-band dechirping lfmcw radar receiver with high image rejection using microwave photonic i/q mixer," *Optics express*, vol. 25, no. 18, pp. 22055–22065, 2017.

[11] H. Mohapatra and Rath A. K. Detection and, "and avoidance of water loss through municipality taps in india by using smart taps and ict," *IET wireless sensor systems*, vol. 9, no. 6, pp. 447–457, 2019.

[12] Z. Sabouri, A. Akbari, H. A. Hosseini, et al., "Eco-friendly biosynthesis of nickel oxide nanoparticles mediated by okra plant extract and investigation of their photocatalytic, magnetic, cytotoxicity, and antibacterial properties," *Journal of Cluster Science*, vol. 30, no. 6, pp. 1425–1434, 2019.

[13] A. Seri, G. Corrielli, D. Lago-Rivera, et al., "Laser-written integrated platform for quantum storage of heralded single photons," *Optica*, vol. 5, no. 8, pp. 934–941, 2018.

[14] Imdad Ali Shah, NZ Jhanjhi, and Areeba Laraib. "Cybersecurity and blockchain usage in contemporary business,". in *Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications*, pp. 49–64. IGI Global, 2023.

[15] L. J. Xu, X. Lin, Q. He, et al., "Highly efficient eco-friendly x-ray scintillators based on an organic manganese halide," *Nature communications*, vol. 11, no. 1, pp. 1–7, 2020.

[16] F. Zhang, Q. Guo, and S. Pan, "Photonics-based real-time ultra-high-range-resolution radar with broadband signal generation and processing," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

[17] T. Zhong and P. Goldner, "Emerging rare-earth doped material platforms for quantum nanophotonics," *Nanophotonics*, vol. 8, no. 11, pp. 2003–2015, 2019.

[18] T. Zhong, J. M. Kindem, J. G. Bartholomew, et al., "Nanophotonic rare-earth quantum memory with

optically controlled retrieval," *Science*, vol. 357, no. 6358, pp. 1392–1395, 2017.

Biography

Yi Li was born in Henan, China, in 1980. From 1999 to 2003, she studied in Zhengzhou University of Light Industry and received her bachelor's degree in 2003. From 2006 to 2009, she studied in Huazhong University of Science and Technology and received her Master's degree. Currently, she worked in Henan Judicial Police Vocational College, Zhengzhou. Her research fields include computer education, Big data visualization and application and digital graphic image processing.

Multi-user Keyword Searchable Signcryption Scheme in Heterogeneous 5G Network Slicings

Ming Luo[1,3], Qibang Zhan[3], Minrong Qiu[2], and Li Cen[1]
(Corresponding author: Ming Luo)
Post-doctoral Scientific Research Workstation of Aheadsoft Software Co., Ltd, Nanchang, China[1]
GongQing Institute of Science and Technology, Gongqingcheng, China[2]
School of Software, Nanchang University, Nanchang, China [3] Email: lmhappy21@163.com
(Received July 29, 2023; Revised and Accepted Mar. 4, 2024; First Online June 22, 2024)

Abstract

The technology of 5G network slicings (5GNS) offers convenience for storing and sharing data in the cloud. To ensure data security and availability, the cloud data necessitates a signcryption scheme with keyword search. Furthermore, diverse 5GNS typically incorporate distinct cryptographic systems. Nevertheless, there are limited heterogeneous signcryption schemes that enable multiple users to conduct keyword searches. To address this issue, this paper introduces a multi-user keyword searchable signcryption (MUKSS) scheme within a heterogeneous 5GNS environment. In the MUKSS scheme, a data owner in a certificateless cryptographic system (CLC) can signcrypt the keyword for multiple users using an identity-based cryptographic system (IBC). Furthermore, a rigorous security analysis demonstrates that the MUKSS scheme ensures confidentiality, integrity, and unforgeability. Moreover, the MUKSS scheme is demonstrated to be resistant to keyword guessing attacks. Based on the efficiency analysis, the MUKSS shceme proves to be better suited for the heterogeneous 5GNS environment when compared to the seven most recent schemes.

Keywords: 5G Network Slicings; Heterogeneous; Keyword Searchable; Signcryption

1 Introduction

As the fifth generation (5G) communication technology, 5G is an upgrade and improvement of 4G (fourth generation) technology, which aims to provide higher data transmission speed, lower latency and greater network capacity. In order to promote the development of 5G network technology, various countries and regions are promoting the research of 5G network technology. In 2012, the European Commission funded the METIS (Mobile and Wireless Communications Enables for the 2020 Information Society) project [1]. In 2013, China established the IMT-2020 (5G) promotion Group [2]. In 2014, Japan established the 5th generation mobile communication promotion Forum 5GMF [3]. In 2021, the United States signed a bill to increase investment and formulate policies to promote the further development of 5G technology.With the rapid development of 5G technology in recent years,5G has been applied to Internet of Vehicles (IoV) [4], Industrial Internet of Things (IIoT) [5],Virtual Reality (VR), Cloud Storage [6] and other industries. Furthermore, it introduces several innovative technologies, among which 5G Network Slicing (5GNS) is recognized as a pivotal component of 5G networks.

5GNS enables the partitioning and optimization of a single network into multiple independent logical networks, facilitating the provision of diverse services, including data sharing. In scenarios where data needs to be shared among multiple users, it is commonly uploaded to a cloud server [7]. However, there are frequent instances in which cloud data leaks occur, resulting in significant losses for both individuals and enterprises [8]. To address the challenge, Elkhalil *et al.* [9] designs a signcryption scheme to the data before it is transmitted to the cloud server, which enhances the security of cloud data. The introduction of signcryption ensures confidentiality, unforgeability, and integrity while reducing the computation cost.

While Elkhalil *et al.*'s scheme guarantees the confidentiality and unforgeability of cloud data, users have encountered difficulties in searching for the signcrypted data. The rapid development of cloud data sharing applications has led to a significant increase in the demand for encrypted or signcrypted data search. Therefore, there is an urgent need to design a cryptographic primitive that incorporates search functionality. In response to this demand, Boneh *et al.* [10] proposed a public key encryption scheme with keyword search (PEKS), which enables the searching of encrypted keywords without the need for decryption. Setting the keyword appropriately is crucial as it allows the server to match the corresponding data and enhance accuracy. Consequently, several homogeneous PEKS schemes have been developed [11–18]. Additionally, Omala *et al.* [19] constructed a heterogeneous signcryption scheme with keyword search designed specifically

for a single user. However, in Omala *et al.*'s scheme, the data owner faces limitations in distributing the data to multiple users, thus reducing its practicality. Furthermore, the cloud-assisted 5GNS environment requires critical considerations for reducing computing costs and enhancing data security.

Motivated by the need to address this problem, this paper presents a designed multi-user keyword searchable signcryption (MUKSS) scheme specifically tailored for the heterogeneous 5GNS environment. In summary, our MUKSS scheme offers the following contributions:

1) For the first time, the MUKSS scheme implements heterogeneous signcryption with keyword search, transitioning from CLC to IBC. Furthermore, the MUKSS scheme supports multiple users to perform keyword search by uploading a user list to the cloud server.

2) The analysis demonstrates that the MUKSS scheme achieves confidentiality, integrity, and unforgeability, all under the discrete logarithm problem and computational Diffie-Hellman problem in the random oracle model. Additionally, the MUKSS scheme provides security against keyword guessing attacks.

3) Based on the efficiency analysis, the MUKSS scheme has been found to be more suitable for the heterogeneous 5GNS environment compared to [11, 15, 17–19, 33, 34].

1.1 Related Work

The PEKS scheme was initially introduced by Boneh *et al.* [10]. The core concept of PEKS revolves around enabling a server to search for encrypted data using an encrypted keyword and trapdoor. Byun *et al.* [20] point out that Boneh's scheme is insecure, as it allows both internal and external individuals to retrieve information related to specific keywords from intercepted query messages. Abdalla *et al.* [21]. showed that robust PEKS schemes should satisfy consistency. Jeong *et al.* [22] pointed out that it is impossible to construct a secure and consistent PEKS scheme that is resistant to keyword guessing attacks when the number of possible keywords is bounded by some polynomial. Subsequently, Lu [16] designed a PEKS scheme that provides protection against adaptively-chosen-target adversaries. A PEKS scheme was presented by Pan [11], which incorporates multi-ciphertext and trapdoor indistinguishability. To address the challenge of certificate management in the public key infrastructure (PKI), researchers have developed multiple PEKS schemes based on identity-based cryptographic systems (IBC) [14,15,23]. In the context of identity-based cryptographic systems (IBC), the private key of an entity is generated by the private key generator (PKG), which subsequently retains the private keys of all users. However, if the PKG becomes malicious, the security of the data will be compromised. The solution to this problem is to combine PEKS with certificateless encryption system (CLC) [24–26]. In the context of certificateless cryptographic systems (CLC), a user actively participates in the selection of a portion of their private key, while the key generation center (KGC) generates another complementary part. An important advantage of CLC is that even in the event of the KGC becoming malicious, the security of the data remains intact. This is because the KGC does not possess a portion of the user's private key, ensuring the data's confidentiality and integrity. Additionally, Liu [27] proposed a searchable attribute-based signcryption scheme specifically designed for securing personal electronic health records. Eltayieb *et al.* [28] introduced a PKKS construction that incorporates decryption verification using attributed-based encryption.

The concept of signcryption, a cryptographic primitive that guarantees confidentiality, unforgeability, and integrity while offering low computation cost, was introduced by Zheng [29]. Liu [30] designde a hybrid blockchain-enabled scheme for searchable proxy signcryption. Varri [31] formulated an attributed-based signcryption scheme with keyword search. The rapid development of heterogeneous environments has created an urgent need for heterogeneous searchable signcryption schemes. Omala [19] present a heterogeneous searchable signcryption scheme that enables data owners in certificate-less cryptography (CLC) to share their data securely with trusted receivers in the public key infrastructure (PKI). The multi-user setting is practical for the PEKS scheme. Nair [12] formulated a multi-user searchable encryption scheme with access control. Olakanmi [18] designed a certificateless PEKS scheme with a multi-user setting in the fog-enhanced Industrial Internet of Things (IIoT) network. Yu *et al.* [32] presented a lattice-based encryption scheme with a multi-user setting.

1.2 Organization

In the following section, we introduce the model description of the MUKSS scheme along with preliminary information. The third section describes the concrete construction of the MUKSS scheme, while the fourth section presents the security analysis. In the fifth section, we provide an efficiency analysis of the MUKSS scheme. Finally, the paper concludes in the last section.

2 Preliminaries

In this section, we will introduce several preliminary concepts and components utilized in the scheme.

2.1 Keyword Guessing Attacks

Most PEKS schemes do not perform keyword selection in a keyword space that is super-polynomially large. Consequently, adversaries can attempt all possible keywords to determine which one is encapsulated within the trapdoor. These types of attacks are commonly known as keyword guessing attacks (KGA) [13].

2.2 Bilinear Pairing

Suppose G_1 is an additive cyclic group and G_2 is a multiplicative cyclic group with the same prime order q. The properties of a bilinear pairing $\hat{e} : G_1 \times G_1 \to G_2$ are as follows:

1) Bilinearity. For any $D, E \in G$ and $a, b \in Z_q^*$, $\hat{e}(aD, bE) = \hat{e}(D, E)^{ab}$

2) Non-degeneracy. There exists a $F \in G_1$ that $\hat{e}(F, F) \neq 1_{G_2}$,where 1_{G_2} is a generator of G_2.

3) Computability. Given any $D, E \in G_1$, $\hat{e}(D, E)$ can be calculated.

Definition 1. *Computational Diffie-Hellman problem (CDHP): Given a tuple (aP, bP),where $a, b \in Z_q^*$, the value abP is difficult to calculate.*

Definition 2. *Discrete Logarithm Problem (DLP): Given a tuple (P, aP), where $a \in Z_q^*$, it is difficult to obtain the value a.*

2.3 System Model

Figure 1 depicts the system model of the MUKSS scheme. The system model comprises four types of entities: a 5GNS data owner in CLC, multiple users, a cloud server in IBC, and a network manager (NM). It is important to note that NM fulfills the roles of both the Key Generation Center (KGC) in CLC and the Public Key Generator (PKG) in IBC.During the initialization phase, NM establishes the cryptographic parameters, and users register with NM by providing their identities. Subsequently, a 5GNS data owner in CLC creates a user list and uploads it, along with the associated data, to the cloud server in IBC. The data owner has the ability to add or remove users from the user list.Users in IBC generate their trapdoors based on the keyword and submit them to the cloud server. Once the cloud server searches the user list and verifies the correctness of the users' trapdoors, it grants access to the data for users with appropriate access rights.

2.4 Generic Model

The symbolic used of this article are listed in Table 1. The following ten algorithms are available in the generic MUKSS scheme:

1) *Setup*. KGC and PKG perform this setup algorithm. KGC selects its master private key t and generates its cryptographic parameters. Meanwhile, PKG generates its cryptographic parameters similarly.

2) $CL-PPKG$. Given the identity ID of a data owner in CLC, the owner's partial private key θ_{ID} and partial public key U_{ID} are computed by KGC.

3) $CL-PKG$. Given the cryptographic parameters, a data owner in CLC picks its secret value e_{ID} and outputs its public key PK_{ID}.

4) CUL. The data owner in CLC creates a user list that records the association between its data and multiple users, and then uploads it to the cloud server.

5) $IB-KG$. Given the identity ID of a user in IBC, PKG outputs the user's private key SK_{ID} and public key Q_{ID}.

6) SC. Given the public key Q_{ID} of a user in IBC, a data owner in CLC performs this algorithm to signcrypt a keyword w.

7) $Add-user$. The data owner performs this algorithm to grant access rights to multiple users.

8) $Remove-user$. A data owner in CLC performs this algorithm to revoke multiple users' access rights.

9) $Trapdoor$. Given a keyword w, each user generates its trapdoor TD_{ID}.

10) *Search*. Given the trapdoors and user list, if the trapdoors are right and multiple users have access rights, NM returns the data. Otherwise, NM returns $\perp$.

Table 1: Symbolic used

Symbol	Explanation
G_1	A cyclic addition group
G_2	A cyclic multiplication group
t	The KGC's master secrete key
P_{pub}	The KGC's public key
α	The security parameter chosen by KGC
θ	A sender's partial private key
U_{ID}	A sender's partial public key
e_{ID}	A sender's secret value
PK_{ID}	Another part of the sender's public key
SK_{ID}	A receiver's private key
Q_{ID}	A receiver's public key
w	A keyword
T_{ID}	A keyword trapdoor
$H_{i(i=1,2,3,4)}$	Four one-way hash functions
UL_{ID}	User mapping list
σ_i	A keyword ciphertext
C	A challenger
ID_s, ID_r	A sender's identity and a receiver's identity

2.5 Security Model

According to Definitions 3 and 4, the confidentiality and unforgeability are described respectively. Assume that C is the challenger. and $A_{i(i=1,2)}$ are the two kinds of adversaries. KGC's master private key cannot be obtained by A_1, but A_1 can replace the public key. Meanwhile, KGC's master private key can be obtained by A_2, but A_2 cannot replace the public key. The basic queries in Game 1 and Game 2 are defined as follows:

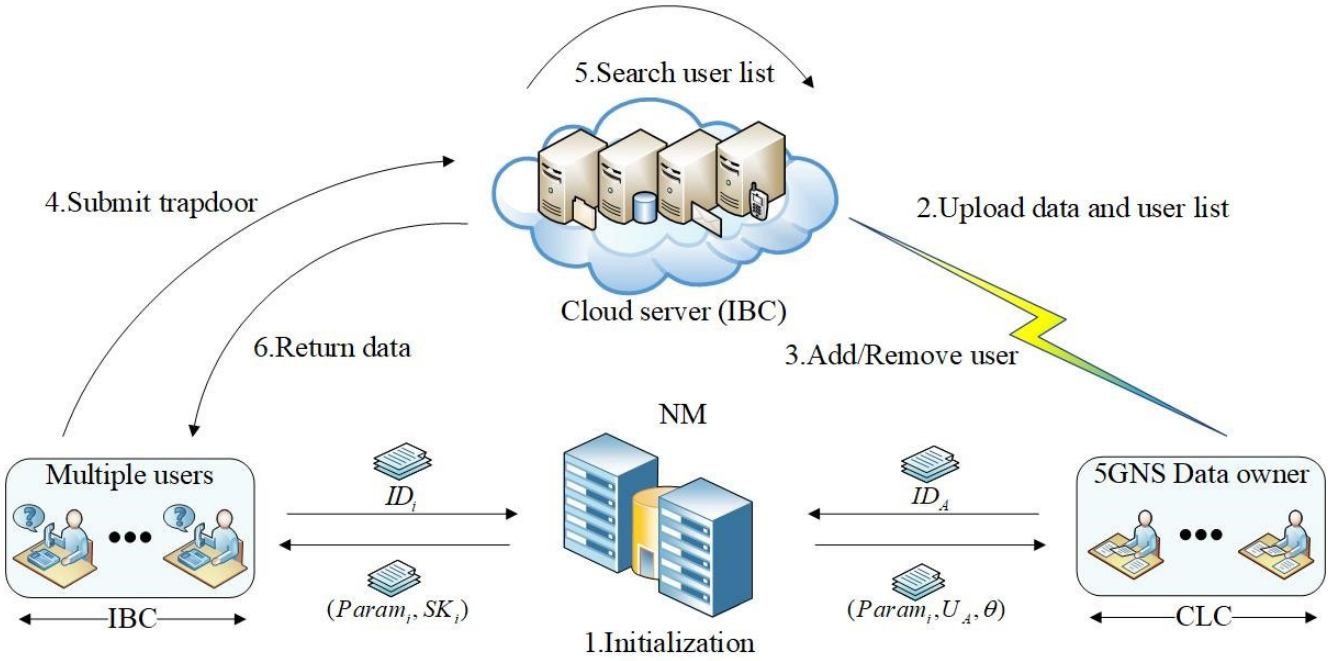


Figure 1: The system model of the MUKSS scheme

1) $CL-PK-queries$. Given the identity as input, the CL-PPG, CL-SVS and CL-PKG algorithms are performed by C to return the corresponding public key (U_{ID}, PK_{ID}).

2) $CL-SKE-queries$. Given the identity as input, the CL-SVS and CL-PPKG algorithms are performed by C to return the corresponding private key (e_{ID}, θ_{ID}).

3) $CL-RPK-queries$. Given the public key of ID as input, if the public key is valid, C replaces the related public key.

4) $IB-PK-queries$. Given the identity ID as input, C performs the IB-KG algorithm. The public key Q_{ID} will be returned.

5) $IB-SKE-queries$. Given the identity as input, C performs the IB-KG algorithm. The private key SK_{ID} will be returned.

6) $SC-queries$. Assumed that (ID_s, ID_r) is the identity of the sender and receiver, upon getting the keyword w and (ID_s, ID_r) as input, C performs the SC algorithm and returns the signcryption σ. The e_{ID} also needs to be updated when the corresponding PK_{ID} has been replaced.

7) $TD-queries$. Given the keyword w and ID as input, C executes the Trapdoor algorithm to generate the trapdoor $T_{w_{ID}}$.

8) $Search-queries$. Given a signcryption σ and trapdoor $T_{w_{ID}}$, C validates whether $T_{w_{ID}}$ is correct. If so, C outputs 1. Otherwise, C outputs $\perp$.

Definition 3. *(Confidentiality). If in Game 1, every polynomially bounded adversary A_1 only owns a negligible success advantage then the MUKSS scheme is indistinguishable against adaptive chosen keyword attacks (IND-MUKSS-CCA2).*

Game 1:

Initialization. First of all, the challenger C executes the Setup algorithm. Then the master private key and the cryptographic parameters are delivered to A_1.

Phase 1. A_1 probes C with CL-PK, CL-SKE, CL-RPK, IB-PK, IB-SKE, TD and Search queries.

Challenge. A_1 outputs the sender and receiver's identity (ID_s^*, ID_r^*). Besides, A_1 selects two keywords (w_0, w_1), which cannot be queried for the corresponding trapdoor. C picks a bit $\beta \in \{0, 1\}$, and returns a trapdoor $T_{w_\beta}^*$ with the keyword W_β.

Phase 2. After receiving the challenged trapdoor $T_{w_\beta}^*$, performs the same queries as phase 1. However, the TD queries to $T_{W_\beta}^*$ are not allowed.

Guess. A_1 tries to guess β'. A_1 's winning condition is $\beta' = \beta$.

Definition 4. *(Unforgeability). If in Game 2, every polynomially bounded adversary F only owns a negligible advantage then the MUKSS scheme is existentially unforgeable against adaptive chosen keyword attacks (EUF-MUKSS-CCA).*

Game 2:

Initialization. Setup algorithm is performed by Challenger C. Then the cryptographic parameters and master private key are delivered to F.

Probing. F makes the same queries as those in Game 1 except TD and Search queries. In addition, F is able to make SC queries.

Forgery. F outputs the sender and receiver's identity (ID_s^*, ID_r^*) and a signcryption σ^*. F's winning conditions are as follows:

1) The private key of ID_s^* and ID_r^* has not been asked.
2) σ^* cannot be generated from (ID_s^*, ID_r^*, w) through SC query.

3 The MUKSS Scheme

In this part, the concrete construction of the MUKSS scheme is given. The multi-user keyword search functionality of the MUKSS scheme is provided for heterogeneous 5GNS applications. Firstly, a user list is created and uploaded to the cloud server by the data owner in CLC. Secondly, the owner can signcrypt keywords that corresponds to his data, and grant or revoke access rights to multiple users. Finally, multiple users in IBC submit trapdoors to the cloud server, and if they have access rights, they receive the corresponding data. Figure 2 illustrates the MUKSS scheme, and the algorithms are as follows:

Setup. First of all, NM selects the groups G_1, G_2 with the same prime order q. Then NM defines a bilinear pairing $\hat{e} : G_1 \times G_1 \to G_2$. Then NM defines four hash functions:$H_1 : \{0,1\}^* \times G_1 \to Z_q^*$, $H_2 : \{0,1\}^* \to Z_q^*$, $H_3 : \{0,1\}^* \to Z_q^*$ and$H_4 : \{0,1\}^* \times G_1^3 \to Z_q^*$. Assume that P is the generator of G_1, NM selects its secret key $t \in Z_q^*$ and calculates its public key P_{pub}=tP. Finally, NM outputs the cryptographic parameters $Param = \{G_1, G_2, q, \hat{e}, P, P_{pub}, H_1, H_2, H_3, H_4\}$.

CL-PPKG. Given the identity ID_A of a data owner in CLC, NM selects $\alpha \in Z_q^*$, calculates $U = \alpha P$ and $\beta = H_1(ID_A, U)$, then outputs the owner's partial public key $U_A = (U + \beta P)$ and partial private key $\theta = t^{-1}(\alpha + \beta)(\mathrm{mod} q)$.

CL-PKG. Given the cryptographic parameters $Param$, a data owner in CLC picks its secret value $e_A \in Z_q^*$ randomly, and calculates the other part of its public key $PK_A = e_A P$.

CUL. The data owner creates a user list UL_A that records the association between its data D_A and multiple users in IBC, and then uploads it to the cloud server.

IB-KG. Given the identity ID_i of a user in IBC, NM computes the user's private key $SK_i = (t + H_2(ID_i))^{-1} P_{pub}$ and public key $Q_i = (t + H_2(ID_i))P$.

SC. For a data owner with identity ID_A in CLC, it uses the keyword w, its partial private key σ, its secret value e_A, its public key (U_A, PK_A) and each user's public key Q_i to calculate the signcrypted data as follows:

1) Pick $x_i \in Z_q^*$.
2) Compute $f_w = H_3(w)$.
3) Compute $V_i = x_i f_w e_A Q_i$.
4) Compute $h_i = H_4(x_i, ID_A, U_A, Pk_A, V_i)$.
5) Compute $y_i = h_i^{-1} e_A (x_i + \theta)(\mathrm{mod} q)$.
6) Output the ciphertext $\sigma_i = (V_i, y_i)$ for keyword w.

Add-user. When the data owner intends to grant the access right to a new user ID_i, it first inserts $(ID_i, \sigma_i, U_A, PK_A, D_A)$ into UL_A and then uploads UL_A to the cloud server.

Remove-user. When the data owner intends to revoke the access right of a user ID_i, it first deletes the corresponding tuple $(ID_i, \sigma_i, U_A, PK_A, D_A)$ in UL_A and then uploads UL_A to the cloud server.

Trapdoor. The user ID_i who intends the search for the keyword outputs its trapdoor T_{wi} by performing the following steps:

1) Compute $f_w = H_3(w)$.
2) Compute $T_{w_i} = f_w^{-1} SK_i$.

Search. Upon getting the trapdoor T_{w_i} corresponding to the keyword w from user ID_i, the cloud server searches $(ID_i, \sigma_i, U_A, PK_A, D_A)$ from UL_A, computes $h_i = H_4(ID_A, U_A, PK_A, V_i)$ and checks whether $\hat{e}(U_A, PK_A) = \hat{e}(P, P_{pub})^{h_i y_i} / \hat{e}(T_{w_i}, V_i)$. If the equation holds, the data D_A is returned by the cloud server. Otherwise, the cloud server returns $\perp$.

The correctness of $\hat{e}(U_A, PK_A) = \frac{\hat{e}(P, P_{pub})^{h_i y_i}}{\hat{e}(T_{w_i}, V_i)}$ is given as follows:

$$\begin{aligned} \hat{e}(U_A, PK_A) &= \hat{e}(P, P_{pub})^{h_i y_i} / \hat{e}(T_{w_i}, V_i) \\ &= \hat{e}(PK_A, P_{pub})^{x_i + \theta} / \hat{e}(PK_A, P_{pub})^{x_i} \\ &= \hat{e}(U_A, PK_A). \end{aligned}$$

4 Security Analysis

This section provides the security proof of the MUKSS scheme.

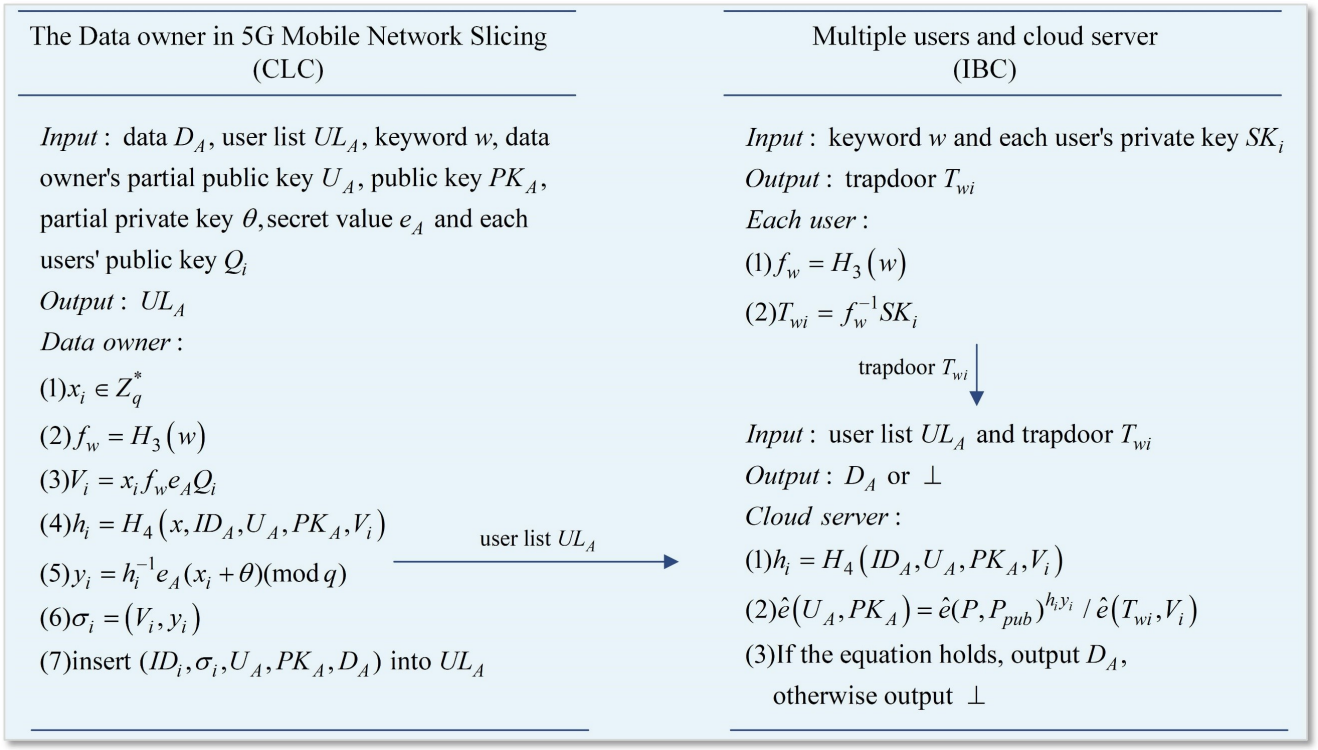


Figure 2: The MUKSS scheme

Theorem 1. *Assume that CDHP is intractable, under the random oracle model (ROM), the MUKSS scheme is indistinguishable against the IND-MUKSS-CCA2 adversary* A_1.

Proof. Assume that the CDHP instance is (aP, bP), a challenger C uses A_1 as a subroutine to distinguish abP as follows:

Game 1:

Initialization: First of all, C executes the Setup algorithm to generate the cryptographic parameters and secret key $t \in Z_q^*$. Then C sends $Param = \{G_1, G_2, q, \hat{e}, P, P_{pub}, H_1, H_2, H_3, H_4$ to $A_1\}$. Meanwhile, C maintains four empty lists $L_{i(i=1-4)}$ to record $H_{i(i=1-4)}$ queries. In addition, C maintains three lists LK_c, LK_b and L_t to record CL-PK, IB-PK and TD queries, respectively.

Phase 1: A_1 asks C the following queries, and C answers in the following way:

- H_1 *queries* : Given a H_1 query with ID_i, U_i as input, C first searches L_1 for (ID_i, U_i, β_i). If L_1 contains the element, C returns β_i to A_1. Otherwise, C picks a value $\beta_i \in Z_q^*$, inserts the tuple (ID_i, U_i, β_i) to L_1 and answers β_i to A_1.
- H_2 *queries* : Given a H_2 query with ID_i as input, C first searches L_2 for $(ID_i, h_{2,i})$. If L_2 contains the element, C returns $h_{2,i}$ to A_1. Otherwise, C picks a value $h_{2,i} \in Z_q^*$, inserts the tuple $(ID_i, h_{2,i})$ to L_2, and answers $h_{2,i}$to A_1.
- H_3 *queries* : Given a H_3 query with w as input, C first searches L_3 for (w, f_w). If L_3 contains the element, C returns f_w to A_1. Otherwise, C picks $f_w \in Z_q^*$, inserts the tuple (w, f_w) to L_3, and answers f_w to A_1.
- H_4 *queries* : Upon getting $(x_i, ID_i, U_i, PK_i, V_i)$ as input, C first searches L_4 for $(x_i, ID_i, U_i, PK_i, V_i, h_{4,i})$. If L_4 contains the element, C returns $h_{4,i}$ to A_1. Otherwise, C picks a value $h_{4,i} \in Z_q^*$, inserts the tuple $(x_i, ID_i, U_i, PK_i, V_i, h_{4,i})$ to L_4, and answers $h_{4,i}$ to A_1.
- $CL-PK-queries$: Suppose A_1 makes this query $q_b > 0$ times at most, C chooses a challenge identity $ID_e (e \in \{1, 2, ..., q_b\})$. Upon getting ID_i as input, if $ID_i = ID_e$, C selects θ_i, U_i randomly, inserts $(ID_i, \theta_i, U_i, \perp, aP)$ to LK_c and answers the public key (U_i, PK_i) to A_1. Otherwise, C first searches LK_c for $(ID_i, \theta_i, U_i, e_i, PK_i)$. If Lk_c contains the element, C returns (U_i, PK_i) to A_1. Otherwise, C selects $\alpha, e_i \in Z_q^*$, searches β_i from L_1, calculates $\theta_i = t(\alpha + \beta_i) (\mathrm{mod}\, q), U_i = (\alpha + \beta_i) P, PK_i = e_i P$, adds $(ID_i, \theta_i, U_i, e_i, PK_i)$ to LK_c, and then re-

turns (U_i, PK_i) to A_1.

- $CL-SKE-queries$: Assume that A_1 has made a CL-PK query on ID_i before this query, so LK_c contains $(ID_i, \theta_i, U_i, e_i, PK_i)$. Upon getting ID_i as input, C searches LK_c for $(ID_i, \theta_i, U_i, e_i, PK_i)$ and answers the private key (θ_i, e_i) to A_1.
- $CL-RPK\ queries$: Upon getting a valid public key PK_i^* of ID_i as input, C uses the given valid public key PK_i^* to update the tuple $(ID_i, \theta_i, U_i, e_i, PK_i)$ with $(ID_i, \theta_i, U_i, K_i^*, \perp, P)$.
- $IB-PK\ queries$: Suppose A_1 makes this query $q_b > 0$ times at most, C chooses a challenge identity ID_e $(e \in \{1, 2, \ldots, q_b\})$. Upon getting ID_i as input, if $ID_i = ID_e$, C sets $Q_i = bP, SK_i = \perp$. If $ID_i \neq ID_e$, C searches the public key Q_i from LK_b. If LK_b does not contain the tuple, C searches $h_{2,i}$ from L_2 and computes $Q_i = (t + h_{2,i})\, P,$, $SK_i = (t + h_{2,i})^{-1} P_{pub}$. Finally, C answers the public key Q_i to A_i and adds (ID_i, SK_i, Q_i) to LK_b.
- $IB-SKE-queries$: Assume that A_1 has made an IB-PK query on ID_i before this query, so LK_b contains (ID_i, SK_i, Q_i). Upon getting ID_i as input, if $ID_i = ID_e, C$ aborts the simulation. Otherwise, C searches the tuple (ID_i, SK_i, Q_i) from LK_b and returns the private key SK_i to A_1.
- $TD-queries$: Upon getting (ID_i, w) as input, C searches T_{wi} from L_T. If L_T does not contain the tuple, and $ID_i = ID_e, C$ selects $T_{wi} \in G_1$. If $ID_i \neq ID_e, C$ searches (SK_i, f_w) from LK_b and L_3, and computes $T_{wi} = f_w^{-1} SK_i$. Then C inserts (ID_i, w, T_{wi}) to L_T. Finally, C returns the trapdoor T_{wi} to A_1.
- $Search-queries$: Given a signcrypted keyword $\sigma = (V, y)$ and a trapdoor T_{wi} as input, C executes the Search algorithm. If σ is valid, C returns 1. Otherwise, C returns $\perp$.

Challenge: A_1 outputs the sender and receiver's identities (ID_s^*, ID_r^*) and a pair of keywords (w_0^*, w_1^*). Note that TD queries have not been made for the keywords. If $ID_r^* \neq ID_e$, C aborts the game. Otherwise, C selects $\beta \in \{0, 1\}$ and $Z^* \in G_1, y^* \in Z_q^*$. Then C sets $V^* = Z^*$. Finally, C returns $\sigma^* = (V^*, y^*)$ to A_1.

Phase 2: A_1 performs the same queries as phase 1. But C rejects TD queries on challenged keywords.

Guess: A_1 outputs its guess β'. If $\beta' = \beta, A_1$ wins the game. Then, C searches (f_w^*, x^*) from L_3 and L_4, respectively. Finally, C can get the solution of CDHP as $Z^* / (x^* f_w^*) = abP$. However, there is no algorithm to break CDHP so far. Thus, the MUKSS scheme is able to achieve confidentiality.

□

Theorem 2. *Assume that CDHP is intractable, under the ROM, the MUKSS scheme is indistinguishable against the IND-MUKSS-CCA2 adversary A_2.*

Proof. Similar to that of Theorem 1, A_2 and B play a game, but A_2 cannot make CL-RPK or CL-SKE queries. Assume that $x_s{}^*, U_s^*$ and $PK_s{}^*$ are the corresponding temoporary key and public key of the sender ID_s^*. If A_2 wants to obtain β, it needs to calculate $h^* = H_4(x_s^*, ID_s^*, U_s^*, PK_s^*, V^*)$. Because A_2 does not know x_s^*, it is also facing CDHP. Therefore, the MUKSS scheme is indistinguishable against the IND-MUKSS-CCA2 adversary A_2. □

Theorem 3. *Assume that DLP is intractable, under ROM, the MUKSS scheme is existentially unforgeable against the EUF-MUKSS-CMA adversary F.*

Proof. Assume that the DLP instance is (P, aP), the process of the challenger C uses F as a subroutine to obtain the value a is as follows:

Game 2:

Initialization: First of all, C executes the Setup algorithm to generate the cryptographic parameters Param and secret key $t \in Z_q^*$. Then C sends Param $= \{G_1, G_2, q, \hat{e}, P, P_{pub}, H_1, H_2, H_3, H_4\}$ to F. Meanwhile, C maintains four empty lists $L_{i(i=1-4)}$ to record $H_{i(i=1-4)}$ queries. In addition, C maintains three lists LK_c, LK_b and L_t to record CL-PK, IB-PK and TD queries, respectively.

Probing: The H_1, H_2, H_3 and H_4 queries are the same as that of Theorem 1. The other queries are defined in the following way:

- $CL-PK-queries$: Suppose F makes this query $q_c > 0$ times at most, C chooses a challenge identity ID_e $(e \in \{1, 2, \ldots, q_c\})$ randomly. Upon getting ID_i as input, if $ID_i = ID_e$, C sets $e_i = \perp, PK_i = aP$. If $ID_i \neq ID_e, C$ searches the tuple $(ID_i, \theta_i, U_i, e_i, PK_i)$ from LK_c. If LK_c does not contain the tuple, C picks $\alpha, e_i \in Z_q^*$ randomly, searches β_i from L_1, computes $\theta_i = t^{-1}(\alpha + \beta_i) \pmod q, U_i = (\alpha + \beta_i) P, PK_i = e_i P$, answers the public key PK_i to F and inserts $(ID_i, \theta_i, U_i, e_i, PK_i)$ to LK_c.
- $CL-SKE-queries$: Assume that F has made a CL-PK query on ID_i before this query, so LK_c contains

$(ID_i, \theta_i, U_i, e_i, PK_i)$. If $ID_i = ID_e, C$ aborts the simulation. Else, C searches $(ID_i, \theta_i, U_i, e_i, PK_i)$ from LK_c and returns the private key (θ_i, e_i) to F.

- $IB - PK - queries$: Upon getting ID_i as input, C searches (ID_i, SK_i, Q_i) from LK_b. If LK_b contains the element, C returns the public key Q_i to F. Otherwise, C searches $h_{2,i}$ from L_2, calculates $Q_i = (t + h_{2,i}) P, SK_i = (t + h_{2,i})^{-1} P_{pub}$, returns the public key Q_i to F and inserts (ID_i, SK_i, Q_i) to LK_b.
- $IB - SKE - queries$: Assume that F has made an IB-PK query on ID_i before this query, so LK_b contains (ID_i, SK_i, Q_i). Upon getting ID_i as input, C searches the list LK_b for (ID_i, SK_i, Q_i) and answers the private key SK_i to F.
- $SC - queries$: Given the identity (ID_s, ID_r) of the sender and receiver, and a keyword w. If $ID_s \neq ID_e, C$ searches $(\theta_s, U_s, e_s, PK_s, Q_r)$ from LK_c and LK_b, and executes SC algorithm to output the signcryption $\sigma = (V, y)$. Otherwise, C stops the simulation.

Forgery: F outputs a signcryption $\sigma^* = (V^*, y^*)$ with the identities (ID_s^*, ID_r^*) of the sender and receiver. Note that F cannot query the private key of ID_s^*, and σ^* cannot be generated by asking SC queries on (ID_s^*, ID_r^*). Similar to [19], C can obtain $\left((ID^*, w), h_{4,i}^*, y^* = h_{4,i}^{*-1} e_A (x_i + \theta) (\mathrm{mod} q)\right)$, and searches (U^*, T^*) from LK_c and L_T as the input of Search query. If Search query returns 1, C uses the forking lemma in [13] to get a valid signcrypion $\sigma' = (V', y')$ and $\left((ID^*, w), h'_{4,i}, y^* = h'^{-1}_{4,i} e_A (x_i + \theta) (\mathrm{mod} q)\right)$ with (ID_s^*, ID_r^*), where $h_{4,i}^* \neq h'_{4,i}$. Finally, C can get the solution of DLP as: $a = (y^* - y) / \left(\left(h_{4,i}^{*1} - h_{4,i}'^{-1}\right) (x_i + \theta)\right)$.

Therefore, F owns a non-negligible success advantage in breaking DLP. However, so far there is no efficient algorithm to break DLP. Thus, the MUKSS scheme is able to achieve unforgeability. □

In conclusion, the MUKSS scheme is proved to realize confidentiality and unforgeability. From [19], a signcryption scheme with keyword search can resist KGA if it achieves both confidentiality and unforgeability. Hence, the MUKSS scheme is secure against KGA.

5 Efficiency Analysis

This section compares the performance and security of our MUKSS scheme with Pan [11], Zhang *et al.* [15], Ma *et al.* [17], Olakanmi [18], Yang *et al.* [33], Zhang *et al.* [34], and Omala [19]. Similar to [18], the experiment platform is as follows: the MIRACL library on a PC with an Intel Core i7-7700HQ 2.80 GHz CPU, and 16 of GB memory. In Table 2, the meaning of each notation is illustrated.

Table 2: Notation

Notation	Meaning
E	Exponentiation operation
M	Point multiplication operation
P	Bilinear pairing operation
$\lvert Z_q^* \rvert$	Element in Z_q^*
$\lvert G_1 \rvert$	Element in G_1
$\lvert G_2 \rvert$	Element in G_2
CT	Length of ciphertext
TD	Length of trapdoor
CTG	Ciphertext generation
TGT	Trapdoor generation and test
TCC	Total computation cost
RKGA	Resist keyword guessing attacks
NCMP	No certificate management problem
MU	Support multiple users to perform keyword search
SL	Security level
CS	Cryptographic system

The comparison of security is shown in Table 3. Note that "Y" means the corresponding attribute is satisfied and "N" means not. Compared with [11,15,17–19,33,34], our scheme not only achieves a 160-bit security level and RKGA, but also supports multiple users to perform keyword search. Besides, our scheme does not suffer from the certificate management problem when compared with Pan [11]. Moreover, only our scheme realizes signcryption in a heterogeneous environment and supports multiple users to perform keyword search.

The computation cost and communication overhead of different schemes are compared separately in Table 4 and Table 5. From [18], an exponentiation operation takes 1.067 ms miliseconds and is represented using E, a point multiplication operation needs 2.165 ms and is represented using M, and a bilinear pairing operation spends 5.427 ms and is represented using P. The computation cost of different schemes can be seen in Figure 3. During the ciphertext generation phase, compared with Pan [11], Zhang *et al.* [15], Ma *et al.* [17], Olakanmi [18], Yang *et al.* [33], Zhang *et al.* [34], and Omala [19], our scheme saves the computation cost by 19.30%, 33.33%, 77.77%, 80.91%, 63.68%, 63.68%, and 55.34% respectively. In the trapdoor generation and test phase, our scheme has a lower computation cost than Zhang *et al.* [15], Ma *et al.* [17], Yang *et al.* [33], Zhang *et al.* [34], and Omala [19], but has a slight increase of 0.15% and 0.30% when compared with Pan [11]

Table 3: Comparison of Securit

Schemes	Attribute							
	Confidentiality	Integrity	Unforgeability	RKGA	NCMP	MU	SL	CS
Pan [11]	Y	Y	N	Y	N	N	160bits	PKI
Zhang *et al.* [15]	Y	Y	Y	Y	Y	N	160bits	IBC
Ma *et al.* [17]	Y	Y	N	Y	Y	N	160bits	CLC
Olakanmi [18]	Y	Y	N	Y	Y	Y	128bits	CLC
Yang *et al.* [33]	Y	Y	N	N	Y	Y	160bits	CLC
Zhang *et al.* [34]	Y	Y	N	Y	Y	Y	160bits	CLC
Omala [19]	Y	Y	Y	Y	Y	N	160bits	CLC-PKI
Ours	Y	Y	Y	Y	Y	Y	160bits	CLC-IBC

Table 4: Computational Cost Comparison

Schemes	CTG	TGT	TCC
Pan [11]	$3E + M = 5.366ms$	$3E + 3P = 19.482ms$	$6E + M + 3P = 24.848ms$
Zhang *et al.* [15]	$3M = 6.495ms$	$2M + 3P = 20.611ms$	$5M + 3P = 27.106ms$
Ma *et al.* [17]	$3E + 3P = 19.482ms$	$3E + M + 4P = 27.074ms$	$6E + M + 7P = 46.556ms$
Olakanmi [18]	$4E + 6M + P = 22.685ms$	$2E + 8M = 19.454ms$	$6E + 14M + P = 42.139ms$
Yang *et al.* [33]	$3M + P = 11.922ms$	$3M + 3P = 22.776ms$	$6M + 4P = 34.698ms$
Zhang *et al.* [34]	$3M + P = 11.922ms$	$5M + 3P = 27.106ms$	$8M + 4P = 39.028ms$
Omala [19]	$3E + 3M = 9.696ms$	$3M + 3P = 22.776ms$	$3E + 6M + 3P = 32.471ms$
Ours	$2M = 4.33ms$	$E + M + 3P = 19.513ms$	$E + 3M + 3P = 23.843ms$

Table 5: Communication Over Comparison

Schemes	CT	TD
Pan [11]	$2\|G_1\| = 2048bits$	$2\|G_1\| + \|G_2\| = 3072bits$
Zhang *et al.* [15]	$2\|G_1\| = 2048bits$	$\|G_2\| = 1024bits$
Ma *et al.* [17]	$\|G_1\| + 2\|G_2\| = 3072bits$	$\|G_1\| + 2\|Z_q^*\| = 1344bits$
Olakanmi [18]	$2\|G_1\| = 2048bits$	$\|G_1\| = 1024bits$
Yang *et al.* [33]	$2\|G_1\| + \|G_2\| = 3072bits$	$3\|G_1\| = 3072bits$
Zhang *et al.* [34]	$2\|G_1\| = 2048bits$	$\|G_2\| = 1024bits$
Omala [19]	$3\|G_1\| = 7072bits$	$\|G_1\| = 1024bits$
Ours	$\|G_1\| + \|Z_q^*\| = 1184bits$	$\|G_1\| = 1024bits$

and Olakanmi [18]. However, Pan [11] suffers from the certificate management problem, and neither Pan [11] nor Olakanmi [18] supports keyword searchable signcryption for multiple users. Furthermore, Pan [11] and Olakanmi [18] are not suitable for the heterogeneous 5GNS environment.

In conclusion, our scheme reduces the total computation cost by at least 4.04% when compared with [11, 15, 17–19, 33, 34].

From [18], an element in G_1 or G_2 is 1024 bits, and an element in Z_q^* is 160 bits. The lengths of ciphertext and trapdoor for different schemes are shown in Table 5. It shows that our scheme has a shorter ciphertext length than [11, 15, 17–19, 33, 34], and a shorter trapdoor length than [11, 17, 33]. Although our scheme has the same trapdoor length as [15, 18, 19, 34], only our scheme supports multi-user keyword searchable signcryption in a heterogeneous environment. In conclusion, our scheme reduces the communication burdens and enhances the security of

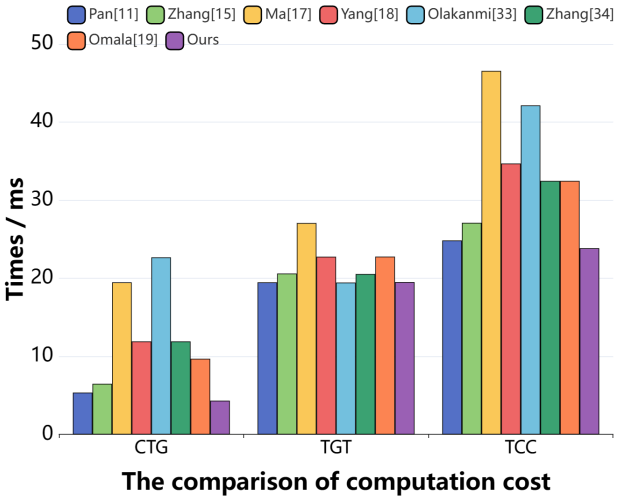


Figure 3: The comparison of computation cost

data, so it is more suitable for the heterogeneous 5GNS applications.

6 Conclusion

This paper proposes a multi-user heterogeneous keyword searchable signcryption scheme for 5GNS. Our MUKSS scheme has achieved the keyword search functionality for the heterogeneous 5GNS environment. In the MUKSS scheme, a data owner in CLC can share its data with multiple users in IBC. Through the security analysis, in ROM, it is proved that the MUKSS scheme has realized confidentiality under the CDH problem, and the MUKKS scheme has realized unforgeability under the DLP problem. In addition, the MUKSS scheme is secure against KGA. According to the efficiency analysis, when compared to schemes [11,15,17–19,33,34], the MUKSS scheme is more suitable for the heterngeneous 5GNS environment.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China through Grant No.62262041, and Graduate Innovation Special Fund of Jiangxi Province under Grant No. YC2023-S013.

References

[1] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka *et al.*, "Scenarios for 5g mobile and wireless communications: the vision of the metis project," *IEEE communications magazine*, vol. 52, no. 5, pp. 26–35, 2014.

[2] L. Dong, H. Zhao, Y. Chen, D. Chen, T. Wang, L. Lu, B. Zhang, L. Hu, L. Gu, B. Li *et al.*, "Introduction on imt-2020 5g trials in china," *IEEE journal on selected areas in communications*, vol. 35, no. 8, pp. 1849–1866, 2017.

[3] T. Nakamura, "Research activities of the fifth generation mobile communication promotion forum radio access technologies towards the fifth generation mobile communications system," in *2015 21st Asia-Pacific Conference on Communications (APCC)*. IEEE, 2015, pp. 169–173.

[4] C. R. Storck and F. Duarte-Figueiredo, "A survey of 5g technology evolution, standards, and infrastructure associated with vehicle-to-everything communications by internet of vehicles," *IEEE access*, vol. 8, pp. 117 593–117 614, 2020.

[5] J. Cheng, W. Chen, F. Tao, and C.-L. Lin, "Industrial iot in 5g environment towards smart manufacturing," *Journal of Industrial Information Integration*, vol. 10, pp. 10–19, 2018.

[6] R. Siddavaatam, I. Woungang, G. H. Carvalho, and A. Anpalagan, "Mobile cloud storage over 5g: A mechanism design approach," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4060–4071, 2019.

[7] H. Khan and K. M. Martin, "A survey of subscription privacy on the 5g radio interface - the past, present and future," *Journal of Information Security and Applications*, vol. 53, p. 102537, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212620300235

[8] S. Wijethilaka and M. Liyanage, "Survey on network slicing for internet of things realization in 5g networks," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 957–994, 2021.

[9] A. Elkhalil, J. zhang, R. Elhabob, and N. Eltayieb, "An efficient signcryption of heterogeneous systems for internet of vehicles," *Journal of Systems Architecture*, vol. 113, p. 101885, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1383762120301594

[10] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in *Advances in Cryptology - EUROCRYPT 2004*, C. Cachin and J. L. Camenisch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 506–522.

[11] X. Pan and F. Li, "Public-key authenticated encryption with keyword search achieving both multi-ciphertext and multi-trapdoor indistinguishability," *Journal of Systems Architecture*, vol. 115, p. 102075, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1383762121000643

[12] M. S. Nair and R. M.S, "Fine-grained search and access control in multi-user searchable encryption without shared keys," *Journal of Information Security and Applications*, vol. 41, pp. 124–133, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212618300486

[13] C. Li, C. Xu, S. Li, K. Chen, and Y. Miao, "On the security of verifiable searchable encryption schemes," *IEEE Transactions on Cloud Computing*, vol. 10, no. 4, pp. 2977–2978, 2022.

[14] X. Zhang, Y. Tang, H. Wang, C. Xu, Y. Miao, and H. Cheng, "Lattice-based proxy-oriented identity-based encryption with keyword search for cloud storage," *Information Sciences*, vol. 494, pp. 193–207, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519303706

[15] X. Zhang, Y. Tang, S. Cao, C. Huang, and S. Zheng, "Enabling identity-based authorized encrypted diagnostic data sharing for cloud-assisted e-health information systems," *Journal of Information Security and Applications*, vol. 54, p. 102568, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221421261930852X

[16] Y. Lu and J. Li, "Lightweight public key authenticated encryption with keyword search against

adaptively-chosen-targets adversaries for mobile devices," *IEEE Transactions on Mobile Computing*, vol. 21, no. 12, pp. 4397–4409, 2022.

[17] M. Ma, D. He, S. Fan, and D. Feng, "Certificateless searchable public key encryption scheme secure against keyword guessing attacks for smart healthcare," *Journal of Information Security and Applications*, vol. 50, p. 102429, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212619307203

[18] O. O. Olakanmi and K. O. Odeyemi, "A certificateless keyword searchable encryption scheme in multi-user setting for fog-enhanced industrial internet of things," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 4, p. e4257, 2022.

[19] A. A. Omala, I. Ali, and F. Li, "Heterogeneous signcryption with keyword search for wireless body area network," *Security and Privacy*, vol. 1, no. 5, p. e25, 2018.

[20] J. W. Byun, H. S. Rhee, H.-A. Park, and D. H. Lee, "Off-line keyword guessing attacks on recent keyword search schemes over encrypted data," in *Secure Data Management*, W. Jonker and M. Petković, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 75–83.

[21] M. Abdalla, M. Bellare, D. Catalano, E. Kiltz, T. Kohno, T. Lange, J. Malone-Lee, G. Neven, P. Paillier, and H. Shi, "Searchable encryption revisited: consistency properties, relation to anonymous ibe, and extensions," in *Advances in Cryptology – CRYPTO 2005*, V. Shoup, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 205–222.

[22] I. R. Jeong, J. O. Kwon, D. Hong, and D. H. Lee, "Constructing peks schemes secure against keyword guessing attacks is possible?" *Computer Communications*, vol. 32, no. 2, pp. 394–396, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366408005768

[23] P. Jiang, F. Guo, and Y. Mu, "Efficient identity-based broadcast encryption with keyword search against insider attacks for database systems," *Theoretical Computer Science*, vol. 767, pp. 51–72, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304397518306017

[24] N. Pakniat, D. Shiraly, and Z. Eslami, "Certificateless authenticated encryption with keyword search: Enhanced security model and a concrete construction for industrial iot," *Journal of Information Security and Applications*, vol. 53, p. 102525, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212619309032

[25] X. Yang, G. Chen, M. Wang, T. Li, and C. Wang, "Multi-keyword certificateless searchable public key authenticated encryption scheme based on blockchain," *IEEE Access*, vol. 8, pp. 158 765–158 777, 2020.

[26] A. Karati, C.-I. Fan, and E.-S. Zhuang, "Reliable data sharing by certificateless encryption supporting keyword search against vulnerable kgc in industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3661–3669, 2022.

[27] Z. Liu, Y. Liu, and Y. Fan, "Searchable attribute-based signcryption scheme for electronic personal health record," *IEEE Access*, vol. 6, pp. 76 381–76 394, 2018.

[28] N. Eltayieb, R. Elhabob, A. Hassan, and F. Li, "Secure mobile health system supporting search function and decryption verification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 2221–2231, 2021.

[29] Y. Zheng, "Digital signcryption or how to achieve cost(signature & encryption) ≪ cost(signature) + cost(encryption)," in *Advances in Cryptology — CRYPTO '97*, B. S. Kaliski, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 165–179.

[30] S. Liu, L. Chen, G. Wu, H. Wang, and H. Yu, "Blockchain-backed searchable proxy signcryption for cloud personal health records," *IEEE Transactions on Services Computing*, pp. 1–14, 2023.

[31] U. S. Varri, S. K. Pasupuleti, and K. Kadambari, "Practical verifiable multi-keyword attribute-based searchable signcryption in cloud storage," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2022.

[32] X. Yu, C. Xu, B. Dou, and Y. Wang, "Multi-user search on the encrypted multimedia database: lattice-based searchable encryption scheme with time-controlled proxy re-encryption," *Multimedia Tools and Applications*, vol. 80, pp. 3193–3211, 2021.

[33] X. Yang, J. Wang, T. Tian, and X. Wang, "Multi-user fuzzy keyword searchable encryption scheme based on certificateless cryptosystem," in *Journal of Physics: Conference Series*, vol. 1828, no. 1. IOP Publishing, 2021, p. 012118.

[34] Y. Zhang, L. Wen, H. Wang, Y. Zhang, and C. Wang, "Certificateless authentication searchable encryption scheme for multi-user," p. 1094, 2020. [Online]. Available: //article/id/2ded9cea-a9ee-46fb-8075-2ba7f7ea62a1

Biography

Ming Luo [corresponding author] received the Ph.D degree in computer application technology from Northeastern University, Shenyang, China in 2010. He is currently a professor and the deputy dean of the School of Software, Nanchang University, Nanchang, China. His research interests cover Internet of Things, Cyberspace Security and Cryptography.

Qibang Zhan received the B. S. degree from Institute of Technology East China Jiao Tong University. He is currently pursuing the M. S. degree with the School of Software of Nanchang University, Nanchang, China. His research interests cover Internet of Things, Cyberspace

Security and Cryptography.

Minrong Qiu received the Ph.D degree in industrial economics from Wuhan University of Technology, Wuhan, China in 2018. She is currently a associate professor of the GongQing Institute of Science and Technology, Gongqingcheng, China. Her research interests cover Internet of Things, Cyberspace Security and Information System Management.

Li Cen is currently pursuing the Ph.D degree with the School of Computer Science of National University of Malaysia. She is currently the dean of the School of International Education, Jiangxi University of Software Professional Technology, Nanchang, China. Her research interests cover Cyberspace Security and Cryptography.

Enterprise Accounting Management Reform of Industrial Integration Under Intelligent Information Dissemination

Tang Min

(Corresponding author: Tang Min)

School Of Accounting And Finance, Wuxi Vocational Institute of Commerce, Wuxi 214000, Jiangsu, China

Email: Sjysyt2023@163.com

(Received Oct. 28, 2023; Revised and Accepted May 10, 2024; First Online June 22, 2024)

Abstract

The intense competition among enterprises is transforming the traditional financial management mode, which is limited by outdated information technology and weak data analysis abilities. This results in passive recording of business status, making financial statement analysis inaccurate. The digital economy drives the transformation of financial management to the entire value management process, with financial personnel transforming from passive data receivers to business partners. This shift in industry finance integration thinking focuses on optimizing enterprise financial management mode, aiming to create value for enterprises and improve their financial management. This study examines the issues surrounding integrating blockchain technology with enterprise financial information management. It identifies shortcomings and focuses explicitly on the necessity of integrating blockchain with enterprise financial information management to address vulnerabilities such as the internal mechanism of 51% double flower attacks in blockchain financial ledgers. The model's efficacy is tested using MATLAB simulation, and recommendations for risk mitigation and control strategies are proposed. The experimental results prove that the research and construction of the enterprise financial and accounting management reform model system based on industrial integration under intelligent information dissemination enriches the mining and application of economic data, provides basis and support for the formulation of operation and management decisions of group companies, and is conducive to improving the market competitiveness of domestic group companies.

Keywords: Accounting Management; Game Model Evolution; Industrial Integration; Simulation and Verification

1 Introduction

Traditional financial management focuses on accounting; that is, the operation of the enterprise is presented in the form of reports. However, with the development of the economy, this model of the current enterprise. Finance needs to flexibly take advantage of the convenience of obtaining data, processing data in combination with business work, and providing decision-making opinions for managers. The financial management model in the new era requires value creation as the orientation, from accounting to strategic management, from functional to service-oriented, and from result-oriented static accounting to comprehensive and dynamic supervision. Modern financial management should be supported by a detailed and comprehensive information management system, supplemented by standardized business processes to respond to the rapidly changing market in time and enable enterprises to continue to develop strongly through pre-prediction, in-process supervision, and post-assessment.

In the past, financial management produced financial statements through daily financial accounting. It is only the statistics and presentation of relevant data by accounting requirements and the need for more thinking, thereby weakening the role that finance can bring to enterprise management. In terms of time, traditional financial management focuses more on post-event management. It does not pay attention to the causes of data generation and business-related connections, which limits the additional functions of finance — monitoring, forecasting, and decision-making. For the purpose of protecting financial data and guaranteeing data integrity, network security is essential in business accounting management reform.

In addition to help for digital transformation, it offers confidentiality assurance, transaction integrity, defence against cyber-attacks, regulatory compliance, business continuity, fraud protection, and reputation preservation. Reliability and continuity of company operations are ensured, unauthorized activity is stopped, and an organization's reputation is preserved via effective network security. Network security becomes a basis for the safe integration of technologies like blockchain, artificial intelligence, and cloud computing as businesses embrace digital

transformation. Under the traditional model, there are many barriers to communication between business and finance, and there is little communication between the two departments.

Under the industrial integration model, the new financial management needs the support of the new business management model, especially in large enterprises, which usually face complex business models. The previous business models mainly focused on operating their modules, ignoring the various departments. The synergy between them disassembles the enterprise into unit operation. The work skills of business personnel are mainly limited to their business functions, and they are concentrated in a particular link in the value chain. This will lead to the waste of resources and more operational risks in the traditional business management and financial management models, which will quickly lead to the short-term interests of the department, ignoring the role of long-term sustainable development. If an enterprise wants to maintain its leading position, it must improve its core competitiveness and operating efficiency, which requires strategic transformation.

Based on the principle of business-finance integration, a comprehensive and efficient information management system is developed, and various departments work together to support enterprise decision-making through extensive data analysis and then provide weapons for enterprises to compete and survive in the market. Financial management in the data age must be efficient to support the enterprise's strategic development, encourage financial personnel to actively understand the essence of business, understand the process of enterprise value chain proliferation, and engage in pre-budget management and in-process cost management. Post-assessment performance management forms a closed-loop business scenario, thereby assisting the realization of corporate strategic planning.

The accumulative total illegal occupation of Chenlong sawing machine funds was 120 million yuan. The fundamental reason for the fraud of corporate financial information is that the reliability of financial information cannot be guaranteed. The source of financial information cannot be checked and the responsibility can be investigated. In addition, the financial statements of many enterprises need help with non-standard preparation and lack of data on critical subjects, and financial statements are an essential basis for enterprises to reflect financial status and evaluate business performance.

This paper focuses on integrating enterprise business and enterprise dissemination, especially the establishment of distributed financial ledgers, the accurate storage of financial statement data, and the reliable tracing of financial information. It analyzes the current situation and shortcomings of integrating blockchain and enterprise financial information management from these three aspects. On this basis, it proposes the need for the integration of blockchain and enterprise financial information management.

2 Related Works

Under the relatively loose market environment after the 21st century, the rapid development of various enterprises has intensified competition, and many enterprises have begun to seek breakthroughs by improving operational efficiency. Researchers began to gradually implement the theoretical knowledge of business-finance integration into enterprise management. Literature [4, 28] focused on system optimization in enterprise operations and proposed the construction of a business-finance integration system to realize business-finance integration. The financial function is embedded with the business process. Literature [26, 31] Based on the sample survey method, collecting samples of different enterprises confirmed that the enterprises with sound business and financial integration are better than others in production efficiency. Based on the analysis of 178 enterprise samples, the literature [17] concluded that most business and financial departments need more communication, which hinders the exchange of information and affects the efficiency of business operations. The improvement of the operating efficiency of enterprises requires good channels and mechanisms for information sharing and exchange between business and finance.

As early as 1985, literature [8] put forward a qualitative analysis from several primary characteristics of enterprise risk analysis, risk control, risk finance, risk management decision-making, and risk management, which has created a precedent for the study of financial risks of Chinese enterprises. Literature [12, 25] combines realistic, complex economic problems with the method of econometrics, adopts the process of establishing foreign simultaneous equation models, and builds an enterprise financial simulation model. The direction of risk forecasting opens up new frontiers.

Literature [3,21] used the principal component analysis method to construct a Y-score model for judging the financial risk of listed companies in my country. It divided the financial risk levels of the sample companies by analyzing the sample indicators, which provided the financial risk prediction research of listed companies in my country. Literature [1,7] first proposed the use of the KMV model, and the power curve was introduced for comparison. The research shows that the KMV model, as a mature theoretical method for option pricing research, also applies to the direction of early warning corporate financial crisis. Literature [18,22] used the BP neural network model, Z-score discriminant method, and logistic regression analysis method to conduct simulation tests with 150 listed manufacturing companies as samples. The results show that the prediction accuracy of different models is 86%, 76%, and 82%, respectively, which further indicates that the BP neural network model has a good effect in discriminating the financial crisis of enterprises and has practical value.

Literature [6,24] used a factor analysis model to study the early financial warning of listed companies in China.

Research shows that this model can predict the financial status of enterprises, and the closer the years are to economic failure, the higher the accuracy of prediction. Literature [5, 15] uses the idea of system dynamics, defines the modeling purpose and system boundary in sequence according to the modeling steps of system dynamics, establishes the causality diagram, flow diagram, and quantity equation through system structure analysis, and establishes the financial risk prediction model of the automobile manufacturing industry. The research shows that the model has a good effect in judging the financial status of enterprises.

Literature [20] uses the proportional advantage model to study the financial distress of listed companies. Financial accounting has undergone a transformation thanks to the accounting industry's incorporation of computer and internet technology. Accurate and quick data transmission is made feasible by fusing computerized information technology with conventional means. By doing this, the efficiency of accounting job is maximized, freeing up accountants to concentrate on the analysis of business financial data. But integrating network technology also brings new difficulties, thus remedies to enhance the expertise of financial staff members must be developed [11]. Due to China's advances in information technology, artificial intelligence is now widely used in homes, and businesses are more competitive. This essay explores the rationale for integrating AI with management accounting and the degree to which this revolution has improved financial management skills [29].

Literature [30] believes that the integration of industry and finance refers to the integration of the work behavior of the business department and financial departments work behavior. This view implies that in the operation of an enterprise, although business and finance are in different departments, their purposes and behaviors should be consistent. Business provides development impetus for financial work, and finance guides business. Literature [14] believes that business management and financial management based on the same strategic objective should complement each other, and the integration of industry and finance can further enhance the competitiveness of enterprises.

Literature [19] believes that the separation of the business department and the financial department will cause both parties to immerse themselves in the organizational goals of their departments and then lose the basis for departmental cooperation, which is not conducive to implementing the enterprise strategy. Literature [27] points out that the value activities of enterprises can be divided into basic activities dominated by business processes and supporting activities participated by functional departments. As the core Department of enterprise operation and management, the financial department must carry out operational management and decision-making support for the production and operation of the business department, and the strategic development of the financial department needs the cooperation of the business department.

Literature [13] uses big data technologies to study how Chinese firms integrate audit and cost accounting. It concludes that big data prioritizes scientific advancement and thorough application rather than changing the causal link between accounting and audit transactions. Literature [10] examines the development of intelligent finance in Chinese firms, a digital trend driven by artificial intelligence. By offering a BI-based management accounting platform for analysis and decision-making, it provides multi-level management assistance and flips the "information island" concept.

Other scholars proposed that the essence of industry finance integration is value integration based on the above two viewpoints. Under the market economy system, enterprises aim to maximize value. Enterprises create value through business activities and then realize value through functional departments. Literature [23] points out that the essence of industry finance integration is the comprehensive integration of enterprise value chains and business chains. Through real-time sharing of information, business, and capital, the management activities of enterprise decision-making, execution, and evaluation can be further optimized. This is the essence of industry finance integration.

Literature [16] proposes that finance should break through department boundaries, go deep into the front end of the business, and strengthen, thereby improving operational efficiency. This point of view emphasizes that integrating business and finance maximizes creating and realizing enterprise value. These are all to improve financial management, serve the business, and create value. The business chain must abide by creating the basic principle of value that makes value throughout the business process. Business and financial integration is the comprehensive business chain and value chain integration. The study focuses on optimizing production efficiency, operating environment, and profitability to ensure sustainable development in intense market competition. It designs and optimizes enterprise financial management models within the Internet framework using big data and data mining technologies [2].

To sum up, a logical succession and internal connection have been formed from the perspective of the essence of the integration of business and wealth. The essence of business-finance integration is to create value and realize value. However, there needs to be more interaction between financial information and business information, separating the value chain and the business chain. Therefore, the integration of business and finance must be based on an all-round value integration perspective, not only to solve the problem of separation of functions but also to improve information interaction. Based on the above research foundation, this research deeply explores the ideas and innovative paths of enterprise financial and accounting management reform under this macro background.

3 Building a Blockchain Model Based on the Integration of Industry and Finance

The value chain theory of business-finance integration provides a theoretical basis for enterprise financial management. A strong network security mechanism is necessary for the blockchain technology's integration of finance and industry. A decentralized architecture, strong encryption methods, stringent access control, smart contract security, strong identity management, and an immutable ledger are necessary to accomplish this. By taking these precautions, the network becomes more resistant to assaults and lowers the possibility of a single point of failure. Network traffic is monitored by firewalls and intrusion detection systems, and any vulnerabilities are found and fixed through routine security audits. To reduce the possibility of malicious attacks causing service disruptions, distributed denial of service (DDoS) security is also deployed. Regulatory compliance guarantees that the blockchain network conforms to applicable financial and industrial rules. Through the use of these procedures, establishments may augment the system's general security and reliability, cultivating a safe milieu for monetary exchanges and cooperative endeavours. Financial management can be disassembled through value chain analysis into multiple step-by-step links, data analysis of the value-added process, and an advanced management method that integrates business flow, capital chain, and information flow into one. Improve the value-added efficiency of the enterprise's value chain. In the context of intelligent information dissemination, the "virtual value chain" concept shows the importance of information data to the operation of enterprises. It can encourage enterprises to improve the information system and realize timely and accurate data transmission. Therefore, the entire value chain theory has a theoretical guiding role in optimizing financial management. In this section, by studying the value-added process of enterprises in procurement, inventory, production, sales, and other links, the monetary value is assigned to each link to optimize the waste of resources and play the role of financial supervision, analysis, and optimization before and after the event. As shown in Figure 1.

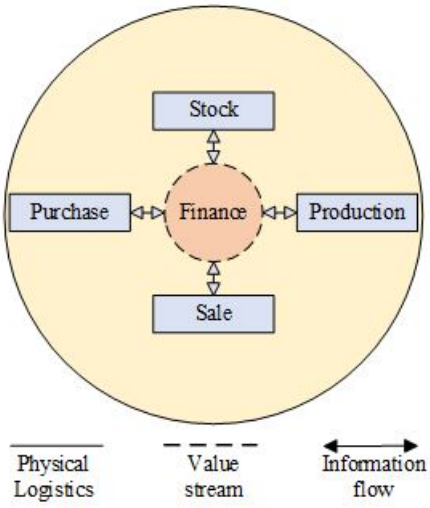


Figure 1: Internal financial value chain model

However, enterprise competition is becoming increasingly fierce, product updates and iteration speeds are increasing, and the corresponding process needs to be continuously improved and redesigned. The theory of business process reengineering emphasizes that the organizational form of enterprises should develop from "function-oriented" to "process-oriented," and its content is shown in Table 1.

Through process reengineering, the job responsibilities of each link in the business process can be clarified, and financial control and analysis can be introduced in the critical links of enterprise production and operation to realize business Cost reduction and efficiency increase in all aspects of the process. Process transformation can simplify the communication process, promote the integration of business and financial value, adjust the flow of each link, improve the efficiency of business approval, and integrate business processes into a value-added process with a front-to-back relationship.

To sum up, this paper conducts a systematic study to integrate business and finance as the optimization direction of financial management. On the one hand, it uses the value chain theory to discuss the company's value realization process and analyzes the advantages and problems of its financial management model. On the other hand, use the theory of business process reengineering to create value, discuss the extension of financial tentacles in business circulation, and propose optimization plans and safeguard measures. Integrating industry and finance provides target guidance for optimizing financial management models. The optimization process of financial management models of some listed companies also enriches the practical experience of the theory of industry and finance integration in enterprise applications. Based on reading a lot of literature and understanding the corresponding theoretical basis and finance, as shown in Figure 2.

In the context of the integration of business and finance, investment organizations or individuals pay more attention to the security of funds, and more and more investment organizations and investors begin to pay attention to the financial analysis of the companies they invest in. Enterprise financial analysis is not only an indispensable part of the daily financial management of the enterprise but also a fundamental decision-making basis for the enterprise's stakeholders to control and make decisions on the enterprise's financial affairs. Corporate financial analysis usually provides the necessary decision-making basis for the company's senior management to manage the company and formulate operating plans and economic policies by analyzing and evaluating the company's history, recent

Table 1: "Process Orientation" of Business-Financial Integration

Stage	Meaning
Business process demand analysis	Analyze the existing business process innovation needs and diagnose the problems
Redesign of business process	According to the results of analysis and diagnosis, redesign the existing process and realize the standardization of the business process.
Implementation of business flow group reorganization	Truly implement the redesigned process into the production and operation process of the enterprise.

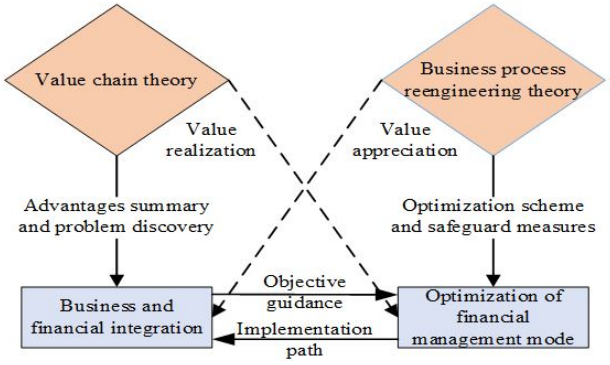


Figure 2: The theoretical analysis framework of financial management model optimization under the background of business-finance integration

operating conditions, and financial conditions. Therefore, if the main body of enterprise financial analysis is different, and the degree of interest inclination is different, the purpose of financial analysis will be other, as shown in Figure 3.

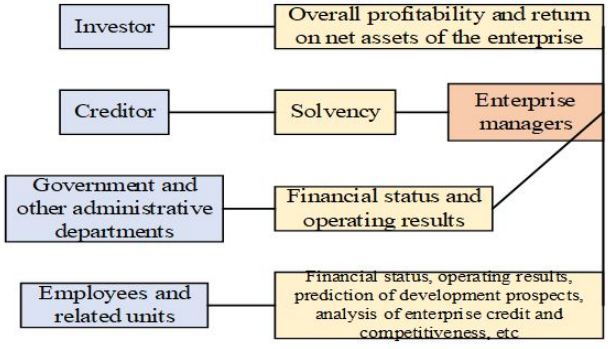


Figure 3: Demand analysis of financial management model optimization under the background of business-finance integration

Under intelligent information dissemination, the core of using data mining methods for enterprise financial analysis is data processing. The data generated by modern enterprises, giant enterprises, is massive every day. It is not enough to send valuable data from these gigantic data and obtain the status of enterprises. We should not only discover problems from massive data but also make corresponding responses according to the problems we discover. This is a process from data to valuable data, then to reaction actions, and then to data. It is a cyclic process. This requires us to build a model to simulate and process this process, as shown in Figure 4.

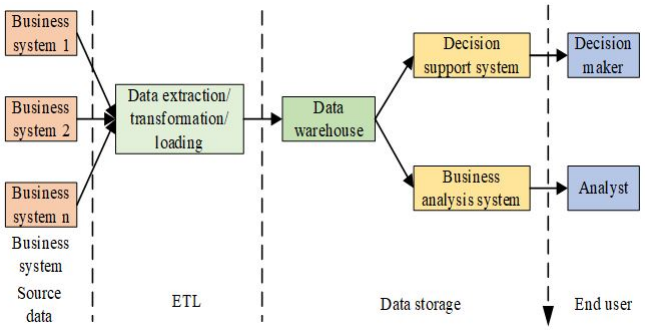


Figure 4: Data mining optimization mode of enterprise financial analysis

The financial management mode's blockchain architecture is a technical data storage system. Let the information system (U, A) be as shown in Table 2, x_1, x_2, x_3 and x_4 are four objects, and a_1, a_2, a_3 and a_4 are four attributes. If $B = a_1$ in equivalence relation $R(B)$, the two particles are: $X_2 = \{x_3, x_4\}$;If $B = a_2$, be $X_1 = \{x_1, x_3\}, X_2 = \{x_2, x_4\}$; If $B = a_3$, be $X_1 = \{x_1, x_4\}, X_2 = \{x_2, x_3\}$; If $B = a_4$, be $X_1 = \{x_1, x_3, x_4\}, X_2 = \{x_2\}$.

Table 2: Equivalence relation of information system architecture

U	A			
	a_1	a_2	a_3	a_4
x_1	2	0	2	0
x_2	0	1	0	1
x_3	1	2	0	0
x_4	1	1	1	2

4 Methods

This paper designs the idea chart of the integration of enterprise business and enterprise financial information management, as shown in Figure 5.

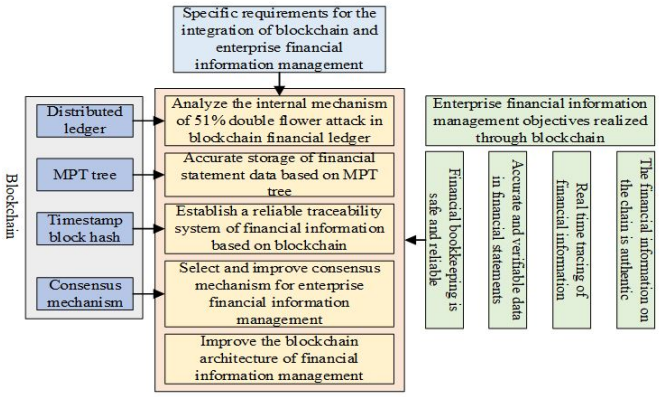


Figure 5: The road map of business-finance integration

In relational databases, the core of association rules mining is to generate association rules that meet certain conditions. These association rules must meet the following conditions: (1) The support degree of the generated association rules is greater than or equal to the minimum support degree; (2) The credibility of the generated association rules is greater than or equal to the minimum credibility. The minimum support and minimum reliability involved here are set in advance by the user according to their needs. The physical meaning of the minimum support threshold is the minimum importance of the data item selected in the statistical sense; the physical meaning of the minimum reliability threshold is the minimum reliability that the association rules must satisfy. In the previous section, we have detailed the theoretical basis of association rules. Here is a data model with item set I and transaction set D. Among them, the item set I = A, B, C, D, E, as shown in Table 3.

Table 3: Collection of associations between business, finance, and accounting reforms

TID	ITEM
100	ACD
200	BCE
300	ABCE
400	BE

In short, an efficient data mining joint model for enterprise financial analysis can be divided into four steps: data preprocessing using ratio analysis, clustering of massive preprocessed data using clustering algorithms, and association rules. The mining algorithm extracts association rules for each cluster. Since the clustering has been carried out, the amount of data processed by the algorithm is reduced, the mining time of association rules is significantly reduced, and the algorithm's efficiency is improved; the extracted data is calculated using the decision tree algorithm. In a decision tree algorithm, network security refers to putting policies in place to safeguard the algorithm, its training set, and the decision-making process against possible online attacks. Data encryption, safe data transmission, access control, model validation and verification, safe model deployment, defence against hostile attacks, frequent security audits, firewalls and intrusion detection systems, safe system integration, user authentication and authorization, and data privacy compliance are important factors to take into account. Access restrictions prohibit unwanted access, secure communication channels stop eavesdropping, and encryption techniques guarantee data secrecy. Model integrity is ensured by processes for model validation and verification, while attack protection is provided via secure model deployment. The algorithm functions safely inside the networked environment thanks to regular security audits, firewalls, intrusion detection systems, secure interaction with other systems, strong user authentication and authorization procedures, and data privacy compliance. The characteristic probability of association rules is used to clarify the index characteristics of the financial status of enterprises. Therefore, the core of an enterprise financial management analysis system is an efficient data mining joint model. Its specific structure is shown in Figure 6.

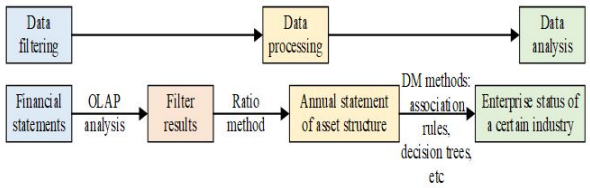


Figure 6: Architecture of enterprise financial management system based on efficient data mining joint model

First of all, based on the enterprise balance sheet, enterprise income statement, and enterprise cash flow statement, use the slicing function of OLAP to filter the original data, use the ratio analysis method in enterprise financial analysis to preprocess the original data, and establish a practical annual statement of corporate asset structure. Secondly, use the efficient data mining joint model shown in Chapter 3 to analyze the preprocessed enterprise financial data, including cluster analysis, association rule mining, and decision tree analysis. Finally, the analysis results are displayed in charts to assist enterprises in financial analysis and evaluation, providing a reliable and accurate decision-making basis for enterprise decision-makers and stakeholders.

The enterprise financial analysis system has five categories of functional rights: security management rights, dimension management rights, data input rights, report rights, and superuser rights. In role management, super users can create different roles and then assign additional permissions to these roles, and a role can have multiple attributes. User management: A superuser can create

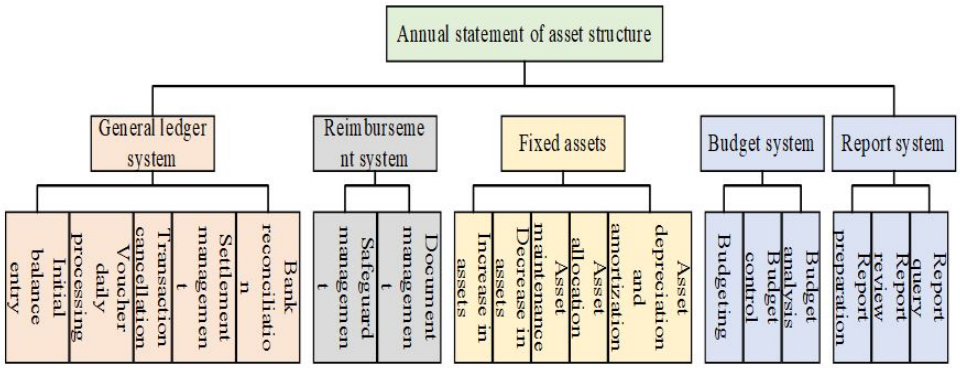


Figure 7: The framework of the enterprise accounting management model of industrial integration under intelligent information dissemination

many users, delete many, and post different roles to each user. However, a user can only belong to one role. Menu management: In this section, decide which roles can have which menu functions, such as a role that can enter the dimension management menu.

4.1 Data Acquisition

Take iron and steel enterprises as an example of conducting financial management analysis. The original financial data comes from the quarterly and annual financial reports of 36 iron and steel companies that Zhongcai.com has published. Based on the balance sheet of each enterprise, form a steel industry enterprise balance sheet, put the data into the SQL Server database, and use the OLAP tool analysis service to create a three-dimensional space including time dimension, index dimension, and enterprise dimension Cubes, as shown in Table 4, Table 5, and Table 6.

Table 4: Time dimension table

Time ID	Year	Moon	Day	Time
1	2006	200609	20060931	2006-9-31
2	2006	200612	200161231	2006-12-31
25	2012	201209	20120931	2012-9-31
26	2012	201212	20121231	2012-12-31
27	2013	201303	20130331	2013-3-31

In the dimension of the enterprise financial analysis system, there is a member named steer under the organization dimension, and there is a total of 36 members A1-A36 under it; under the dimension account, there are current assets members, long-term investment members, etc. Create a new form on the form setting page, including the following dimensions: dimension organization, time dimension, dimension account, and dimension currency. Enter the newly created form input page, and make the following settings: select the steer members "A1 to A9" in the organization dimension, select "all" in the account dimension, select "RMB" in the currency dimension, and set the time to December 30, 2012, scenario Set to actual. Next, enter data in the cell. If the data has been saved in the system fact table, the data can be displayed, as shown in Figure 7.

According to the funding situation of the headquarters and each subsidiary company, make scientific and reasonable budgets and estimates, and realize the overall financial budget method formulation support for the entire group company through the cost control module. In this module, various economic indicators can be set. The financial data related to the daily operation of the group company can be dynamically monitored, and the implementation of the budget plan can be comprehensively monitored. Figure 8 shows the information transfer process between the functional modules of the group company's financial data analysis and decision-making system.

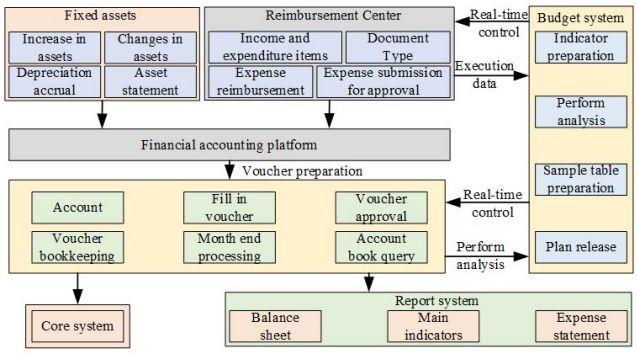


Figure 8: Flow chart of information transfer between modules

Activity diagrams can be used to interpret financial data in the actual management of financial data. Taking the user's application for reimbursement as an example, the process of applying for reimbursement using the activity diagram in UML technology is shown in Figure 9.

Table 5: Data dimension table

Company	Time	Current assets	Long-term investment	Fixed assets	Intangible assets and others	Current liabilities	Long term liabilities	Equity	Capital reserve	Surplus reserve	Undistributed profits
1	29	695,353.88	3,	248.16	318,314.61	497,191.56	11,857.3	111,223.16	162,318.71	69,226.01	85,505.98

Table 6: Indicator dimension table

Company ID	Company Name	Company code
1	Xinxing cast pipe	000778
2	TISCO stainless	000825
3	Angang Steel Co., Ltd	000898
4	Shougang service	000959
5	Wuhan Iron & Steel	600005
6	Baosteel Co., Ltd	600010
7	Hangzhou Steel Service	600126
8	Anyang Iron and steel	600569
9	Xinhua shares	600782

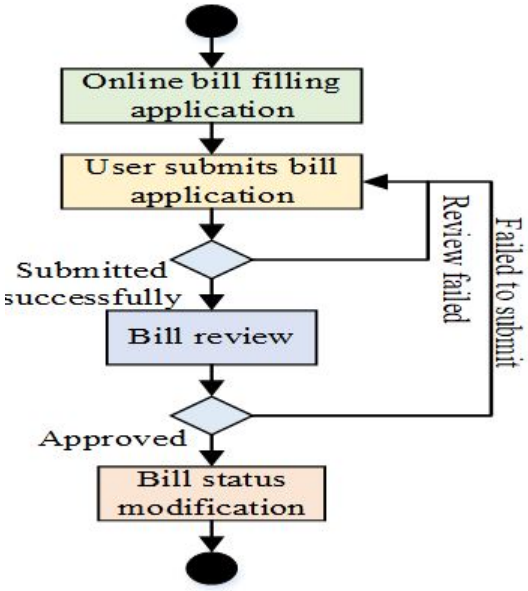


Figure 9: Actual business flow chart

Table 7: Simulation parameter assignment

Calculation power cost h (yuan)	New currency reward b (yuan)	Transaction price p (yuan)	Transaction cost f (yuan)
200	80	60	200
			170
			100
		200	100
			70
			20
		280	60
			20
		400	100
			50

4.2 Simulation Analysis

The evolution process of the willingness to attack the 51% double-spending node of the financial and accounting fusion node, this paper uses MATLAB software for simulation analysis. First, assign values to p, b, h, and f. According to the conclusions obtained in the evolutionary game model and the actual situation, this paper sets the computing power cost h and the new currency reward b to a fixed value of 200 yuan and 80 yuan, respectively, and then sets the transaction price p divided into four cases. In each case, the transaction fee f has multiple values. The specific assignments are shown in Table 7.

In the simulation experiment, the time stamp is complete and verifiable data, which indicates that there is data at a specific point in time. Each block in the blockchain contains a timestamp indicating the writing time of data, so the blocks are connected in order. This ensures that the information in the blockchain cannot be tampered with, and the longer it runs, the more difficult it will be to tamper with. In addition, it also enables events that occurred at a particular time in history to be traced and verified from the blockchain. A block hash is a digital signature of a block generated by a hash algorithm. This linkage mechanism prevents block information from being artificially tampered with and forged and enables the information in the block to be traced through the block hash. This paper combines the two methods of time stamp tracing based on the blockchain and hash tracing based on the blockchain, considering the time attribute of the block information and the one-to-one correspondence between the block information and the hash generated by the encryption algorithm.

Enterprises are competing fiercely, which creates new

development chances. Due to inadequate data analysis and antiquated technology, traditional financial management techniques cannot fairly represent an organization's worth. Financial staff become business partners instead of just passive data recipients due to the shift to value management brought about by the digital economy. Enterprise financial management mode optimization is the main subject of research. Finding and addressing essential areas for improvement is the goal of the study. The analysis focuses on the crucial challenge of dissecting the underlying workings of blockchain financial ledgers' 51% double flower assaults. Building on this analysis, the study sets the conditions for the smooth integration of blockchain technology and business financial information management. Acknowledging the need to closely examine potential weaknesses like the double flower attacks stated earlier, the study attempts to strengthen the integration process by thoroughly grasping its possible dangers and obstacles. This study uses the SQL server database and OLAP tool analysis to put the data and generate a 3D space, such as time, index, and enterprise dimension cubes, respectively. The study uses MATLAB simulation techniques to put the integration framework through a rigorous testing environment to assess the proposed model and its effectiveness. The outcomes of these simulations provide a crucial reference point that enables evaluation of the model's performance and resilience against possible dangers like double blossom attacks.

5 Case Study

5.1 Reasonableness Verification

This paper stores the original vouchers, accounting vouchers, and the information content in the financial statements in the blockchain after inputting and adds the corresponding information categories to the blockchain to store the blocks of financial statement data. The MPT tree root hash is added to the chain, and then each block is stamped with a timestamp to record the entry time of the information. Each timestamp corresponds to the content in the block one-to-one, and different blocks are connected in chronological order, as shown in Figure 10.

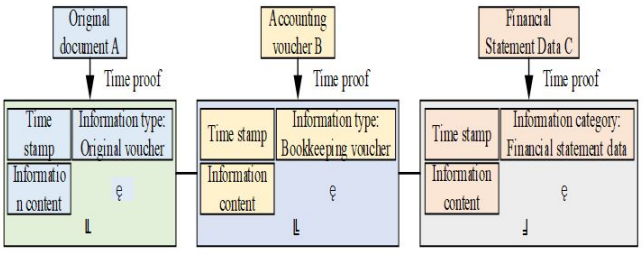


Figure 10: Results of the financial management blockchain simulation experiment

For the original voucher, the financial matters reflected on the actual voucher can be queried and traced through the time stamp so that the ins and outs of all economic businesses can be proved; for the accounting voucher, the time stamp is equivalent to automatically generating a serial number and a serial number for the accounting voucher. The compilation date saves the cost of manual numbering and sorting; for the financial statement data, each piece is located through the timestamp, thereby clarifying the source of the financial statement data.

5.2 Traceability Verification

Connects the blocks in sequence through the relationship between the block hashes, as shown in Figure 11. The block hash encodes the timestamp, MPT root hash, financial information category, and financial information content in the block, and the information in each block corresponds to a unique block hash value.

Then, based on the block hash, locate and trace the information category and content. For the original voucher, you can accurately understand each economic business, verify all the information in the voucher, and use it as an accounting voucher for auditing. For the accounting vouchers, you can check whether the content is consistent with the original vouchers, whether the listed accounting subjects, loan directions, and amounts are correct, whether the date, summary, and other information are filled in, and whether the accounting supervisor, bookkeeping personnel and other relevant responsible persons can be checked. Tracing and querying the corresponding single data under the specific subject of the financial statement is more accurate and efficient than the retrospective query of the entire report file [9].

5.3 Security Verification

The demanders of the enterprise financial information traceability function mainly include the internal enterprise, relevant enterprise stakeholders, and the audit department. Among them, the relevant departments within the enterprise should supervise its financial status in real time and make appropriate management decisions. Relevant stakeholders mainly include investors and creditors, who are required to ensure the authenticity and reliability of financial statements and to understand the ability to maintain and increase the value of corporate capital and the solvency of the company. The auditing department includes the internal auditing department of the enterprise, the government auditing agency, and the social auditing agency, which is responsible for checking the financial affairs, financial revenue and expenditure, and operation and management activities of the enterprise and holding them accountable for related illegal acts. This paper builds a security traceability system for corporate financial information, connects the corporate financial department, the blockchain that stores financial information, and the demander of the financial information traceability function in one system, and provides real-time financial data based on timestamp block hashes. The tasks

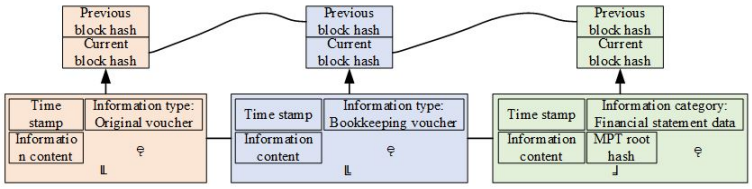


Figure 11: Schematic diagram of enterprise financial information blockchain based on business and economic integration

of query and traceback are shown in Figure 12.

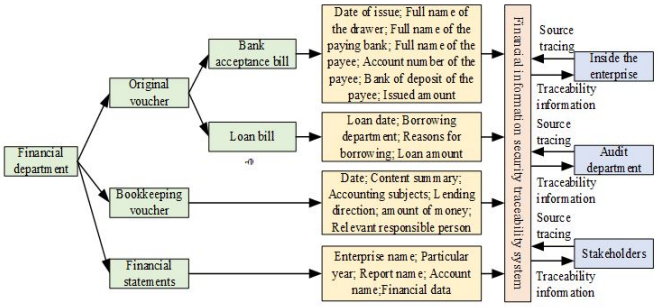


Figure 12: Safety traceability system verification

First of all, after completing the filling and review of the original vouchers, accounting vouchers, financial statements, and other materials, the financial department of the enterprise enters the financial information into the blockchain network. All the information in the blockchain automatically enters the traceability system of the enterprise's financial information. The internal audit department and relevant stakeholders of the enterprise can send traceability requests to the financial information in the blockchain. Specifically, the stored time stamp and block hash are used as the tracing source code to send a request to the secure tracing system of financial information, and the system background will automatically trace back to the corresponding block, providing the person who sends the tracing request with the financial information category and financial information content in the block. For example, suppose the audit department finds any problems during the audit of the accounts. In that case, it can quickly start tracing with the time stamp or block hash of the block where the accounts in question are located as the tracing source, trace the person in charge of the accounts, investigate relevant economic matters, and carry out targeted accountability.

6 Conclusion

The financial management mode of recording and supervising enterprise business based on data accounting can no longer meet the needs of enterprise development. Under the new situation, financial personnel must go deep into the business process and expand from ex-post reflection to ex ante prediction and in-process supervision. Finance should identify and control risks from the front end of the business, ensure operation quality in the business process, and conduct data analysis afterward to support the enterprise's further plans. Under the new situation, it is necessary to break the communication barriers between departments, take products as the main route, take customers as the center, cooperate, support each other, and help the healthy development of enterprises. A multi-layer granular structure model is built based on a rough set for the demand of enterprise financial report data based on the MPT tree. Then, the improved MPT tree completes the accurate storage of report data. To establish a trusted traceability system of enterprise financial information based on time stamps and block hashes, the traceability mode of enterprise financial information is designed based on time stamps and block hashes, respectively.

References

[1] F. Alexander, A. Almgren, J. Bell, A. Bhattacharjee, J. Chen, P. Colella, D. Daniel, J. DeSlippe, L. Diachin, E. Draeger, and A. Dubey, "Exascale applications: skin in the game," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, p. 20190056, 2020.

[2] X. Bao, "Construction of financial management system model based on internet technology," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–10, 2022.

[3] T. Berger, J. P. Steghöfer, T. Ziadi, J. Robin, and J. Martinez, "The state of adoption and the challenges of systematic variability management in industry," *Empirical Software Engineering*, vol. 25, no. 3, pp. 1755–1797, 2020.

[4] D. Carlucci, P. Renna, S. Materi, and G. Schiuma, "Intelligent decision-making model based on minority game for resource allocation in cloud manufacturing," *Management decision*, vol. 58, no. 11, pp. 2305–2325, 2020.

[5] L. Du, Y. Feng, W.Lu, L. Kong, and Z. Yang, "Evolutionary game analysis of stakeholders' decision-making behaviors in construction and demolition waste management," *Environmental Impact Assessment Review*, vol. 84, p. 106408, 2020.

[6] Y. Du, H. Zhou, Y. Yuan, and H. Xue, "Exploring the moral hazard evolutionary mechanism for bim

implementation in an integrated project team," *Sustainability*, vol. 11, no. 20, p. 5719, 2019.

[7] A. Gorkhali, L. Li, and A. Shrestha, "Blockchain: A literature review," *Journal of Management Analytics*, vol. 7, no. 3, pp. 321–343, 2020.

[8] A. P. G.Scheidegger, T. F.Pereira, M. L. M. de Oliveira, A. Banerjee, and J. A. B.Montevechi, "An introductory guide for hybrid simulation modelers on the primary simulation methods in industrial engineering identified through a systematic review of the literature," *Computers & Industrial Engineering*, vol. 124, pp. 474–492, 2018.

[9] J. Leng, G. Ruan, P. Jiang, K. Xu amd Q. Liu, X. Zhou, and C. Liu, "Blockchain-empowered sustainable manufacturing and product lifecycle management in industry 4.0: A survey," *Renewable and sustainable energy reviews*, vol. 132, p. 110112, 2020.

[10] G. Li, "Research on innovation of enterprise management accounting informatization platform based on intelligent finance," in *In 1st International Symposium on Economic Development and Management Innovation (EDMI 2019)*, pp. 286–291, Atlantis Press, August 2019.

[11] Q. Li, "Analytical study of financial accounting and management trends based on the internet era," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, 2022.

[12] X. Li, J. Li, Y. Huang, J. He, X. Liu, J. Dai, and Q. Shen, "Construction enterprises' adoption of green development behaviors: An agent-based modeling approach," *Humanities and Social Sciences Communications*, vol. 9, no. 1, pp. 1–11, 2022.

[13] Y. Li, "Integration analysis of enterprise cost accounting management and audit based on big data information technology," in *In 2022 3rd International Conference on Big Data and Social Sciences (ICBDSS 2022)*, pp. 1133–1140, Atlantis Press, December 2022.

[14] K. Y. H. Lim, P. Zheng, and C. H. Chen, "A state-of-the-art survey of digital twin: techniques, engineering product lifecycle management and business innovation perspectives," *Journal of Intelligent Manufacturing*, vol. 31, no. 6, pp. 1313–1337, 2020.

[15] A. F. Mashaly and A.G. Fernald, "Identifying capabilities and potentials of system dynamics in hydrology and water resources as a promising modeling approach for water management," *Water*, vol. 15, no. 5, p. 1432, 2020.

[16] T.D. Phan, E. Bertone, and R.A. Stewart, "Critical review of system dynamics modelling applications for water resources planning and management," *Cleaner Environmental Systems*, vol. 2, p. 100031, 2021.

[17] T. Rebs, M. Brandenburg, and S.Seuring, "System dynamics modeling for sustainable supply chain management: A literature review and systems thinking approach," *Journal of cleaner production*, vol. 208, pp. 1265–1280, 2019.

[18] A. Sharma, N.P. Rana, and R. Nunkoo, "Fifty years of information management research: A conceptual structure analysis using structural topic modeling," *International Journal of Information Management*, vol. 58, p. 102316, 2021.

[19] C. Sillaber, B. Waltl, H. Treiblmaier, U. Gallersdörfer, and M. Felderer, "Does whistleblowing work for air pollution control in china? a study based on three-party evolutionary game model under incomplete information," *Sustainability*, vol. 11, no. 2, p. 324, 2019.

[20] C. Sillaber, B. Waltl, H. Treiblmaier, U. Gallersdörfer, and M. Felderer, "Laying the foundation for smart contract development: an integrated engineering process model," *Information Systems and e-Business Management*, vol. 19, no. 3, pp. 863–882, 2021.

[21] F. Taghikhah, A. Voinov, N. Shukla, T. Filatova, and M. Anufriev, "Integrated modeling of extended agro-food supply chains: A systems approach," *European journal of operational research*, vol. 288, no. 3, pp. 852–868, 2021.

[22] A. Talwariya, P. Singh, and M. Kolhe, "A stepwise power tariff model with game theory based on monte-carlo simulation and its applications for household, agricultural, commercial and industrial consumers," *International Journal of Electrical Power & Energy Systems*, vol. 111, pp. 14–24, 2019.

[23] G. Wang, G. Zhang, X. Guo, and Y. Zhang, "Digital twin-driven service model and optimal allocation of manufacturing resources in shared manufacturing," *Journal of Manufacturing Systems*, vol. 59, pp. 165–179, 2021.

[24] M. Wang, Y.Li, Z. Cheng, C. Zhong, and W. Ma, "Evolution and equilibrium of a green technological innovation system: Simulation of a tripartite game model," *Journal of Cleaner Production*, vol. 278, p. 123944, 2021.

[25] Q. Wang, L. Kong, J. Li, B. Li, and F.Wang, "Behavioral evolutionary analysis between the government and uncertified recycler in china's e-waste recycling management," *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, p. 7221, 2020.

[26] Z. Wang, Q. Wang, B. Chen, and Y.Wang, "Evolutionary game analysis on behavioral strategies of multiple stakeholders in e-waste recycling industry," *Resources, Conservation and Recycling*, vol. 115, p. 104618, 2020.

[27] L. Xu, Z. Di, and J. Chen, "Evolutionary game of inland shipping pollution control under government co-supervision," *Marine Pollution Bulletin*, vol. 171, p. 112730, 2021.

[28] D. M. Yazan, V. Yazdanpanah, and L. Fraccascia, "Learning strategic cooperative behavior in industrial symbiosis: A game-theoretic approach integrated with agent-based simulation," *Business strategy and the environment*, vol. 29, no. 5, pp. 2078–2091, 2020.

[29] W. Ye, "Intelligent decision-making model based on minority game for resource allocation in cloud manufacturing," in *In 1st International Conference on Business, Economics, Management Science (BEMS 2019)*, pp. 265–268, Atlantis Press, January 2019.

[30] J. Zhou, J. Sun, W. Zhang, and Z. Lin, "Multi-view underwater image enhancement method via embedded fusion mechanism," *Engineering Applications of Artificial Intelligence*, vol. 121, p. 105946, 2023.

[31] M. Zomorodian, S.H. Lai, M. Homayounfar, S. Ibrahim, S. E.Fatemi, and A. El-Shafie, "The state-of-the-art system dynamics application in integrated water resources modeling," *Journal of environmental management*, vol. 227, pp. 294–304, 2018.

Biography

Tang Min was Graduated from Shandong Normal University. Tang received Undergraduate and Master of Fudan University. He is a lecture with Wuxi Vocational Institute of Commerce. His research directions are accounting and auditing.

Cryptanalysis and Improvement of a Fast Hash Family for Memory Integrity

Chengbo Xu[1] and Shuying Yang[2]
(Corresponding author: Chengbo Xu)
School of Mathematical Sciences, University of Jinan[1]
No. 336, Nanxinzhuang West Road, Jinan 250022, Shandong, P. R. China
School of Data and Computer Science, Shandong Women's University[2]
No. 2399, University Road, Jinan 250300, Shandong, P. R. China
Email: cbqysy@163.com
(Received July 27, 2023; Revised and Accepted Nov. 22, 2023; First Online June 22, 2024)

Abstract

Universal hash functions are important building blocks for unconditionally secure message authentication codes. Recently, Li and Sovio proposed a family of matrix polynomial hash functions for memory integrity that extends the classical polynomial hash functions. In this paper, we provide a cryptanalysis of the family of hash functions and demonstrate that the family does not satisfy universal and balanced properties as the authors claimed. We also present an attack strategy to show that it is easy to construct a collision pair with probability 1. To avoid the attack, we propose an improved family of matrix polynomial hash functions and show that it is ϵ-Δ-universal and ϵ-balanced. Furthermore, by experiments, we demonstrate that the improved hash algorithm performs better than Li and Sovio's hash with respect to security. However, a little efficiency loss is taken as the cost.

Keywords: Hash Function; Message Authentication; Universal Hashing

1 Introduction

Universal hash functions, introduced in the seminal paper by Carter and Wegman [9], replace the unfounded assumption of the random choice of the input set by a random choice of the hash function from a suitable family of functions that gives for any input set a good distribution of the assigned values [8]. Later, Stinson developed the concept further and proposed several relevant definitions of various types of hash families, such as ϵ-universal hash family, ϵ-Δ-universal hash family, and ϵ-strongly universal hash family, etc. [15].

Over the years, universal hash functions have many applications in cryptography, error-correcting codes, complexity theory, randomized algorithms, data structures, etc. [3–6, 10, 13, 16]. A direct application of universal hash functions is to build an unconditionally secure MAC which can be constructed by first applying a universal hash function on the input messages and encrypting the result. MACs of this kind have been standardized in ISO/IEC 9797-3-2011/Amd 1-2020 [1] which includes UMAC [7], Badger [14], Poly1305-AES [2] and GMAC [14].

Meanwhile, various universal hash function families are constructed [4, 5, 9, 11, 13, 15–18]. One of the most widely used universal hash function families is polynomial hash functions [9], such as Galois/Counter Mode [12] used in IPsec, SSH and TLS, and Poly 1305 [2] used in Google Chrome's TLS and OpenSSL. Recently, Bibak generalized the polynomial hash functions to multivariate polynomial hash functions and applied them to construct quantum key distribution [3]. In another direction, Li and Sovio propose a family of matrix polynomial hash functions for memory integrity that also extend the classical polynomial hash functions [13].

Contributions. In this paper, we cryptanalyse and improve the family of matrix polynomial hash functions proposed by Li and Sovio [13].

1) We first point out a flaw in the proof which try to show that the family of matrix polynomial hash functions is ϵ-Δ-universal and ϵ-balanced.

2) Based on the flaw, we design an attack executed to break the security of Li and Sovio's hash functions.

3) In order to avoid the attack, we propose an improved family of matrix polynomial hash functions, and show that is ϵ-Δ-universal and ϵ-balanced.

2 Preliminaries

Notation. We use $\mathbb{Z}$ for the set of integers and $\mathbb{Z}_n$ for the ring of integers modulo n defined as $\mathbb{Z}_n = \{0, \ldots, n-1\}$.

Elements of these sets are denoted by lowercase letters. We use uppercase letters in bold font (such as $\mathbf{M}$) to denote matrices, lowercase letters in bold font (such as $\mathbf{v}$) to denote vectors. Matrices enclosed by square brackets, such as $[\mathbf{M_1}\ \mathbf{M_2}]$, refer to an augmented matrix formed by horizontally joining the columns of $\mathbf{M_1}$ and $\mathbf{M_2}$, assuming the number of rows in these two matrices are the same. Matrices (or vectors) enclosed by parenthesis, such as $(\mathbf{v_1}, \mathbf{v_2})$, refer to an matrix or vector formed by vertically joining the rows of $\mathbf{v_1}$ and $\mathbf{v_2}$. For a set S, we write $s \leftarrow S$ to denote that s is chosen uniformly at random from S.

2.1 Universal Hash Functions

In this subsection, we review several related definitions of universal hash functions.

Definition 1. (Keyed hash functions) *Let $\mathcal{K}$, $\mathcal{M}$ and $\mathcal{T}$ be finite, non-empty sets. A keyed hash function $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ is a function that takes two inputs: a key k and a message m, and outputs a digest $t = H(k, x)$, where $\mathcal{K}$, $\mathcal{M}$ and $\mathcal{T}$ are called keyspace, message space and digest space, respectively.*

Definition 2. (Universal hash functions) *Let $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ be a keyed hash function. H is universal if for any two distinct $x, y \in \mathcal{M}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) = H(k, y)] \leq \frac{1}{|\mathcal{T}|}$. Furthermore, H is an ϵ-almost universal (ϵ-AU) hash function if for any two distinct $x, y \in \mathcal{M}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) = H(k, y)] \leq \epsilon$.*

Definition 3. (Δ-Universal hash functions) *Suppose $\mathcal{T}$ is a finite additive Abelian group. Let $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ be a keyed hash function. H is Δ-universal if for any two distinct $x, y \in \mathcal{M}$ and all $b \in \mathcal{T}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) - H(k, y) = b] = \frac{1}{|\mathcal{T}|}$, where '$-$' denotes the group subtraction operation. Furthermore, H is an ϵ-almost-Δ-universal (ϵ-AΔU) hash function if for any two distinct $x, y \in \mathcal{M}$ and all $b \in \mathcal{T}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) - H(k, y) = b] \leq \epsilon$. When $\mathcal{T} = \mathbb{Z}_2^n = \{0, 1\}^n$ for some integer n, the operation '$-$' can be replaced by '$\oplus$' (XOR), and H is also called ϵ-almost XOR universal hash function.*

Remark 1. By Definitions 2 and 3, it's easy to see that Δ-Universal hash functions are also universal hash function, ϵ-AΔU hash functions are also ϵ-AU hash functions since $\mathcal{T}$ is a finite additive Abelian group in Definition 3 and b could be equal to 0.

Definition 4. *Let $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ be a keyed hash function. H is ϵ-balanced if for any nonzero $x \in \mathcal{M}$ and any $y \in \mathcal{T}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) = y] \leq \epsilon$.*

2.2 Polynomial Hash Functions

Universal hash functions constructed using polynomials modulo a prime is widely attributed to Wegman and Carter [9]. In this subsection, we review this construction.

Definition 5. *Given an integer n and a prime p. Let $\mathcal{K} = \mathbb{Z}_p$, $\mathcal{M} = \mathbb{Z}_p^{d+1}$, and $\mathcal{T} = \mathbb{Z}_p$. Define hash function $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ as*

$$H(k, m) = \sum_{i=0}^{d} m_i k^i (mod p),$$

for every message $\mathbf{m} = (m_0, m_1, \ldots, m_d) \in \mathcal{M}$ and every key $k \in \mathbb{Z}_p$.

Lemma 1. *[3] The hash function defined above is $\frac{d}{p}$-almost-Δ-universal.*

3 Li and Sovio's Hash Family for Memory Integrity

In this section, we revisit the fast hash family for memory integrity proposed by Li and Sovio [13]. In detail, the authors first proved that a natural matrix correspondence (Definition 6) of polynomial hash functions with fixed length inputs is $\frac{1}{p^n-1}$-almost-Δ-universal and $\frac{1}{p^n-1}$-balanced. Then, based on this, they proposed a variable size imputs hash family (Definition 7) and tried to prove it is also $\frac{1}{p^n-1}$-almost-Δ-universal and $\frac{1}{p^n-1}$-balanced.

Definition 6. *Given integers n, r and a prime p. Let $\mathcal{K} = \{$all n-by-n invertible matrices over finite field $\mathbb{Z}_p\}$, $\mathcal{M} = \mathbb{Z}_p^{nr}$ and $\mathcal{T} = \mathbb{Z}_p^n$. Define hash function $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ as*

$$H_r(\mathbf{K}, \mathbf{m}) = [\mathbf{K}^r\ \mathbf{K}^{r-1}\ \ldots\ \mathbf{K}]\mathbf{m}$$

where the multiplications and additions are done over $\mathbb{Z}_p$.

If we parse the messge vector $\mathbf{m}$ as r length-n blocks and write $\mathbf{m} = (\mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_r})$, then the hash function defined above is equivalently represented by

$$H_r(\mathbf{K}, \mathbf{m}) = \mathbf{K}^r\mathbf{m_1} + \mathbf{K}^{r-1}\mathbf{m_2} + \cdots + \mathbf{K}\mathbf{m_r}$$

The representation is very similar in form to the representation in Definition 5. However, the proof of the universal property of the hash function is far less straightforward and easy than the proof of the universal property of the function defined in Definition 5. We only repeat the conclusion here. See the original for proof.

Lemma 2. *The hash function defined in Definition 6 is $\frac{1}{p^n-1}$-almost-Δ-universal and $\frac{1}{p^n-1}$-balanced.*

Definition 7. *Given integers $n \geq 1$, a prime p and a constant nonzero initial state vector $\mathbf{s_0} \in \mathbb{Z}_p^n$. Let $\mathcal{K} = \{$all n-by-n invertible matrices over finite field $\mathbb{Z}_p\}$, $\mathcal{M} = \mathbb{Z}_p^{n*}$ and $\mathcal{T} = \mathbb{Z}_p^n$. Define hash function $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ such that, for any message $\mathbf{m} \in \mathcal{M}$, let $r = |\mathbf{m}|_n$ be the number of blocks in m, the hash value of $\mathbf{m}$ is given by*

$$H(\mathbf{K}, \mathbf{m}) = [\mathbf{K}^r\ \mathbf{K}^{r-1}\ \ldots\ \mathbf{K}]\mathbf{m} + \mathbf{K}^r\mathbf{s_0}$$

where the multiplications and additions are done over $\mathbb{Z}_p$.

If we parse the messge vector $\mathbf{m}$ as r length-n blocks and write $\mathbf{m} = (\mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_r})$, then the hash function defined above is equivalently represented by

$$H(\mathbf{K}, \mathbf{m}) = \mathbf{K}^r(\mathbf{m_1} + \mathbf{s_0}) + \mathbf{K}^{r-1}\mathbf{m_2} + \cdots + \mathbf{K}\mathbf{m_r}$$

See Algorithm 1 for a fast pseudo-code implementation using Horner's rule.

Algorithm 1 Li and Sovio's Hash Functions

Input:
initial vector $\mathbf{s_0}$, key $\mathbf{K}$ and message $\mathbf{m}$

Output:

1: $\mathbf{s} \leftarrow \mathbf{s_0}$;
2: for $i = 1, 2, \ldots, r$ do
3: $\mathbf{s} \leftarrow \mathbf{K}(\mathbf{s} + \mathbf{m_i})$
4: end for
5: **return** $\mathbf{s}$

Theorem 1. *The hash function defined in Definition 7 is $\frac{1}{p^n-1}$-almost-Δ-universal and $\frac{1}{p^n-1}$-balanced.*

For the convenience of analysis, the proof of this theorem is restated as follows.

Proof. Since $\mathbf{K}$ is invertible, $\mathbf{v} = \mathbf{K^{-1}s_0}$ is a nonzero block. For any r-block message $\mathbf{x}$, let $\hat{\mathbf{x}} = (\mathbf{v}, \mathbf{x})$ be the $(r+1)$-block vector obtained by prepending $\mathbf{v}$ to $\mathbf{x}$, we have

$$\begin{aligned} H(\mathbf{K}, \mathbf{x}) &= [\mathbf{K}^r\ \mathbf{K}^{r-1}\ \ldots\ \mathbf{K}]\mathbf{x} + \mathbf{K}^{r+1}\mathbf{v} \\ &= [\mathbf{K}^{r+1}\ \mathbf{K}^r\ \mathbf{K}^{r-1}\ \ldots\ \mathbf{K}](\mathbf{v}, \mathbf{x}) \\ &= H_{r+1}(\mathbf{K}, \hat{\mathbf{x}}) \end{aligned}$$

where H_{r+1} is as defined in Definition 6.

Hence, for any $\mathbf{y} \in \mathbb{Z}_\mathbf{p}^\mathbf{n}$, since $\hat{\mathbf{x}} \neq \mathbf{0}$, by Lemma 2, we have

$$Pr[H(\mathbf{K}, \mathbf{x}) = \mathbf{y}] = Pr[H(\mathbf{K}, \hat{\mathbf{x}}) = \mathbf{y}] \leq \frac{1}{p^n - 1}$$

Furthermore, without loss of generality, let $w \in \mathbb{Z}_p^{sn}$ be another s-block message where $1 \leq s \leq r$, such that either $\mathbf{w}$ is a shorter message(i.e., $s < r$), or $\mathbf{x}$ and $\mathbf{w}$ are of the same length ($s = r$) and $\mathbf{x} \neq \mathbf{w}$.

Now, let $\hat{\mathbf{w}} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{v}, \mathbf{w})$ be the result of prepending $r - s$ zero blocks (if $s < r$) and $\mathbf{v}$ to $\mathbf{w}$. In this case, $\hat{\mathbf{x}}$ and $\hat{\mathbf{w}}$ are of the same length $(r+1)n$, and $\hat{x} \neq \hat{w}$. Also, by construction $H(\mathbf{K}, \mathbf{w}) = H_{r+1}(\mathbf{K}, \hat{\mathbf{w}})$. Therefore, from Lemma 2, we have

$$\begin{aligned} Pr[H(\mathbf{K}, \mathbf{x}) = H(\mathbf{K}, \mathbf{w})] &= Pr[H_{r+1}(\mathbf{K}, \hat{\mathbf{x}}) = H_{r+1}(\mathbf{K}, \hat{\mathbf{w}})] \\ &\leq \frac{1}{p^n - 1} \end{aligned}$$

Hence, the hash family H is $\frac{1}{p^n-1}$-balanced and $\frac{1}{p^n-1}$-almost-Δ-universal. □

4 Cryptanalysis of Li and Sovio's Hash Functions

In this section, we point out a flaw in the proof of Theorem 1. Furthermore, based on the flaw, we design an attack executed to break the security of Li and Sovio's Hash Functions.

4.1 Flaw in Li and Sovio's Hash Functions

By Definition 4, a hash function H is called ϵ-balanced if for any nonzero $x \in \mathcal{M}$ and any $y \in \mathcal{T}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) = y] \leq \epsilon$, where k is uniformly chosen from $\mathcal{K}$ at random and both x and y are independent from the chosen of k.

However, in the proof of Theorem 1, $\hat{\mathbf{x}} = (\mathbf{v}, \mathbf{x})$ where $\mathbf{v} = \mathbf{K^{-1}s_0}$ is determined by $\mathbf{K}$ and $\mathbf{s_0}$ and therefore is different depending on the choice of $\mathbf{K}$. Hence, the message $\hat{\mathbf{x}}$ constructed like this does not satisfy the premise of the definition of ϵ-balanced hash functions, the inequality

$$Pr[H(\mathbf{K}, \hat{\mathbf{x}}) = \mathbf{y}] \leq \frac{1}{p^n - 1}$$

cannot be deduced from Lemma 2.

Similarly, according to Definition 3, a hash function H is called ϵ-almost-Δ-universal hash function if for any two distinct $x, y \in \mathcal{M}$ and all $b \in \mathcal{T}$, we have $Pr_{k \leftarrow \mathcal{K}}[H(k, x) - H(k, y) = b] \leq \epsilon$, where k is uniformly chosen from $\mathcal{K}$ at random and both x and y are independent from the chosen of k.

However, in the proof of Theorem 1, $\hat{\mathbf{w}} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{v}, \mathbf{w})$ where $\hat{\mathbf{w}}$ depends on $\mathbf{v}$ and further depends on $\mathbf{K}$. Hence, both the message $\hat{\mathbf{x}}$ and $\hat{\mathbf{w}}$ constructed like this does not satisfy the premise of the definition of the ϵ–almost-Δ-universal hash functions, the inequality

$$Pr[H_{r+1}(\mathbf{K}, \hat{\mathbf{x}}) = H_{r+1}(\mathbf{K}, \hat{\mathbf{w}})] \leq \frac{1}{p^n - 1}$$

also cannot be deduced from Lemma 2.

4.2 Attack on Li and Sovio's Hash Functions

The flaw illustrated above just means that the proof of Theorem 2 does not prove the theorem. Is the theorem true or not? We must either propose a correct proof, or give a counterexample. In this subsection, we design a pair of collision messages that shows Li and Sovio's Hash Functions is neither $\frac{1}{p^n-1}$-almost-Δ-universal nor $\frac{1}{p^n-1}$-balanced.

Let $\mathbf{m} = (\mathbf{m_1}, \mathbf{m_2})$ and $\mathbf{m}' = (-\mathbf{s_0}, \mathbf{m_1} + \mathbf{s_0}, \mathbf{m_2})$ where $\mathbf{m_1}$ $\mathbf{m_2}$ are two arbitrary length-n blocks and $\mathbf{s_0}$ is a constant nonzero initial state vector. Obviously, $\mathbf{m} \neq \mathbf{m}'$.

But we have

$$\begin{aligned} H(\mathbf{K},\mathbf{m}) &= [\mathbf{K}^2 \ \ \mathbf{K}]\mathbf{m} + \mathbf{K}^2\mathbf{s_0} \\ &= \mathbf{K}^2(\mathbf{m_1}+\mathbf{s_0}) + \mathbf{K}\mathbf{m_2} \\ H(\mathbf{K},\mathbf{m}') &= [\mathbf{K}^3 \ \ \mathbf{K}^2 \ \ \mathbf{K}]\mathbf{m}' + \mathbf{K}^3\mathbf{s_0} \\ &= \mathbf{K}^3(-\mathbf{s_0}+\mathbf{s_0}) + \mathbf{K}^2(\mathbf{m_1}+\mathbf{m_0}) + \mathbf{K}\mathbf{m_2} \\ &= \mathbf{K}^2(\mathbf{m_1}+\mathbf{s_0}) + \mathbf{K}\mathbf{m_2} \end{aligned}$$

Therefore,

$$Pr[H(\mathbf{K},\mathbf{m}) = H(\mathbf{K},\mathbf{m}')] = 1 \geq \frac{1}{p^n-1}$$

Hence, Li and Sovio's hash functions are not $\frac{1}{p^n-1}$-almost-Δ-universal. Similarly, denote $\mathbf{y} = \mathbf{H}(\mathbf{K},\mathbf{m}')$. Then,

$$Pr[H(\mathbf{K},\mathbf{m}) = \mathbf{y}] = 1 \geq \frac{1}{p^n-1}$$

In this case, Li and Sovio's hash functions are also not $\frac{1}{p^n-1}$-balanced.

Remark 2. The messages $\mathbf{m} = (\mathbf{m_1},\mathbf{m_2})$ and $\mathbf{m}'' = \mathbf{K}(\mathbf{m_1}+\mathbf{s_0}) + \mathbf{m_2} - \mathbf{s_0}$ are easily verified to also satisfy $H(\mathbf{K},\mathbf{m}) = H(\mathbf{K},\mathbf{m}'')$. But these two messages do not form a collision pair because $\mathbf{m}''$ depends on $\mathbf{K}$.

5 Improvement of Li and Sovio's Hash Functions

In this section, we propose an improved hash family based on the construction of Li and Sovio's Hash functions, in order to avoid the flaw and attack illustrated above.

Definition 8. *Given integers $n \geq 1$, a prime p and a constant nonzero initial state vector $s_0 \in \mathbb{Z}_p^n$. Let $\mathcal{K} = \{$all n-by-n invertible matrices over finite field $\mathbb{Z}_p\}$, $\mathcal{M} = \mathbb{Z}_p^{n*}$ and $\mathcal{T} = \mathbb{Z}_p^n$. Define hash function $H : \mathcal{K} \times \mathcal{M} \to \mathcal{T}$ such that, for any message $\mathbf{m} \in \mathcal{M}$, let $r = |\mathbf{m}|_n$ be the number of blocks in m, the hash value of $\mathbf{m}$ is given by*

$$H(\mathbf{K},\mathbf{m}) = [\mathbf{K}^r \ \mathbf{K}^{r-1} \ \ldots \ \mathbf{K}]\mathbf{m} + \mathbf{K}^{r+1}\mathbf{s_0}$$

where the multiplications and additions are done over $\mathbb{Z}_p$.

If we parse the messge vector $\mathbf{m}$ as r length-n blocks and write $\mathbf{m} = (\mathbf{m_1},\mathbf{m_2},\ldots,\mathbf{m_r})$, then the hash function defined above is equivalently represented by

$$H(\mathbf{K},\mathbf{m}) = \mathbf{K}^{r+1}\mathbf{s_0} + \mathbf{K}^r\mathbf{m_1} + \mathbf{K}^{r-1}\mathbf{m_2} + \cdots + \mathbf{K}\mathbf{m_r}$$

See Algorithm 2 for a fast pseudo-code implementation using Horner's rule.

Theorem 2. *The hash function defined in Definition 8 is $\frac{1}{p^n-1}$-almost-Δ-universal and $\frac{1}{p^n-1}$-balanced.*

Proof. Let $\mathbf{m}' = (\mathbf{m'_1},\mathbf{m'_2},\ldots,\mathbf{m'_r})$ and $\mathbf{m}'' = (\mathbf{m''_1},\mathbf{m''_2},\ldots,\mathbf{m''_s})$ be two arbitrary distinct messages,

Algorithm 2 Our improved Hash Functions

Input:
initial vector $\mathbf{s_0}$, key $\mathbf{K}$ and message $\mathbf{m}$
Output:
1: $\mathbf{s} \leftarrow \mathbf{K}\mathbf{s_0}$;
2: for $i = 1, 2, \ldots, r$ do
3: $\mathbf{s} \leftarrow \mathbf{K}(\mathbf{s} + \mathbf{m_i})$
4: end for
5: **return** $\mathbf{s}$

where $\mathbf{m'_i}$ and $\mathbf{m''_j}$ are length-n blocks. By Definition 8, we have

$$\begin{aligned} H(\mathbf{K},\mathbf{m}') &= [\mathbf{K}^r \ \mathbf{K}^{r-1} \ \ldots \ \mathbf{K}]\mathbf{m}' + \mathbf{K}^{r+1}\mathbf{s_0} \\ &= \mathbf{K}^{r+1}\mathbf{s_0} + \mathbf{K}^r\mathbf{m'_1} + \mathbf{K}^{r-1}\mathbf{m'_2} + \cdots + \mathbf{K}\mathbf{m'_r} \\ H(\mathbf{K},\mathbf{m}'') &= [\mathbf{K}^s \ \mathbf{K}^{s-1} \ \ldots \ \mathbf{K}]\mathbf{m}'' + \mathbf{K}^{s+1}\mathbf{s_0} \\ &= \mathbf{K}^{s+1}\mathbf{s_0} + \mathbf{K}^s\mathbf{m''_1} + \mathbf{K}^{s-1}\mathbf{m''_2} + \cdots + \mathbf{K}\mathbf{m''_s} \end{aligned}$$

According to the distinction between $|\mathbf{m}'|$ and $|\mathbf{m}''|$, we can prove it in the following two cases.

Case 1: $|\mathbf{m}'| = |\mathbf{m}''|$, i.e. $r = s$.

since $r = s$, we have

$$\begin{aligned} H_{r+1}(\mathbf{K}, H(\mathbf{K},\mathbf{m}') &= H_{s+1}(\mathbf{K},\mathbf{m}'') \\ \iff H(\mathbf{K},\mathbf{m}') &= H(\mathbf{K},\mathbf{m}'') \end{aligned}$$

where H_{r+1} and H_{s+1} are defined in Definition 6. By Lemma 2,

$$\begin{aligned} & Pr[H(\mathbf{K},\mathbf{m}') = H(\mathbf{K},\mathbf{m}'')] \\ = \ & Pr[H_{r+1}(\mathbf{K},\mathbf{m}') = H_{s+1}(\mathbf{K},\mathbf{m}'')] \\ \leq \ & \frac{1}{p^n-1} \end{aligned}$$

Hence, the hash function is $\frac{1}{p^n-1}$-almost-Δ-universal. In a similar way, it can be easily obtained that the hash function is also $\frac{1}{p^n-1}$-balanced.

Case 2: $|\mathbf{m}'| \neq |\mathbf{m}''|$, i.e. $r \neq s$.

Without loss of generality, we assume that $r < s$. Then,

$$\begin{aligned} & H(\mathbf{K},\mathbf{m}'') - H(\mathbf{K},\mathbf{m}') \\ = \ & (\mathbf{K}^{s+1}\mathbf{s_0} + \mathbf{K}^s\mathbf{m''_1} + \mathbf{K}^{s-1}\mathbf{m''_2} + \cdots + \mathbf{K}\mathbf{m''_s}) \\ & - (\mathbf{K}^{r+1}\mathbf{s_0} + \mathbf{K}^r\mathbf{m'_1} + \mathbf{K}^{r-1}\mathbf{m'_2} + \cdots + \mathbf{K}\mathbf{m'_r}) \\ = \ & \mathbf{K}^{s+1}\mathbf{s_0} + \mathbf{K}^s\mathbf{m''_1} + \cdots + \mathbf{K}^{r+1}(\mathbf{m''_{s-r}} - \mathbf{s_0}) \\ & + \cdots + \mathbf{K}(\mathbf{m''_s} - \mathbf{m'_r}) \\ = \ & H_{s+1}(\mathbf{K},\mathbf{m}''') \end{aligned}$$

where $\mathbf{m}''' = (\mathbf{s_0},\mathbf{m''_1},\ldots,\mathbf{m''_{s-r}} - \mathbf{s_0},\ldots,\mathbf{m''_s} - \mathbf{m'_r})$ is a non-zero vector since the first coordinates $\mathbf{s_0}$ is non-zero. By Lemma 2, we have

$$\begin{aligned} & Pr[H(\mathbf{K},\mathbf{m}') = H(\mathbf{K},\mathbf{m}'')] \\ = \ & Pr[H(\mathbf{K},\mathbf{m}'' - H(\mathbf{K},\mathbf{m}') = 0] \\ = \ & Pr[H_{s+1}(\mathbf{K},\mathbf{m}''') = 0] \\ \leq \ & \frac{1}{p^n-1} \end{aligned}$$

Hence, the hash function is $\frac{1}{p^n-1}$-almost-Δ-universal. Smilarly, it can be easily proved that the hash function is also $\frac{1}{p^n-1}$-balanced. □

Remark 3. The hash functions defined in Definitions 6 and 8 are very similar in form. The latter has been shown to be a $\frac{1}{p^n-1}$-almost-Δ-universal hash family for variable size inputs in Theorem 2. However, the former can only be proven to be a $\frac{1}{p^n-1}$-almost-Δ-universal hash family for fixed length messages, not for variable length inputs. For example, the two messages $\mathbf{m}' = (\mathbf{m_1}, \mathbf{m_2})$ and $\mathbf{m}'' = (\mathbf{0}, \mathbf{m_1}, \mathbf{m_2})$ are a collision for the hash functions defined in Definition 6 under all keys $\mathbf{K}$.

6 Experiments

In this section, we compare and analyze Li and Sovio's hash algorithm and our improved algorithm in two aspects of security and performance through experiments. Our experiments are done on a Windows desktop PC with an Intel Core i5-7200U CPU running at 2.50 GHz. The algorithms are implemented in Matlab R2016b.

6.1 Security Evaluation

In Subsection 4.2, a pair of collision messages are constructed to show Li and Sovio's hash function is neither $\frac{1}{p^n-1}$-almost-Δ-universal nor $\frac{1}{p^n-1}$-balanced. Here, we verify experimentally that the constructed messages is indeed a pair of collisions with probability 1. In contrast, the pair of messages occur collision with probability only $\frac{1}{p^n-1}$ in our improved hash algorithm, which shows that the improved algorithm is $\frac{1}{p^n-1}$-almost-Δ-universal nor $\frac{1}{p^n-1}$-balanced and therefore more secure. For simplicity, the prime p is always set to 2 in our experiments.

Without loss of generality, we randomly generate two length-n blocks $\mathbf{m_1}$ $\mathbf{m_2}$ and concatenate them to construct length-$2n$ message $\mathbf{m} = (\mathbf{m_1}, \mathbf{m_2})$. Then, in the manner described in Subsection 4.2, we construct the message $\mathbf{m}' = (-\mathbf{s_0}, \mathbf{m_1} + \mathbf{s_0}, \mathbf{m_2})$ where $\mathbf{s_0}$ is a constant nonzero initial state vector which also generated randomly in our implementation.

In both Figures 1 and 2, we fix the message length n and generate $\mathbf{m}$ and $\mathbf{m}'$ discribed as above, and then calculate the collision frequency of $\mathbf{m}$ and $\mathbf{m}'$ using Li and Sovio's hash algorithm and our improved algorithm, respectively. Figure 1 illustrates the case of $n = 5$. Obviously, in the basic hash algorithm (i.e. Li and Sovio's algorithm), the collision frequency of $\mathbf{m}$ and $\mathbf{m}'$ is always identical to 1, that is, the collision occurs with probability 1. In contrast, in the improved algorithm, the collision frequency quickly converges to 0.2183. Figure 2 reflects the n=10 situation, which is very similar to the situation in Figure 1.The difference is that in the improved algorithm, the collision frequency rapidly converges to 0.0195. Although the frequency convergence value obtained by the experiment is somewhat higher than the theoretical value in Theorem 2 due to the deviation of the algorithm used to generate the random matrix $\mathbf{K}$, it does not affect that our improved algorithm has higher security compared with Li and Sovio's hash algorithm, because the collision frequency in Li and Sovio's hash algorithm is always 1.

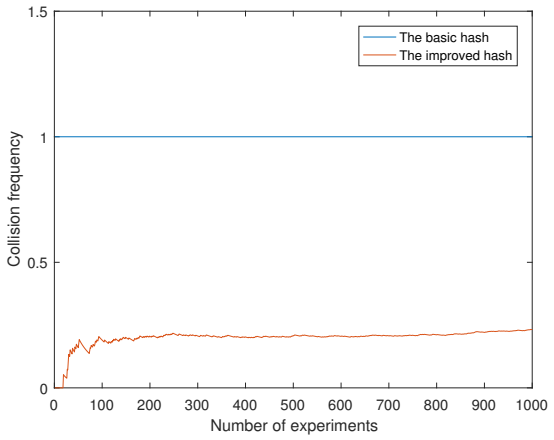


Figure 1. The security comparison between Li and Sovio's hash algorithm and our improved algorithm in case of $n = 5$

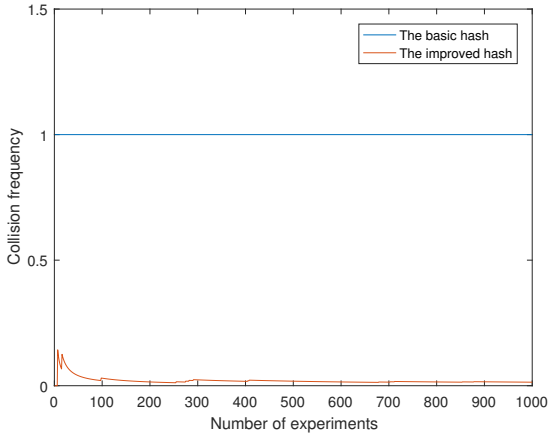


Figure 2. The security comparison between Li and Sovio's hash algorithm and our improved algorithm in case of $n = 10$

6.2 Performance Evaluation

To avoid complicating the evaluation process, the experiments are done by repeatedly hashing messages of different sizes and gathering the average processing time per megabyte (MB) of data. The messages are randomly generated and are always multiples of the block size n.

In Figures 3 to 5, we present a comparison of Li and Sovio's hash algorithm and our improved algorithm in terms of running time and average time by setting n=10, 100, 500, respectively.

As can be seen from these figures, The average running time of the improved algorithm is generally a little longer than that of Li and Sovio's hash algorithm, which

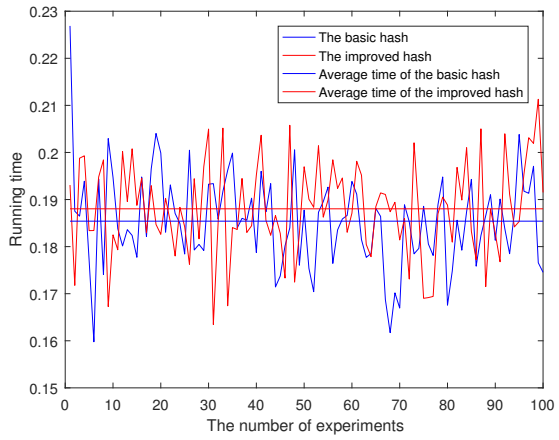


Figure 3. The performance comparison between Li and Sovio's hash algorithm and our improved algorithm in case of $n = 10$

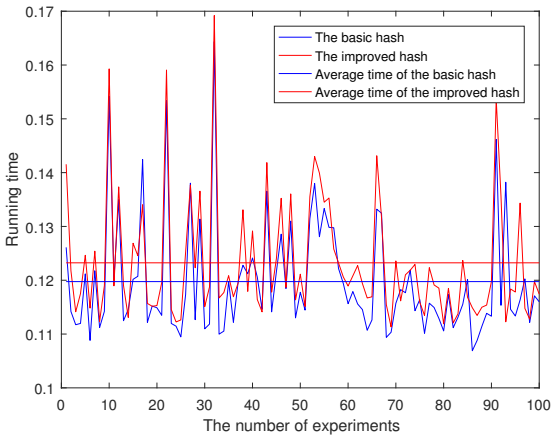


Figure 5. The performance comparison between Li and Sovio's hash algorithm and our improved algorithm in case of $n = 500$

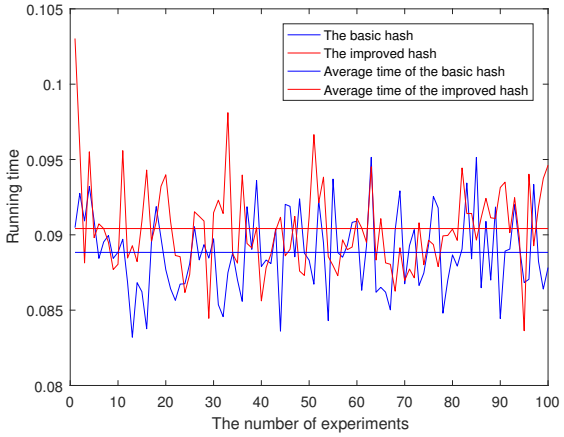


Figure 4. The performance comparison between Li and Sovio's hash algorithm and our improved algorithm in case of $n = 100$

is caused by the fact that in order to make up for the security defect of Li and Sovio's hash algorithm, our algorithm uses an extra matrix multiplication operation as the cost. This overhead, since it is one-time, will become insignificant as the length of the hashed message increases.

Another observation from these figures is that the throughput of the algorithm in Figure 4 is significantly better than that in Figures 3 and 5.This shows that there is an optimal value for the block length n, which is not about as large as possible or as small as possible, and needs to be selected in engineering practice.

7 Conclusions

In this paper, we provide a cryptanalysis of the family of hash functions proposed by Li and Sovio and demonstrate that the family of hash functions does not satisfy universal and balanced properties. Based on the observation, we present an attack strategy to show that it is easily to construct a collision pair. Finally, we propose an improved family of matrix polynomial hash functions, and show that is ϵ-Δ-universal and ϵ-balanced.

Acknowledgments

This work was supported by the Doctoral Fund of University of Jinan (Granted No. XBS1455), and National Science Foundation of Shandong Province (No. ZR2018LF006).

References

[1] "Iso/iec 9797-3-2011/amd 1-2020: Information technology - security techniques - message authentication codes (macs) - part 3: Mechanisms using a universal hash-function," 2020.

[2] D. J. Bernstein, "The poly1305-aes message authentication code," in *Proceedings of Fast Software Encryption (FSE'05)*, pp. 32–49, Heidelberg, 2005.

[3] K. Bibak, "Quantum key distribution using universal hash functions over finite fields," *Quantum Inf. Process*, vol. 21, no. 121, 2022.

[4] K. Bibak, B. M. Kapron, and V. Srinivasan, "Authentication of variable length messages in quantum key distribution," *EPJ Quantum Technol.*, vol. 9, no. 8, 2022.

[5] K. Bibak and R. Ritchie, "Quantum key distribution with prf(hash, nonce) achieves everlasting security," *Quantum Inf. Process*, vol. 20, no. 228, 2021.

[6] K. Bibak, R. Ritchie, and B. Zolfaghari, "Everlasting security of quantum key distribution with 1k-dwcdm and quadratic hash," *Quantum Inf. Comput.*, vol. 21, no. 3&4, pp. 181–202, 2021.

[7] J. Black, S. Halevi, H. Krawczyk, T. Krovetz, and P. Rogaway, "Umac: fast and secure message au-

thentication," in *Advances in Cryptology - CRYPT0 '99, LNCS 1666*, pp. 216–233, Heidelberg, 1999.

[8] P. Brass, "Universal hash functions for an infinite universe and hash trees," *Information Processing Letters*, vol. 109, p. 461–462, 2009.

[9] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *J. Comput. Syst. Sci.*, vol. 18, no. 2, pp. 143–154, 1979.

[10] M. Ciampi and Y. Xia, "Multi-theorem fiat-shamir transform from correlation-intractable hash functions," *IACR Cryptology ePrint Archive*, vol. 2023, no. 1677, 2023.

[11] A. G. Davoodi, S. Chang, H. G. Yoo, A. Baweja, M. Mongia, and H. Mohimani, "Forestdsh: a universal hash design for discrete probability distributions," *Data Mining and Knowledge Discovery*, vol. 35, no. 3, pp. 748–795, 2021.

[12] F. Grasselli, H. Kampermann, and D. Bruß, "Conference key agreement with single-photon interference," *New J. Phys.*, vol. 21, no. 123002, 2019.

[13] Q. Li and S. Sovio, "A fast hash family for memory integrity," *IACR Cryptology ePrint Archive*, vol. 2022, no. 1378, 2022.

[14] D. A. McGrew and J. Viega, "The galois/counter mode of operation (gcm)," in *http://csrc.nist.gov/groups/ST/toolkit/BCM/*, 2004.

[15] D. R. Stinson, "On the connections between universal hashing, combinatorial designs and error-correcting codes," *Congressus Numerantium*, vol. 114, pp. 7–27, 1996.

[16] D. R. Stinson, "Universal hash families and the leftover hash lemma, and applications to cryptography and computing," *J. Combin. Math. Combin. Comput.*, vol. 42, pp. 3–31, 2002.

[17] P. Wang, Y. Li, L. Zhang, and K. Zheng, "Related-key almost universal hash functions: Definitions, constructions and applications," in *Proceedings of Fast Software Encryption (FSE'16)*, pp. 514–532, 2016.

[18] S. Yang, "A new family of universal hash functions for quantum key distribution," *International Journal of Network Security*, vol. 25, no. 6, pp. 1059–1063, 2023.

Biography

Chengbo Xu received the B.S. degree in Mathematics from the Liaocheng University, China, in 2002, the M.S. degree in Cryptology from the Hubei University, China, in 2005, and the Ph.D. degree in Computer Science from the Beijing University of Post and Telecommunication, China, in 2014, respectively. Currently, he is a Lecture in the School of Mathimatical Sciences at University of Jinan. His research interests include information security and cryptology. Dr. Xu may be reached at cbqysy@163.com.

Shuying Yang received the B.S. and M.S. degree in Mathematics from Shandong Normal University, China, in 2004 and 2007, respectively. Currently, she is an associate professor in department of data and computer science at Shandong Women's University. Her research interests include information security and cryptology. Prof. Yang may be reached at ysystudy2005@163.com.

Image Encryption Based on Pixel Decomposition

Chunming Xu and Yong Zhang
(Corresponding author: Chunming Xu)
School of Mathematics and Statistics, Yancheng Teachers University
No.50, Kaifang Avenue, Yancheng 224002, China
Email:ycxcm@126.com
(Received Aug. 3, 2023; Revised and Accepted Apr. 20, 2024; First Online June 22, 2024)

Abstract

Decomposing an image into smaller elements for encryption can make the image encryption algorithm more complex, resulting in better encryption effects. This paper proposes an image encryption algorithm based on image decomposition. The proposed algorithm decomposes the pixel values of the plain image into smaller units and performs scrambling and diffusion operations based on them. For an image of size, $M \times N$, the algorithm first decomposes the image pixel values with a bit depth of 8 into two 4-bit binary bits, which are then converted into two integers between 0 and 15, so the original $M \times N$-sized image matrix is transformed into an $M \times 2N$-sized data matrix. Then, the data matrix is scrambled using the Zigzag transform and subjected to XOR diffusion operations. Finally, the scrambled matrix is reversely transformed into a matrix with $M \times N$ size, and another round of diffusion operation is performed to obtain the cipher image. Experiments were conducted on three classic images. Experiment results showed that the proposed algorithm has good security and high effective ness, meeting the requirements of secure transmission of images in networks.

Keywords: Chaotic System; Image Encryption; Image Scrambling; Pixel Decomposition; ZigZag Transform

1 Introduction

With the advent of the big data era and the rapid development of internet communication technology, various digital image data has exploded in growth and has been widely used in fields such as business, military, and medicine. On the internet, many images are often publicly visible. However, due to security or privacy reasons, we may want to transmit certain sensitive image information without making it public. Therefore, image encryption technology has emerged and has been the subject of extensive research. Image encryption technology transforms image data into a form similar to noise image, making it impossible for individuals without the decryption key to restore the original image content. This ensures the security of images during transmission or storage, providing effective protection for data security in various fields [1, 4, 5, 7, 8, 13].

Chaos exhibits characteristics such as ergodicity, nonperiodicity, high sensitivity to initial conditions and control parameters, and pseudo randomness. These characteristics make chaos very suited for image encryption. Chaos-based image encryption schemes have gained much recognition and have become mainstream methods for image encryption and many effective chaotic encryption algorithms have been proposed [2, 6, 12, 15, 18, 20].

In terms of implementation methods, image encryption algorithms can be divided into two types: pixel-based and bit-based encryption. Pixel-based image encryption algorithms operate directly on pixel values, while bit-based image encryption algorithms decompose image pixels into bit sequences and then scramble these bit sequences to achieve encryption. If the image is reconstructed using the scrambled bit streams, the pixel values of the image are also changed [17]. Therefore, bit-based encryption enhances the strength and security of encryption, making it more difficult to be decrypted.

Inspired by the above discussions, this paper proposes an image encryption method based on image pixel decomposition. Image pixel values are decomposed into two smaller units and then scrambling and diffusion operations are performed on them to obtain the encrypted image which will further be tested to evaluate the image encryption performance.

2 Fundamental Knowledge

2.1 Hindmarsh-Rose Neuron Model

In 2021, Zhang et al. presented a three-dimensional no-equilibrium Hindmarsh-Rose (HR) neuron model with memristive electromagnetic induction which has hidden homogeneous extreme multistability and is described

by [19]:

$$\begin{cases} \dot{x} = y - ax^3 + bx^2 + I + kcos(z)x \\ \dot{y} = c - dx^2 - y \\ \dot{z} = x \end{cases} \tag{1}$$

where x, y, z are state variables, and a, b, c, d, k, I are system parameters. When the system parameters are $a = 1, b = 3, c = 1, d = 5, k = 0.95, I = 1$, System (1) exhibits complex chaotic behavior. The state space plots for System (1) are shown in Figure 1.

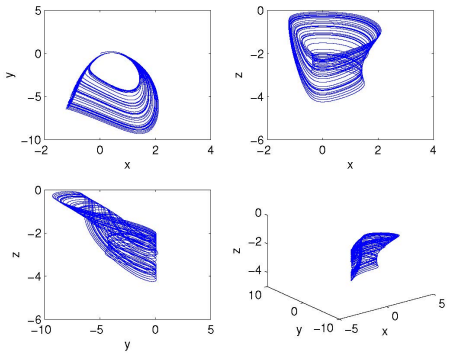


Figure 1: Typical dynamical behaviors of the Hindmarsh-Rose neuron model.

2.2 Image Pixel Decomposition

A pixel is the smallest unit of image and it is represented by a grayscale value which reflects the brightness or color information of that point. For grayscale images, the pixel depth is usually 8 bits, which means that the value of each pixel is typically between 0 and 255, where 0 represents black and 255 represents white. For color image, each pixel consists of three components: red, green, and blue. Similarly, the value of each component is typically between 0 and 255, representing the intensity of that pixel in the corresponding color channel. By combining the intensities of different channels, various colors can be obtained.

In this article, we decompose the 8-bit image pixel values into two binary bits of length 4, which are then converted into two integers between 0 and 16. For example, for a pixel value of 179, convert it into binary sequence 10110011 firstly then divide it into two binary sequences 1011 and 0011 with the length of 4 and finally convert them into decimal numbers 11 and 3 respectively. Assuming the width of the image is M and the height is N, we can convert it into an $M \times 2N$ matrix by pixel decomposition.

Similarly, we can provide the inverse transformation of pixel decomposition process. For example, if there are two adjacent elements 6 and 9 which can be transformed into binary numbers 0110 and 1001, connect 0110 and 1001 will result in a binary number 01101001 with a length of 8, and we can further convert the binary number 01101001 to obtain a decimal number 105.

2.3 Zigzag Transform

Zigzag transform is a classical method for scanning matrix elements and can be used for image scrambling [14]. The matrix elements are scanned in a "Z" shape order, and then the scanned elements are sequentially stored in a vector. Finally, the vector is transformed back into a two-dimensional matrix. The specific process of Zigzag transform is shown in Figure 2.

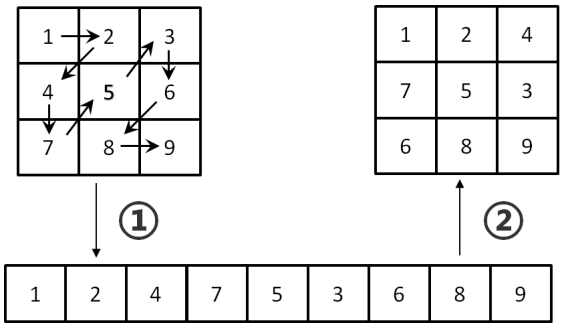


Figure 2: Zigzag Transform.

2.4 Image Pixel Decomposition based Zigzag Transform

To illustrate the effect of Zigzag transform based on pixel decomposition, we provide an example here. Assuming there exists an image matrix A of size 3×3:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \tag{2}$$

Matrix A is first transformed into matrix B of size 3×6 using pixel decomposition:

$$B = \begin{bmatrix} 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 4 & 0 & 5 & 0 & 6 \\ 0 & 7 & 0 & 8 & 0 & 9 \end{bmatrix} \tag{3}$$

Then we perform Zigzag transformation on matrix B and get the matrix C:

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 4 & 0 \\ 2 & 0 & 7 & 0 & 5 & 0 \\ 3 & 0 & 8 & 0 & 6 & 9 \end{bmatrix} \tag{4}$$

By using the inverse pixel decomposition transformation, we can combine the adjacent elements in matrix C and construct matrix D:

$$D = \begin{bmatrix} 1 & 0 & 64 \\ 32 & 112 & 80 \\ 48 & 128 & 105 \end{bmatrix} \tag{5}$$

Comparing matrices C and D, it can be found that the values of the vast majority of elements in matrix D are completely different from those corresponding to matrix A. This indicates that pixel decomposition based Zigzag transform can not only achieve the effect of scrambling, but also change the pixel values.

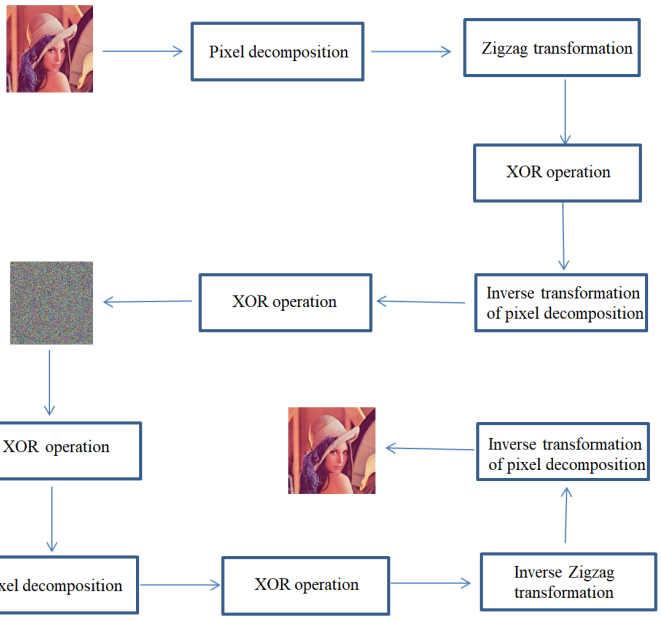


Figure 3: Flow chart of the image encryption and decryption algorithm.

2.5 Generation of the Chaotic Encryption Sequences

In this paper, $SHA256$ hash function is utilized to produce the initial values of the chaotic system. The $SHA256$ hash value is a hexadecimal number sequence with 64 digits long [9]. If the input message is slightly different, it will output completely different results.

Step 1: Calculate the $SHA256$ hash value of the plain image and denote it as $q = q_1 q_2 ... q_{64}$.

Step 2: Calculate x_0, y_0, z_0 and K as follows:

$$\begin{cases} x_0 = \sum_{i=1}^{20} hex2dec(q_i) \\ y_0 = \sum_{i=21}^{40} hex2dec(q_i) \\ z_0 = \sum_{i=41}^{60} hex2dec(q_i) \\ K = \sum_{i=61}^{64} hex2dec(q_i) \end{cases} \tag{6}$$

Step 3: The initial values x_0, y_0 and z_0 for System (1) is set as

$$\begin{cases} x_0 = \frac{x_0 + MN}{2^9 + MN} \\ y_0 = \frac{y_0 + MN}{2^9 + MN} \\ z_0 = \frac{z_0 + MN}{2^9 + MN} \end{cases} \tag{7}$$

where M and N are the height and width of the image.

Step 4: The parameter K which is used to set the number of zigzag transformations is calculated as $K = mod(K, 5) + 5$.

Step 5: Choose the system control parameters a, b, c, d, k, I of System (1).

Step 6: Iterate System (1) for $2L + 2000$ times with the initial values x_0, y_0, z_0, remove the former 2000 values and three chaotic sequences x_s, y_s, z_s of length $2L$ can be gotten, where $L = M \times N$. Calculate two sequences S_1, S_2 which will be used in the following encryption process as $S_1 = |x_s| \times 10^{15} \mod 16$, $S_2 = |y_s(1:L)| \times 10^{15} \mod 256$.

2.6 Image Encryption Method

The flow chart of the proposed image encryption and decryption algorithm is illustrated in Figure 3. When encrypting a gray scale image P_1 of size $M \times N$, the specific steps of the encryption algorithm can be described as follows:

Step 1: Transformed image matrix P_1 to the matrix P_2 with the size of $M \times 2N$ using pixel decomposition method described in subsection 2.2.

Step 2: Apply Zigzag transformation to P_2 K times and then we can obtain the corresponding scrambled matrix P_3.

Step 3: Transform matrix P_3 into the 1D vector V_1 whose length is $2L$. Perform XOR operation on V_1 using random sequence S_1, as shown in Formula (8):

$$\begin{cases} C_{V1}(1,1) = V_1(1,1) \oplus S_1(1,1) \\ C_{V1}(1,i) = V_1(1,i) \oplus (C_{V1}(1,i-1) \\ \qquad\qquad + S_1(1,i)) \ mod \ 16 \end{cases} \tag{8}$$

where $i = 2, 3, \cdots, 2L$ and symbol "$\oplus$" is the bitwise exclusive or operator.

Step 4: Reshape vector C_{V1} to get the matrix P_4. Combine the adjacent elements in matrix P_4 and construct the matrix P_5 using the inverse transformation

of pixel decomposition, where P_4 is of size $M \times 2N$ while P_5 is of size $M \times N$.

Step 5: Transform matrixes P_5 into the 1D vector V_2 and perform the XOR operation on V_2 using the random sequences S_2:

$$\begin{cases} C_{V2}(1,1) = V_2(1,1) \oplus S_2(1,1) \\ C_{V2}(1,i) = V_2(1,i) \oplus (C_{V2}(1,i-1) \\ \qquad\qquad +S_2(1,i)) \ mod\ 256 \end{cases} \tag{9}$$

where $i = 2, 3, \cdots, L$ and C_{V2} is the cipher vector.

Step 6: Reshape vector C_{V2} to get the cipher image matrix P_6.

When encrypting a color image of size $M \times N \times 3$, we can encrypt its R, G, B components separately according to the method of encrypting gray scale images, then combine them to get the ciphered color image.

2.7 Image Decryption Method

Decryption is the inverse process of encryption using the same parameters and keystreams, which mainly consists of XOR operation, inverse Zigzag transform, pixel decomposition and inverse pixel decomposition. The image decryption process mainly contains the following steps:

Step 1: Transform the cipher image C into one-dimensional pixel vector C_{V2} and performance Xor operation on it using the random sequences S_2 to obtain the vector V_2.

Step 2: Reshape vector V_2 to get matrix P_5. Transformed P_5 to get matrix P_4 utilizing pixel decomposition transformation and further transform P_4 into vector C_{V1}. Perform XOR operation on C_{V1} using the random sequences S_1 to get vector V_1.

Step 3: Reshape vector V_1 to get matrix P_3 and apply inverse Zigzag transformation to P_3 K times and then we can obtain the corresponding matrix P_2.

Step 4: Transform P_2 to matrix P_1 utilizing inverse pixel decomposition transformation then we can get the plain image.

3 Test and Analysis of the Proposed Scheme

In this experiment, three standard color images from the USC-SIPI image database of size $256 \times 256 \times 3$ were used for test. They are Lena, Peppers and House. The control parameters of the chaotic system were set as follows: $a = 1, b = 5, c = 1, d = 1, k = 0.95$ and $I = 1$. Figure 4 shows the encrypted and decrypted images corresponding to the plain images. From Figure 4, it can be observed that the cipher images appear as disordered, snowflake-like noise images. Therefore, the encrypted images effectively conceal the information of the original plain images. Furthermore, the decrypted images are identical to the original images, indicating the effectiveness of the decryption algorithm.

Figure 4: The experimental results of the encrypted images. (a) The plain images. (b) The encrypted images. (c) The decrypted images.

3.1 Key Space Analysis

The encryption key of the proposed algorithm consists of several components, including the initial values x_0, y_0, z_0, the control parameters a, b, c, d, k, I of the chaotic system, and the number of scanning rounds K. Assuming that each parameter is represented with double precision up to 15 decimal places, the key space of the encryption algorithm is more than 10^{135}. Therefore, this algorithm has a sufficiently large key space, and the proposed method can effectively resist brute-force attacks.

3.2 Histogram Analysis

The histogram reflects the number of pixels for each gray scale level. A good encryption algorithm requires attackers will be unable to extract meaningful information from the statistical analysis of gray scale values. The first and second lines of Figure 5 provide histograms of R, G, B components of the Lena plain image and its cipher image. As can be seen from Figure 5, the distribution of the pixel values of the cipher image is very uniform so it has a good resistance to statistical analysis.

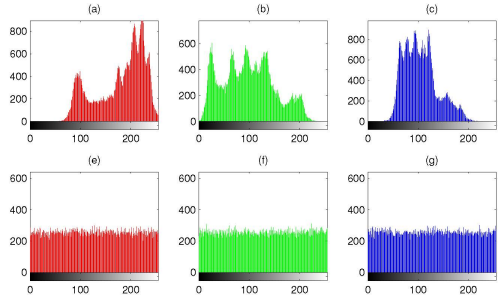


Figure 5: Histograms of Lena image in red, green, and blue components. Histograms of plain and cipher images are shown in rows 1 and 2, respectively.

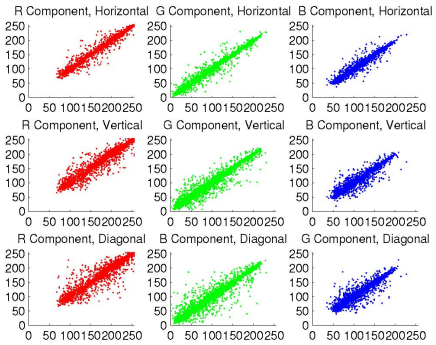


Figure 6: Correlation distributions of plain image in each direction.

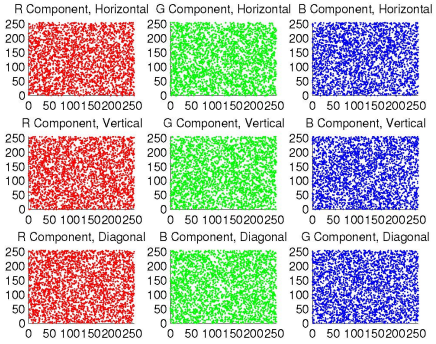


Figure 7: Correlation distributions of cipher image in each direction.

3.3 Correlation Analysis

An effective encryption algorithm must reduce the correlation between adjacent pixels in the plain image, making the pixel values unpredictable. Formula (10) provides a method for calculating the correlation coefficient between adjacent pixels in an image [22]:

$$r_{xy} = \frac{\sum_{i=1}^{L}((x_i - E(x))(y_i - E(y)))}{\sqrt{(\sum_{i=1}^{N}(x_i - E(x))^2)(\sum_{i=1}^{N}(y_i - E(y))^2)}} \quad (10)$$

where $E(x) = \sum_{i=1}^{L} x_i$, $E(y) = \sum_{i=1}^{N} y_i$, x_i and y_i are gray-level values of the selected adjacent pixels, and L is the number of sample pixels.

To test the correlation between adjacent pixels in the plain images and the corresponding encrypted images, we randomly selected 3000 pairs of adjacent pixels in three different directions (horizontal, vertical, and diagonal) from the plain and cipher images, and the correlation coefficients in R, G, B components are calculated. Tables 1 and 2 present the specific values of the correlation coefficients. Additionally, Figure 6 and Figure 7 show the distribution of correlation between adjacent pixels in the plain and cipher image of Lena, respectively. From Tables 1 and 2 and Figures 6 and 7, it can be observed that the adjacent pixels in the plaint images exhibit high correlation, whereas the correlation between adjacent pixels in the cipher images approaches zero, which indicates that the encryption algorithm possesses strong decorrelation capabilities.

Table 1: Correlation coefficients of the R, G and B components of the plain color image of Lena.

Component	Horizontal	Vertical	Diagonal
R component	0.9475	0.9436	0.9086
G component	0.9153	0.9112	0.8540
B component	0.9498	0.9495	0.9142

Table 2: Correlation coefficients of the R, G and B components of the encrypted color image of Lena.

Component	Horizontal	Vertical	Diagonal
R component	0.0220	-0.0122	-0.0347
G component	-0.0323	-0.0175	0.0123
B component	0.0092	0.0300	0.0053

3.4 Information Entropy Analysis

Information entropy reflects the uncertainty of information. The higher the entropy of an encrypted image, the greater the randomness in the distribution of pixel

Table 3: Information Entropy Analysis

Image	Entropy	Component		
		R	G	B
Lena	Plain	7.2594	7.5696	6.9698
	Cipher	7.9969	7.9972	7.9976
Peppers	Plain	7.3920	7.6150	7.1738
	Cipher	7.9970	7.9968	7.9971
House	Plain	6.4510	6.6208	6.5657
	Cipher	7.9975	7.9974	7.9974

Table 4: Analysis of Differential Attack Resistance

Image	Index	Component		
		R	G	B
Lena	NPCR	99.59%	99.61%	99.62%
	UACI	33.34%	33.53%	33.49%
Peppers	NPCR	99.55%	99.60%	99.61%
	UACI	33.42%	33.43%	33.43%
House	NPCR	99.62%	99.65%	99.60%
	UACI	33.31%	33.52%	33.49%

values, indicating higher security [10]. The formula for information entropy is defined as follows:

$$H(m) = -\sum_{i=0}^{255} P(m_i) \log_2 P(m_i) \tag{11}$$

where m_i is the ith gray level for the digital image and $P(m_i)$ represents the probability of m_i.

Table 3 presents the information entropy of the three test images before and after encryption in R, G, B channels. From Table 3, it can be observed that the information entropy of the encrypted images is very close to the ideal value of 8, indicating that the cipher images are random and can resist entropy-based attacks.

3.5 Analysis of Differential Attack Resistance

The resistance against differential attacks is an important indicator of the effectiveness of image encryption. It is commonly measured using two metrics: the Number of Pixels Change Rate (NPCR) and the Unified Average Changing Intensity (UACI) [11]. NPCR describes the proportion of the number of changes in the pixel values of the ciphertext images caused by a small change in the pixel value of any pixel in a plaintext image to the total number of pixels, and UACI provides the degree of change in the pixel values of the ciphertext images. The formulas for calculating NPCR and UACI are as follows:

$$NPCR = \frac{\sum_{ij} D_{ij}}{M \times N} \times 100\% \tag{12}$$

$$UACI = \frac{\sum_{ij} (C_1(i,j) - C_2(i,j))}{255 \times M \times N} \times 100\% \tag{13}$$

where C_1 and C_2 are cipher images of two plain images which have only one-pixel difference and D_{ij} is defined by

$$D_{ij} = \begin{cases} 0 & \text{if } C_1(i,j) = C_2(i,j) \\ 1 & \text{if } C_1(i,j) \neq C_2(i,j) \end{cases} \tag{14}$$

The theoretical value of NPCR is 99.609375%, and the theoretical value of UACI is 33.463541%. Table 4 provides the NPCR and UACI values for each image. From Table 4, it can be observed that a small change in a single pixel value in the plaint image results in almost all pixel values in the cipher image be altered, and the values are very close to the theoretical values. This indicates that the encryption algorithm exhibits excellent resistance against differential attacks.

3.6 Performance Comparison with Other Methods

To further show the effectiveness of the proposed algorithm, we compare the proposed method with the other three image encryption methods proposed in Ref. [3,16,21] from the aspects of correlation analysis, NPCR, UACI, and information entropy for the encrypted Lena image. The specific experimental results are listed in Table 5.

From Table 5, it can been see that the performance of the proposed method is competitive compared with Ref. [3, 16, 21]. In addition, the proposed method is very simplified so that it is an effective tool for image encryption.

4 Conclusions

In this paper, a novel image encryption algorithm is proposed which decompose pixels into smaller units and further utilize the chaotic system, Zigzag transform to encrypt images. The pixel decomposition based Zigzag transform can not only disorder but also change the pixel values to increase the randomness of the encryption algorithm; twice XOR operations are further used for diffusion to increase the encryption effect. Experiments were carried out on three color images. The experimental results show that the presented algorithm has better security and higher effectiveness.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No.11871417) and the Fundamental Science (Natural Science) Foundation of the Jiangsu Higher Education Institutions of China (Grant

Table 5: Performance comparison with other methods of the encrypted Lena image

Index	Chanel	Ref. [3]	Ref. [21]	Ref. [16]	Proposed
Horizontal Correlation	R	-0.0154	0.0013	0.0137	0.0220
	G	-0.0096	0.0032	-0.0246	-0.0323
	B	-0.0030	0.0020	-0.0137	0.0092
Vertical Correlation	R	-0.0102	0.0047	-0.0237	-0.0122
	G	0.0027	-0.0005	-0.0170	-0.0175
	B	0.0117	0.0001	0.0023	0.0300
Diagonal Correlation	R	0.0159	0.0002	0.0109	-0.0347
	G	- 0.0162	0.0048	-0.0133	0.0123
	B	-0.0026	- 0.0040	-0.0013	0.0053
NPCR(%)	R	99.60%	99.62%	99.61%	99.59%
	G	99.60%	99.61%	99.60%	99.61%
	B	99.63%	99.62%	99.60%	99.62%
UACI(%)	R	33.41%	33.42%	33.46%	33.34%
	G	33.49%	33.43%	33.47%	33.53%
	B	33.49%	33.46%	33.47%	33.49%
Information entropy	R	7.9974	7.9917	7.9892	7.9969
	G	7.9971	7.9912	7.9898	7.9972
	B	7.9973	7.9917	7.9899	7.9976

No.23KJA120004). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] H. Dong, E. Bai, X. Q. Jiang, "Color image compression-encryption using fractional-order hyperchaotic system and DNA coding", *IEEE Access*, vol. 8, pp. 163524-163540, 2020.

[2] H. Gao, X. Y. Wang, "An image encryption algorithm based on dynamic row scrambling and ZigZag transform", *Chaos, Solitons and Fractals*, vol. 147, no. 6, 110962, 2021.

[3] K. M. Hosny, S. T. Kamal, M. M. Darwish, " Novel encryption for color images using fractional-order hyperchaotic system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 973-988, 2022.

[4] H. Huang, D. Cheng, "3-image bit-level encryption algorithm based on 3d nonequilateral arnold transform and hyperchaotic system", *Security and Communication Networks*, vol. 7, pp. 1-13, 2020.

[5] M. Naim, A. A. Pacha, " A novel image encryption algorithm based on advanced hill cipher and 6D hyperchaotic system", *International Journal of Network Security*, vol. 25, no. 5, pp. 829-840, 2023.

[6] M. Naim, A. A. Pacha, C. Serief, "A novel satellite image encryption algorithm based on hyperchaotic systems and Josephus Problem", *Progress in space research*, vol. 67, no. 7, pp. 2077-2103, 2021.

[7] S. Yan, L. Li, B. Gu, Y. Cui, J. Wang, J. Song, "Design of hyperchaotic system based on multi-scroll and its encryption algorithm in color image," *Integration*, vol. 88, pp. 203-221, 2023.

[8] G. Qu, X. Meng, Y. Yin, "Optical color image encryption based on Hadamard single-pixel imaging and Arnold transform", *Optics and Lasers in Engineering*, vol. 137, no. 20, 106392, 2021.

[9] A. Soni, P. Netam, B. Acharya, "Multiple grayscale image encryption using cross-coupled chaotic maps," *Journal of Information Security and Applications*, vol. 52, 102470, 2020.

[10] S. Suri, R. Vijay, "A coupled map lattice-based image encryption approach using DNA and bi-objective genetic algorithm," *International Journal of Information and Computer Security*, vol. 12, no. 2, pp. 199-216, 2020.

[11] J. Wang, X. Zhi, X. Chai, Y. Lu, " Chaos-based image encryption strategy based on random number embedding and DNA-level self-adaptive permutation and diffusion," *Multimedia Tools And Applications*, vol. 80, no.01, pp. 16087-122, 2021.

[12] L. Wang, " Image encryption based on hyperchaotic systems And DNA encoding", *International Journal of Network Security*, vol. 25, no. 3, pp.515-521, 2023.

[13] X. Wang, Y. Su, " Image encryption based on compressed sensing and DNA encoding", *Signal Processing Image Communication*, vol. 12, 116246, 2021.

[14] X. Y. Wang, X. Cheng, "An image encryption algorithm based on dynamic row scrambling and Zigzag transformation," *Chaos, Solitons and Fractals*, vol. 147, 110962, 2021.

[15] X. Y. Wang, X. L. Wang, L. Teng, D. H. Jiang, Y. Xian, " Lossless embedding: A visually meaningful image encryption algorithm based on hyperchaos and compressive sensing," *Chinese Physics B*, no. 2, 020503, 2023.

[16] X. Wu, K. Wang, X. Wang, H. Kan, J. Kurths, "Color image DNA encryption using NCA map-based CML and one-time keys," *Signal Processing*, vol. 148, no. 7, pp. 272-287, 2018.

[17] G. Q. Xiong, Z. C. Cai, S. F. Zhao, "A bit-plane encryption algorithm for RGB image based on modulo negabinary code and chaotic system," *Digital Signal Processing*, vol. 17, no. 7, 2023.

[18] S. Zhang, X. Wang, Z. Zeng, "A simple no-equilibrium chaotic system with only one signum function for generating multidirectional variable hidden attractors and its hardware implementation", *Chaos*, vol. 30, no.5, 53129, 2020.

[19] S. Zhang, J. H. Zheng, X. P. Wang, "A novel no-equilibrium HR neuron model with hidden homogeneous extreme multistability," *Chaos, Solitons and Fractals*, vol. 145, 110761, 2021.

[20] X. Zhang, L. Wang, Y. Wang, Y. Niu, Y. Li, " An image encryption algorithm based on hyperchaotic system and variable-step josephus problem", *International Journal of Optics*, vol. 4, pp. 1-15, 2020.

[21] Y. Q. Zhang, Y. He, P. Li, X. Y. Wang," A new color image encryption scheme based on 2DNLCML system and genetic operations," *Optics and Lasers in Engineering*, vol. 128, 106040, 2020.

[22] S. Zhou, Y. Qiu, X. Wang, *et al.*, "Novel image cryptosystem based on new 2d hyperchaotic map and dynamical chaotic s-box," *Nonlinear dynamics*, vol. 111, pp. 9571-9589, 2023.

Biography

Chunming Xu is an associate professor at the mathematics and statistics from Yancheng Teachers University, P.R. China. His main research interests include image processing and artificial intelligence.

Yong Zhang is a professor at the mathematics and statistics from Yancheng Teachers University, P.R. China. His main research interests include cryptography and optimization.

An Improved CNN for Intrusion Detection Method Based on ResNet

Zengyu Cai[1], Pengrong Li[1], Jianwei Zhang[2,3], Yajie Si[1], and Yuan Feng[1]
(Corresponding author: Jianwei Zhang)
College of Computer and Communication Engineering, Zhengzhou University of Light Industry[1]
College of Software Engineering, Zhengzhou University of Light Industry[2]
ZZULI Research Institute of Industrial Technology[3]
Zhengzhou, Henan 450000,China
Email: mailzjw@163.com
(Received Aug. 14, 2023; Revised and Accepted Feb. 22, 2024; First Online June 22, 2024)

Abstract

With the popularization and application of the Internet, the importance of network security has become more and more prominent, and intrusion detection technology plays an important role in protecting the security. Aiming at the problem that the existing intrusion detection methods are not ideal for detecting new intrusions in the network. This paper proposes an improved CNN intrusion detection method based on ResNet. Based on the traditional convolutional neural network, the residual structure is added to the model by using the identity mapping principle of residuals, which can efficiently process data features in multi-layer convolution and avoid the weakening of model training effect. Compared with other commonly used neural network models, this model improves the prediction accuracy while avoiding over-fitting of the network model. Finally, this paper uses the NSL-KDD dataset to test the performance of the proposed model and ablation experiments. The results show the effectiveness and robustness of the proposed model.

Keywords: CNN; Intrusion Detection; Residual Neural Network

1 Introduction

With the rapid development of science and technology, the Internet has gradually integrated into all aspects of people's lives, work and learning, and has become an indispensable and important tool. As a new resource, big data has also brought many far-reaching impacts on our lives. The advent of the era of big data has also become a hot topic of concern in today's society. Among them, the most concerned is the issue of network security. At present, there are many ways to defend against network attacks, such as firewall, intrusion detection system, vulnerability scanning technology and so on. The Intrusion detection system (IDS) is an active defense method used to ensure network security. It collects and analyzes data information in the host or network, identifies normal behaviors in the network and abnormal behaviors that threaten network security in real time, and responds to monitored abnormal data in the time [10].

Before entering the era of big data, machine learning performs probabilistic inference or fuzzy matching on the characteristics of network packets by learning existing intrusion or normal modes, so as to discover unknown intrusions. The traditional network intrusion detection method using machine learning to identify network traffic is popular and effective [8]. The mainstream intrusion detection methods based on machine learning include Naive Bayes, Decision Tree [6], Random Forest [16], Support Vector Machine [11], Kmeans [3], etc. In the era of big data, a large number of new attack methods are generated, and new vulnerabilities are released every day. The explosion of data volume and the concealment of network intrusion means make traditional intrusion detection require higher storage space and data computing ability. Due to factors such as fewer captured samples and shorter training time, traditional intrusion detection methods will have a higher false alarm rate and false negative rate. Traditional intrusion detection methods are only suitable for classification and extraction of small data sets, which cannot meet the current network security requirements [19]. At present, deep learning with autonomous learning ability has become the main development direction in the field of intrusion detection, and scholars have also achieved many results. References [4, 15, 17] and a large number of studies have shown that using recurrent neural network (DNN), convolutional neural network (CNN), autoencoder (AE) and other commonly used neural networks to build models, the prediction accuracy is between 76 % and 95%, and the accuracy needs to be improved. In Reference [9], the long-term short-term memory (LSTM) recurrent neural network was used to train the KDD Cup

'99 data set for the intrusion detection model, and the correlation between the information data was effectively obtained, which improved the detection accuracy and reduced the false alarm rate to a certain extent. However, there are some limitations in dealing with the deep feature structure, and the credibility of the data set is not high.

Synthesize the above research, this paper proposes an improved CNN for intrusion detection method based on ResNet, which adds residual structure on the basis of CNN. The model fully exploits the interdependence between network traffic characteristics. Compared with other commonly used neural network models, the model improves the prediction accuracy while minimizing the gradient disappearance caused by the deepening of the network depth, effectively avoiding the over-fitting of the network model and making the model have strong robustness. The method uses the NSL-KDD dataset, and the test shows that the prediction accuracy of the model proposed in this paper has been effectively improved.

2 Related Work

2.1 Intrusion Detection Based on Deep Learning

At present, intrusion detection has become a major means of network security technology protection. Its concept was first proposed in 1980 by Anderson [12], who is effective for the National Security Agency of the United States, and has opened up a pioneer in intrusion detection research. Intrusion detection based on machine learning algorithms was a popular research in the era of small data volume, low network complexity, and simplified attack types, and a large number of meaningful research results have been achieved [2-6]. In today's era of massive data, deep learning intrusion detection methods have made effective improvements in prediction accuracy, network performance, avoidance of gradient disappearance, gradient explosion, and automated multidimensional learning. Due to the characteristics of its method, the intrusion detection method of deep learning has made breakthrough progress in the fields of speech semantic recognition, context awareness, prediction traffic, target recognition, image processing and so on.

2.2 Convolutional Neural Network

The Convolutional neural network (CNN) is an efficient feedforward neural network model in the field of deep learning. Reference [1] compares the performance of CNN and RNN in intrusion detection systems. In most of the anomaly-based IDS studies using artificial intelligence, machine learning methods are being used to build models for detecting intrusions. However, since deep learning is expected to provide higher performance and can automatically process feature selection, deep learning technology continues to be widely used in intrusion detection systems, which is consistent with the widespread use of deep learning in different fields [20].The basic structure of CNN consists of input layer, convolution layer, pooling layer, fully connected layer and output layer. First, the data is input into the convolutional layer for feature extraction. Then, the incoming pooling layer reduces the dimension of the extracted features, compresses the amount of data parameters, and reduces overfitting. After that, the fully connected layer is responsible for connecting all the data features after pooling. Finally, the results are obtained by the softmax classifier and output.

The intrusion detection problem is essentially a classification problem. The classification model trained by supervised learning is applied to network data for prediction. When CNN is used, the weight sharing and local connection used in the convolution and pooling process greatly reduce the number of parameters and weights in the convolution kernel, reduce the complexity of the network, accelerate the detection speed, and face a large number of network data. Pressure-free processing is an important advantage of CNN's application in intrusion detection. However, CNN will produce a degradation phenomenon as the network depth deepens, and the constant mapping-linear transformation of the residual network could improve this phenomenon.

2.3 Residual Network

The residual network (ResNet) [13, 18] was proposed by He Kaiming in 2015, which solved the bottleneck of CNN at that time-the ' network degradation ' phenomenon caused by the deepening of the network layer. The main idea of ResNet is to add a quick link to ensure that the performance of the n+1 th network layer is better than that of the previous network layer through identity mapping (y = x). The gradient of the high layer can be directly transmitted to the low layer, which effectively prevents the disappearance of the gradient and improves the overall performance of the network. The residual network is added to the convolution, and the convolution layer jump is realized by identity mapping. The multi-layer convolution efficiently processes the data features while avoiding the weakening of the model training effect. In [2], an intrusion detection model based on residual neural network is proposed. The residual structure is used to mine the interdependence between network traffic characteristics and predict the current network state. The accuracy is 98.27 %. Compared with the traditional neural network, this method improves the prediction accuracy to a certain extent, but the current network security situation needs to be improved.

3 An Improved CNN Intrusion Detection Model Based on ResNet

In network traffic detection, the traditional CNN algorithm extracts the traffic characteristics in the network nonlinearly in a shared way due to its special hierarchical characteristics, which reduces the complexity of the network model. However, CNN is a neural network with multiple hidden layers. With the deepening of the network layer, overfitting will occur, which will lead to the degradation of the network model.

An improved CNN intrusion detection model based on ResNet proposed in this paper solves the above problems by integrating ResNet on the basis of CNN. The model classifies the network traffic characteristics to determine whether there is an attack and its attack type.

3.1 Data Preprocessing

Firstly, this paper uses one-hot one-hot coding to convert all discrete features in the NSL-KDD dataset into numerical features. Since an improved CNN intrusion detection model based on ResNet can only deal with numerical features, and the data set itself contains character features, converting character data into numerical data is conducive to improving the accuracy of the prediction model, more accurately describing the relationship between data, and easier mathematical calculation and modeling to ensure the reliability and validity of the analysis results. Each sample in the NSL-KDD dataset has 41 eigenvalues, of which the three types of values of protocol type , service and flag are character types. The one-hot coding is mainly used to numerically process the above three eigenvalues, which avoids the influence of the size connection between integers on the model training. For example, the protocol has three values, namely tcp , udp and icmp , and the corresponding values are [0,0,1], [0,1,0] and [1,0,0]. After processing, the eigenvalue length of each network data increases from the original 41 dimensions to 122 dimensions.

Then, the min-max normalization method is used to normalize the eigenvalues. After numerical processing, in order to avoid the adverse effects of unequal contribution to the results caused by the excessive range of some eigenvalues, it is necessary to normalize the eigenvalues, so that the values of each feature fall within the range of 0 and 1, and the values of all data are in a unified order of magnitude, reducing the impact on the results of model training. The formula is as follows :

$$x_i' = \frac{x_i - min\left(x_0, x_{1,\ldots}x_n\right)}{max\left(x_0, x_{1,\ldots}x_n\right) - min\left(x_0, x_{1,\ldots}x_n\right)} \quad (1)$$

In the formula, x_i represents an eigenvalue, $min()$ represents the minimum value of the data set, $max()$ represents the maximum value of the data set, and x_i' represents the normalized eigenvalues.

3.2 Construction of an Improved CNN Intrusion Detection Model Based on ResNet

In order to solve the problem of network degradation caused by the deepening of network structure, this paper constructs an improved CNN intrusion detection model based on ResNet. An improved convolutional neural network model based on ResNet inputs a 122-dimensional data. After five convolutional layers, residual connections connecting the second and fourth convolutional layers, and pooling layers. The adaptive average pooling reduces the spatial dimension of the tensor to 1x1, effectively compressing the tensor into a one-dimensional vector. The fully connected layer receives a one-dimensional vector as an input, where each element in the vector corresponds to the activation of the first few layers of the network, and then the softmax function maps this one-dimensional vector to a specified number of five types of result outputs, which are judged as normal or abnormal intrusion. The overall architecture of the model is shown in Figure 1.

3.2.1 ConResNet

In the feature extraction stage, this model uses a five-layer convolution and a linear residual connection connecting the second and fourth layer convolution blocks to extract the deep features of network traffic. The input of the previous layer convolution is the input of the latter layer convolution, and the latter layer convolution integrates the data features extracted from the previous layer to continue deep analysis. However, due to the addition of residual connection, the input of the fifth layer convolution is the sum of the output vectors of the second and fourth layers. In the five-layer convolution block, the convolution kernel size is 3, stride and padding are set to 1. According to the size characteristics of the training data, the input channel of convolution layer 1 is 1, and the output channel of convolution layer 5 is 8. The data is output from the original 122-dimensional to 5-dimensional data.

The specific process of convolution is as follows :

$$Output^{x(i,j)} = \sum_{m=0}^{n-1} W_i^{x(m)} \bullet y^{x(j+m)} \quad (2)$$

In the formula, n is the width of the convolution kernel, $W_i^{x(m)}$ is the m-th weight of the i-th convolution kernel in the x-layer, $y^{x(j+m)}$ is the j-th convolution local area in the x-layer, and $Output^{x(i,j)}$ is the corresponding convolution output.

Each layer of convolution of the network data requires relu function activation and regularization to add nonlinear elements to the model. Ensure that each layer of convolution has the ability to learn and simulate other complex data types, enhance the expression ability of the neural network, make the input value of the next convolution layer more stable, and stabilize the network performance.

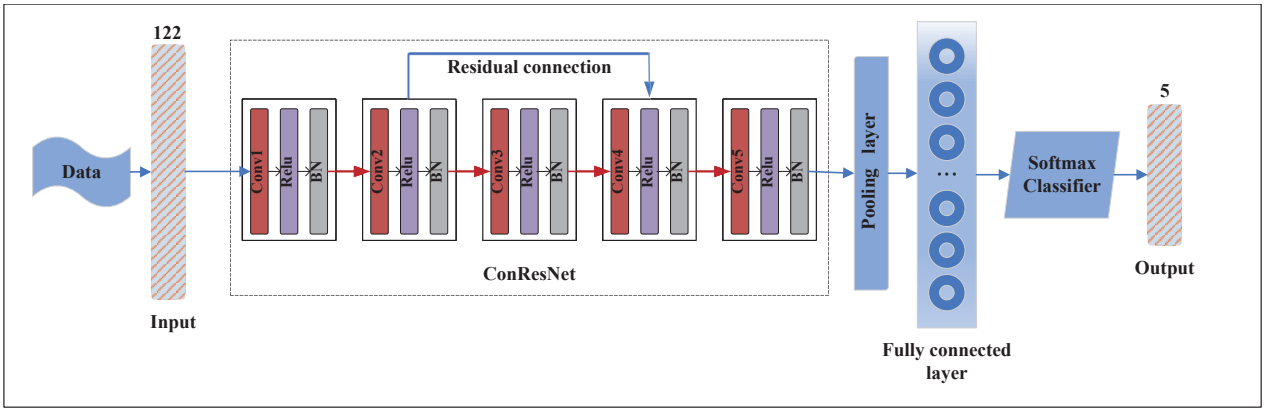


Figure 1: The overall architecture of the model

Table 1: The impact of Activation and Optimizer on model evaluation indicators

Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Activation	Optimizer
99.48	99.32	99.19	99.26	ReLU	Adam
97.38	98.11	94.83	96.44	ReLU	SGD
99.00	99.40	99.20	99.30	ReLU	NAdam
98.97	99.52	98.99	99.25	Sigmoid	Adam
96.04	96.75	87.64	91.97	Sigmoid	SGD
98.95	99.31	99.10	99.21	Sigmoid	NAdam
98.94	99.38	99.03	99.20	Tanh	Adam
93.53	99.10	79.02	86.45	Tanh	SGD
98.92	98.58	99.19	98.89	Tanh	NAdam

After $Output^{x(i,j)}$ is activated by relu, the expression is :

$$Y^{x(i,j)} = f(Output^{x(i,j)}) = max0, Output^{x(i,j)} \quad (3)$$

3.2.2 Adaptive Average Pooling

After convolution and residual of data features, two-dimensional adaptive average pooling is performed on the output vector of the fifth layer convolution to reduce the offset of the estimated mean, improve the robustness of the model, and obtain effective feature dimension reduction. The pooling layer automatically calculates the average pooling layer size and the moving step size according to the set output signal size and the size of the input signal. The size of each channel after pooling is a 1x1, and the number of channels remains unchanged. That is, the number of input feature channels of the fully connected layer and the number of convolution output channels of the fifth layer are the same as 8.

3.2.3 Fully Connected Layer

After convolution residuals and pooling layers, the spatial dimension of data features is reduced, and the model has initially completed data feature extraction. The fully connected layer aggregates the global information from these simplified features into a compact representation, which allows the model to capture higher-level patterns and relationships between features throughout the input data.

The input data of the fully connected layer is a one-dimensional array composed of scalar eigenvalues. Each element represents an input feature, corresponding to a neuron. The goal of this layer is to identify the category of each data and classify network attacks.

The specific process of the fully connected layer is as follows :

$$Output^{x+1(j)} = \sum_{i=1}^{n} W_{ij}^{x} y^{x(j)} + b_j^x \quad (4)$$

In the formula, W_{ij}^{x} is the weight value between the ith neuron in layer x and the jth neuron in layer x+1, $y^{x(j)}$ represents the value of the jth neuron in layer x, b_j^x is the bias value of all neurons in layer x to the jth neuron in layer x+1, $Output^{x+1(j)}$ represents the vector value of the jth output neuron in layer x+1.

3.2.4 Softmax Classifier

The softmax activation function is applied to the output of the fully connected layer to obtain the class probability,

and the original score is converted into the probability distribution of the output class. This model aims to classify the input data into one of five categories, namely Normal and DoS, probe, U2R, R2L four attack types. The output layer will have five neurons, each neuron represents the probability that the input belongs to a specific class, and the class with the highest probability will be the input prediction class.

3.3 Intrusion Detection Model Training and Testing

In the process of model training, it is mainly divided into two parts: forward propagation and back propagation. Forward propagation is to input data into the model and proceed sequentially from left to right to output the prediction results; back propagation is to carry out the process of gradient solving, and use the chain derivation rule to find out the gradient of the final output of the network. The entire network calculates the gradient once per iteration step, and the parameters are updated according to the corresponding individual components in the gradient; 20 iterations are required in this model.

After training, the model is applied to the test set for classification to get the prediction results, and the accuracy of the model needs to be improved according to the comparison of the prediction results with the actual ones. This model uses the cross-entropy loss function to evaluate the consistency between the model predictions and the actual results, and the weight parameters are updated to minimize the loss function and optimize the performance of the model.

In the model test assume that the data softmax classifier output sample category prediction is $\widetilde{y} \in [0,1]$, the target value of the sample category in the training set is y, the expression of the cross entropy function is:

$$Loss = -\sum_{i=1}^{5} y_i * \log \widetilde{y_i} \tag{5}$$

4 Experimental Results and Analysis

In order to test an improved CNN intrusion detection model based on ResNet proposed in this study, we conducted simulation experiments using python3.7 's PyTorch 1.10 in the Windows environment. All experiments were performed on a PC with eight TITAN RTX systems.

4.1 NSL-KDD Dataset

KDD-CUP99 is a classic dataset in the field of intrusion detection, but it has redundant and repetitive data, which leads to a large error in the evaluation results, while the NSL-KDD dataset improves on the basis of KDD99 dataset, with a more reasonable setup of the training set and the test set, which is more suitable as an effective benchmark dataset for intrusion detection methods. The NSL-KDD dataset contains a total of 125,937 pieces of data, which is just the right size to train the model. In this paper, we use 10-fold cross-validation in our experiments by dividing the dataset into ten parts, nine of which are used for training the model and the other for testing. No redundant records are included in the training set, the model classifier will not be biased towards frequently occurring records due to memory, and the test set is also free of duplicate records, which makes the detection rate of the model more accurate.The number of records in the training and test sets of the NSL-KDD dataset is reasonably distributed, which makes it more accommodating to the classification rate of different learning methods, and its application is wider to make an effective comparison with other models.

4.2 Model Evaluation Indicators

When training the model, in order to evaluate the model more accurately, this paper chooses Accuracy, Precision, Recall, and F1 as the evaluation indexes Table 2 shows the effect of different parameters on the model performance.

Accuracy(Acc) is the ratio of the correctly categorized data to the total number of data.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision(Pre) is also known as checking accuracy, which indicates the proportion of data with positive predictions that are actually positive.

$$Pre = \frac{TP}{TP + FP} \tag{7}$$

Recall indicates the proportion of the actual number of positive data in the data with a positive prediction result to the proportion of positive data in the full data.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F1 is a weighted average of precision and recall.

$$F1 = \frac{2 * (Pre \,.\, \mathrm{Re}\, call)}{Pre + Recall} \tag{9}$$

In the formula, TP is the number of normal data correctly categorized by the model; TN is the number of abnormal data correctly categorized by the model; FP is the number of model abnormal classes for normal data; FN is the number of model categorized normal classes for abnormal data.

4.3 Model Hyperparameter Setting

There are many hyperparameters in the commonly used models for deep learning, and tuning means finding the optimal parameters for the model, which determines the performance of the model. In this paper, we focus on

Table 2: Comparison of different intrusion detection model methods

Paper	Method	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
[7]	ANN And Feature Selection	88.39	85.44	95.95	98.84
[14]	Recurrent Neural Network	98.48	—	96.12	98.25
[5]	Convolutional Neural Networks	97.09	—	93.49	—
[2]	Residual Neural Network	98.27	97.29	99.55	98.39
This paper	An improved CNN based on ResNet	99.48	99.32	99.19	99.26

Table 3: Ablation experiment

Model	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
ResNet-free Model	98.38	98.16	98.61	98.39
CNN-free Model	98.27	97.29	98.55	98.39
An improved CNN Model based on ResNet	99.48	99.32	99.19	99.26

the selection of *batch size*, *epochs*, activation, and optimizer. *Batch size* and *epochs* are closely related. A cycle of training and adjusting network weights is called *epochs*, and *batch size* is the number of data samples captured in a training session. *batch size* is too large or too small, which will affect the convergence and speed of the network. *batch size* increases at the same time, in order to reach the same accuracy, it is necessary to increase *epochs*. *batch size* is generally chosen The choice of *batch size* is generally concentrated between 16 and 256 and is a power of 2 for the best results, in this model the *batch size* is determined to be 64.

The activation function directly affects the performance of the network model, which can introduce nonlinear factors to help the model learn and simulate complex data features, and can also spatially transform the features through linear mapping, so that the data can be better categorized to improve the prediction accuracy of the network model. In this paper we choose the three most commonly used experimental activation functions, namely *Sigmoid*, *ReLU* and *Tanh*.

A suitable optimizer allows deep learning to adjust each parameter during backpropagation, guiding the loss function towards the global minimum approximation, selecting the optimal solution for the model, and avoiding overfitting. In this paper, we choose three optimizers such as *Adam*, *SGD* and *NAdam* for our experiments.

Comparison experiments are conducted on the two hyperparameters Activation and Optimizer which are set differently to determine their effects on the model evaluation indicators, as shown in Table 1.The best performing model is to use the *ReLU* function to activate and select *Adam* as the optimizer.

4.4 Model Performance Test

In each iteration of the model, a batch of training samples is used to calculate the gradient, and the gradient is used to update the model parameters to gradually optimize the model. The appropriate number of iterations needs to be tested and verified. If the number of iterations is too small, the model may not have enough time to learn the complex pattern of data, and too many iterations may lead to overfitting of the model, and the detection effect will be poor. After experimental comparison, the model with the highest performance is activated using the *ReLU* function, and *Adam* is selected as the optimizer. In order to further optimize the model, this paper selects the different values of *Epoch*, *Batch Size* and *LR* as the test of model performance.

(1) The influence of *Epoch* difference on the model performance index.

In the case of unified activation function and optimizer, as shown in the line chart of Figure 2, with the increase of the number of iterations from 1 to 20, the four evaluation indexes of Accuracy, Precision, Recall and F1 are gradually improved, and finally tend to be stable after 20 iterations. In the iterative process, the model is continuously modified and improved. The highest Accuracy is 99.48 %, the precision is 99.32 %, the Recall is 99.19 %, and the F1 is 99.26 %, achieving effective network traffic detection results.

(2) The influence of different *Batch Size* on the model performance index

In the case of unifying the activation function and the optimizer, the determination of the size of the *batch size* also has a certain impact on the training process and results. The smaller *batch size* means more frequent parameter updates, which will enable the model to respond faster to changes in data, but may require more iterations to converge ; a larger *batch size* may cause the model to overfit the training data and affect the model detection accuracy. As shown in Figure 3.

(3) The influence of different *LR* on the model performance index

In the case of unified activation function and optimizer,

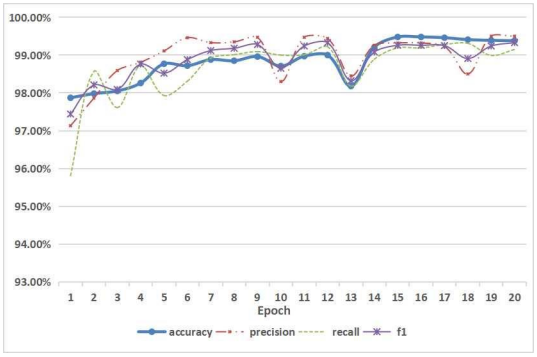


Figure 2: Classification results display

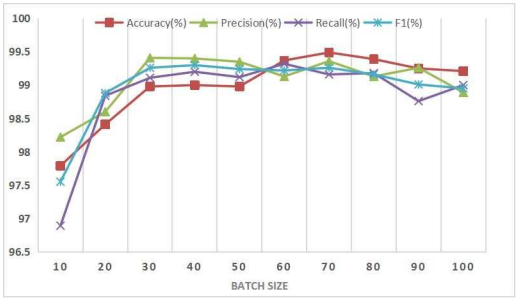


Figure 3: Classification results display

in order to further analyze the influence of learning rate learning rate on the performance of the model, this paper selects six different learning rate tests. If the learning rate is too large or too small, the model cannot achieve local optimum. As shown in Figure 4, when $lr = 0.001$, the overall comprehensive performance is optimal.

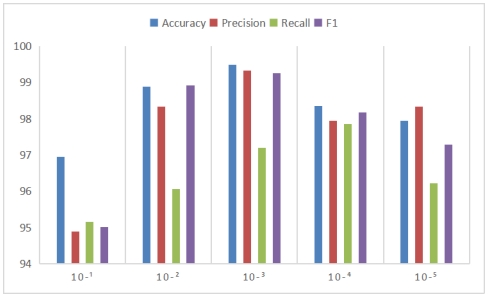


Figure 4: Classification results display

After the above performance comparison experiments, the model hyperparameter settings were determined to be *batch size* = 64, *epochs* = 20, $lr = 0.001$, *optimizer* = *Adam*, *activation* = *ReLU*.

4.5 Comparison of Different Methods of Intrusion Detection Model

The main contribution of this paper is that the detection accuracy of the intrusion detection model has been significantly improved. The accuracy of the improved CNN model proposed in this paper is compared with the existing intrusion detection models constructed using other different methods, highlighting the effectiveness of this research in improving detection accuracy. The results of Table 2 show that the accuracy of the model is improved by 11.09 % compared with the common intrusion detection models based on artificial neural network and feature selection. Compared with RNN, the accuracy of the model is improved by 1 %. Compared with CNN, the accuracy of the model is improved by 2.39 %. Compared with the residual network, the accuracy of the model is improved by 1.21 %. Similarly, the model in this paper has a relative improvement in precision, recall and f1 compared with the other models mentioned above.

4.6 Ablation Experiment

In order to further verify the robustness of our proposed method, we carried out the ablation experiment of the model, as shown in Table 3. For our proposed method : an improved CNN model based on ResNet, one component is removed in turn and compared with the original model to perform ablation test.

The results of Table 3 show that no matter what component is removed, the four performance indicators of the model will decrease. Among them, the accuracy of our proposed model is 99.48 %. When the ResNet structure is removed, the network performance decreases significantly. If the network structure is deepened, the detection effect will be worse. When the CNN structure was removed, the result was the worst, with Precision only 97.29 %. It shows that the convolutional network can effectively learn the corresponding features from a large number of samples, avoiding the complex feature extraction process, and the residual can avoid overfitting on the basis of convolution, so as to make the prediction results more accurate.

5 Conclusions

In this paper, we proposed an intrusion detection method based on ResNet an improved CNN. Our main contribution is to introduce an improved intrusion detection method that enhances the network performance by utilizing the residual learning principle to build residual structures in the neural network model based on the original convolutional neural network. Compared with other methods, this paper makes the model benefit from the

deep network structure, which improves the accuracy of intrusion detection while effectively avoiding the overfitting of the network model. The performance test and ablation experiment of the proposed model are carried out in the NSL-KDD dataset. The results show the effectiveness and robustness of the proposed model.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (62072416), Key Research and Development Special Project of Henan Province (221111210500), and Key Technologies R&D Program of Henan Province (232102211053, 222102210170).

References

[1] M. Arief and S. H. Supangkat, "Comparison of cnn and dnn performance on intrusion detection system," in *2022 International Conference on ICT for Smart Society (ICISS)*, pp. 1–7, Bandung, Indonesia, 2022.

[2] Z. Y. Cai, J. C. Wang, and J. W. Zhang, "Intrusion detection algorithm based on residual neural network," *International Journal of Network Security*, vol. 24, no. 6, pp. 1135–1141, 2022.

[3] H. Ding, L. Chen, and L. Dong, "Imbalanced data classification: A knn and generative adversarial networks-based hybrid approach for intrusion detection," *Future Generation Computer Systems*, vol. 131, pp. 240–254, 2022.

[4] M. Earum, Z. Aneela, and U. Muhammad, "A two stage intrusion detection system with auto encoder and lstms," *Applied Soft Computing*, vol. 121, p. 108768, 2022.

[5] N. Gao, L. Gao, Q. Gao, and H. Wang, "An intrusion detection model based on deep belief networks," in *2014 Second International Conference on Advanced Cloud and Big Data*, pp. 247–252, 2014.

[6] J. Gu and S. Lu, "An effective intrusion detection approach using svm with naïve bayes feature embedding," *Computers & Security*, vol. 103, p. 102158, 2021.

[7] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," *Eai Endorsed Transactions on Security and Safety*, vol. 3, p. e2, 2016.

[8] I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," *Computer Networks*, vol. 188, p. 107840, 2021.

[9] F. E. Laghrissi, S. Douzi, and K. Douzi, "Intrusion detection systems using long short-term memory (lstm)," *Journal of Big Data*, vol. 8, no. 1, 2021.

[10] J. Lansky, S. Ali, and M. Mohammad, "Deep learning-based intrusion detection systems: A systematic review," *IEEE Access*, vol. 9, pp. 101574–101599, 2021.

[11] R. Jaffal M. H. Alshayeji, M. AlSulaimi, "Network intrusion detection with auto-encoder and one-class support vector machine," *International Journal of Information Security and Privacy (IJISP)*, vol. 16, no. 1, pp. 1–18, 2022.

[12] A. A. Salih and A. M. Abdulazeez, "Evaluation of classification algorithms for intrusion detection system: A review," *Soft Computing and Data Mining*, vol. 2, no. 1, pp. 31–40, 2021.

[13] A. Shaikh and P. Gupta, "Real-time intrusion detection based on residual learning through resnet algorithm," *International Journal of System Assurance Engineering and Management*, 2022.

[14] S. Sivamohan, S. S. Sridhar, and S. Krishnaveni, "An effective recurrent neural network (rnn) based intrusion detection via bi-directional long short-term memory," in *2021 International Conference on Intelligent Technologies (CONIT)*, pp. 1–5, 2021.

[15] A. Turaiki and A.Najwa, "A convolutional neural network for improved anomaly-based network intrusion detection," *Big Data*, vol. 9, no. 3, pp. 233–252, 2021.

[16] C. Wang, Y. Sun, and W. Wang, "Hybrid intrusion detection system based on combination of random forest and autoencoder," *Symmetry*, vol. 15, no. 3, 2023.

[17] Z. D. Wang, Y. D. Liu, and D. J. He, "Intrusion detection methods based on integrated deep learning model," *Computers & Security*, vol. 103, p. 102177, 2021.

[18] L. Wen, X. Y. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on resnet-50," *Neural Computing and Applications*, vol. 32, pp. 6111–6124, 2020.

[19] C. Zhang, D. H. Jia, and L. Wang, "Comparative research on network intrusion detection methods based on machine learning," *Computers & Security*, vol. 121, p. 102861, 2022.

[20] J. Zhang, Y. Ling, X. Fu, X. Yang, G. Xiong, and R. Zhang, "Model of the intrusion detection system based on the integration of spatial-temporal features," *Computers & Security*, vol. 89, p. 101681, 2020.

Biography

Zengyu Cai is an associate professor fellow of the College of Computer and Communication Engineering of Zhengzhou University of Light Industry. His research interests include network se-curity and artificial intelligence.

Pengrong Li is a graduate student of School of Computer and Communication Engineering, Zhengzhou University of Light Industry. Her research interests include network security and artificial intelligence.

Jianwei Zhang is a professor fellow of the College of Software Engineering of Zhengzhou University of Light Industry. His research interests include next generation

network and artificial intelligence.

Yajie Si is a graduate student of School of Computer and Communication Engineering, Zhengzhou University of Light Industry. Her research interests include network security and artificial intelligence.

Yuan Feng is an associate professor fellow of the College of Computer and Communication Engineering of Zhengzhou University of Light Industry. Her research interests include cyberspace security and artificial intelligence.

Image Enhancement and Cloud Secure Transmission Based on Reversible Image Information Hiding Technology

Zailin Li

(Corresponding author: Zailin Li)

School of 3D Printing, Xinxiang University, Xinxiang 453003, China

Email: lizailin@126.com

(Received Aug. 15, 2023; Revised and Accepted Apr. 24, 2024; First Online June 22, 2024)

Abstract

Image distortion and poor transmission security are the key problems restricting the reversible image information hiding technique. The existing image enhancement algorithms do not consider the restoration problem, failing to obtain the original image after obtaining secret information, and the object of study is usually the gray image in the medical field. This study proposes an adaptive image enhancement algorithm based on gray-level images and constructs a reversible information-hiding algorithm based on a constant tone plane. In addition, a new reversible information-hiding algorithm based on MSB prediction and error embedding is proposed to enhance the security of image transmission. Compared with the traditional algorithms, the gray image and color image enhancement algorithms proposed in this study improve the image quality by 8% and 15%, respectively. The change rate of image pixels encrypted by the cloud security transmission algorithm is more than 99%. The proposed image enhancement and encryption algorithm can significantly improve image quality, providing a development platform for applying reversible information-hiding technology in images.

Keywords: Cloud Secure Transmission; Error Location; Gray Level Histogram; Image Enhancement; Reversible Information Hiding Technology

1 Introduction

Nowadays, Reversible Data Hiding (RDH) was born. This technique can add secret message into various types of data carriers, which can completely extract secret information [18]. The combination of RDH and image enhancement technology has become a hot research direction recently. RDH technology has practical value, so it is required to study key RDH technologies for image enhancement and cloud secure transmission [8]. Scholars focus on the study of grayscale images in the medical and military fields. However, existing algorithms still have the problem of weak contrast of gray image, and there are few researches on algorithms based on color image enhancement and data encryption [12]. Therefore, three innovations are made in this research: first, an adaptive enhancement algorithm based on gray image is presented; secondly, the image enhancement and encryption algorithms of color images are studied, and the RDH algorithm of color image contrast enhancement with constant tone plane is proposed.

Thirdly, to ensure the security of image transmission during cloud transmission, a new encryption image RDH algorithm is proposed. In response to the problem that existing algorithms cannot simultaneously achieve high embedding capacity and good reconstructed image quality, this paper proposes a reversible information hiding algorithm for encrypted images based on the most significant bit prediction and error embedding. On the one hand, the innovation of the research takes into account all types of prediction errors and can correctly recover all pixels with prediction errors; On the other hand, the algorithm in this paper uses error blocks to mark the location of prediction errors, and uses message blocks to embed secret information.

The contribution of this study is to provide reference for the fusion of reversible information hiding algorithms and image enhancement algorithms, while also exploring research on color images. This research mainly analyzes the image RDH technology from four aspects: The first part is the review and discussion of the current RDH algorithm related literature; the second part is to build RDH algorithm based on gray image, color image and cloud security transmission. The third part is the comparative analysis and application test of the results of image enhancement and encryption algorithms. The last part is the conclusion of the full text.

2 Related Work

Recently, RDH developed fast and formed a series of classic frameworks, such as histogram shift, prediction error extension, etc. Based on these classical algorithms, a large number of scholars have explored more optimized image enhancement techniques [5]. Among them, He et al. applied RDH to color images and constructed an image hiding strategy with 2D histogram translation based on the histogram shift framework, which could enhance the image quality by deleting the number of invalid pixels. The scheme processed images with higher PSNR ratio and smaller image files [6]. Wang and Liu proposed a variational histogram equalization framework to adjust image pixel values (PV) through energy functional. This algorithm had higher convergence and feasibility [21]. Chen et al. transformed the encrypted image with the freely selected image. To ensure image conversion's quality, an algorithm based on gray co-occurrence matrix was proposed, which used medical feature extraction to improve model's efficiency and image quality. Under this method, the RMS value was reduced by 5% [2]. Wu et al. constructed a histogram translation algorithm with checkerboard development for prediction optimization of encrypted images. The image output by this enhancement algorithm has better bits per pixel [24]. Based on the cloud-edge model, Chen and Shiu suggested a new technique with distributed encrypted images. This technology takes edge nodes as Bridges, uses XOR secret sharing as cryptographic tags, and imbeds differential extension for image and processing. This algorithm has good applicability in image RDH [3].

To ensure the security of images in RDH technology, scholars in this field studied a large quantity of models with image cloud security transmission. Among them, Anushiadevi and Amirtharajan introduced elliptic curve cryptography to obtain encryption before embedding, and combined the original image with the confidential data through the addition homomorphism property. The image memory after encryption was unchanged by this method, and 100% reversibility was achieved [1]. Wang et al. suggested a new block-based image encryption method. This method hides the encrypted data into each block through Huffman coding, which ensures the image's security while achieving a higher embedding rate [22]. Qu et al. proposed a reversible data hiding method for encrypted images, which enhanced its embedding capacity and security. This method also ensured the embedding content and security of the image [16]. Qian et al. suggested a hiding algorithm for encrypted color images. In this algorithm, the user could decrypt the original bit stream directly by constructing a mark encrypted image bit stream. This method had a higher embedding rate and easier user-oriented operation [15]. Wang et al. suggested an image cloud transmission encryption technology based on quadtree segmentation and integer wavelet transform. The technology encrypted images by 2×2 blocks to ensure higher security, and introduced integer wavelet transform to transform the encrypted images. This scheme improved the embedding rate of encrypted data [23]. Yin et al. proposed a point target policy algorithm based on adaptive Most Significant Bit (MSB) and Haverman coding, and conducted experimental tests. Compared with other algorithms, this algorithm had a higher embedding rate [25].

To sum up, image enhancement and image encryption based on RDH technology have developed so far, and a large number of algorithms have emerged. There are histogram shift and prediction error based on image enhancement, and MSB encryption algorithm based on cloud security transmission. However, the above algorithms still have many shortcomings. To handle the issue of low contrast and low security of gray image, an adaptive enhancement algorithm of gray image is constructed. To handle the issue of uneven color and saturation in color image enhancement and hiding, an RDH algorithm based on constant tone plane is constructed. To handle the issue of low information security in cloud transmission, an encrypted image RDH algorithm is proposed. The aim of this study is to provide better scientific advice for image RDH technology.

3 Image Enhancement and Encryption in RDH Technology

Currently, the algorithms for image RDH technology are not perfect, and image distortion and transmission security are two prominent problems. Therefore, an adaptive enhancement algorithm with gray-scale image is constructed. Meanwhile, the image enhancement and encryption algorithms of color images are discussed, and a color image contrast enhancement algorithm based on constant tone plane is suggested. In addition, to ensure the security of image transmission, a new RDH algorithm with MSB Prediction and Error Embedding (MSB-EE) is constructed.

3.1 RDH Technology Based on Image Contrast Enhancement

The existing methods for gray image enhancement are mainly concentrated in the military medical field, but these fields have high requirements for image security and quality, and some traditional algorithms can not meet the needs of both. Therefore, an adaptive gray-scale image enhancement algorithm is proposed in this paper by combining RDH technology and image enhancement. The algorithm enhances contrast by performing Pixel concentration ratio (PCR) segmentation for Regions of interest (ROI) and embedding secret information. The gray value of Regions of non-interest (NROI) is lowered to embed more secret information. The algorithm flow is shown in Figure 1.

First, this study used an adaptive threshold detector to determine the threshold and divide the ROI and NROI.

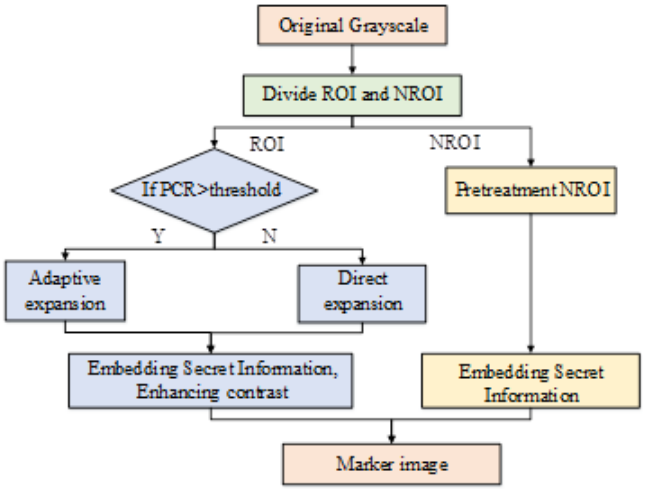


Figure 1: Grayscale enhancement process based on reversible information hiding

The specific partitioning method is as follows: when the PV is greater than the threshold value, it is regarded as the research target, and the unified value is 0, which is displayed in black; when the PV is less than the threshold, the unified value is 255 and is displayed in white. Finally, the outermost PV of 0 is used as the ROI boundary. When histogram stretching and secret information embedding are performed on ROI, PCR needs to be calculated first. The result is the ratio of pixels' number in the target bias [0,127] or [128,255] to the pixels' total number in the ROI. The calculation method of pixel moving distance D_{shift} in ROI region is shown in Equation (1).

$$D_{shift} = \begin{cases} 255 - ROI_{\max}, & PCR_{[0,127]} > r \\ ROI_{\min}, & PCR_{[128,255]} > r \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

In Equation (1), $ROI_{\max}$ and $ROI_{\min}$ are the biggest and smallest PV; r is a ratio between 0 and 1, setting the value to 0.7 based on a large number of calculations. Then you need to stretch the histogram to create more space to add secret message, and the stretching equation is shown in Equation (2).

$$ROI_{stretch}(x,y) = round[(L_{\max} - L_{\min}) \times \frac{ROI_{shift}(x,y) - ROI'_{\min}}{ROI'_{\max} - ROI'_{\min}} + L_{\min}] \quad (2)$$

In Equation (2), $ROI_{stretch}(x,y)$ is the PV of the stretched image; $L_{\max}$ and $L_{\min}$ are the upper and lower boundaries of the stretched pixel; $ROI_{shift}(x,y)$ the PV of the picture after moving the corresponding pixel; $ROI'_{\max}$ and $ROI'_{\min}$ are the biggest and smallest values. By stretching the image, more space is gained. By embedding the secret information into the non-empty PV surrounded by the empty PV until there is no non-empty PV inside, the specific calculation method of embedding is shown in Equation (3).

$$k' = \begin{cases} k + b_i, & \text{if } k = k_{peak} \ \& \ k_{peak} \in [0,126] \\ & \quad \& h(k_{peak}+1) = 0 \\ k - b_i, & \text{if } k = k_{peak} \ \& \ k_{peak} \in [129,255] \\ & \quad \& h(k_{peak}-1) = 0 \\ k, & \text{if } k \neq k_{peak} \end{cases} \quad (3)$$

In Equation (3), k is the value of $ROI_{stretch}(x,y)$; b_i denotes the i position, which has a value of 0 or 1. k_{peak} is the peak pixel, that is, it has the biggest number of pixels and the adjacent PV are empty; $h(k_{peak})$ is the number of pixels of the peak pixel. After multiple embeddings, the histogram distribution before and after the algorithm embeddings is shown in Figure 2.

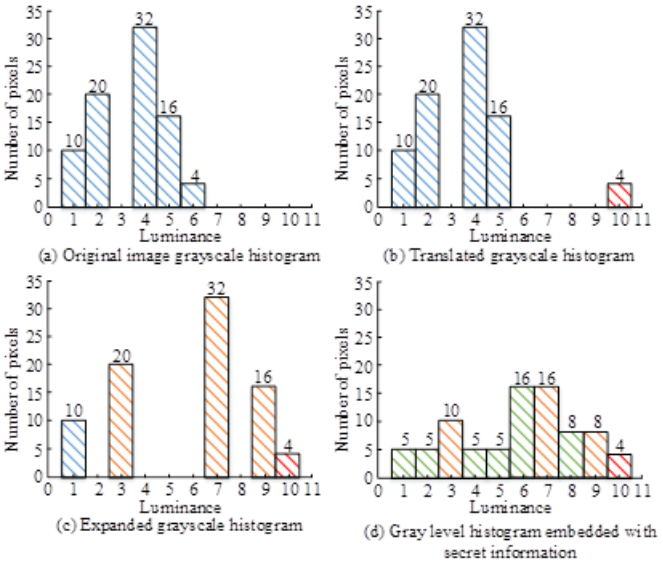


Figure 2: Schematic diagram of histogram distribution before and after embedding

To ensure the quality of the image, the study enhances the internal contrast by reducing the value of NROI pixels.

$$NROI_{preprocess}(x,y) = NROI(x,y) - NROI_{\min} \quad (4)$$

In Equation (4), $NROI_{preprocess}(x,y)$ denotes the background PV after processing; $NROI(x,y)$ refers to the original background PV; $NROI_{\min}$ denotes the minimum PV in the background. When the secret information cannot be fully embedded within the ROI, the remaining information will be embedded in the background. The secret information is replaced by the Least Significant Bit (LSB) of up to three pixels in the background to satisfy that the PV of the background changes within the range of [0,7] after embedding. Use N to denote the number of secret information bits which required to be embedded in the background, and the embedding equation is shown in Equation (5).

$$LB_j = b_i, j \in \{1,2,3\} \quad (5)$$

In Equation (5), LB_j denotes the j-th last bit of NROI. After the user sends the picture with the secret message to the receiver, the user needs to get the secret message and the original picture through image enhancement, which is mainly split into three steps. First, LSB is read to obtain the quantity of secret message bits in NROI, and then Equation (5) is used to reverse solve the secret message in NROI. Secondly, the secret message in ROI is obtained by inverse solution of Equation (3). Finally, Equation (1) and Equation (2) are used to reverse solve the original image.

When encountering the RDH and image enhancement of color images, due to the limitation of hue and saturation in color images, the processing method of gray image is not suitable for this [20]. Therefore, a contrast enhancement and RDH algorithm for color pictures is proposed with constant tone plane. The specific algorithm processing flow is shown in Figure 3.

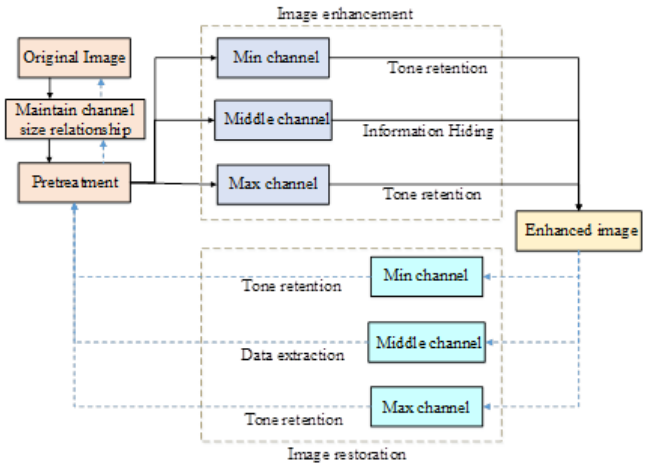


Figure 3: Flowchart of reversible information hiding and enhancement algorithm for color images

In the preprocessing, this study makes space for the secret information hiding by the size relation preserving operation; Then the IP of the pre-processed image is obtained through the mapping relationship, and the mapping method is shown in Equation (6).

$$Value_{mapping} = round[\frac{Value_{maintain}}{255} \times (255 - S \times 2) + S] \quad (6)$$

In Equation (6), $Value_{mapping}$ denotes the PV after mapping; $Value_{maintain}$ denotes the PV maintained after the channel size relationship; S is the PV that needs to be left empty for the secret message. In this paper, the segm-segmed middle channel is selected for secret information hiding, and secret message is added into the picture after preprocessing, as shown in Equation (7).

$$I_{e-mid}(a,b) = \begin{cases} I_{p-mid}(a,b)-1, & \text{if } I_{p-\min d}(a,b) < f_L \\ I_{p-mid}(a,b)-b, & \text{if } I_{p-\min d}(a,b) = f_L \\ I_{p-mid}(a,b), & \text{if } f_L < I_{p-\min d}(a,b) < f_R \\ I_{p-mid}(a,b)+b, & \text{if } I_{p-\min d}(a,b) = f_R \\ I_{p-mid}(a,b)+1, & \text{if } I_{p-\min d}(a,b) < f_R \end{cases} \quad (7)$$

In Equation (7), $I_{e-mid}(a,b)$ denotes the median value of embedded pixels; $I_{p-mid}(a,b)$ denotes the median value of pixels separated after preprocessing; b_i denotes the value of i-th secret information; f_L and f_R are the PV with the most pixels, and $f_L < f_R$. To keep the color image hue constant, the research proposed a constant tone plane, as shown in Figure 4. All on the same plane have the same hue [11].

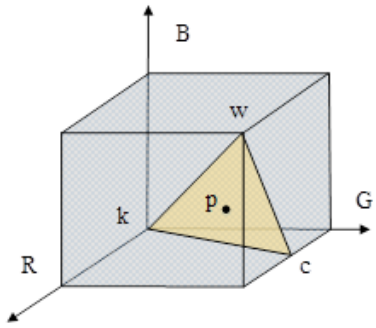


Figure 4: Schematic diagram of a constant tone plane

The value of the same constant tonal plane is set as (p_r, p_g, p_b), then the specific tonal preservation algorithm is shown in Equation (8).

$$\begin{cases} p'_r = p_r + amplitude \\ p'_g = p_g + amplitude \\ p'_b = p_b + amplitude \end{cases} \quad (8)$$

In Equation (8), (p'_r, p'_g, p'_b) denotes the PV after hue preservation; $amplitude$ is the value of the color change. After the tone is maintained, the minimum channel and maximum channel PV are adjusted as stated by the change value of the middle channel pixel to avoid image distortion [13]. Then, the receiver first obtains the PV of the three-channel variable recovery channel to obtain the secret information; then Equation (6) is applied to inversely solve the PV of the picture before mapping, and finally the original image is obtained.

3.2 Cloud Security Transmission Algorithm Based on MSB-EE

LSB is used in the above image encryption algorithm to hide information, but the hiding efficiency is lower than MSB algorithm when the hiding quality is the same [9]. Therefore, based on MSB-EE, a more applicable RDH

technology for image cloud security transmission is proposed. Figure 5 illustrates the encryption and decryption process.

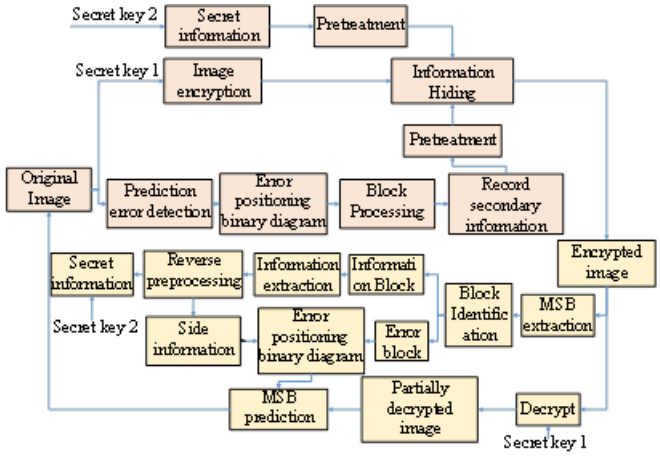


Figure 5: Process of reversible information hiding algorithm for encrypted images based on MSB-EE

First, the prediction error detection is performed on the initial image [7]. By comparing the initial PV, the predicted PV and the inverse MSB PV, the prediction error of the pixel is determined. The MSB is predicted 0 and 1. The calculation method of error location binary graph is shown in Equation (9).

$$M_e(i,j) = \begin{cases} 0, & \text{if } \triangle_I(i,j) < \triangle_{INV}(i,j) \\ 1, & \text{else} \end{cases} \quad (9)$$

In Equation (9), $M_e(i,j)$ denotes the value of each pixel in the error positioning binary graph; $\triangle_I(i,j)$ is the absolute value of the difference between the predicted PV and the initial PV; $\triangle_{INV}(i,j)$ denotes the absolute difference between the inverse MSB PV and the predicted PV. Then, the labeled error location binary graph Mem is obtained by block processing. The specific labeling methods are three kinds of blocks: defining error blocks (including prediction error blocks), message blocks (for embedding information blocks), and marking blocks (distinguishing error blocks and message blocks).

To make the image content unreadable during transmission, the image is encrypted. Then, the pseudo-random sequence and the original image are XOR operation to get the encrypted image, the model is shown in Equation (10).

$$I_e(i,j) = \sum_{k=1}^{8} (I(i,j)_k \oplus S(i,j)_k) \times 2^{k-1} \quad (10)$$

In Equation (10), $I_e(i,j)$ is the PV after encryption; k denotes bit k of the binary stream; $I(i,j)$ denotes the initial PV; $S(i,j)$ denotes a pseudo-random sequence PV. To prevent the receiver from misdiagnosing the content of the marker block after receiving the encrypted image, the study defined the secondary information SI to record the predicted error position, recorded the second piece through Fs, and recorded the last piece with FI. When the value is 0, it means that there is no prediction error in the block, otherwise it is 1. The second part of the secondary information record is the same content as the flag block in the error block. Change the first binary code in the error block to 0 when both are the same. The modified prediction error position was recorded simultaneously by coordinate method and arithmetic coding. The result with the lower median value of the two is taken as the prediction error result SI.

Before embedding secret information and secondary information, the research distinguishes message blocks from marker blocks by preprocessing. Secondary information is processed in the same way as secret information, as shown in Equation (11).

$$T(k) = \begin{cases} 1, & \text{if } k > 5 \text{ and } \sum_{u=0}^{5} SM_e(k-u) = 6 \\ & \quad \text{and } \sum_{u=1}^{5} T(k-u) = 0 \\ 0, & \text{else} \end{cases} \quad (11)$$

In Equation (11), $T(k)$ is the intermediate variable. When it is 1, you need to insert A 0 after bit k of information. When it is 0, no encoding is inserted. SM_e is the embedded secret information with the same value as T. After the initial picture is encrypted, the original MSB value in the picture needs to be replaced with the PV after the secret message is hidden, and the replacement method is shown in Equation (12).

$$I_{em}(i,j) = \begin{cases} I_e(i,j), & \text{if } M_{em}(i,j) \in B_e \\ & \&\triangle_I(i,j) = \triangle_{INV}(i,j) \\ M_{em}(i,j) \times 128 & \\ \quad + \text{mod}\,(I_e(i,j), 128), & \text{else} \end{cases} \quad (12)$$

In Equation (12), $I_{em}(i,j)$ denotes the PV after marking encryption; $M_{em}(i,j)$ denotes the encrypted PV. The specific replacement process is shown in Figure 6.

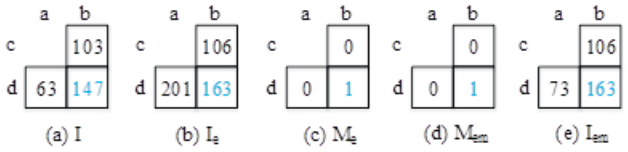


Figure 6: The embedding process of prediction error pixel information

When the receiver gets the encrypted image, the initial picture and secret message are obtained through reverse processing. When extracting information, MSB value is obtained by Equation (13).

$$M_{MSB}(i,j) = [I_{em}(i,j)/128] \quad (13)$$

In Equation (13), $M_{MSB}(i,j)$ denotes the MSB value of the corresponding pixel of I_{em}. Then, the reconstructed image is reversely solved according to Equation (10) to

get the partially decrypted image PV, and the complete $M_e(i,j)$ is obtained according to Equation (14).

$$M_e(i,j) = \begin{cases} 1, & \text{if } M_{MSB}(i,j)=1 \\ & \&M_{MSB}(i,j) \in B_e \\ 0, & \text{else} \end{cases} \quad (14)$$

In Equation (14), B_e is sequence M_e. Finally, according to Equation (15), the image is reconstructed and the original image is obtained.

$$I(a,b) = \begin{cases} I'(a,b)_{MSB=0}, & \text{if } (\triangle_0(a,b) < \triangle_1(a,b) \\ & \&M_e(a,b)=0) \\ & \text{or } (\triangle_0(a,b) > \triangle_1(a,b) \\ & \&M_e(a,b)=1) \\ I'(a,b)_{MSB=1}, & \text{if } (\triangle_0(a,b) > \triangle_1(a,b) \\ & \&M_e(a,b)=0) \\ & \text{or } (\triangle_0(a,b) < \triangle_1(a,b) \\ & \&M_e(a,b)=1) \\ I'(a,b), & \text{else} \end{cases} \quad (15)$$

In Equation (15), $I'(a,b)$ denotes the PV of partially decrypted image; $\triangle_0(a,b)$ denotes the absolute value of the predicted PV and $I'(a,b)_{MSB=0}$; $\triangle_1(a,b)$ denotes the absolute value of the predicted PV and $I'(a,b)_{MSB=1}$. The encrypted image completed by this method can embed more secret information, and it is not easy to steal information by non-recipients.

4 Performance Analysis of Image Enhancement and Encryption Algorithms based on RDH Technology

Firstly, the performance of adaptive enhancement algorithm based on gray image is tested in gray image database. The effect of image quality enhancement is judged by comparing the histogram and visual vision before and after image embedding information, and compared with other popular algorithms. Different image quality evaluation indexes are applied to assess the performance of the proposed algorithm for picture information hiding and quality enhancement. Figure 7 shows the histogram comparison of grayscale images before and after embedding information.

In Figure 7, the expanded histogram leaves more space for secret information to be embedded. After information embedding, the distribution trend is still the same as the initial picture, so as to avoid the situation of picture distortion, and the secret information embedding is more uniform. At the same time, the suggested method is compared with three other methods, including the RDH algorithm based on Receding horizon control Reversible data hiding. RHCRDH), (Automatic Color Equalization Reversible data hiding, ACERDH and Reversible data hiding Multi-box pruning (RDHMBP) algorithm to enhance contrast and maintain original image brightness. The experiments were compared when the embedded secret information was 0.1 bit/pixel, 1 bit/pixel and 2 bit/pixel respectively, and the results were shown in Figure 8.

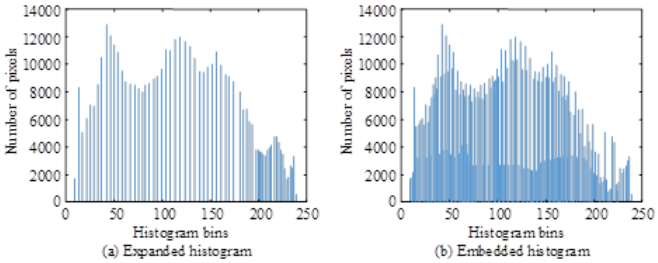


Figure 7: Comparison of Histograms after Expansion and Embedding

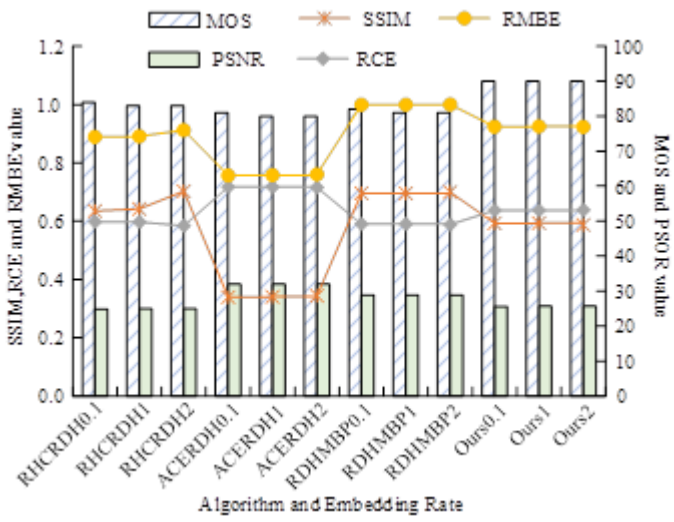


Figure 8: Evaluate results using indicators from different algorithms

In Figure 8, to compare the performance of different algorithms, five indicators are introduced to evaluate image quality, namely MOS and PSNR, SSIM, RCE and RMBE. It can be found that the algorithm proposed in this study has the highest MOS score, and has better image enhancement effect. And the RCE value is greater than 0.6, which indicates that the processed picture's quality is enhanced and the effect is better than that of the initial picture.

To assess the effect of the RDH for color image contrast enhancement based on constant tonal plane proposed in this study, three new indexes are added to the indexes selected for gray image. Hueand Lightness Dependent Correction to Industrial Colour Difference Evaluation, Hu1eand lightness dependent correction to industrial colour difference Evaluation, CIEDE2000 and No reference image spatial quality evaluator (BRISQUE). It is compared with the current popular color image enhancement algorithms, including the contrast enhancement model with deep learning to handle the issue of low illumination image. Contrast limited adaptive his-

togram equalization, Contrast limited adaptive histogram equalization, CLAHE and Multi-Scale Retinex with Color Restoration (MSRCR). Figure 9 shows the results after processing with four algorithms respectively.

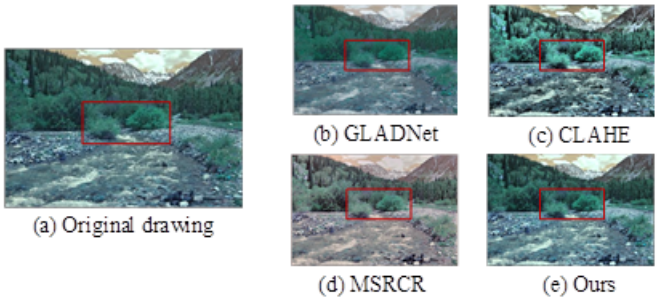


Figure 9: Comparison of Detail Visibility of Different Algorithms

In Figure 9, the detail enhancement result of the suggested model is better. After GLADNet algorithm is used to enhance the image, the retention effect of the details such as leaves is poor. After CLAHE algorithm enhancement, the brightness of the image has been optimized, but the color conversion is rather stiff, resulting in the loss of the original color of the enhanced object. The image distortion is obvious and does not conform to human visual sense. After MSRCR algorithm enhancement, the overall color of the image changes, the red part deepens, and the color distortion occurs. There is a wide range of color distortion, and the overall effect has a large deviation from the initial picture. After the enhancement of color image by the proposed method, it not only ensures the relative balance of the original image hue, contrast and brightness, but also increases the details in the leaves and water, which has a better enhancement effect on the image. At the same time, the above image quality index is applied to assess the effect of the model objectively. Two commonly used color image datasets were selected, denoted by I and II respectively. By controlling the number of empty PV at both ends of the image histogram, the index values of various algorithms are used.

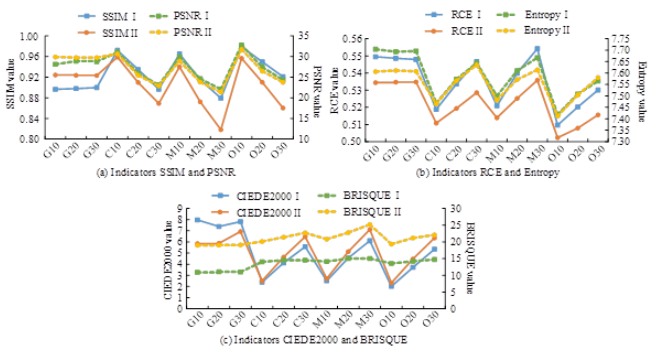


Figure 10: Comparison of indicators of different algorithms on datasets

The comparison of SSIM and PSNR is included in Figure 10. When the empty PV is 10, the SSIM and PSNR results of suggested model are the highest. The SSIM value is about 0.96, and the PSNR value is also above 30, indicating that the enhanced image quality is better than other algorithms. Figure 10(b) contains a comparison of RCE and Entropy. The higher the Entropy value, the more evenly distributed the grayscale histogram of the encrypted image is. Although the RCE result of the algorithm in this paper is only between 0.5 and 0.52, it still means that the image information is more abundant and the image distortion is avoided. As the number of empty PV increases, both the RCE and Entropy of the proposed method tend to rise. A comparison of CIEDE2000 and BRISQUE is included in Figure 10(c). A smaller CIEDE2000 value indicates a smaller difference between the original and the processed graph. The CIEDE2000 of the research method has the smallest value, which reaches 2, 3 and 5 respectively when the empty PV is 10, 20 and 30. The highest BRISQUE value is MSRCR algorithm, followed by the suggested model, which indicates that the picture enhancement quality is better.

Finally, to assess the effectiveness of the RDH with MSB-EE, a large number of images from two famous image libraries were selected for experiments. The two picture databases were replaced with Roman numerals III and IV respectively. Figure 11 respectively shows the size of the secret information and sub-information of the algorithm in the two image libraries.

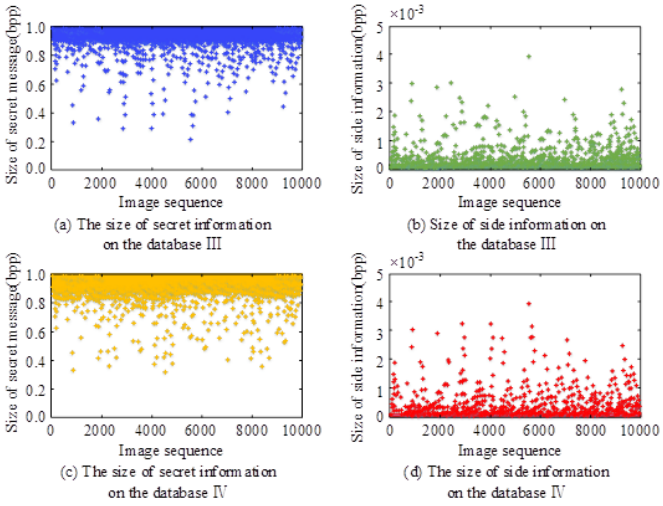


Figure 11: The size of information in the database

Figure 11(a) and Figure 11(b) respectively illustrates the distribution of secret information and secondary information on database III after using the proposed algorithm. Figure 11(c) and Figure 11(d) respectively show the distribution of secret information and secondary information on database IV after calculating by the suggested model. The size of the sub-information ranges from 10-3. To assess the safety of the suggested model, the exper-

iment is based on the above indicators. Two indexes, NPCR (the number of pixels change rate) and UACI (the unified average changing intensity), were added to assess the security of encrypted images. The correlation between horizontal direction and vertical direction is compared to judge the intensity of the algorithm encryption.

Table 1: The evaluation results of image indicators for the algorithm in this article

Index	Database III	Database IV
SSIM	0.008	0.007
PSNR (dB)	8.576	7.986
Entropy (bpp)	8.004	8.003
NPCR	99.548(%)	99.339(%)
UACI	33.567(%)	32.952(%)
Horizontal correlation	0.005	0.00
3 Vertical correlation	0.006	0.007

In Table 1, the NPCR value is above 99%, indicating that the safety of the encrypted picture is guaranteed. The correlation between horizontal direction and vertical direction is weak, indicating that the image encrypted by the suggested model is strong. Meanwhile, the research algorithm was compared with other encryption algorithms, and the PSNR values of these five algorithms were calculated respectively in Airplane, Lena, Peppers and Man, and other algorithms included in literature [14], literature [10], literature [4] and literature [19], as shown in Figure 12.

As shown in Figure 12, PSNR values of the four algorithms all decrease when the embedding rate increases. The most obvious decline is in the algorithm proposed in reference [19]. When the embedding rate reaches 0.8, the PSNR has dropped to 20dB. Compared with other algorithms, the suggested model achieves greater embedding rate and security, and less distortion in the image, and the result of image reconstruction is better.

In order to verify the image encryption effect of the research method, this experiment used ideal value deviation analysis to generate the encoding and decoding effects of seven images in BOWS-2. The specific experimental results are shown in Table 2. Seven images from BOWS-2 contain different texture information, with the main body being colored tiles, colored bricks, houses, mountains, lakes, sky, and crowds. The Dobbed values in the experiment are calculated as $D = \sum_{i=0}^{2} 55|H(C_i) - H(C)|/(M \times N)$, $H(C_i)$, and $H(C)$ representing the histograms of the ideal password image and the obtained password image, respectively. The lower the D value, the better the image encryption effect [17]. The table shows that the D value of most images is less than 0.05, indicating the superior image encryption effect of the research method.

In addition, to verify the effectiveness of the indicators NPCR and UACI, in this experiment, the pixel values of coordinates (20,30) were changed from 12 to 13,

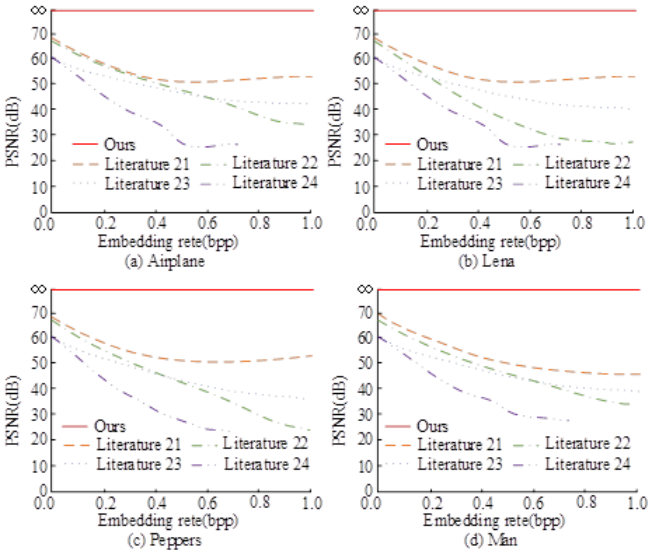


Figure 12: Performance comparison of different algorithms

Table 2: Ideal value deviation analysis

Image	Number and proportion of error marker blocks	Dobbed values
Colored tiles	2209 (6.7413%)	0.04805
Colored bricks	1418 (4.3274%)	0.04937
House	5467 (16.6840%)	0.04835
Mountains	244 (0.7446%)	0.05141
Lake	299 (0.9125%)	0.04864
Sky	107 (0.3265%)	0.04876
Crowds	948 (2.8931%)	0.04911

Table 3: Sensitivity analysis of NPCR indicators and UACI

Evaluating indicator	P(20,30)	P(155,100)	P(200,300)	P(512,512)
NPCR (%)	99.6223	99.6169	99.6169	99.6140
UACI (%)	33.4851	33.4004	33.4051	33.4312

coordinates (155,100) were changed from 14 to 15, coordinates (200,300) were changed from 169 to 170, and coordinates (512,512) were changed from 65 to 66. Analyze the changes in NPCR and UACI under pixel changes, as shown in Table 3. From the table, it can be seen that with minor changes to the pixel values of a coordinate point in the plaintext image, the NPCR and UACI values calculated using the algorithm in this paper are close to or higher than the calculated values of 99%, indicating that the algorithm in this paper has good differential attack resistance performance.

5 Conclusion

To realize better image enhancement and image encryption in image RDH technology, image enhancement and image hiding algorithms based on image RDH are proposed respectively. In this study, a variety of evaluation indexes are used to evaluate the image quality after the algorithm processing. The suggested model is compared with other commonly used models: In the algorithm to achieve gray level image enhancement, the histogram distribution of gray level distribution of the enhanced image is more uniform after information embedding. The suggested model has the highest MOS score and better image enhancement effect. Among the color image enhancement models, the suggested model has better detail enhancement performance. It not only ensures the relative balance between the original image hue and contrast, brightness, etc., but also increases the details in the picture, and has a better enhancement effect on the picture. When the empty PV is 10, the SSIM obtained by the proposed algorithm reaches about 0.96. PSNR value is also above 30; the RCE results were only between 0.5 and 0.52. The CIEDE2000 of the research algorithm has the minimum value. When the empty PV is 10, 20 and 30, it reaches 2, 3 and 5 respectively. The BRISQUE value is between 15 and 25. When assessing the effectiveness of the suggested cloud transmission model, the NPCR value is above 99%. The correlation between horizontal direction and vertical direction is weak, indicating that the proposed algorithm can encrypt images with high intensity. However, there are still some shortcomings in the research. MSB is used in the algorithm to encrypt images, which limits the capacity of embedded information. Therefore, a way is needed to expand the embedding capacity of secret message in the future.

References

[1] R. Anushiadevi, and R. Amirtharajan, "Reversible data hiding in an encrypted image using the homomorphic property of elliptic curve cryptography," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 5, pp. 5583-5594, 2021.

[2] X. Chen, H. Zhong, and Z. Bao, "A GLCM-feature-based approach for reversible image transformation," *Computers, Materials and Continua*, vol. 59, no. 1, pp. 239-255, 2019.

[3] Y. Chen, and C. Shiu, "Distributed encrypted image-based reversible data hiding," *Journal of Internet Technology*, vol. 22, no. 1, pp. 101-107, 2021.

[4] A. Durdu, "Nested two-layer RGB based reversible image steganography method," *Information Technology and Control*, vol. 50, no. 2, pp. 264-283, 2021.

[5] Y. Fu, P. Kong, H. Yao, Z. Tang, and C. Qin, "Effective reversible data hiding in encrypted image with adaptive encoding strategy," *Information Sciences*, vol. 494, no. 1, pp. 21-36, 2019.

[6] B. He, Y. Chen, Y. Zhou, Y. Wang, and Y. Chen, "A novel two-dimensional reversible data hiding scheme based on high-efficiency histogram shifting for JPEG images," *International Journal of Distributed Sensor Networks*, vol. 18, no. 3, pp. 354-362, 2022.

[7] M. S. Hwang, E. F. Cahyadi, Y. C. Chou, C. Y. Yang, "Cryptanalysis of Kumar's Remote User Authentication Scheme with Smart Cards," in *14th International Conference on Computational Intelligence and Security (CIS'18)*, Hangzhou, China, pp. 416-420, 2018.

[8] J. A. Kaw, N. A. Loan, S. A. Parah, K. Muhammad, J. A. Sheikh, and G. M. Bhat, "A reversible and secure patient information hiding system for IoT driven e-health," *International Journal of Information Management*, vol. 45, no. 4, pp. 262-275, 2019.

[9] L. H. Liu and J. Cao, "Analysis of One lightweight authentication and key agreement scheme for internet of drones," *International Journal of Electronics and Information Engineering*, vol. 13, no. 4, pp. 142-148. 2021.

[10] Z. Liu, and C. Pun, "Reversible image reconstruction for reversible data hiding in encrypted images," *Signal Processing*, vol. 161, no. 8, pp. 50-62, 2019.

[11] F. Masood, J. Masood, H. Zahir, K. Driss, N. Mehmood, and H. Farooq, "Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data," *Journal of Com-*

putational and Cognitive Engineering, vol. 2, no. 1, pp. 57-67, 2023.

[12] Q. Mo, H. Yao, F. Cao, Z. Chang, and C. Qin, "Reversible data hiding in encrypted image based on block classification permutation," *Computers, Materials, and Continuum*, vol. 59, no. 1, pp. 119-133, 2019.

[13] M. M. Nabi and F. Nabi, "Cybersecurity mechanism and user authentication security methods," *International Journal of Electronics and Information Engineering*, vol. 14, no. 1, pp. 1-9, 2022.

[14] M. Navetha, "Survey on secured reversible image data hiding techniques," *Research Journal of Engineering and Technology*, vol. 10, no. 1, pp. 4-10, 2019.

[15] Z. Qian, H. Xu, X. Luo, and X. Zhang, "New framework of reversible data hiding in encrypted jpeg bitstreams," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 351-362, 2019.

[16] L. Qu, H. He, S. Zhang, and F. Chen, "Reversible data hiding in encrypted images based on prediction and adaptive classification scrambling," *Computers, Materials, and Continuum*, vol. 63, no. 3, pp. 2623-2638, 2020.

[17] D. Ravichandran, S. Fathima, V. Balasubramanian, A. Banu, Anushiadevi and R. Amirtharajan, "DNA and chaos based confusion-diffusion for color image security," in *International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN'19)*, Vellore, India, pp. 1-6, 2019.

[18] M. Shah, W. Zhang, H. Hu, X. Dong, and N. Yu, "Prediction error expansion based reversible data hiding in encrypted images with public key cryptosystem," *IET Image Processing*, vol. 13, no. 10, pp. 1705-1713, 2019.

[19] F. H. Shajin, and P. Rajesh, "FPGA realization of a reversible data hiding scheme for 5G MIMO-OFDM system by chaotic key generation-based paillier cryptography along with LDPC and its side channel estimation using machine learning technique," *Journal of Circuits, Systems and Computers*, vol. 31, no. 5, pp. 2250093, 2021.

[20] J. Wang, Z. Sun, and G. Li, "High capacity reversible data hiding algorithm based on parabolic interpolation space," *Tien Tzu Hsueh Pao/Acta Electronica Sinica*, vol. 47, no. 1, pp. 137-144, 2019.

[21] W. Wang, and R. Liu, "A saturation-value histogram equalization model for color image enhancement," *Inverse Problems and Imaging*, vol. 17, no. 4, pp. 746-766, 2023.

[22] X. Wang, C. Chang, and C. Lin, "Reversible data hiding in encrypted images with block-based adaptive MSB encoding," *Information Sciences*, vol. 567, no. 8, pp. 375-394, 2021.

[23] X. Wang, C. Chang, C. Lin, and C. Chang, "Privacy-preserving reversible data hiding based on quad-tree block encoding and integer wavelet transform," *Journal of Visual Communication & Image Representation*, vol. 79, no. 8, pp. 103203, 2021.

[24] M. Wu, T. Chang, H. Chen, Z. Yang, and S. Liu, "Reversible information hiding in images based on histogram shift method," *Sensors and Materials: An International Journal on Sensor Technology*, vol. 34, no. 7, pp. 2555-2566, 2022.

[25] Z. Yin, Y. Xiang, and X. Zhang, "Reversible data hiding in encrypted images based on multi-MSB prediction and Huffman coding," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 874-884, 2020.

Biography

Zailin Li obtained his BE in Electronic Information Engineering from Southwest normal university in 2005. He obtained a Master's degree in Electronic and Communication Engineering from Nanjing University of Science and Technology in 2012. Presently, he is working as an associate professor in the school of 3D printing, Xinxiang University. His areas of interest are computer communication technology, electronic information, image processing and network security.

An Improvement of A Robust Authentication Protocol for Multi-server Architecture Using Elliptic Curve Cryptography*

Min-Shiang Hwang[1,2,4], Hou-Wen Li[3], and Cheng-Ying Yang[5]
(Corresponding author: Cheng-Ying Yang)

Department of Computer Science & Information Engineering, Asia University[1]
Fintech and Blockchain Research Center, Asia University[2]
The Ph.D. Program in Artificial Intelligence, Asia University, Taiwan[3]
500, Lioufeng Rd., Wufeng, Taichung 41354, Taichung, Taiwan (R.O.C.)
Department of Medical Research, China Medical University Hospital, China Medical University[4]
Taichung 40402, Taiwan
Department of Computer Science, University of Taipei, Taipei, Taiwan[5]
Email: cyang@utaipei.edu.tw
(Received Mar. 8, 2024; Revised and Accepted June 23, 2024; First Online June 30, 2024)

Abstract

The user authentication scheme of the multi-server architecture is mainly that the user only needs to register once to access many distributed servers. This architecture solves the problem of repeated multiple registrations in many servers and can provide a scalable solution. Recently, Zhang *et al.* proposed a secure, robust remote user authentication scheme for multi-servers using elliptic curve cryptography. However, their proposed scheme can solve many severe flaws and authenticate messages while keeping users anonymous. However, we will show that their scheme couldn't withstand online password-guessing with a user's smart card attack and a denial of service attack. This article proposes improving user authentication protocol against these vulnerabilities in Zhang *et al.*'s protocol.

Keywords: ECC; Multi-server Architecture; Smart Card; User Authentication

1 Introduction

In cyberspace, users can shop online at any time, conduct online banking transactions, and pay TV and other network information services. Authentication is vital in building secure communication channels between information system participants. More robust user authentication schemes are urgently required secure communications among these actors.

The user authentication protocols are designed to authenticate authorization services in a server over an insecure public network. The user and server can mutually authenticate through remote user authentication schemes and then use the server's services [8, 20]. Many user authentication protocols have been proposed [1–4, 9, 11, 13, 14,16–19,21,25]. A good user authentication scheme must meet security, simplicity, and practicality [5, 12, 15].

In 2013, Pippal *et al.* proposed a multi-server user authentication protocol without a verification form [18]. Also, it allows legitimate users to access multiple servers without the help of a registry. That is to say, the user and the service server can finish mutually authentication. However, its verification protocol has a severe problem. We must keep this smart card safe because many sensitive parameters are stored in the user's smart card. In 2013, Li *et al.* demonstrated that Pippal *et al.* scheme is vulnerable to impersonation attacks, offline password-guessing attacks, and privileged insider attacks [13, 14]. Li *et al.* also demonstrated the remedy for flexible registration in many servers. Next, In 2017, Srinivas *et al.* demonstrated Li *et al.*'s scheme was vulnerable to several over-looked security vulnerabilities and proposed a new authentication for multi-servers [19]. Multi-server authentication protocols have recently been released to provide greater robustness and efficiency [2, 3, 21].

Recently, Zhang *et al.* introduced a secure user authentication protocol for multi-servers using ECC (elliptic curve cryptography) [25]. However, their proposed scheme can solve many severe flaws and authenticate messages while keeping users anonymous. However, we will show that their scheme couldn't withstand the denial of

*A portion of the material in this manuscript was presented at the ICMEA'23 conference in Incheon, Korea, in 2023.

service and online password-guessing with a user's smart card attacks. ECC is a high performance public cryptosystem [6, 7, 10, 22–24]. Therefore, this article proposes an improvement of user authentication protocol to remedy these vulnerabilities in Zhang *et al.*'s protocol.

The structure of the article is organized in the following. The review and weaknesses of the Zhang *et al.* scheme are described in Sections 2 and 3. In Section 4, we propose an improved a robust user authentication scheme for multi-server environments using ECC resistant to all possible attacks mentioned in Section 3. Finally, Section 5 concludes the paper.

2 Review of Zhang *et al.*'s Robust Authentication Protocol

The multi-server system of Zhang *et al.* protocol consists of three participants [25], the registration center (RC), the authorization server (S_j), and the end user (U_i). The registration center is a trusted party and manages the entire system. The server S_j is authorized to provide some services. The user U_i can apply to these available services.

This section reviews the authentication scheme for multi-server environments proposed by Zhang *et al.*. The scheme is divided into five phases in total: initialization, server registration, user registration, authentication key negotiation, and password change phases.

After initialization, server registration, and user registration phases, the server will get the secret $s_j = H(SID_j||x)$. Here, SID_j is the identity of the authorization server S_j and x is RC' secret key. The user U_i will get a smart card which stored $\{B_i = K_i \oplus A_i, P, P_{pub} = xP, r, c_i = H(ID_i||PW_i||r)\}$, where $K_i = H(ID_i||x)$, $A_i = H(ID_i||RPW_i)$, r is a random number which generated by the user U_i. ID_i and PW_i are the identity and password of U_i, respectively, and $RPW_i = H(PW_i||r)$.

The process of the authentication key agreement phase of the Zhang *et al.*'s protocol is described as follows. Whenever U_i wants to access R_j's services, the following actions are performed during the authenticated vital agreement phase.

Step A1: U_i inserts his/her smart card into a card reader and enters ID_i, PW_i. The smart-card then calculates $C_i^* = H(ID_i||PW_i||r)$. Next SC checks if C_i^* equals C_i in the stored SC. If no, the smart card terminates the process. Otherwise, SC calculates $X = \alpha \times P$, $X' = \alpha \times P_{pub}$, $K_i = B_i \oplus H(ID_i||H(PW_i||r))$, $D_i = Enc_{H(X||X')}(ID_i, SID_j, H(ID_i||K_i||SID_j))$, and random number α. Next, U_i transmits $M1 = \{D_i, X\}$ as a login require to S_j.

Step A2: After S_j receives $M1$, it generates a random integer β and computes $Y = \beta \times P$ and $V1 = H(D_i||s_j||Y)$. Next, the server S_j transfers $M2 = \{D_i, X, Y, V1\}$ to the register center RC.

Step A3: After RC receives $M2$, it first calculates $X' = x \times X$. Then, RC calculates $Dec_{H(X||X')}(D_i)$. Then, RC calculates $H(ID_i||H(ID_i||x)||SID_j)$. RC compares it with D_i. If $H(ID_i||H(ID_i||x)||SID_j)$ does not exist $Dec_{H(X||X')}(D_i)$, the register center RC terminates the session. Otherwise, the register center RC has successfully verified U_i. Netx, the register center RC will continue to check the validity of S_j. RC calculates $V1^* = H(D_i||H(SID_j||x)||Y)$ using the decrypted value SID_j of the above D_i and verifies $V1^*$ whether equal or not to $V1$. If not, RC rejects the request and stops the login request. Otherwise, the register center RC passes the request and computes $V2 = H(s_j||X||Y)$, $V3 = H(K_i||X'||Y)$. Finally, the register center RC replies a message $M3 = \{V2, V3\}$ to the authorization server S_j.

Step A4: After receiving $M3$, the authorization server S_j calculates $H(s_j||X||Y)$ and verifies it against $V2$. If unequal, S_j will reject the message and terminate the login request. Otherwise, the authorization server S_j successfully verifies the register center RC, then the server computes $SK_j = \beta \times X = \alpha\beta \times P$ and $V4 = H(X||Y||SK_j)$. Afterwards, the authorization server S_j commits $M4 = \{V3, V4, Y\}$ to U_i.

Step A5: After the smart card receives the response $M4$, check whether the equation $V3 = H(K_i||X'||Y)$ is established. If not, S_j terminates the session. Otherwise, the smart card S_j computes $SK_j = \alpha \times Y = \alpha\beta \times P$. The smart card S_j verifies whether $H(X||Y||SK_j)$ equals to $V4$. If not, S_j terminates the login request session. Otherwise, S_j computes $V5 = H(SID_j||Y||SK_j)$. Next, the smart card S_j responses $M5 = \{V5\}$ to S_j.

Step A6: After S_j receives $M5$, calculate and verify $V5$ whether equal to $H(SID_j||Y||SK_j)$. If it holds, the server S_j has successfully verified the user U_i, and thus mutual verification is complete. Otherwise, the session will be terminated.

After the user U_i, the server S_j, and the register center RC complete mutual authentication, U_i and S_j share the session public key $SK = H(SID_j||X||Y||SK_j)$.

3 Weakness of Zhang *et al.*'s Robust Authentication Protocol

This section shows that Zhang *et al.*'s protocol is insecure against online password-guessing and denial-of-service attacks against users' smart cards.

3.1 Online Password Guessing Attack Using User Smart Card

When an adversary steals or picks up a user's smart card, they may have the opportunity to use a password-guessing attack. During the authenticated key agreement phase:

1) The adversary guesses the user password PW_i'. The adversary enters ID_i and PW_i' into the user's smart card.

2) SC computes $C_i^* = H(ID_i||PW_i||r)$ and then compares C_i^* with C_i in SC. If C_i^* is not equal to C_i, SC terminates the service to the user. SC aborts the process. Otherwise, SC calculates $X = \alpha \times P$, $X' = \alpha \times P_{pub}$, $K_i = B_i \oplus H(ID_i||H(PW_i||r))$, $D_i = Enc_{H(X||X')}(ID_i, SID_j, H(ID_i||K_i||SID_j))$ and random number α. Next, the user U_i transmits a login request message $M1 = \{D_i, X\}$ to the authorization server S_j.

3) The attacker monitors the login request session. If a message is sent to the authorization server, it means the password has been guessed. Otherwise, the attacker would repeatedly guess the password and perform Step A1.

3.2 Denial of Service Attack

A hacker can intercept the login request message $M1 = \{D_i, X\}$ from Step A1 of the authentication key agreement phase. Next, the attacker transmits a login request message $M1 = \{D_i, X\}$ to S_j. The server will perform Steps A2, A3, A4, and A6 of the authentication key agreement phase. S_j did not discover this illegal user until Step A6.

For the denial of service attack, the server must calculate one ECC and one Hash function in Step A2. In Step A4, one ECC and two Hash functions need to be calculated. In Step A6, the Hash function needs to be calculated once. The server wasted 2 ECC calculations and 4 Hash function calculations. RC needs to calculate 1 ECC and 3 Hash functions in Step A5. These calculations will reduce the server's performance and cannot provide regular services.

4 The Proposed Robust Authentication Protocol

In the proposed scheme, the initialization, server registration, user registration, and password change phases are the same as Zhang *et al.*'s protocol. The authentication key agreement phase is modified as follows.

Step A1: U_i inserts his SC into a card reader and then enters ID_i and PW_i. SC produces α (a random number) and calculate $X = \alpha \times P$ and $X' = \alpha \times P_{pub}$. Next, SC calculates $C_i^* = H(ID_i||PW_i||r)$ and verifies if C_i^* equals to C_i. If no, SC calculates $D_i = Enc_{H(X||X')}(False||T_u)$ and sends $M1 = \{D_i, X\}$ to S_j. Here, T_u denotes a time stamp. If yes, SC calculates $K_i = B_i \oplus H(ID_i||H(PW_i||r))$, $D_i = Enc_{H(X||X')}(ID_i, SID_j, H(ID_i||K_i||SID_j||True), T_u)$ Next, the user U_i transmits a login request message $M1 = \{D_i, X\}$ to the authorization server S_j.

Step A2: After S_j receives $M1$, it produces a random integer β, computes $Y = \beta \times P$ and $V1 = H(D_i||s_j||Y||T_s)$. Next, the server S_j transfers $M2 = \{D_i, X, Y, V1, T_s\}$ to the register center RC.

Step A3: After RC receives $M2$, it first calculates $X' = x \times X$. Then, RC calculates $Dec_{H(X||X')}(D_i)$. If D_i is the message False, the user is illegal. Otherwise, RC checks the time stamp T_u is a new one. Next, RC calculates $H(ID_i||H(ID_i||x)||SID_j)$ and compares it with D_i to verify the user U_i. If $H(ID_i||H(ID_i||x)||SID_j)$ does not exist $Dec_{H(X||X')}(D_i)$, the register center RC terminates the session. Otherwise, the register center RC has successfully verified the user U_i and RC will continue to check the validity of the server S_j. the register center RC calculates $V1^* = H(D_i||H(SID_j||x)||Y||T_s)$ using the decrypted value SID_j of the above D_i and checks $V1^*$ whether equal to $V1$. If not, RC rejects the login request and stops the login request session. Otherwise, RC accepts the login request and computes $V2 = H(s_j||X||Y||T_r)$, $V3 = H(K_i||X'||Y||T_r)$. Next, the register center RC replies the message $M3 = \{V2, V3, T_r\}$ to the server S_j.

Step A4: After receiving $M3$, the authorization server S_j calculates $H(s_j||X||Y||T_r)$ and verifies it against $V2$. If unequal, S_j will reject the message and terminate the login request session. Otherwise, S_j successfully verifies the register center RC. The server S_j computes $SK_j = \beta \times X = \alpha\beta \times P$ and $V4 = H(X||Y||SK_j||T_r)$. Afterwards, the server S_j commits $M4 = \{V3, V4, Y, T_r\}$ to the user U_i.

Step A5: After the smart card receives the response $M4$, check whether the equation $V3 = H(K_i||X'||Y||T_r)$ is established. If not, the smart card SC terminates the session. Otherwise, SC computes $SK_j = \alpha \times Y = \alpha\beta \times P$ and verifies if $H(X||Y||SK_j||T_r)$ is equal or not to $V4$. If not, SC terminates the login request session. Otherwise, SC calculates $V5 = H(SID_j||Y||SK_j)$. Afterwards, the smart card S_j responses $M5 = \{V5\}$ to S_j.

Step A6: After S_j receives $M5$, calculate and verify $V5$ whether equal to $H(SID_j||Y||SK_j)$. If it holds, the server S_j has successfully verified the user U_i, and thus mutually verification is complete. Otherwise, the login request session will be terminated.

After the user U_i, the authorization server S_j, and the register center RC complete mutual authentication, U_i and S_j share the session public key $SK = H(SID_j||X||Y||SK_j)$.

Table 1: The cost computation for identifying the DoS attack

Steps	Zhange *et al.* [25]	The proposed scheme
Step A2 (Server)	$T_{ECC} + T_H$	$T_{ECC} + T_H$
Step A3 (RC)	$T_{ECC} + 2T_H$	$T_{ECC} + T_H + T_{Dec}$
Step A5 (Server)	$T_{ECC} + 3T_H$	-
Step A6 (Server)	T_H	-
Total	$3T_{ECC} + 7T_H$	$2T_{ECC} + 2T_H + T_{Dec}$

T_{ECC}: The computation of performing an ECC operation.
T_H: The computation of performing a hash function operation.
T_{Dec}: The computation of performing deciphering operation.

5 Security Ananlysis and Performance

The validity proof based on the BAN logic is similar to Zhang *et al.*'s article. Please refer to their article [11]. In this section, we only discuss the improved scheme that could overcome the weaknesses of Zhang *et al.* 's robust authentication protocol discussed in Section 3.

1) Online Password Guessing Attack Using User Smart Card: When an adversary steals or picks up a user's smart card, they may have the opportunity to use a password-guessing attack. During the authenticated key agreement phase, the adversary guesses the user password PW_i' and enters ID_i and PW_i' into the user's smart card. Next, the attacker monitors the login request session. The password is guessed if a message is sent to the authorization server. Otherwise, the attacker repeatedly guesses the password and performs Step A1.

   In the proposed scheme, despite whether the password is correct, the user (attacker) U_i always transmits a login request message $M1 = \{D_i, X\}$ to the authorization server S_j. Therefore, the proposed scheme can be against the online password guessing attack using a user smart card.

2) Denial of Service Attack: In a denial of service attack, an attacker can intercept the login request message $M1 = \{D_i, X\}$ from Step A1 of the authentication key agreement phase. The attacker repeatedly transmits a login request message $M1 = \{D_i, X\}$ to S_j. The authorized server S_j will perform Steps A2, A3, A4, and A6 of the authentication key agreement phase. S_j did not discover this illegal user (attacker) until Step A6.

   In the proposed scheme, the authorized server generates a random integer β and computes $Y = \beta \times P$ and $V1 = H(D_i||s_j||Y)$ after S_j receives $M1$. The authorized server S_j transfers $M2 = \{D_i, X, Y, V1\}$ to the register center RC. After RC receives $M2$, RC calculates $X' = x \times X$. Then, RC calculates $Dec_{H(X||X')}(D_i)$ and $H(ID_i||H(ID_i||x)||SID_j)$. RC compares it with D_i. If $H(ID_i||H(ID_i||x)||SID_j)$ does not exist $Dec_{H(X||X')}(D_i)$, the register center RC terminates the session.

   For the denial of service attack, the cost computation for identifying the DoS attack is shown in Table 1.

   In Zhena *et al.* scheme, the server must calculate one ECC and one Hash function in Step A2. In Step A4, one ECC and two Hash functions must be calculated. In Step A6, the Hash function needs to be calculated once. The server wasted 2 ECC calculations and 4 Hash function calculations. RC needs to calculate 1 ECC and 3 Hash functions in Step A5. These calculations will reduce the server's performance and cannot provide regular services.

   In the proposed scheme, the server must calculate one ECC and one Hash function in Step A2. In Step A3, one ECC, deciphering, and Hash function must be calculated.

6 Conclusion

In conclusion, we have shown the weakness of Zhang *et al.*'s robust authentication protocol. Their protocol is not immune to online password guessing and denial of service attacks using users' smart cards. We also propose an improved robust authentication protocol to defend against these vulnerabilities in the Zhang *et al.* protocol.

References

[1] S. Q. Cao, Q. Sun, L. L. Cao, "Security analysis and enhancements of a remote user authentication scheme, *International Journal of Network Security*, vol. 21, pp. 661–669, 2019.

[2] C. C. Chang, W. Y. Hsueh, T. F. Cheng, "An advanced anonymous and biometrics-based multi-server authentication scheme using smart cards," *International Journal of Network Security*, vol. 18, pp. 1010-1021, 2016.

[3] T. Y. Chen, C. C. Lee, M. S. Hwang, J. K. Jan, "Towards secure and efficient user authentication scheme

using smart card for multi-server environments," *The Journal of Supercomputing*, vol. 66, pp. 1008-1032, 2013.

[4] R. H. Dong, B. B. Ren, Q. Y. Zhang, H. Yuan, "A lightweight user authentication scheme based on fuzzy extraction technology for wireless sensor networks, *International Journal of Network Security*, vol. 23, pp. 157-171, 2021.

[5] T. H. Feng, C. H. Ling, M. S. Hwang, "Cryptanalysis of Tan's improvement on a password authentication scheme for multi-server environments, *International Journal of Network Security*, vol. 16, pp. 318–321, 2014.

[6] L. C. Huang, M. S. Hwang, "Two-party authenticated multiple-key agreement based on elliptic curve discrete logarithm problem, *International Journal of Smart Home*, vol. 7, pp. 9-18, 2013.

[7] M. S. Hwang, C. C. Lee, J. Z. Lee, C. C. Yang, "A secure protocol for bluetooth piconets using elliptic curve cryptography, *Telecommunication Systems*, vol. 29, pp. 165-180, 2005.

[8] M. S. Hwang, L. H. Li, "A new remote user authentication scheme using smart cards," *IEEE Transactions on Consumer Electronics*, vol. 46, pp. 28-30, 2000.

[9] M. S. Hwang, H. W. Li, C. Y. Yang, "An improved of enhancements of a user authentication scheme, *International Journal of Network Security*, vol. 25, pp. 508-514, 2023.

[10] M. S. Hwang, S. F. Tzeng, C. S. Tsai, "Generalization of proxy signature based on elliptic curves, *Computer Standards & Interfaces*, vol. 26, pp. 73-84, 2004.

[11] C. C. Lee, C. H. Liu, M. S. Hwang, "Guessing attacks on strong-password authentication protocol, *International Journal of Network Security*, vol. 15, pp. 64-67, 2013.

[12] L. H. Li, L. C. Lin, M. S. Hwang, "A remote password authentication scheme for multi-server architecture using neural networks, *IEEE Trans on Neural Network*, vol. 12, pp. 1498-1504, 2001.

[13] X. Li, J. Ma, W. D. Wang, Y. P. Xiong, J. Z. Zhang, "A novel smart card and dynamic ID based remote user authentication scheme for multi-server environments," *Mathematical and Computer Modelling*, vol. 58, pp. 85-95, 2013.

[14] X. Li, J. Niu, S. Kumari, J. Liao, W. Liang, "An enhancement of a smart card authentication scheme for multi-server architecture," *Wireless Personal Communications*, vol. 80, pp. 175-192, 2015.

[15] I. C. Lin, M. S. Hwang, L. H. Li, "A new remote user authentication scheme for multi-server architecture, *Future Generation Computer System*, vol. 19, pp. 13-22, 2003.

[16] T. W. Lin, C. L. Hsu, "Chaotic maps-based privacy-preserved three-factor authentication scheme for telemedicine systems, *International Journal of Network Security*, vol. 25, pp. 194-200, 2023.

[17] M. Mao, J. Yu, Y. Wang, "A more secure and revocable anonymous authentication scheme for IoT, *International Journal of Network Security*, vol. 25, pp. 595-602, 2023.

[18] R. S. Pippal, C. D. Jaidhar, S. Tapaswi, "Robust smart card authentication scheme for multi-server architecture," *Wireless Personal Communications*, vol. 72, pp. 729-745, 2013.

[19] J. Srinivas, S. Mukhopadhyay, D. Mishra, "A self-verifiable password based authentication scheme for multi-server architecture using smart card," *Wireless Personal Communications*, vol. 96, pp. 6273-6297, 2017.

[20] C. S. Tsai, C. C. Lee, M. S. Hwang, "Password authentication schemes: Current status and key issues," *International Journal of Network Security*, vol. 3, pp. 101-115, 2006.

[21] T. Wan, X. Liu, W. Liao, N. Jiang, "Cryptanalysis and improvement of a smart card based authentication scheme for multi-server architecture using ECC," *International Journal of Network Security*, vol. 21, pp. 993-1002, 2019.

[22] C. C. Yang, T. Y. Chang, M. S. Hwang, "A new anonymous conference key distribution system based on the elliptic curve discrete logarithm problem, *Computer Standards & Interfaces*, vol. 25, pp. 141-145, 2003.

[23] S. L. Yin, H. Li, S. Karim, Y. Sun, "ECID: Elliptic curve identity-based blind signature scheme, *International Journal of Network Security*, vol. 23, pp. 9-13, 2021.

[24] M. Zhang, B. Yang, H. Hou, M. Huang, Y. Zhou, "A practical and efficient two-part edwards curve digital signature for mobile networks, *International Journal of Network Security*, vol. 23, pp. 558-568, 2021.

[25] X. Zhang, B. Wang, W. Zhang, "A robust authentication protocol for multi-server architecture using elliptic curve cryptography," *International Journal of Network Security*, vol. 21, pp. 191-198, 2019.

Biography

Min-Shiang Hwang received M.S. in industrial engineering from National Tsing Hua University, Taiwan in 1988, and a Ph.D. degree in computer and information science from National Chiao Tung University, Taiwan, in 1995. He was a professor and Chairman of the Department of Management Information Systems, NCHU, during 2003-2009. He was also a visiting professor at the University of California (UC), Riverside, and UC. Davis (USA) during 2009-2010. He was a distinguished professor of the Department of Management Information Systems, NCHU, during 2007-2011. He obtained the 1997, 1998, 1999, 2000, and 2001 Excellent Research Award of National Science Council (Taiwan). Dr. Hwang was a dean of the College of Computer Science, Asia University (AU), Taichung, Taiwan. He is currently a chair professor with the Department of Computer

Science and Information Engineering, AU. His current research interests include cryptography, Steganography, and mobile computing. Dr. Hwang has published over 300+ articles on the above research fields in international journals.

Hou-Wen Li received a bachelor's degree in business administration from National Cheng Kung University in 1992 and a master's degree in law from Tunghai University in Taiwan in 2009.He worked as a teacher at the New Taipei Municipal New Taipei Industrial Vocational High School from 1992 to 2002, and at the National Taiwan University of Sport since 2002. He had worked on the Regulations Committee and is a professional investigator of gender equality for the Ministry of Education. He is currently pursuing his Ph.D. in the Artificial Intelligence Ph.D. program at Asia University. His research interests include natural language processing and the application of artificial intelligence in law, etc.

Cheng-Ying Yang (Member, IEEE) received the M.S. degree in Electronic Engineering from Monmouth University, New Jersey, in 1991, and Ph.D. degree from the University of Toledo, Ohio, in 1999. He is a member of the IEEE Satellite & Space Communication Society. Currently, he is a Professor with the Department of Computer Science, University of Taipei, Taiwan. His research interests include performance analysis of digital communication systems, signal processing, error control coding, Petri net applications and computer security.

Guide for Authors
International Journal of Network Security

IJNS will be committed to the timely publication of very high-quality, peer-reviewed, original papers that advance the state-of-the art and applications of network security. Topics will include, but not be limited to, the following: Biometric Security, Communications and Networks Security, Cryptography, Database Security, Electronic Commerce Security, Multimedia Security, System Security, etc.

1. Submission Procedure

Authors are strongly encouraged to submit their papers electronically by using online manuscript submission at http://ijns.jalaxy.com.tw/.

2. General

Articles must be written in good English. Submission of an article implies that the work described has not been published previously, that it is not under consideration for publication elsewhere. It will not be published elsewhere in the same form, in English or in any other language, without the written consent of the Publisher.

2.1 Length Limitation:

All papers should be concisely written and be no longer than 30 double-spaced pages (12-point font, approximately 26 lines/page) including figures.

2.2 Title page

The title page should contain the article title, author(s) names and affiliations, address, an abstract not exceeding 100 words, and a list of three to five keywords.

2.3 Corresponding author

Clearly indicate who is willing to handle correspondence at all stages of refereeing and publication. Ensure that telephone and fax numbers (with country and area code) are provided in addition to the e-mail address and the complete postal address.

2.4 References

References should be listed alphabetically, in the same way as follows:
For a paper in a journal: M. S. Hwang, C. C. Chang, and K. F. Hwang, ``An ElGamal-like cryptosystem for enciphering large messages," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 445--446, 2002.

For a book: Dorothy E. R. Denning, *Cryptography and Data Security*. Massachusetts: Addison-Wesley, 1982.

For a paper in a proceeding: M. S. Hwang, C. C. Lee, and Y. L. Tang, ``Two simple batch verifying multiple digital signatures," in *The Third International Conference on Information and Communication Security (ICICS2001)*, pp. 13--16, Xian, China, 2001.

In text, references should be indicated by [number].

Subscription Information

Individual subscriptions to IJNS are available at the annual rate of US$ 200.00 or NT 7,000 (Taiwan). The rate is US$1000.00 or NT 30,000 (Taiwan) for institutional subscriptions. Price includes surface postage, packing and handling charges worldwide. Please make your payment payable to "Jalaxy Technique Co., LTD." For detailed information, please refer to http://ijns.jalaxy.com.tw or Email to ijns.publishing@gmail.com.