IJNS
International Journal of Network Security

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

Vol. 23, No. 5 (September 2021)

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: 23, No: 5 (September 1, 2021)

International Journal of Network Security

Research on E-book Text Copyright Protection and Anti-tampering Technology

Yung-Chen Chou[1], Kurnia Anggriani[2,3], Nan-I Wu[4], and Min-Shiang Hwang[2,5]
(Corresponding author: Min-Shiang Hwang)

iSchool, Feng Chia University[1]
Department of Computer Science & Information Engineering, Asia University[2]
500, Lioufeng Rd., Wufeng, Taichung 41354, Taichung, Taiwan, ROC
Faculty of Engineering, University of Bengkulu, Indonesia[3]
Department of Digital Multimedia, Lee-Ming Institute of Technology[4]
No.2-2, Lijhuan Rd., Taishan Township, Taipei County 243, Taiwan, ROC
Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan[5]
Email: mshwang@asia.edu.tw
(Invited Paper; First Online Aug. 20, 2021)

Abstract

The emergence of e-books makes reading more colorful. Unlike paper books, e-books provide multimedia supplementary content such as hyperlinks, music, videos, and pictures. Readers can have a lot of fun through e-book reading. At present, the development and use of e-books have not yet become widespread. One of the most important factors is the challenge of maintaining the copyright of e-book content. How to ensure that the copyright of e-books is well controlled has become a significant issue. First of all, after understanding the presentation of e-book text, we will develop and use text encoding and HTML attribute settings to hide the digital watermark in the e-book text. Secondly, we also use different text encoding fonts to encode symbols and strengthen the watermark through various embedding methods; third, we use digital signature technology to generate verification codes for the text content of the e-book. And we use confidential information sign-in technology to sign the digital signature into the content of the e-book file to facilitate the verification of the integrity of the text content in the future. Since e-books use XML to integrate multimedia resources, we have developed validation information that effectively generates XML and multimedia resources. Furthermore, we develop effective embedding technology to embed the verification code into the XML of the e-book. In addition, we also use the embeddable Javascript function of EPub 3 to develop technologies to prevent tampering with e-book content.

Keywords: Anti-tampering Technology; Copyright Protection; Data Hiding; E-Book; XML

1 Introduction

Since the Internet has flourished, our daily lives have been closely integrated with information technology. Nowadays, daily life events such as social interaction, sports, shopping, etc., are closely related to information technology. Even reading and learning are closely related to information technology and mobile devices. E-books rich content and the advantages of integrating multimedia resources will make reading more fun and choices. However, there is still a lot of room for the research and development of e-book technology. The main factor is that the publisher wants to be billed once for each e-book. Therefore, users cannot freely use e-books on their mobile devices or computer devices after the authorization key is sold but can only install e-book devices. Another reason that hinders significant progress in the e-book market is the copyright issue of e-books [4, 6, 23, 31]. Because e-books are easier to copy and deliver than paper books. At present, there is no perfect mechanism or copyright protection method, which restricts the promotion of e-books. Therefore, ensuring the copyright protection of selling or renting e-book materials has become an issue that it must pay attention to in the digital age.

At present, there is no strong mainstream technology for the development of e-books. The current e-book format can be roughly divided into six categories: Amazon Kindle, Adobe PDF, Microsoft Reader, Mobipocket, IDEF EPUB, and Palm doc (see Figure 1). IDEF EPUB e-books are more popular, and readers can browse ePub e-books on different computer platforms or mobile devices. Although the display's appearance is not necessarily the same, the content can be used across platforms, and there is no need to make separate e-books for different platforms. Among the many e-book formats, the more conve-

nient format is to use HTML and XML formats. HTML is used to compile the content of e-books, and XML is used to mark multimedia resources used by e-books. In addition to the locking mechanism of the e-book software itself, the copyright protection of e-books is also an important issue in protecting e-book content.

To solve the above problems, we can extend steganography technology to the copyright protection problem. The main concept is to encrypt the watermark and then embed it in the cover media to form an e-book with a digital watermark. In this way, the publishers will promote e-book confidently because the e-book with a hidden watermark is almost the same as the original e-book. Therefore, users can maintain good reading quality when using e-books with digital watermarks. Furthermore, when the protected media has copyright doubts, publishers can confirm the copyright by capturing the digital watermark information in the media [1,2,9,13,34]. On the other hand, protecting the original copyright of digital content is also a significant research topic. If the digital content signature can be embedded in the media through the information hiding technology, the digital signature hidden in the media can be extracted and compared to distinguish whether the digital content has been tampered with.

The biggest difference between e-books and paper books is that they also contain multimedia materials in addition to ordinary text. For example, voice (Audio), video (Video), 3D models and images (Images), etc. These multimedia materials have been carefully designed and arranged as part of the e-book. Therefore, protecting these multimedia data, the content, and information in e-books is a significant issue. Under these problems, we will study the possibility of embedding watermark in the text content of the e-book, at the same time, protect the integrity of the text content of the e-book. In addition, you can also use XML digital signatures to achieve the integrity of e-book resources.

2 Related Works

Due to the sensitivity of human senses and the format of multimedia files, digital multimedia files (such as images, videos, and audio) are easier to achieve copyright protection. The main method is to modify the multimedia data or adjust the coefficient data in the multimedia file to hide the digital watermark information [3, 5, 14, 15, 32, 33, 35, 36]. The human eye is unlikely to perceive changes in the image, and it is easier to manipulate than audio and video. Therefore, images are most often used as a payload medium for transmitting confidential information [1, 2, 9, 13, 26, 30, 34, 38]. Many scholars have further developed various copyright protection technologies (such as digital watermarking technology) based on the characteristics of images. Although the watermark data is hidden in the multimedia file, it is difficult for users without professional training to determine whether the watermark information is hidden in the multimedia file. Furthermore, when digital images use invisible digital watermarking technology to hide watermark data, it is difficult for users to detect anomalies directly from the image itself. This achieves the purpose of copyright protection and preserves the visual quality of the image.

Due to the masking effect in the audio, the human ear cannot hear the sound under the masking sound. Many audio compression technologies use this feature to delete data to facilitate compression. This method provides great convenience for audio digital watermarking technology. As long as the watermark data is converted into the data form under the masked sound and added to the audio file, it can achieve the purpose of the audio watermark. Users of this method will not feel any difference when listening to the audio. On the other hand, data hiding technology using text as a loading medium is also booming. Data hiding technologies that secrets are hidden in the text as a loading medium can be roughly divided into Microsoft Word [7,8,18,24,25,27,29], Portable Document Format (PDF) [21,39], hypertext markup language (HTML) [11,12,20,28,37], email [19], and program source code files [22].

Various information technologies applications have given modern people a completely different lifestyle from the digital age. Social networking sites have changed the traditional way of communicating with people face to face. An email has changed the traditional way of mailing letters. The Internet has become a new place for individuals to express themselves and a new channel for corporate marketing or knowledge transfer. E-books are gradually changing the way modern people read and learn. After entering the Web 2.0 era, under the demand of cross-platform, Web technology is further applied to the production of e-books. In addition to ordinary text, CSS (Cascading Style Sheet) makes web pages' layout and color matching richer. XML (eXtensible Markup Language) makes it easier for people to read the document's content. More importantly, it is a language format and grammar that computer programs can easily recognize. But in any case, the most important thing about e-books is the presentation of text and multimedia.

Sui and Luo proposed a method to modify the capitalization of HTML tags to hide confidential information [28]. Since HTML is composed of many tags, these tags include <HTML>, </HTML>, <BODY>, </BODY>, and <OL>. The case of these tags does not affect the difference in the content of the webpage displayed by the browser. For example, there is no difference between <HTML> and <HTml> on the browser. Therefore, Sui and Luo et al. use this convenience to hide confidential information. For example, uppercase represents the confidential bit '1', the lower case represents the confidential bit '0', etc., so <HTml> represents the confidential information "1100". Another example is listed in Figure 2.

On the other hand, we can use different quotation marks in the HTML tag attribute settings. Yang and Yang proposed a different "quotation mark" method to

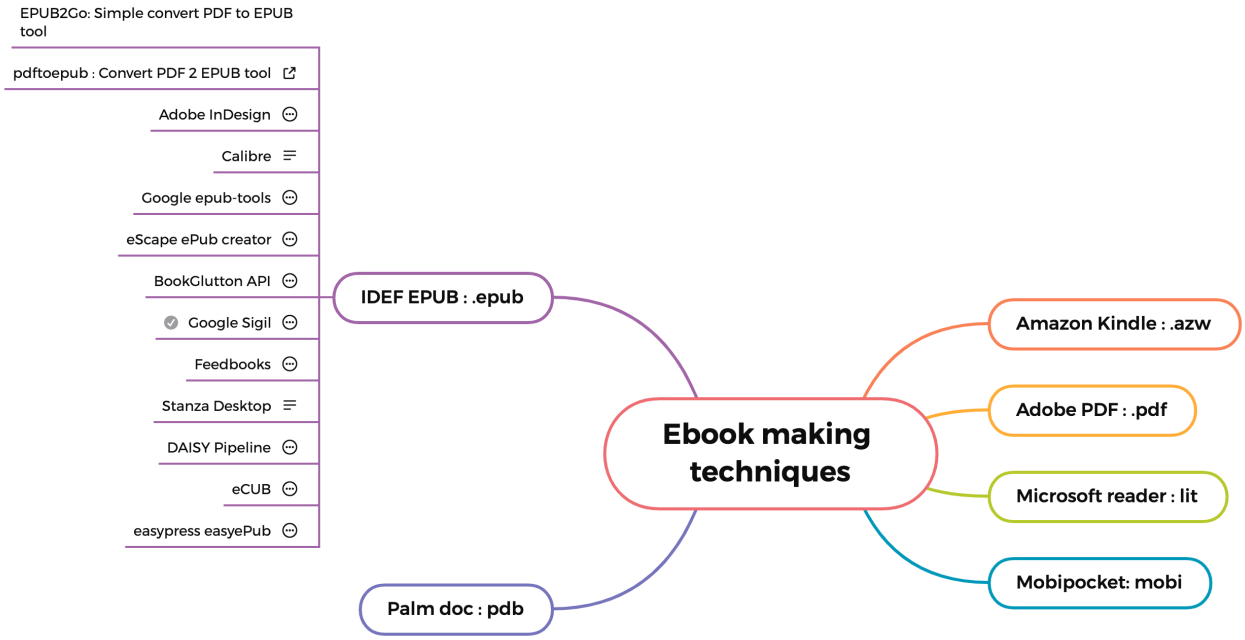
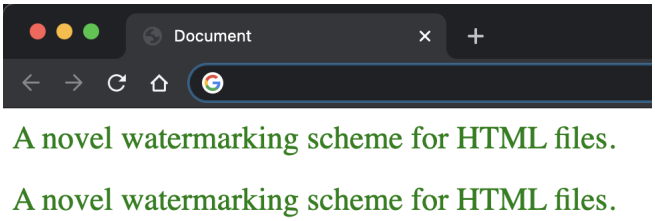


Figure 1: E-book Technologies



(a)

demo1.html > ...

```
1  <!DOCTYPE html>
2  <html lang="en">
3  <head>
4      <meta charset="UTF-8">
5      <meta http-equiv="X-UA-Compatible" content="IE=edge">
6      <meta name="viewport" content="width=device-width, initial-scale=1.0">
7      <title>Document</title>
8  </head>
9  <body>
10     <font size="5" color="green">A novel watermarking scheme for HTML files.</font>
11     <br/><br/>
12     <Font size="5" color="green">A novel watermarking scheme for HTML files.</Font>
13 </body>
14 </html>
```

(b)

Figure 2: An illustration of Sui and Luo method [28]. (a) The result after the second line of characters is hidden is "10001000". (b) The source code of the web page in Figure 2(a).

hide confidential information [17]. HTML tags are the most important element in the structure of the displayed results of a web page. Many of these tags are parameters that can be set for attributes to enhance the characteristics of the web page. For example, to set the font color displayed on a web page, you can use the <font color = blue> tag and its attribute settings to make the text on the web page appear blue. But in order to meet this requirement, there can also be wordings like <font color = "blue"> and <font color ='blue'>. Therefore, options are provided for users to imply their needs in different places to hide secret information. For example, the parameter with double quotation marks is set to confidential information '1'. Single quotation marks indicate confidential information '0'. Such as <font color="blue" size = '5'> hide confidential information "10". Another example is listed in Figure 3.

Huang *et al.* proposed a method to hide secret information using the attribute sequence in HTML tags [11, 12]. Because the same tag may contain multiple attribute settings, such as <HTML face="Times New Roman" color="blue" size="5">. These attributes can be zeroed in alphabetical order, and then the order of appearance can be adjusted to hide secret information. First, face, color, and size have a total of 6 sorting orders: <color, face, size>, <color, size, face>, <size, color, face>, <size, face, color>, <face, color, size>, and <face , size, color>. Therefore, it can be used to hide up to $\lfloor \log_2 6 \rfloor = 2$ bits of information. For example, <font size="5" face="Times New Roman" color="green"> represents secret information "11", because the order <color, face, size> means "00", <color, size, face> Represents "01", and so on. Another example is listed in Figure 4.

Katzenbeisser and Petitcolas uses invisible special characters as secret information and embeds it in the web page's source code [17]. Since invisible special characters will not be displayed on the webpage, no matter how many special characters are added, it will not affect the normal display of the webpage. Chen *et al.* proposed to modify the expression of sentences in web pages to hide secrets [8]. For example, use active or passive grammar or use synonyms to metaphor secret information. Due to the matching of relevant rules and secret information, only legitimate users know. Therefore, unauthorized users will not be able to know what the real secret information is.

Lee and Tsai proposed an effective information hiding technique that uses English sentences composed of many words and accompanied by many blank characters [20]. The main method is to replace the original "blank" with a variety of special codes used to represent "blank" to achieve the purpose of hiding confidential information. The human eye cannot distinguish these special blank character codes in the web page display because of these special blank character codes. Therefore, the content of the web page cannot be seen to be any strange.

Inoue *et al.* proposed a method to hide confidential information in XML [16]. Since XML is much stricter than HTML, a relatively flexible approach is better. Inoue *et al.* proposed five different information hiding techniques.

1) First, XML allows users to customize tags. Users of each type of label can also define their own attributes. Therefore, Inoue used a pair of labels to represent '0' and a single label to represent '1'. For example, <img></img> represents confidential information '0'. And <img/> is used to represent the confidential information '1'.

2) The second method is to use the ">" symbol in the XML tags to allow blanks to hide confidential information. For example, <tab>, </tab> or <tag/> are used to represent confidential information '0'. And <tab>, </tab> or <tag /> are used to represent confidential information '1'.

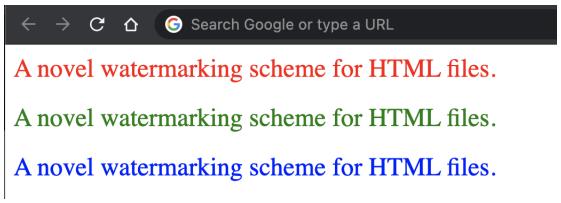
3) The third method is to use the different appearance order of the elements contained in each object in XML to represent confidential information. For example, <user><name></name><id></id></user> represents confidential information '0'. And <user><id></id><name></name></user> is used to represent confidential information '1'.

4) The fourth method is to represent confidential information in the order in which different attributes appear in the label. For example, <event month="OCT" date="24">EVENT</event> represents confidential information '0', and <event date="24" month="OCT">EVENT</event> represents confidential information '1'.

5) The fifth method is to use the differences contained inside and outside of different elements to hide confidential information. For example, <favorite> <fruit> Apple </fruit> </favorite> represents secrt information '0', and <fruit> <favorite> Apple </favorite> </fruit> represents confidential information '1'.

3 Research Issues

In this research, we used HTML and XML-based e-book content copyright protection technology. To be widely used on various mobile devices, e-books must be cross-platform. Therefore, HTML and XML technologies must be some of the most important technologies for producing e-books. On the other hand, considering the aesthetics and convenience of the e-book content, it is obviously necessary to hide the invisible watermark. Therefore, this research embeds the digital watermark of an e-book publisher into HTML. To realize the resilience of watermarking, we develop and hid watermarking within many times technology. In addition, to ensure that the e-book content has not been modified, we have also developed a verification code generation mechanism for the e-book text content and hide it in the HTML code of the e-book.



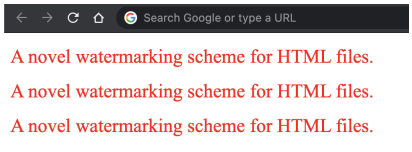
(a)

demo2.html > ...

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Document</title>
</head>
<body>
    <font size="5" color=red>A novel watermarking scheme for HTML files.</font>
    <br/><br/>
    <Font size="5" color='green'>A novel watermarking scheme for HTML files.</Font>
    <br/><br/>
    <Font size="5" color="blue">A novel watermarking scheme for HTML files.</Font>
</body>
</html>
```

(b)

Figure 3: An illustration of Yang and Yang method [17]. (a) The result of using different ”quotation marks” in the color attribute of the ¡font¿ tag. (b) Source code of Figure 3(a).



(a)

demo3.html > ...

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Document</title>
</head>
<body>
    <p><font face="Times New Roman" size="5" color="red">
        A novel watermarking scheme for HTML files.</font></p>
    <p><font color="red" size="5" face="Times New Roman">
        A novel watermarking scheme for HTML files.</font></p>
    <p><font color="red" face="Times New Roman" size="5">
        A novel watermarking scheme for HTML files.</font></p>
</body>
</html>
```

(b)

Figure 4: An illustration of Huang *et al.* data hiding method [11, 12]. (a) Results are presented in different order of attributes. (b) Source code of Figure 4(a).

In addition, the advantage of e-books over paper books is that e-book can add multimedia resources such as images, videos, and audio to e-book files simultaneously. As far as e-book files are concerned, various multimedia resources and data used can be controlled and organized through XML. To prevent unauthorized users from arbitrarily modifying or adjusting the e-book content structure or multimedia resources, this research proposes an integrity verification code for the final e-book content XML. Furthermore, it hides the verification code in the XML. In this way, it will further guarantee the integrity of e-book multimedia resources.

This article proposes the following two major research topics in e-book text copyright protection and anti-tampering technology: (1) Research on watermarking technology for e-book text content; (2) Research on e-book text content prevention technology.

3.1 Research on Watermarking Technology for E-book Text Content

The main focus of this research is to hide a digital watermark in the HTML of e-books. To achieve the purpose of robustness, we propose a variety of different hiding methods to hide the same watermark data multiple times. When there is a copyright dispute, the owner of the e-book can use a watermark extraction program to prove the copyright. The main idea is to use at least three watermark embedding techniques to hide the watermark in the e-book code. Thus, the watermark's e-book will have the same browsing quality as the original e-book content. When copyright disputes occur, the judge will extract the watermark data from the three embedding modes. The judge will use the majority decision-making mechanism to determine the final extracted watermark data.

HTML is a markup language that is currently a common practice for making cross-platform e-books. The users can read the completed e-book through a browser, *e.g.*, Firefox installs EPUBReader plug-in, Chrome installs the Reading extension, or e-book browsing software (*e.g.*, iPad, iPhone, and iOS devices can use iBook App, Android devices can use iBook App). Through the interpretation and display of the browser, e-books can present diversified and rich content. Provide users with the enjoyment of e-book content. Figure 5 is a simple e-book example. We can see from Figure 5(b) that the HTML source code of an e-book is composed of many tags. Figure 5(a) results from previewing the e-book using the free e-book making software Sigil.

When embedding the watermark, we use the CSS style setting modification method and the character encoding embedding method to hide the watermark in the HTML file and style setting of the e-book. In the watermark extraction stage, we extract the watermark through the reverse operation of the original watermark embedding technology. It is worth noting that although we have used three watermark embedding technologies, all three embedding technologies hide the same watermark. In this way, three watermarks will be obtained when the watermark is extracted, and finally, the final data is determined through a voting mechanism. Figures 6 and 7 are the watermarks embedding flow chart and watermark extraction flow chart conceived in this project.

Next, we will outline the concepts and practices of hiding various watermark data.

Embedding Method 1: Use CSS edge margin to set the embedding method

From the flexibility of CSS margin settings (See Table 1), it is found that the same display effect can be obtained by using different setting methods. For example, leave 1em of space around the block of the tag. There are four setting ways are , , , or . Because of the same effect, there are four setting methods, we can use it to represent two-digit watermark data. They are " 00", "01", "10", and "11". In this way, the watermark can be hidden in the code of the e-book.

Table 1: CSS margin setting

Margin attribute setting	Description
<span style= "margin: 1em;>	Top margin is 1em Right margin is 1em Bottom margin is 1em Left margin is 1em
<span style= "margin: 1em 1em;>	
<span style= "margin: 1em 1em 1em 1em;>	
<span style= "margin-top: 1em; margin-bottom: 1em; margin-left: 1em; margin-right: 1em;>	
<span style= "margin: 1em 1.5em;>	Top margin is 1em Right margin is 1.5em Bottom margin is 1em Left margins is 1.5em
<span style= "margin: 1em 1.5em 1em 1.5em;>	
<span style= "margin-top: 1em; margin-bottom: 1em; margin-right: 1.5em; margin-left: 1.5em;>	
<span style= "margin: 1em 1.5em 2em;>	Top margin is 1em Right margin is 1.5em Bottom margin is 2em Left margin is 1.5em
<span style= "margin: 1em 1.5em 2em 1.5em;>	
<span style= "margin-top: 1em; margin-bottom: 2em; margin-right: 1.5em; margin-left: 1.5em;>	

The following is our hidden watermark method as follows:

Step 1: Disrupt and mix the watermark data;

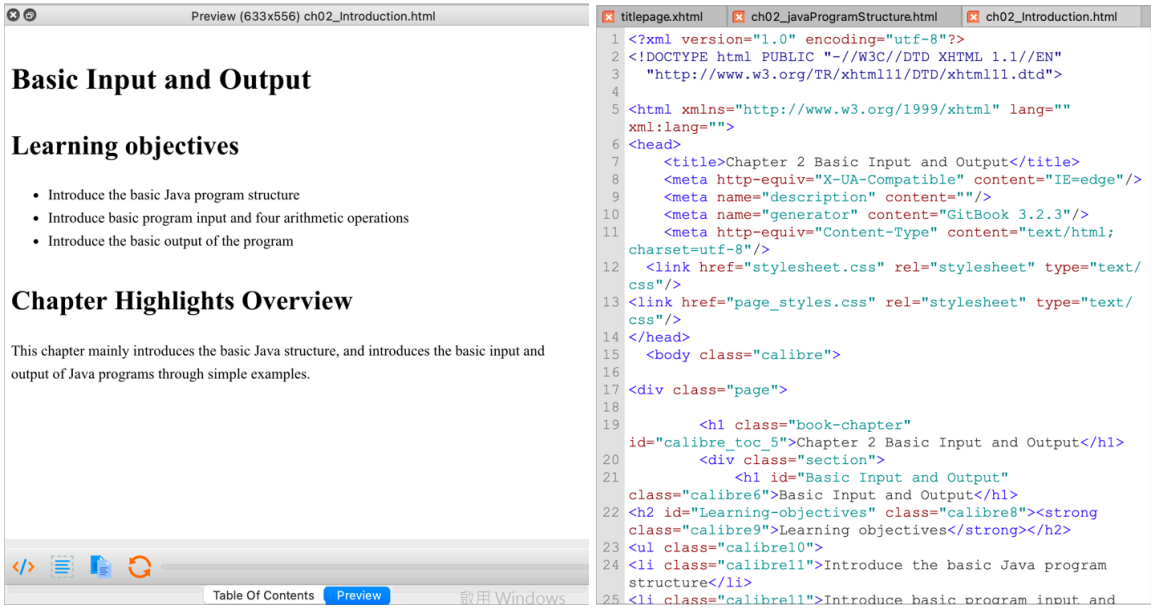
Step 2: Analyze CSS style settings;

Step 3: Modify the CSS settings according to the secret information hiding rules;

Step 4: Repeat Steps 2 to 4 until all watermarks are hidden.

Embedding Method 2: Use CSS to set the font size and embedding method

E-book writing in EPub format can be set through CSS, so that the content of the e-book can be more diversified. In the CSS style settings, you



(a) (b)

Figure 5: An E-book example. (a) Examples of e-books. (b) E-book HTML source code.

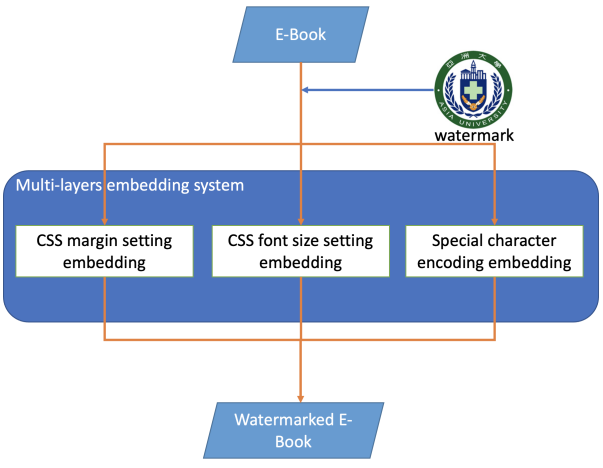


Figure 6: E-book watermark embedding flowchart

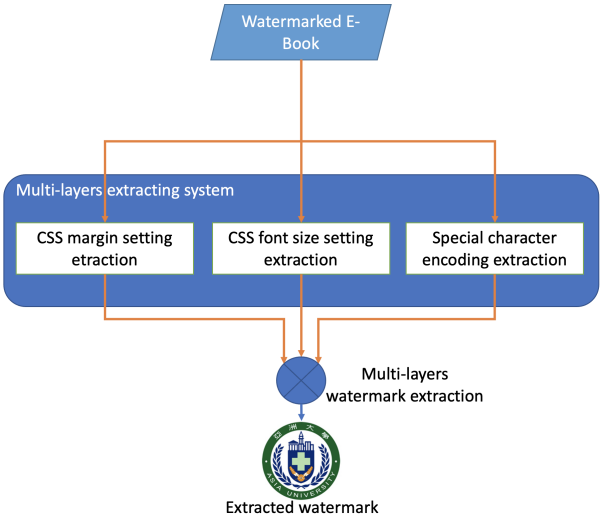


Figure 7: E-book watermark extraction flowchart

can use different units such as pt, px, %, em, *etc.* to adjust. For example, to adjust the text font size to 10 (pt), there are four setting ways: , </ span>, , or . Therefore, we can use this flexible setting method to hide the watermark and text integrity verification code. In addition, we found that with different settings, the font size displayed on the browser looks the same. Table 2 summarizes the correspondence between the font-size attribute value in the tag and other symbols. Therefore, this feature provides us with a good opportunity to hide the watermark data. In other words, assuming that the watermark data we want to embed is "10", then we can use to represent it.

Table 2: Correspondence table of different CSS font units [10]

em	%	px	Pt
0.63	63%	10.08	8
0.82	82%	13.12	10
1	100%	16	12
1.13	113%	18.08	13.5
1.5	150%	24	18
2	200%	32	24
3	300%	48	36

The following is our hidden watermark method as follows:

Step 1: Take out a paragraph of text from the e-book and set the text size;

Step 2: Take 2 watermark bits to form ws;

Step 3: Embed watermark:

- If $ws ==$ "00", use the size attribute setting of the <font> tag;
- If $ws ==$ "01", use css to set the font size;
- If $ws ==$ "10", use css to set the font size;
- If $ws ==$ "11", use css to set the font size.

Step 4: Repeat Steps 2 to 3 until all watermarks are hidden.

Embedding Method 3: Use characters

In producing e-books and writing the content directly in words, we can also use different code words instead. For example, '<' and '>' can be displayed together with "<" and ">" can be coded separately. We can also use the "‹" and "&#8250" codes to display. Therefore, we only need to match the key among multiple characters, randomly select some characters and their corresponding codes to embed the watermark and verification code. For example, we can use '.' to indicate that the watermark bit is '0', and '․' to indicate that the watermark bit is '1'. In this way, the user will see the character '.' when reading the e-book. From the e-book, it is no different from the human eye. Table 3 is an example of characters and their corresponding codes.

Table 3: Correspondence table of character encoding

Characters	Codes	Characters	Codes
<	<	'	’
>	>	,	‚
'	‘	'	‛
"	“	<	‹
"	”	>	›
.	․	...	…

The following is our hidden watermark method as follows:

Step 1: Use the key to randomly select the characters that the e-book wants to embed the watermark information;

Step 2: Get the digital watermark;

Step 3: Embed watermark:

- If the watermark bit is '0', the original characters are used for e-book writing;
- If the watermark bit is '1', use the code corresponding to the character to compile the e-book.

Step 4: Repeat Steps 1 to 3 until all watermarks are hidden.

3.2 Research on E-book Text Content Prevention Technology

Compared with HTML, XML has stricter writing format requirements. XML is used to represent the content of the data, while HTML focuses on how to display the data, making the presentation of the data easy to browse. In the XML specification, there are many elements under the root tag. And these elements form a tree structure. At the same time, the tags used for each element must appear in pairs, and both are indispensable. For example, <data></data>, <product></product>, <pname></pname>, or <price></price>, etc. But for some special cases, XML tags can allow a single tag. But the terminator of this tag is not ">", but "/>". For example, <student id="20140234" name="jack" />. If the element contains other elements, these tags must be arranged in a nested form and cannot overlap or

cross each other. For example, <A><B></A></B>, such label arrangement is illegal. Must be arranged in <A><B></B></A> to be legal. In addition, the naming of tags is different. For example, the <A> tag and the <a> tag are different tags. The attribute value in the tag in HTML can be without quotation marks, but it is illegal not to include quotation marks in XML. Therefore, the tag attribute value in XML must use single quotation marks or double quotation marks. It is also important that special characters in XML (for example: <, >, &, etc.) must use entity reference, that is, we must use <, >, and & to represent symbols such as <, >, and &, respectively. On the other hand, in standards after EPub 3.0, Javascript script commands are embedded in e-books. In this way, the e-book is rich in content and has the function of interacting with users. We use embedded Javascript to verify the integrity of the e-book content to prevent unauthorized users from modifying the e-book content without authorization. Next, we will outline the concepts and practices of hiding various watermark.

Embedding Method 1: Element label letter difference embedding method

Modify <student></student> to capitalize the letters in the label to hide the watermark. For example, <sTudEnt></sTudEnt> is used to indicate that the watermark information is "0100100". However, the names of the front and rear labels must be consistent. In addition, the publisher can embed the watermark and integrity verification code through the capitalization of the attribute name. For example, <student sTudID='29099234'></student>. Among them, "sTudID" is used to represent the watermark data "010011". In addition, we can also hide the watermark by adding other connection symbols such as "_". For example, <student st_u_d_id='29099234'></student> is used to represent the watermark data "01110".

Embedding Method 2: Tag attribute quotation mark difference embedding method

As XML stipulates that tag attribute values must be framed with "single quotation marks" or "double quotation marks". Therefore, we will use "'single quotation marks" to represent the watermark bit '0' and "double quotation marks" to represent the watermark bit '1'. For example, <student id='29099234' sex="M"></student> is used to represent the watermark data "01".

Embedding Method 3: Inserting the attribute value space into the embedding method

According to XML regulations, the setting of tag attribute values is given by "=". However, there are no restrictions on the blank spaces on both sides of the "=". In addition, the terminator (for example, ">", "/>") in each tag is also allowed to be blank. Therefore, we use these blank arrangements to represent different watermark information. For example, we define 0 means there is no blank, and 1 means blank. Then use <student sid='29099234' sex="M" ></student><product pID="0001' price="293" /> to represent the watermark "0001101101".

Javascript's integrity protection mechanism for e-book content

Javascript is a commonly used front-end interactive language today. Its advantage is that it can provide an interactive interface for users. For example, through Javascript, an e-book can have an instant test function. After the user completes the test, the e-book can also immediately check the correctness of the answer. If we encounter the wrong problem, we can also provide users with reference solutions. Through this function, we propose a mechanism to embed the integrity verification code of the e-book content into the e-book. At the same time, the function of locking editing content has been added. If we want to edit the content, we must have an authorization code for the editing function and verify it through the Javascript script in the e-book. The user can edit only when the verification is passed as legal editing. In this way, we can realize the integrity protection of the e-book and provide the flexibility of legally authorized editing.

4 Conclusion

Using electronic devices with rich content and simple operation to read or learn knowledge has become an indispensable and important way of reading in the information age. Therefore, the development of e-book technology with good copyright protection and the realization of copyright protection will help enterprises be more confident to develop more distinctive and innovative e-books, provide novel interactive functions, and achieve the purpose of teaching or teaching-learning and knowledge transfer.

At present, many digital watermarking and confidential information hiding technologies have been proposed. However, there are many practical technical challenges in applying these technologies to the copyright protection of e-books. Therefore, this article provides a different idea. In addition to embedding the digital watermark, we can also combine it with other effective operation strategies to enhance the strength of the digital watermark.

Acknowledgments

The Ministry of Science and Technology partially supported this research, Taiwan (ROC), under contract no.: MOST 109-2221-E-468-011-MY3, MOST 108-2410-H-468-023, and MOST 108-2622-8-468-001-TM1.

References

[1] A. M. Alattar, "Reversible watermark using the difference expansion of a generalized integer transform," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1147-1156, 2004.

[2] M. U. Celik, G. Sharma, A. M. Tekalp, E. Saber, "Lossless generalized-LSB data embedding," *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 253-266, 2005.

[3] C. C. Chang, K. F. Hwang, M. S. Hwang, "A digital watermarking scheme using human visual effects", *Informatics*, vol. 24, no. 4, pp. 505–511, Dec. 2000.

[4] C. C. Chang, K. F. Hwang, M. S. Hwang, "A block based digital watermarks for copy protection of images," in *Fifth Asia-Pacific Conference on ... and Fourth Optoelectronics and Communications Conference on Communications*, vol. 2, pp. 977-980, 1999.

[5] C. C. Chang, K. F. Hwang, M. S. Hwang, "A feature-oriented copyright owner proving technique for still images," *International Journal of Software Engineering and Knowledge Engineering*, vol. 12, no. 3, pp. 317-330, 2002.

[6] C. C. Chang, K. F. Hwang, M. S. Hwang, "Robust authentication scheme for protecting copyrights of images and graphics", *IEE Proceedings-Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 43-50, 2002.

[7] C. C. Chang, C. C. Wu, I. C. Lin, "A data hiding method for text documents using multiple-base encoding," *Communications in Computer and Information Science*, vol. 66, Springer, pp. 101-109, 2010.

[8] C. Chen, S. Z. Wnag, X. P. Zhang, "Information hiding in text using typesetting tools with stego-encoding," in *Proceedings of the First International Conference on Innovative Computing, Information and Control*, Beijing, China, vol. 1, pp. 459-462, 2006.

[9] S. F. Chiou, I-En Liao, and M. S. Hwang, "A capacity-enhanced reversible data hiding scheme based on SMVQ", *Imaging Science Journal*, vol. 59, no. 1, pp. 17–24, 2011.

[10] Y. C. Chou, H. C. Liao, "A webpage data hiding method by using tag and CSS attribute setting," in *Proceedings of the Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP'14)*, Kitakyushu, Japan, pp. 122-125, 2014.

[11] H. J. Huang, X. M. Sun, Z. S. Li, G. Sun, "Detection of hidden information in webpage," in *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, Haikou, China, August, vol. 4, pp. 317-321, 2007.

[12] H. J. Huang, S. H. Zhong, X. M. Sun, "An algorithm of webpage information hiding based on attributes permutation," in *Proceedings of the Fourth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Harbin, China, pp. 257-260, 2008.

[13] L. C. Huang, L. Y. Tseng, M. S. Hwang, "The study on data hiding in medical images", *International Journal of Network Security*, vol. 14, no. 6, pp. 301–309, 2012.

[14] M. S. Hwang, K. F. Hwang, C. C. Chang, "A time-stamping protocol for digital watermarking", *Applied Mathematics and Computation*, vol. 169, pp. 1276–1284, 2005.

[15] M. S. Hwang, C. C. Chang, K. F. Hwang, "Digital watermarking of images using neural networks", *Journal of Electronic Imaging*, vol. 9, no. 4, pp. 548–555, 2000.

[16] S. Inoue, K. Makino,I. Murase, O. Takizawa, T. Matsumoto, H. Nakagawa, *A Proposal on Information Hiding Methods using XML*, Aug. 13, 2021. (`https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.2099&rep=rep1&type=pdf`)

[17] S. Katzenbeisser, F. A. P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*, Boston, MA: Artech House, 2000.

[18] Y. W. Kim, K. A. Moon, I. S. Oh, "A text watermarking algorithm based on word classification and inter-word space statistics," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, pp. 775-779, 2003.

[19] I. S. Lee, W. H. Tsai, "Data hiding in emails and applications using unused ASCII control codes," *Journal of Information Technology and Applications*, vol. 3, no. 1, pp. 13-24, 2008.

[20] I. S. Lee, W. H. Tsai, "Secret communication through web pages using special space codes in HTML files," *International Journal of Applied Science and Engineering*, vol. 6, no. 2, pp. 141-149, 2008.

[21] I. S. Lee, W. H. Tsai, "A new approach to covert communication via pdf files," *Signal Processing*, vol. 90, no. 2, pp. 557-565, 2010.

[22] I. S. Lee, W. H. Tsai, "Security protection of software programs by information sharing and authentication techniques using invisible ASCII control codes," *International Journal of Network Security*, vol. 10, no. 1, pp. 1-10, 2010.

[23] C. Y. Lin, C. C. Wu, M. S. Hwang, "Research on e-book security tracking schemes," *International Journal of Network Security*, vol. 23, no. 4, pp. 549-557, 2021.

[24] I. C. Lin, P. K. Hsu, "A data hiding scheme on word documents using multiple-base notation system," in *Proceedings of the Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Darmstadt, Germany, pp. 31-33, 2010.

[25] T. Y. Liu, W. H. Tsai, "A new steganographic method for data hiding in microsoft word documents by a change tracking technique," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 1, pp. 24-30, 2007.

[26] Z. Ni, Y. Q. Shi, N. Ansari, W. Su, "Reversible data hiding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 354-362, 2006.

[27] M. A. Qadir, I. Ahmad, "Digital text watermarking: secure content delivery and data hiding in digital documents," *IEEE Aerospace and Electronic Systems Magazine*, vol. 21, no. 11, pp. 18-21, 2006.

[28] X. G. Sui, H. Luo, "A new steganography method based on hypertext," in *Proceedings of Asia-Pacific Radio Science Conference*, pp. 181-184, 2004.

[29] X. M. Sun, G. Lou, H. J. Huang, "Component-based digital watermarking of chinese texts," in *Proceedings of the Third International Conference on Information Security*, Shanghai, China, November, vol. 85, pp. 76–81, 2004.

[30] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 890-896, 2003.

[31] C. Y. Tsai, C. Y. Yang, I. C. Lin, M. S. Hwang, "A survey of e-book digital right management," *International Journal of Network Security*, vol. 20, no. 5, pp. 998-1004, 2018.

[32] C. C. Wu, S. J. Kao, W. C. Kuo, M. S. Hwang, "A robust-fragile watermarking scheme for image authentication," in *3rd International Conference on Innovative Computing Information and Control*, pp. 176, 2008.

[33] C. C. Wu, S. J. Kao, W. C. Kuo, M. S. Hwang, "A digital watermarking scheme using human visual effects," *Informatics*, vol. 24, no. 4, 2000.

[34] N. I. Wu, M. S. Hwang, "A novel LSB data hiding scheme with the lowest distortion", *The Imaging Science Journal*, vol. 65, no. 6, pp. 371–378, 2017.

[35] M. R. Xie, C. C. Wu, J. J. Shen, M. S. Hwang, "A survey of data distortion watermarking relational databases", *International Journal of Network Security*, vol. 18, no. 6, pp. 1022-1033, 2016.

[36] M. R. Xie, C. C. Wu, J. J. Shen, M. S. Hwang, "A survey of data distortion watermarking relational databases", *International Journal of Network Security*, vol. 18, no. 6, pp. 1022-1033, 2016.

[37] X. P. Zhang, S. Z. Wang, "Steganography using multiple-base notational system and human vision sensitivity," *IEEE Signal Processing Letters*, vol. 12, no. 1, pp. 67-70, 2005.

[38] X. P. Zhang, S. Z. Wang, "Efficient steganographic embedding by exploiting modification direction," *IEEE Communications Letters*, vol. 10, no. 11, pp. 1-3, 2006.

[39] S. P. Zhong, X. Q. Cheng, T. R. Chen, "Data hiding in a kind of pdf texts for secret communication," *International Journal of Network Security*, vol. 4, no. 1, pp. 17-26, 2007.

Biography

Yung-Chen Chou received the BS degree in Management Information Systems from National Pingtung University of Science & Technology, Pingtung, Taiwan, Republic of China, in 1998, and the MS degree in Information Management from Chaoyang University of Technology, Taichung, Taiwan, in 2002. He received Ph.D. degree in Computer Science and Information Engineering in 2008 from the National Chung Cheng University, Chiayi, Taiwan. From February 2009 to July 2021, he was an Associate Professor of Asia University, Taichung, Taiwan. Since August 2021 he has been an Associate Professor of iSchool, Feng Chia University, Taiwan. His current research interests include steganography, watermarking, and image processing.

Kurnia Anggriani received BS degree in Informatics from University of Bengkulu, Indonesia in 2011, and the MS degree in Informatics from Bandung Institute of Technology, Indonesia in 2014. Currently she is taking Ph.D degree in Asia University, Taiwan. Her current research interests include steganography and image processing.

Nan-I Wu received a Ph.D. degree in the Institute of Computer Science and Engineering from Nation Chung Hsing University (NCHU), Taichung, Taiwan, in 2009. From 2010 to 2011, he was a post-doctoral research fellow at the Academia Sinica Institute of information science. He was an assistant professor at the Department of Animation and Game Design, TOKO University (Taiwan), during 2011-2018 and an associate professor during 2018-2019. Now he is an associate professor at the Department of Digital Multimedia, Lee-Ming Institute of Technology (Taiwan) since 2019 and also the Director of the eSports Training Centre since 2020. His current research interests include game design, eSports training/magagement, multimedia processing, multimedia security, data hiding, and privacy-preserving. He published more than 10 international journal papers (SCI) and conference papers.

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 distinguished professor and Chairman of the Department of Management Information Systems, NCHU, during 2003-2011. He obtained 1997, 1998, 1999, 2000, and 2001 Excellent Research Awards from the 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 information security, electronic commerce, database, data security, cryptography, image compression, and mobile computing. Dr. Hwang has published over 300+ articles on the above research fields in international journals.

Efficient and Secure Elliptic Curve Scalar Multiplication Based on Quadruple-and-Add

Shuang-Gen Liu, Si-Jia An, and Yi-Wei Du
(Corresponding author: Shuang-Gen Liu)

School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
Email: liusgxupt@163.com

(Received Jan. 12, 2020; Revised and Accepted Aug. 6, 2020; First Online Aug. 14, 2021)

Abstract

Elliptic curve cryptosystem is one of the main directions of public-key cryptography. Because the short key and efficient arithmetic has attracted increasing attention, particularly in resource-limited hardware environments such as smart cards and phone cards. Scalar multiplication is the most core operation in the elliptic curve cryptosystems, and its operating speed affects the efficiency of the entire cryptosystem. Previous studies have researched how to improve the efficiency of scalar multiplication. In this paper, we propose a new efficient and secure elliptic curve scalar multiplication algorithm. Based on the generalized Fibonacci sequence, a new addition chain is proposed. The new algorithm iterates the "$4P+Q$" operation every time and has the powerful ability to resist SPA (Simple Power Attack) naturally. Compared with the existing chains, the new addition chain proposed in this study has a shorter chain length, and combine with the new affine coordinates, can further improve the efficiency. The experimental results indicate that the new algorithm is 27.9% faster than Fibonacci-and-add and 13.1% faster than GRAC (Golden Ratio Addition Chain).

Keywords: Addition Chain; Elliptic Curve; Fibonacci Sequence; Scalar Multiplication; Simple Power Analysis

1 Introduction

The Elliptic Curve Cryptogram (ECC) is a public key cryptosystem based on elliptic curve mathematics. Firstly, it was introduced independently by Neal Koblitz [16] and Victor Miller [20] in around 1985. Compared to the RSA public key cryptosystem, the Elliptic curve cryptosystem can provide shorter keys at the same security level [22]. For example, elliptic curve with the key of 160 bits is competitive with RSA with the key of 1024 bits. Shorter keys mean smaller power consumption, computational costs and storage space. Scalar multiplication kP is the most time-consuming operation in ECC, where P is a point on the elliptic curve and k is an arbitrary integer [1], which plays the role of a secret key.

In the optimization of scalar multiplication algorithm, it is usually carried out from two aspects: on the one hand, it improves the effective expression of integer k, shortens the length of the chain, and reduces the upper level of computation; on the other one hand, to optimize the underlying domain operations and reduce the underlying computing times.

Side Channel Attacks (SCA) [17] was first proposed by Kocher in 1996. It is based on an attack idea that obtains the key by analyzing the time or the energy consumption in the encryption process. It can basically be divided into two categories: Simple Power Analysis (SPA) [30, 34] and Differential Power Analysis (DPA) [5, 31]. The SPA is a method that analyzes the leaked information in one operation of the algorithm, and analyzes the key by analyzing the different energy required for the execution of the point doubling and point addition in the algorithm. The DPA is a method that performs the algorithm several times, collects information, and analyzes the correlation between key bits, ciphertext, and power consumption to obtain the key. In contrast, the DPA is more advanced, but the SPA is more convenient to be implemented.

There are two main measures which are usually adopted to resist SPA attacks. One class of methods are to use the indistinguishable point addition or point doubling algorithm in scalar multiplication algorithm to smooth the energy curve. For example, Gold Ratio Addition Chain (GRAC) [9] and Montgomery Ladder Algorithm [23]. The other type of methods are to normalize the scalar multiplication algorithm. No matter what information is processed, the information detected by the instrument is regular and consistent. Two typical approaches are Montgomery method [13] and DoubleandAdd method [29].

In recent years, many studies have been focused on multiple points. Literature [15] proposed the idea of the point halfing operation, which greatly improved the computational efficiency of scalar multiplication on elliptic curves. In the study [3], on the basis of the half point operation and the multi-base representation, combined with the Extended Double Base Number System (Extended

DBNS) [4], a scalar multiplication extension algorithm based on the half point and the multi-base representation was proposed to improve the efficiency of scalar multiplication. The work in [2] implements the skill of transforming inversion into multiplication operation, and can strengthen the operations at the bottom of the elliptic curve such as $2P+Q$, $3P$, $3P+Q$, $4P$, $4P+Q$, *etc.*

In this paper, on the basis of the generalized the Fibonacci sequence, we propose a new scalar multiplication addition chain, which has shorter chain length and good universality, At the same time, it can improve efficiency of the previous ones and have security naturally. The rest of this paper is organized as follows. In Section 2, we give a brief overview on elliptic curve cryptography, with some classic definitions on addition chains and Fibonacci sequence. In Section 3, we introduce the new addition chain and the explicit scalar multiplication algorithm based on the new addition chain. In Section 4, we discuss and compare our results with the previous ones in the literatures. Finally, in Section 5, we draw some concluding remarks.

2 Background

In this section, we give a brief overview on elliptic curve cryptography, stating some classic definitions on addition chains and Fibonacci sequence.

2.1 Elliptic Curve Cryptography

We start with a practical definition of the concept of an elliptic curve and review the formula for point quadrupling (QPL) and combined quadruple-and-add (QA) in even characteristic. More details could be cited from [12, 18, 21, 24, 28].

Definition 1. *An elliptic curve E over a finite field K is defined by an equation:*

$$E: y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_5 \tag{1}$$

where a_1, a_2, a_3, a_4, $a_5 \in K$, and $\Delta \neq 0$, while Δ is the discriminant of E.

In practice, the Weierstrass Equation (1) can be greatly simplified by applying admissible changes of variables. If the characteristic of K is not equal to 2 and 3, then (1) can be rewritten as:

$$E: y^2 = x^3 + ax + b \tag{2}$$

where $a, b \in K$, and $\Delta = 4a^3 + 27b^2 \neq 0$.

When the characteristic of K is equal to 2, we use the non-supersingular form of an elliptic curve, given for $a \neq 0$ by

$$E: y^2 + xy = x^3 + ax^2 + b \tag{3}$$

where $a, b \in K$, and $\Delta = b \neq 0$.

The set $E(K)$ of rational points on an elliptic curve E defined over a finite field K is an abelian group.

In the study [4], it remarks that the trick used in [6] by Eisentrager et al., which consisted in evaluating only the x-coordinate of $2P$ when computing $2P \pm Q$, can also be applied to speed-up the quadrupling (QPL) primitive. Indeed, given $P = (x_1, y_1)$, where $P \neq -P$, we have $2P = (x_3, y_3)$,

$$\begin{cases} x_3 = \lambda_1^2 + \lambda_1 + a, \\ y_3 = \lambda_1(x_1 + x_3) + x_3 + y_1. \end{cases} \tag{4}$$

where $\lambda_1 = x_1 + \dfrac{y_1}{x_1}$.

And $4P = 2(2P) = (x_4, y_4)$,

$$\begin{cases} x_4 = \lambda_2^2 + \lambda_2 + a, \\ y_4 = \lambda_2(x_1 + x_4) + x_4 + y_1 \end{cases} \tag{5}$$

where $\lambda_2 = x_3 + \dfrac{y_3}{x_3}$.

It can be observed that the computation of y_3 can be avoided by evaluating λ_2 as $\lambda_2 = \dfrac{x_1^2}{x_3} + \lambda_1 + x_3 + 1$.

As a result, computing $4P$ over binary fields requires $2I + 3S + 3M$. Compared to two consecutive doublings, it saves one field multiplication at the extra cost of one field squaring. It should be noted that they are working in characteristic two and thus squarings are free (normal basis) or of negligible cost (linear operation in binary fields).

For the QA operation, in the paper [4], evaluates $4P \pm Q$, as $2(2P) \pm Q$ using one doubling (DBL) and one double-and-add (DA), resulting in $3I + 3S + 5M$. This is always better than applying the previous trick one more time by computing $(((P+Q)+P)+P)+P)$ in $4I + 4S + 5M$; or evaluating $3P + (P+Q)$ which requires $4I + 4S + 6M$.

In the work [2], Ciet et al. have improved the algorithm proposed by Guajardo and Paar [10] for the computation of $4P$, and their new method requires $1I+5S+8M$. Based on their costs, QA is best evaluated as $(4P) \pm Q$ using one quadrupling (QPL) followed by one addition (ADD) in $2I + 6S + 10M$. In Table 1 below, it summarizes the costs of $4P$ and $4P \pm Q$ in the study [4], where it uses the break-even points as $I/M = 5$.

Table 1: Costs comparisons for QPL and QA in even characteristic using affine coordinate

$4P$	$2I + 3S + 3M$
$4P \pm Q$	$3I + 3S + 5M$

2.2 Review on Addition Chains and Fibonacci Sequence

Here, we briefly state some classic definitions used in the study of the Fibonacci sequence and addition chains. More details could be cited from [8, 32].

Definition 2. *An addition chain computing an integer k is given by two sequences $v = (v_0, v_1, ..., v_l)$ and $w = (w_1, w_2, ..., w_l)$ such that $v_0 = 1$, $v = k$, $v_i = v_r + v_s$, for all $1 \le i \le l$ with respect to $w_i = (r, s)$ and $0 \le r$, $s \le i - 1$. The length of the addition chain is l.*

Definition 3. *An Euclidean addition chain (EAC) is an addition chain which satisfies $v_1 = 1$, $v_2 = 2$, $v_3 = v_2 + v_1$ and for all $3 \le i \le l-1$, if $v_i = v_{i-1} + v_j$ for some $j \le i-1$, then $v_{i+1} = v_i + v_{i-1}$ or $v_{i+1} = v_i + v_j$.*

Definition 4. *The Fibonacci sequence is defined as $F_n = F_{n-1} + F_{n-2}$ for $n \ge 2$ where $F_0 = 0$ and $F_1 = 1$.*

According to the researches in recent years, we have found that the Fibonacci sequence has many properties. According to the reference [8], the generalized Fibonacci series is defined as follows:

$$\begin{cases} f_1 = 0, f_2 = 1, \\ f_{n+1} = X f_n + Y f_{n-1}, (n \ge 2). \end{cases} \tag{6}$$

and the general formula of the Fibonacci sequence is defined as follows:

$$f_n = \frac{\alpha^n - \beta^n}{\alpha - \beta} \tag{7}$$

where $\alpha, \beta = (X \pm \sqrt{X^2 + 4Y})/2$.

The recursive formula of the Fibonacci sequence is $F_n = F_{n-1} + F_{n-2} (n \ge 3)$, it can be deformated as $F_n - F_{n-1} - F_{n-2} = 0$, it corresponds to the unary quadratic equation $x^2 - x - 1 = 0$, which is known as the characteristic square of the Fibonacci sequence, the root of the equation, is called the characteristic root.

Definition 5. *The recursive formula of the new sequence is $F_n = 4F_{n-1} + F_{n-2} (n \ge 3)$, it can be deformated as $F_n - 4F_{n-1} - F_{n-2} = 0$, it corresponds to the unary quadratic equation $x^2 - 4x - 1 = 0$, which is known as the characteristic square of the new sequence, the root of the equation, is called the characteristic root.*

As we have seen that the famous formula for the partition ratio of precious metals is $1 : (n + \sqrt{n^2 + 4})/2$. If $n = 1$, it is $1 : (1 + \sqrt{5})/2 \approx 0.618$, it is known as golden ratio. If $n = 2$, it is $1 : (1 + \sqrt{2}) \approx 0.414$, it is known as silver ratio. If $n = 3$, it is $1 : (3 + \sqrt{13})/2 \approx 0.30277$, it is known as bronze ratio. These ratios have been respectively used in the literatures [9,25,27]. And in this paper we use $n = 4$, the partition ratio is $1 : (2 + \sqrt{5}) \approx 0.236$. We use the ratio to calculate addition chain and obtain new scalar multiplication algorithm.

3 Proposed Explicit Algorithm

In this section, we use the new ratio for finding a new shorter addition chain for an arbitrary positive integer, and later we propose an explicit algorithm for it.

3.1 New Ratio Addition Chain

In order to reduce the length of addition chain, we propose the addition chain based on the new ratio.

Firstly, input a big integer number, and then use the two equations to iterate in different situation:

$$F_n = F_{n-1} \times \frac{1}{\delta} \tag{8}$$

$$F_n = 4F_{n-1} + F_{n-2} (n \ge 3) \tag{9}$$

Finally, we can get a new addition chain. The Algorithm 1 describes the detailed process to obtain a new addition chain.

Algorithm 1 New ratio addition chain algorithm

Input: A positive k
Output: $e = \{e_1, e_2, ..., e_n\}$, $s = \{g_1, g_2, ..., g_m, u_{i-1}, u_i\}$

$\delta \leftarrow 2 + \sqrt{5}$
$u_0 \leftarrow k$
$u_1 \leftarrow \lceil k \times \frac{1}{\delta} \rceil$
$u_2 \leftarrow u_0 - 4 \times u_1$
$e = \{\}$
$s = \{\}$
$i \leftarrow 1$
$j \leftarrow 1$
while $u_i \ge 0$ **do**
    $m_{i+1} \leftarrow u_{i-1} - 4 \times u_i$
    if $m_{i+1} \le 0$ **then**
        break
    else
        if $m_{i+1} \times 4 \le u_i \le m_{i+1} \times 5$ **then**
            $u_{i+1} \leftarrow m_{i+1}$
            $e \leftarrow e \cup \{0\}$
        else
            $u_{i+1} \leftarrow \lceil u_i \times \frac{1}{\delta} \rceil$
            $e \leftarrow e \cup \{1\}$
            $g_j \leftarrow m_{i+1}$
            $s \leftarrow s \cup \{g_j\}$
            $j \leftarrow j + 1$
        end if
        $i \leftarrow i + 1$
    end if
end while
$s \leftarrow s \cup \{u_{i-1}, u_i\}$
$e \leftarrow$ reverse the arrangements in e and rename the elements in increasing order starting with numeral 1 to $n + 1$
$s \leftarrow$ reverse the arrangements in g and rename the elements in increasing order starting with numeral 1 to $n + 1$
return $e = \{e_1, e_2, ..., e_n\}$, $s = \{g_1, g_2, ..., g_m, u_{i-1}, u_i\}$
end

To perform Algorithm 1, we give the Example 1 here.

Example 1. *Perform Algorithm 1 for input $k = 207062$.*

We begin by letting,

$$\begin{aligned} u_0 &= k = 207062, \\ u_1 &= \lceil u_0 \times \frac{1}{\delta} \rceil = 48867, \\ u_2 &= u_0 - 4u_1 = 11594, e_1 = 0; \\ u_3 &= u_1 - 4u_2 = 2491, e_2 = 0. \end{aligned}$$

Since $m_4 = u_2 - 4u_3 = 1630$, *and* $1630 \times 4 = 6520 > 2491$, *so* $g_1 = 1630$,

$$\begin{aligned} u_4 &= \lceil 3 \times \frac{1}{\delta} \rceil = 588, e_3 = 1; \\ u_5 &= u_3 - 4u_4 = 139, e_4 = 0; \\ u_6 &= u_4 - 4u_5 = 32, e_5 = 0. \end{aligned}$$

Since $m_7 = u_5 - 4u_6 = 11$, *and* $11 \times 4 = 44 > 32$, *so* $g_2 = 11$,

$$\begin{aligned} u_7 &= \lceil u_6 \times \frac{1}{\delta} \rceil = 8, e_6 = 1; \\ u_8 &= u_6 - 4u_7 = 0, e_7 = 0. \end{aligned}$$

We stop the above continuous procedure at u_8, *since* $u_8 = 0$. *We obtained the following storages.*

$$\begin{aligned} e &= \{0, 0, 1, 0, 0, 1, 0\} \\ s &= \{g_1 = 1630, g_2 = 11, u_7 = 8, u_8 = 0\} \end{aligned}$$

Then we reverse the arrangements of the elements in the set e *and* s *and rename it in increasing order starting from* e_1 *and* g_1. *Thus, it results in the following new representation.*

$$\begin{aligned} e &= \{0, 1, 0, 0, 1, 0, 0\} \\ s &= \{g_1 = 11, g_2 = 1630, u_7 = 8, u_8 = 0\} \end{aligned}$$

3.2 New Scalar Mutiplication

In this section, we propose a SPA (Simple Power Attack) resistant scalar multiplication algorithm. **Algorithm 2** illustrates how to apply the new addition chain on elliptic curves. Set the initial value $T_1 = u_i P$, $T_2 = u_{i-1} P$, and then according to the value of e_i for subsequent calculations. Finally, the kP can be computed.

Algorithm 2 New Scalar Mutiplication

Input: $e = \{\}$, $s = \{\}$
Output: kP
for $m = 1$ To m **do**
  $G_m \leftarrow g_m P$
  $G \leftarrow G \cup \{G_m\}$
  $T_1 \leftarrow u_i P$
  $T_2 \leftarrow u_{i-1} P$
end for
for $i = 1$ To n **do**
  if $e_i = 0$ **then**
    $T_{i+1} \leftarrow 4T_i + T_{i-1}$
  end if
  if $e_i = 0$ **then**
    $T_{i+1} \leftarrow 4T_i + G_m$
    $m \leftarrow m + 1$
  end if
end for
return T_{n+1}
end

To perform Algorithm 2, we give the Example 2 here.

Example 2. *Perform Algorithm 2.*

INPUT:

$$\begin{aligned} e &= \{0, 1, 0, 0, 1, 0, 0\}, \\ s &= \{g_1 = 11, g_2 = 1630, u_7 = 8, u_8 = 0\}, \\ G_1 &= 11P, G_2 = 1630P, \\ T_1 &= 0P, T_2 = 8P, \\ e_1 &= 0, T_3 = 4T_2 + T_1 = 32P; \\ e_2 &= 1, T_4 = 4T_3 + G_1 = 139P; \\ e_3 &= 0, T_5 = 4T_4 + T_3 = 588P; \\ e_4 &= 0, T_6 = 4T_5 + T_4 = 2491P; \\ e_5 &= 1, T_7 = 4T_6 + G_2 = 11594P; \\ e_6 &= 0, T_8 = 4T_7 + T_6 = 48867P; \\ e_7 &= 0, T_9 = 4T_8 + T_7 = 207062P. \end{aligned}$$

OUTPUT: $Q = 207062P$.

4 Analysis of Algorithm

In this section, firstly, we analyze the security of the new algorithm. Then, in order to make the efficiency analysis more accurately, we compare the new algorithm with previous algorithms in the literatures [7, 9, 25–27]. In addition, some practical comparison results are listed in the tables and figures below.

4.1 The Chain Length Analysis of the New Algorithm

Algorithm 3 describes the details to compute the chain length. Randomly select 10000 integers of 160bit, 192bit, 224bit and 256bit respectively, and make statistics on the chain length of these numbers by Algorithm 3.

Algorithm 3 The algorithm of the chain length analysis

length $\leftarrow \{\}$
for $i = 1$ To 10000 **do**
  $l \leftarrow$ RandomInteger[2^{159},FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF]
  $m\{\} \leftarrow 2(l)$
  length $\leftarrow$ count(m)
end for
return length
end

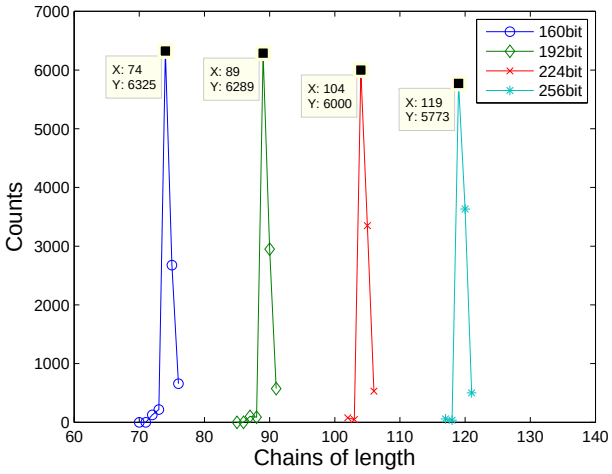


Figure 1: Chain length analysis diagram

The bit conversion relation between ternary and binary is given in Table 2.

The specific experimental results obtained are represented in line graph as shown in Figure 2.

It can be clearly seen from Figure **??** that according to the chain length statistics of 10000 integers with 160bit, the maximum length of chains is concentrated in 74, 75 and 76, and the maximum length is 74. Similarly, we can use the same methods find that the chain length of 192bit is 89, the chain length of 224bit is 104, and the chain length of 256bit is 119.

Table 2: Conversion relationship of bits between binary and ternary

Binary	160	192	224	256
Ternary	101	122	142	162

Table 3 shows the addition chain length of different addition chain, from which we can see that the chain length of the new proposed algorithm is respectively 71.3%, 67.8%, 40.8% and 20.4% shorter than GRAC sequences, Fibonacci sequences, Pell sequences and BRAC sequences, indicating the efficiency of the proposed algorithm.

Table 3: Comparsion of addition chain length

Algorithm	**Length of Chains**
GRAC [9]	258
Fibonacci [7]	230
MADC [26]	160
Pell [27]	125
BRAC [25]	93
New Algorithm	74

4.2 Security Analysis

The scalar multiplication in variable execution time is vulnerable to time attack. Common attacks on ECC include Simple Power Analysis (SPA) and Differential Power Analysis (DPA). In general, preventive measures against SPA can be effiectively converted into measures against DPA [14]. For this reason, we focus on the SPA attack in our security discussion.

SPA is a technology which analyzes the key by analyzing the different energy required for the executions in the algorithm. The system consumption of energy is different that mainly depending on the instructions executed by the microprocessor. When the microprocessor operation performed at different part of the encryption algorithm, some of the energy consumption of the system is very obvious. With this feature, the attacker can distinguish a single instruction to achieve the purpose of breaking the algorithm.

The Algorithm 2 describes the calculation process of scalar multiplication, where the new algorithm performs a quadruple point operation and point addition operation each time. Furthermore, according to the data in Table 1, the algorithm operational capacity is $l(3I + 3S + 5M)$, where l is the chain length. According to the computation amount in the domain, let $1S = 0.8M$, $1I = 5M$, therefore, the total computation amount of the algorithm is $l(15M + 2.4M + 5M) = 22.4lM$.

Since each scalar multiplication operation is fixed at $4P + Q$, it can blur the energy difference of energy release and prevent the opponent from analyzing the key based on the power consumption attack trajectory. In other words, the new proposed chain with $4P+Q$ can resist SPA attack effectively, which can ensure the security of the key.

4.3 The Efficiency of the New Algorithm

To evaluate the efficiency performance of our new proposed algorithm, five existing algorithms including EAC-320 [19], SAC-260 [19], Fibonacci-and-add [7], Window Fib-and-add [19] and GRAC-258 [9] are compared with the proposed algorithm in terms of computation computuation.

Table 4: Efficiency comparison results of various algorithms

Algorithm	**Computation Costs**
EAC-320 [19]	2939M
SAC-260 [19]	2387M
Fibonacci-and-add [7]	2311M
Window Fib-and-add [19]	1960M
GRAC-258 [9]	1907M
New Algorithm	1657.6M

The efficiency comparison results are shown in Table 4, from which we can draw the following conclusion

Table 5: Efficiency comparison results of various algorithms with different length bits

Length of Bits	Algorithms	Computation Costs
160	Reference [33]	3140.8M
	Referencce [11]	2390M
	New Algorithm	1657.6M
192	Reference [33]	3692.8M
	Referencce [11]	2887M
	New Algorithm	1993.6M
224	Reference [33]	4275.2M
	Referencce [11]	3361M
	New Algorithm	2352M
256	Reference [33]	4856.8M
	Referencce [11]	3834M
	New Algorithm	2665.6M

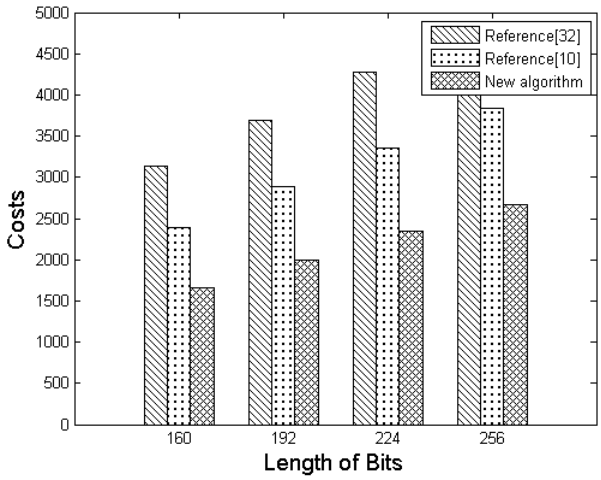


Figure 2: Computation costs analysis diagram

that our proposed algorithm can outperform EAC-320 by about 30.6%, respectively 28.3%, 27.9%, 15.5% and 13.1% improvements for SAC-260, Fibonacci-and-add, Window Fib-and-add and GRAC-258.

The efficiencies of different algorithms are further evaluated in the same bits, and the comparison results are show in Table 5, from which we can see that compared with other two algorithms, our proposed algorithm takes lower computation time with all the different length bits. For a clearer comparison, we plot the data in Figure **??**. The results in Table 4 and Table 5 illustrate that our proposed algorithm is superior to other algorithms in terms of efficiency.

5 Conclusion

In this paper, to overcome the SPA attack in the key study, we have proposed a SPA resistant scalar multiplication algorithm. We have proposed an explicit algorithm for the new addition chains and presented an elegant gant SPA resistant scalar multiplication algorithm. Our new proposed algorithm based scalar multiplication has respectively outperformed EAC-320, SAC-260, Fibonacciand-add, Window Fib-and-add and GRAC-258 by about 30.6%, 28.3%, 27.9%, 15.5% and 13.1%.

Further work may include finding chains of much shorter lengths in order to reduce the computational cost of the scalar multiplication algorithm and research the operation of "$4P + Q$" to reduce the underlying computation. Furthermore, if one could reduce the storage content, new proposed algorithm based algorithm could be more applicable to elliptic curve cryptosystems where constraint memory devices such as smart cards needs to be implemented.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61872058), the Key Research and Development Program of Shaanxi (Program No.2021NY-211). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] M. Boudabra and A. Nitaj, "A new public key cryptosystem based on edwards curves," *Journal of Applied Mathematics & Computing*, no. 1–2, pp. 1–20, 2019.

[2] M. Ciet, M. Joye, K. Lauter, and P. L. Montgomery, "Trading inversions for multiplications in elliptic curve cryptography," *Designs Codes & Cryptography*, vol. 39, no. 2, pp. 189–206, 2006.

[3] V. S. Dimitrov and G. A. Jullien, "Loading the bases: A new number representation with applications," *Circuits & Systems Magazine IEEE*, vol. 3, no. 2, pp. 6–23, 2003.

[4] V. S. Dimitrov and P. K. Mishra, "Efficient and secure elliptic curve point multiplication using double-base chains," in *International Conference on the Theory and Application of Cryptology and Information Security*, vol. 3788, pp. 59–78, 2005.

[5] A. Dubey, R. Cammarota, and A. Aysu, "MaskedNet: The first hardware inference engine aiming power side-channel protection," *Cryptography and Security*, 2019. arXiv:1910.13063.

[6] K. Eisentraeger, K. Lauter, and P. L. Montgomery, "Fast elliptic curve arithmetic and improved weil pairing evaluation," in *RSA Conference on the Cryptographers Track*, 2003. arXiv:math/0208038.

[7] Z. Fuling, "The finite sum formula for generalized fibonacci numbers," *Journal of Chongqing Normal University (Natural Science)*, vol. 28, no. 5, pp. 45–48, 2011.

[8] X. Gaowen and L. Maixue, "The some sum formula for generalized fibonacci numbers," *Chinese Quarterly Journal of Mathematics*, vol. 22, no. 2, pp. 258–265, 2007.

[9] R. R. Goundar, S. Ken-Ichi, and T. Masahiko, "Spa resistant scalar multiplication using golden ratio addition chain method," *Iaeng International Journal of Applied Mathematics*, vol. 38, no. 2, pp. 83–88, 2008.

[10] J. Guajardo and C. Paar, "Efficient algorithms for elliptic curve cryptosystems," in *Annual International Cryptology Conference*, vol. 1294, pp. 342–356, 1997.

[11] L. Hengzhuang, D. Qianhui, and L. Yibing, "Efficient ECC scalar multiplication algorithm based on symmetric ternary in wireless sensor networks," in *Progress in Electromagnetics Research Symposium-Fall*, 2017. DOI: 10.1109/PIERS-FALL.2017.8293258.

[12] M. S. Hwang, S. F. Tzeng, C. S. Tsai, "Generalization of proxy signature based on elliptic curves", *Computer Standards & Interfaces*, vol. 26, no. 2, pp. 73–84, 2004.

[13] I. Kabin, Z. Dyka, K. Dan, and P. Langendoerfer, "Horizontal address-bit DPA against montgomery kP implementation," in *International Conference on ReConFigurable Computing and FPGAs (ReConFig)*, 2017. DOI: 10.1109/RECONFIG.2017.8279800.

[14] E. Karimi, Z. H. Jiang, Y. Fei, and D. Kaeli, "A timing side-channel attack on a mobile GPU," in *IEEE 36th International Conference on Computer Design (ICCD'18)*, 2018. DOI: 10.1109/ICCD.2018.00020.

[15] E. W. Knudsen, "Elliptic scalar multiplication using point halving," in *International Conference on the Theory & Applications of Cryptology & Information Security: Advances in Cryptology*, vol. 1716, pp. 135–149, 1999.

[16] N. Koblitz, "Elliptic curve cryptosystems," *Mathematics of Computation*, vol. 48, no. 177, 1987.

[17] P. C. Kocher, "Timing attacks on implementations of diffie-hellman, RSA, DSS, and other systems," in *Proceedings of the 16th Annual International Cryptology Conference on Advances in Cryptology*, vol. 1109, pp. 104–113, 1999.

[18] Litasari and B. Rahadjo, "Design and implementation stegocrypto based on elgamal elliptic curve," in *International Conferences on Information Technology*, 2017. DOI: 10.1109/ICITISEE.2017.8285567.

[19] N. Meloni, "Fast and secure elliptic curve scalar multiplication over prime fields using special addition chains," *IACR Cryptology Eprint Archive*, 2006. (`https://eprint.iacr.org/2006/216.pdf`)

[20] V. S. Miller, "Use of elliptic curves in cryptography," in *Conference on the Theory and Application of Cryptographic Techniques*, vol. 218, pp. 417–426, 1985.

[21] A. Mullai and K. Mani, "Enhancing the security in RSA and elliptic curve cryptography based on addition chain using simplified swarm optimization and particle swarm optimization for mobile devices," *International Journal of Information Technology*, pp. 1–14, 2020.

[22] Neelappa and N. G. Kurahatti, "Design and Implementation of ECC-Based RFID Tag for Wireless Communications on FPGAs," *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, Advances in Intelligent Systems and Computing, vol. 628, pp 415-430, 2017.

[23] T. Oliveira, J. López, and F. Rodríguez-Henríquez, "The montgomery ladder on binary elliptic curves," *Journal of Cryptographic Engineering*, vol. 8, no. 5, pp. 1–18, 2017.

[24] L. Parrilla, J. A. Alvarez-Bermejo, E. Castillo, J. A. Lopez-Ramos, D. P. Morales-Santos, and A. Garcia, "Elliptic curve cryptography hardware accelerator for high-performance secure servers," *Journal of Supercomputing*, vol. 75, no. 3, pp. 1107–1122, 2019.

[25] L. Shuanggen, L. DanDan, and L. Xiao, "Elliptic curve scalar multiplication algorithm based on bronze ratio addition chain," *Journal of Shandong University (Natural Science)*, vol. 54, no. 11, pp. 12–19, 2019.

[26] L. Shuanggen, Y. Huatong, and L. Fagen, "SPA resistant scalar multiplication on edwards curve," *Computer Engineering & Applications*, vol. 53, no. 1, pp. 103-106, 2017.

[27] L. Shuanggen and Z. Hui, "Fast and secure scalar multiplication based on pell sequence," *Computer Engineering and Applications*, vol. 55, no. 4, pp. 125–129, 2019.

[28] W. N. Wan, H. Chen, J. Chen, S. B. Zhang, and School of Cybersecurity, "Side channel security analysis of elliptic curve cryptography of blockchain," *Journal of Applied Sciences*, 2019.

[29] W. Wang and S. Fan, "Attacking openSSL ECDSA with a small amount of side-channel information," *Science China (Information Sciences)*, vol. 061, no. 003, pp. 53–66, 2018.

[30] W. Wunan, C. Hao, and C. Jun, "The attack case of ECDSA on blockchain based on improved simple power analysis," in *Artificial Intelligence and Security*, pp. 120-132, 2019.

[31] Z. Yaxin, T. Yerong, and L. Yinglong, "An optimization for differential power analysis based on time series verification," in *Proceedings of International Conference on Wireless Communication, Network and Multimedia Engineering*, 2019. (`https://doi.org/10.2991/wcnme-19.2019.38`)

[32] D. Yong, "Research of quasi-cyclic matrices based on generalized fibonacci and lucas numbers," *Journal of Chongqing Normal University (Natural Science)*, vol. 32, no. 6, pp. 72–76, 2015.

[33] Z. Youqiao, Z. Wuneng, and S. Ye, "Improvement scheme for scalar multiplication against power analysis attacks in elliptic curve cryptography," *Computer Engineering & Science*, vol. 36, no. 4, pp. 644–648, 2014.

[34] M. Zhao and G. E. Suh, "FPGA-based remote power side-channel attacks," in *FPGA-based Remote Power Side-channel Attacks*, 2018. DOI: 10.1109/SP.2018.00049.

Biography

Shuang-Gen Liu, received the PH.D. degree in cryptography form Xidian University in 2008. He is currently an associate professor with the school of cyber security, Xi'an University of Posts and Telecommunications, Xi'an, China. He is a member of the China Computer Federation, and a member of the Chinese Association for Cryptologic Research. His recent research interests include crptography and information security.

Si-Jia An, is a graduate student of Xi'an University of posts and telecommunications. She is mainly engaged in the research of elliptic curve cryptosystem.

Yi-Wei Du, is a undergraduate student of Xi'an University of posts and telecommunications. She is mainly engaged in the research of elliptic curve cryptosystem.

Exploitation of The Distributed Network Protocol in ICS with Improved D-Y Model Based on Petri Net

Ye Lu[1] and Wei-Bin Ou[2]

(Corresponding author: Wei-Bin Ou)

School of computer Science, Baoji University of Arts and Sciences[1]
Baoji Shaanxi 721000, China
College of Electronic and Electrical engineering, Baoji University of Arts and Sciences[2]
Email: luye528@126.com

(Received June 14, 2020; Revised and Accepted Feb. 11, 2021; First Online Aug. 8, 2021)

Abstract

Whether the communication protocol is safe or not is related to the stable and reliable operation of the industrial system. As a typical application layer protocol in the industrial field, Distributed Network Protocol has attracted much attention. We use CPN-Tool to present a formal study of DNP3 security, integrity, and authentication in this work. We introduce an improved Delov-Yao attacker model to reduce the size of the state space. Furthermore, we carry out the security evaluation of the protocol in the full attack state and give the vulnerability exploitation path according to the evaluation results. The exploited vulnerabilities reflect that the DNP3 protocol cannot resist three attacks listed in IEEE standards: replay, tampering, and spoofing. The CPN model developed provides a security assessment method for DNP3 that can clarify the methodology for achieving secrecy, integrity, and authentication for designers and developers interested in other protocols.

Keywords: Distributed Network Protocol; Exploitation; Delov-Yao; Petri-Net

1 Introduction

With the rapid development of intelligent manufacturing, big data of industry and the Internet of things, the communication protocols of industrial control systems tend to be open and standardized, especially the introduction of industrial Ethernet, TCP/IP and other open communication protocol standards, which greatly increase the risk of network attacks on industrial control systems. According to ICS-CERT Advisories, there had been 1350 cyber attacks on industrial control systems by May 2020, including critical infrastructure such as power grids, water conservancy facilities and transportation systems. According to the ICS-CNVD of CNCERT in China, a total of 1120 vulnerabilities were reported from 2018 to May 2020, accounting for 44 percent of the total number of vulnerabilities in the past.

In order to deal with more and more network security threats, researchers have proposed a series of security regulations and communication standards for industrial control system network protocols [13], such as CIP-Safety, OPC-UA, DNP3-Sec, DNP3-SA and so on. DNP3-SA protocol is the first industrial Ethernet security protocol with authentication attribute in industrial networked control systems, and it is widely used in DCS systems, although its security loopholes have been found. It is mainly used in industrial infrastructure fields such as electric power automation system, oil and gas system and so on. The protocol ensures the secure transmission of data by generating message authentication code (MAC).

The security of DNP3-SA protocol is described informally in IEEE Electrical Standard [10]. Although the informal description method can guide the design of DNP3-SA protocol, it has the following defects:

1) It is easy to cause ambiguity in the purpose of the agreement, difficult to understand and inaccurate description [12].

2) Security protocols run in an open environment, and attackers can attack protocol entities through replay, spoofing, denial of service and other methods, but manual identification of their threats has a large workload and low accuracy.

3) Due to the lack of formal verification methods, the accuracy of protocol security analysis is low.

Formal description and analysis methods can give the detailed model of the protocol, with the help of formal methods and tools to evaluate the security of the protocol and can ensure that the embedded security mechanism does not affect the functionality of the protocol itself, and

can will not attract new errors. The formal method uses strict mathematical semantics to describe the protocol, and computer-aided verification tools are used to verify the availability and security of the protocol. For example, the typical Delov-Yao attack model [8] is a typical formal attack model. This model can help protocol designers to reduce the difficulty of manually finding protocol vulnerabilities. At present, a lot of research work [19, 24, 26] shows that formal methods can make protocol security improvement more effective.

At present, the main methods for security assessment of protocol state are functional verification based on protocol finite state machine model and security evaluation based on stochastic process model. However, in terms of protocol engineering modeling, protocol performance analysis based on two different models has the following shortcomings: First, the protocol model based on performance analysis is generally unable to accurately simulate the protocol behavior. Second, the use of two types of protocol formal models for protocol design and analysis will lead to too large state space of the model, and even lead to state space explosion. Colored Petri net [16] is a formal analysis theory that integrates protocol behavior verification and performance analysis. Based on this theory, the CPN behavior model of the protocol can be constructed to verify the functional consistency of the protocol, and the attacker model is added to analyze and evaluate the security, so as to overcome the above two kinds of defects and ensure that the state transition in the behavior CPN model can be associated with man-in-the-middle attacks.

Vulnerability disclosure is the basis for the formulation of defense strategy in Industrial Networked Control System, and its accuracy ultimately determines the stable operation of the system. Based on the research of literature [1], this paper further uses CPN modeling tools and state space analysis tools to formally describe the behavior process and security attributes of DNP3-SA protocol, and establish a security evaluation model of the protocol. The main improvements are about the Delov-Yao attacker model, which further reduces the state space, and three improved attacker models of replay, deception and tampering are introduced, and final vulnerability exploitation path of attack. This, in turn, can clarify the methodology for achieving secrecy, integrity, and authentication for designers and developers interested in these Distributed Network Protocols. We believe that our model and discussion of the protocol security properties are beneficial for both researchers and practitioners. To the best of our knowledge, this is the first work that presents vulnerability exploitation path of DNP3-SAv5.

The remainder of this paper is organized follows. section 2 reviews the research progress of DNP3 protocol security. Section 3 gives the adversary model and security attributes of the protocol according to the IEEE specifications. Section 4 reviews the improvement scheme of D-Y attacker model made by Bai, and further proposes an improved parameterized attack model based on CPN. Section 5 uses cpn tools to establish the attacker model of DNP3-SA protocol and verifies the consistency of the model behavior. Section 6 analyzes the attack behavior and excavates the vulnerability based on the state space tool, and gives the attack path. Finally, conclusions and future work are presented in Section 7.

2 Review the Security of DNP3 Protocol

DNP3 protocol has been running in industrial networked control systems (ICS) for decades, such as Supervisory Control And Data Acquisition systems (SCADA), Distributed Control Systems (DCS), and Process Control Systems (PCS). However, as revealed by a number of attacks on critical infrastructure in recent years, there are many security vulnerabilities in the protocols in these systems. A large number of protocols are transmitted in clear text without authentication and integrity checks, such as the shocking Stuxnet worm and Black-Energy. Therefore, for a long time, a large number of researchers have devoted themselves to propose secure TCP/IP-based transport protocols, which have certain authenticity, integrity, confidentiality and non-repudiation by using symmetric or asymmetric cryptography technology. Such as improved DNP3 protocol, DNP3-SA protocol, DNP3-SEC protocol, ICCP-SEC protocol and so on. However, these schemes have some limitations, which only focus on the implementation of the security function of the protocol, but lack of formal methods for the analysis and verification of protocol security. The formal method can study the security of the protocol based on the security strength of the encryption algorithm itself.

At present, there are two secure versions of DNP3 protocol, DNP3-Sec and DNP3-SAv5. The former focuses on link layer security, while the latter focuses on application layer security. Although authentication, encryption, authorization, integrity check and other security mechanisms are used in the two types of security protocols, they still face some security threats. Literature [6] uses a variety of attack scenarios for penetration testing in a simulated environment including DNP3 protocol. Vulnerability analysis and penetration testing show that there is a man-in-the-middle (MITM) attack. Literature [20] uses SCAPY to send a large number of forged data to DNP3 communication links, verifying that DNP3 protocol has security vulnerabilities such as data tampering, replay and spoofing. In the literature [18], the response behavior of DNP3 protocol under replay attack, rogue intrusion attack and flooding attack is studied, and its vulnerability is verified.

Literature [15] validates a series of attacks on a small test bed and proposes DNP3 improvement measures based on encryption and authentication. The literature [22] reviews some vulnerabilities that have been found in DNP3 and makes security improvements to the protocol in the application layer. Literature [23] analyzes a large number of industrial Ethernet protocols from the

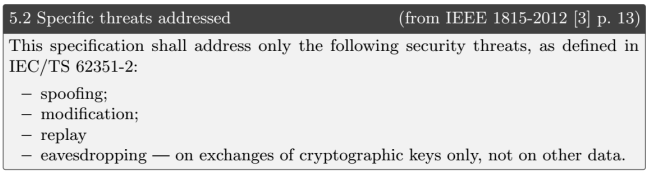
aspects of consistency, integrity and availability, and gives examples of the related studies of DNP3, DNP3SEC and dnp3sa, and gives some suggestions for security improvement. Literature [25] proposes an intrusion detection algorithm based on machine learning algorithm for smart grid to help dnp3 protocol detect attacks in time.

The literature [17] reviews the attacks in the smart grid, analyzes the protocols involved according to the types of attacks, and gives examples of possible improvements. Based on four supervised learning algorithms, Literature [9] has found a Peekaboo attack in a large number of substations running DNP3 protocol. Literature [14] analyzes the fragility of the transport layer of DNP3 protocol, proposes an intrusion detection technology of RNN, and describes the verification process of the proposed model through a DNP3 message of an actual substation. Literature [7] uses Bayesian analysis to model the likelihood distribution of Rttd of legitimate information and information attacked by hackers, and then proposes an intrusion detection model for DNP3 protocol. Based on the colored Petri net, literatures [1, 2] analyzes the security of the protocol, finds that replaying the previously authenticated commands can remotely control the slave station, and puts forward an improved scheme of the protocol. But literature [5] makes a complete analysis of the DNP3-SA protocol, and makes a complete state machine model for the complex behavior of the protocol.

Based on TAMARIN, it is proved that the security of the protocol is almost consistent with that declared by the standard. Literature [21] comprehensively analyzes the advantages and disadvantages and encryption structure of DNP3-SA protocol, and discusses the impact of encryption operation on the performance of the protocol. Literature [11] discusses all kinds of attacks in the application layer of DNP3 protocol. By extracting the traffic characteristics of four substations, a set of lightweight security improvement scheme is proposed. Literature [4] points out that the implementation complexity and analytical fuzziness of the protocol are related to the coupling state, and proposes a security enhancement scheme to avoid similar attacks. From the above literature, it can be concluded that DNP3 protocol and other variants mainly exist the following attack vectors: man-in-the-middle, replay, eavesdropping, data tampering, denial of service, buffer overflow and so on.

3 Adversary Model and Security Attributes

By checking whether unauthorized commands are executed by the slave station, we can verify whether the authentication mechanism of the DNP3-SA protocol achieves the security claimed in the standard. The DNP3-SA protocol complies with the requirements of the IEC62351 specification, which can ensure that the protocol is not affected by attacks such as spoofing, modification, replay and eavesdropping. For related descriptions, see the IEEE-1815-2012 and IEC62351 standards (Figures 1, 2, 3, and4).

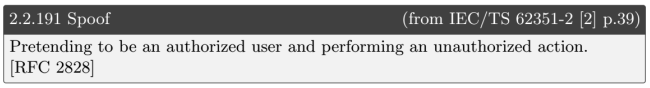

5.2 Specific threats addressed (from IEEE 1815-2012 [3] p. 13)

This specification shall address only the following security threats, as defined in IEC/TS 62351-2:

- spoofing;
- modification;
- replay
- eavesdropping — on exchanges of cryptographic keys only, not on other data.

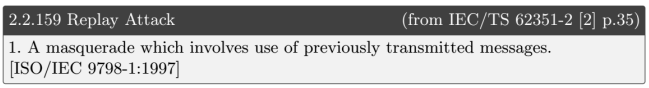
Figure 1: Types of attacks in the IEEE1815-2012 standard

The attack behaviors illustrated in Figures 1 are all described informally in the standard IEC62351. At the same time, the attack types such as masquerade and Perfect forward secrecy are further listed in the reference [5].


2.2.191 Spoof (from IEC/TS 62351-2 [2] p.39)

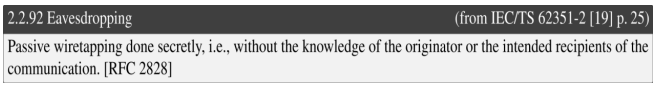
Pretending to be an authorized user and performing an unauthorized action. [RFC 2828]

Figure 2: Spoofing attack in the IEC62351 standard


2.2.159 Replay Attack (from IEC/TS 62351-2 [2] p.35)

1. A masquerade which involves use of previously transmitted messages. [ISO/IEC 9798-1:1997]

Figure 3: Replay attack in the IEC62351 standard


2.2.92 Eavesdropping (from IEC/TS 62351-2 [19] p. 25)

Passive wiretapping done secretly, i.e., without the knowledge of the originator or the intended recipients of the communication. [RFC 2828]

Figure 4: Eavesdropping attacks in the IEC62351 standard

However, for modification, the IEC62351 standard is not clearly stated. For the sake of integrity and with reference to other security standards, this paper defines tampering attacks that specifically refer to the ability of attackers to modify messages in transmission arbitrarily. The attacker models and capability assumptions of the above three man-in-the-middle attacks in our protocol attack-model are given below:

MD_ATK: Modify attack, the attacker has the ability to modify the messages in the NET subpages, including request messages and challenge messages.

RP_ATK: Replay attack, the attacker has the ability to intercept the message sent by the master station M and resend it to the slave station O.

SP_ATK: Spoofing attack, the attacker has the ability to impersonate the master station M to send messages to the slave station O.

MRS_ARK: Attackers launch Modify, Replay and Spoofing attacks at the same time.

0_ATK: An attacker cannot launch an attack without enabling any attack parameters.

4 Improved Attacker Model of Delov-Yao

Delov and Yao published an important paper [16] which has a profound impact on the research of protocol security. There are two main contributions: the first contribution is to discuss the security properties of the protocol itself on the assumption that the cryptosystem is "perfect", which can help researchers to concentrate on the inherent security properties of the protocol without discussing the security of cryptographic algorithms. The second contribution is that Delov and Yao proposed the attacker model and proposed that the attacker has stronger computing power than the real participants of the protocol. The attacker can eavesdrop, intercept, tamper, replay the information exchanged between real entities during the operation of the protocol, encrypt, decrypt, split and combine the original message, and forge the message content. However, the attacker model can combine "arbitrary" messages, which will cause a large increase of invalid messages, make the messages form an infinite loop, cause the model cannot be terminated normally, and finally cause the system state space to explode.

This section improves the Delov-Yao attacker model, on the one hand, applies the attack in the form of parameterization to the arc expression to reduce the state space; on the other hand, it effectively limits the messages that the attacker splits and combines, and only splits the key messages that are valid to prevent the attack from entering a disordered state and preventing the state space from exploding.

4.1 Review Bai's Study

In reference [3], Bai divides the Delov-Yao attacker model into message splitting stage and message combination stage. In the message splitting stage, the attacker splits all intercepted messages atomically, and stores the messages that need to be split and the split atoms in the element set DB. In the message composition phase, attackers extract the atomic information from DB and reassemble them into valid messages according to the protocol specification and store them in the set CB, which reduces the state space while maintaining the attack capability.

The buffer unit AB set of the basic elements is used as the pre-set of the composite element set CB, which is used to store the atomic information of all kinds of split. In order to avoid the cycle caused by split and combination operations, it is necessary to split the intercepted messages first, and then reassemble the messages according to the required capabilities of the attackers, and finally generate a set of combination elements CB, and send the generated elements to the communication link.

Compared with the original model, the biggest advantage of Bai's model is following two aspects. First, serialization defines the logical relationship between split and combination operations, while in the original model, split and combination operations occur randomly, which may lead to infinite loops and state explosions. Secondly, by limiting the data type and key fields, the generated message can not only be effectively received by the receiver, but also closely related to the security attributes, which effectively reduces the state space of the model.

Specifically, it is assumed that entity A and entity B are entity objects participating in the protocol interaction, and the original message transmitted by A and B is m. The encryption key used by both sides of the communication is k, the decryption key is k', and the communication link between the two parties is Channel. If "a" represents the basic elements obtained from the split of all the original messages, the formal description of the transformation rules for message splitting and message combination in the Delov-Yao attacker model in Bai's scheme is shown in Figure 5.

Split rule	
channel(A,m,B) $\xrightarrow{resolve(A,m,B)}$ DB(m), DB(A), DB(B)	(1)
DB(m1•$m2$) $\xrightarrow{resolve(m1,m2)}$ DB(m1), DB($m2$)	(2)
DB({m}k), AB(k') $\xrightarrow{resolve(\{m1\}k,k')}$ DB(m), AB(k ')	(3)
Combination rule:	
CB(m), AB(A), AB(B) $\xrightarrow{assemble(m,A,B)}$ $channel$(m, A, B)	(4)
CB(m), AB(A) $\xrightarrow{assemble(m,A)}$ $channel$(m, A)	(5)
CB($m1$), CB($m2$) $\xrightarrow{assemble(m1,m2)}$ CB($m1$, $m2$)	(6)
CB(m), AB(k) $\xrightarrow{assemble(m,k)}$ CB({m} k), AB(k)	(7)
Transformation rule	
DB(a) $\xrightarrow{change(a)}$ AB(a)	(8)
AB(a) $\xrightarrow{change(a)}$ CB(a)	(9)
DB({m} k), ¬AB(k') $\xrightarrow{change(\{m\}k)}$ CB({m} k)	(10)

Figure 5: Formal description of message split and combination rules

Different from the random split or combination operation of the original model, the improved attacker model will be executed in three phases according to the above rules. Step 1, The split operation is performed in the order of Rules 1 to 3. Step 2, uses the combination operation in the order of Rules 4 to 7 to generate only the messages that can be effectively received by the recipient and with critical value to the security assessment. Step 3, converts according to Rules 8 to10. The improved attacker model can still generate messages output from the original attacker model, which can effectively reduce the state space without weakening the ability of the original Delov-Yao attacker.

4.2 An Extension of the Bai's Scheme

The CPN model checking tool has a parameterized method, through which researchers can build the response behavior of the protocol under different parameters according to their interests, which helps us to dynamically analyze the established model, and is especially suitable for studying the response behavior of the protocol under different attack parameters. Parameterization methods are usually implemented by arc expressions, transition guards and functions.

Enable or disable attack behavior by setting the true and false values of the above Boolean expression. Figure 6 shows an example of the parameterized attack model used later. When the arc expression is false, the attack function Mattack1 can tamper with the function code. In this paper, a variety of attack functions are defined to simulate tampering, spoofing and replay attacks on the protocol.

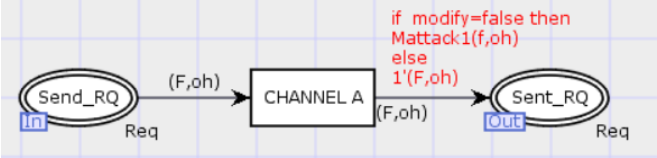


Figure 6: Example of parameterized attack model

4.3 Efficiency Analysis of Improved Attacker Model

This section takes Needham-Schroeder (NS) protocol as an example to verify the improved attacker model and analyze the actual effect of the scheme. As the first real authentication protocol, NS protocol is composed of initiator and responder and the network between them. Its security goal is to ensure that both sides of the communication can confirm each other's true identity and that the protocol entity will not be impersonated.

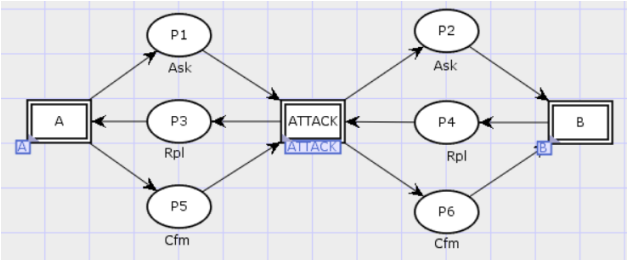


Figure 7: Top-level CPN model of the NS protocol

Figure 7 shows the top-level CPN model of the NS protocol. The layer model consists of entities A and B involved in communication and the attacker "ATTACK". The hierarchical method of the HCPN model uses alternative transitions to simulate each participating entity of the protocol, and the internal model of the transition simulates the detailed behavior of each function in the entity layer. Other message repositories in the top-level model simulate the communication channel of the network. The attacker in the communication channel can intercept all the data packets in the communication network and attack according to the improved attacker model.

Figure 8 shows the entity layer model of NS protocol replacing transition "ATTACK" without attack. In this model, according to the requirements of the implementation specification of NS protocol, the attack behavior of each stage of the protocol is simulated by alternative transition Int_Ask, Int_Rpl and Int_Cfm, respectively. The subscript of the place is used to assign the type of port place, In is the input port place and OUT is the output port place. Taking the Int_ ask subpage as an example, the original Delov-Yao attacker model and the improved Delov-Yao attacker model are analyzed. Considering that the introduction of the attacker model will lead to a large state space of the protocol model, the attacker is limited to attack the protocol only as a middleman, and the replay attack is taken as an example to simulate the execution of a single session of the protocol.

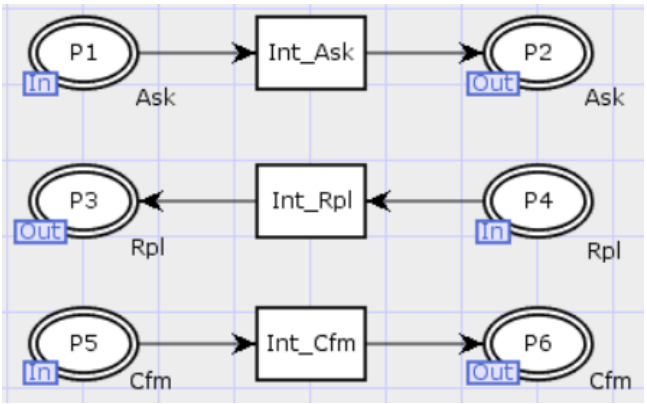


Figure 8: The CPN Model of substituting transition "ATTACK"

Figure 9 shows the subpage model corresponding to the transition Int_Ask under the original Delov-Yao attack model. The knowledge possessed by the attacker is stored in repository P3'. The transition T1 is used to simulate the decryption of the intercepted cipher text by the attacker attack. Transition T2 simulates the process that the message sequence msg obtained after decryption in the previous step is divided into random number n and identity element id. The library P3' is used to store the split and reassembled message sequence; the message sequence is finally replayed to the library P2 through transition T1.

Figure 10 depicts the attacker subpage "Int_Ask" based on the improved attacker model. The red part of the graph is the transition guard and arc expression, which constitutes the parameterized attack model. The request message sent by the protocol entity is intercepted by tran-

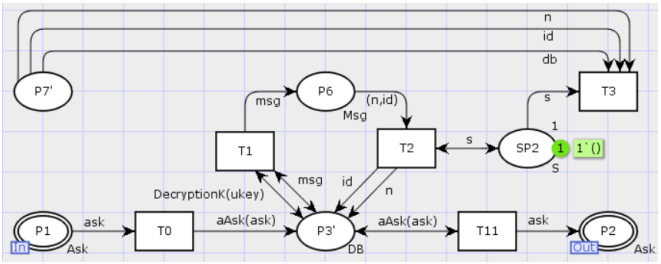


Figure 9: CPN model of the original Int_Ask

sition T0, and the intercepted message is split and stored in the repository "resolve". Transition T11 applies the transition rules defined in Figure 5 to store undecrypted, to be combined and combined messages in the repository P5'. Transition T2 decomposes the message based on the splitting rules listed in Figure 5 and obtains the key message elements and stores them in the repository "element". Transition T3 simulates the combined operation and generates valid data messages that can be identified by the receiver, then stores them in the message repository P5'. The function of SP is to restrict the combination function of transition T3 and prevent multiple messages from concurrency. Transition T4 finally sends the generated attack message to the port library P2 connected to the network channel.

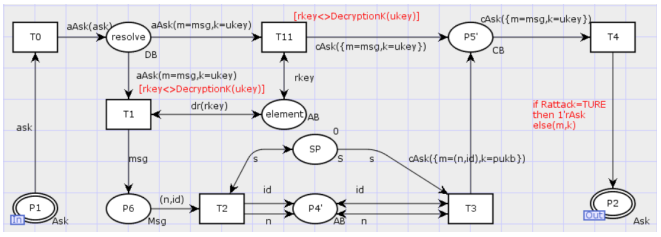


Figure 10: Int_Ask attackers model based on improved Delov-Yao model

Table 1 makes statistics on the state space according to the state space analysis report of the above two models. Within 50 seconds of computing time, the number of state space nodes and directed arcs of NS protocol under the original attacker model far exceeds the state space formed by the improved attacker model under the same assumption. In addition, the original attacker model does not have dead nodes and dead transitions in the simulation time, which violates common sense and may lead to an infinite cycle of protocol behavior. The above comparison results show that the improved attacker model can avoid generating a large number of message data that are not recognized by the receiver. On the premise of ensuring the attack ability, it can significantly improve the efficiency of the attack model and reduce the size of state space nodes.

Table 1: State space comparison of two attacker models

Attack Model	Node	Arc	Dead Node	Dead Transition	Live Transition
Bai's	337	482	0	0	0
Ours	108	170	1	T4	0

5 Attacker Model

5.1 The Establishment of Attacker Model

The CPN model of the protocol is a concrete simulation of the whole communication protocol, including the communication sides of the protocol, the communication network and the messages transmitted. As shown in Figures 11, the double-line rectangle in the diagram is the alternative transition and the ellipse is the message repository. The left alternative transition M represents the communication master station Master Station, while the middle alternative transition NET subpage represents the communication network, and the rightmost alternative transition O represents the communication slave OutStation. The top-level model completely simulates the session process of the protocol, including request-reply mode (NACR) and active mode (AGM), and the process of dealing with key information.

Figure 12 shows the NET subpage attacker model based on the top-level model of the protocol, where all simulated attacks will be carried out. The NET subpage simulates the network channel. According to the assumption of the Delov-Yao attack, the attacker has the powerful ability to eavesdrop, replay and tamper with the messages in the network channel, and then launch all kinds of man-in-the-middle attacksAs shown in Figures 12, the red part of the transition and place simulates RP_ATK attacks, including transition REPLAY, port place send_AGRQ and related transformation rules. The arc expression marked in blue simulates the MD_ATK attack, including the functions: mattack (), attackseq1 () and attackseq2 () in the arc expression on transition CHANNEL A, CHANNEL C and CHANNEL C'. The transition guard of purple part tags simulates SP_ATK attacks, including transition CHANNEL A, CHANNEL B, CHANNEL C, CHANNEL C', CHANNEL D, CHANNEL AGM.

5.2 State Space Analysis of Attacker Model

The running environment of the attacker model: CPU is i5-6200u, main frequency is 2.3GHz, memory is 8GB. By the environment,the state space report (Figures 13) of three full attack modes of man-in-the-middle attack at the same time shows that the number of nodes in the state space is 63344, the number of connecting arcs is 911557, and the time to construct the state space is 3194 seconds.

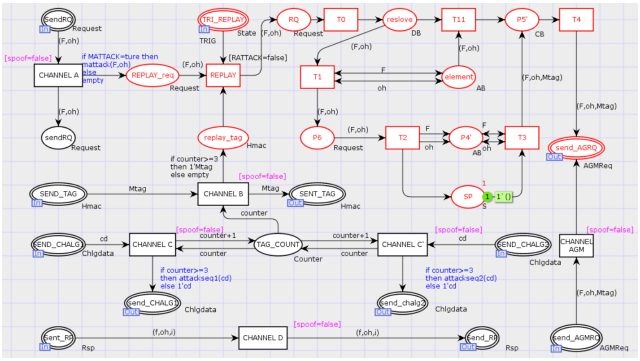


Figure 11: Top-level CPN model of the protocol

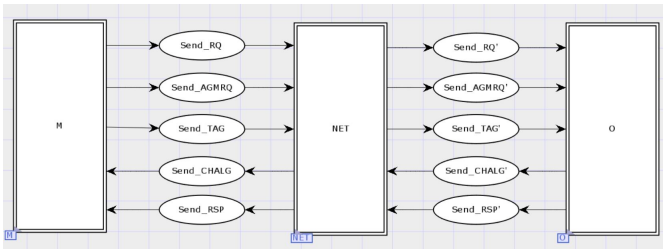


Figure 12: Attacker model in NET subpage

```
Statistics
------------------------------------------------------------------------

  State Space
     Nodes:  63344
     Arcs:   911557
     Secs:   3194
     Status: Full

  Scc Graph
     Nodes:  63344
     Arcs:   911557
     Secs:   1975

  Home Properties
------------------------------------------------------------------------

  Home Markings
     None
  Liveness Properties
------------------------------------------------------------------------

  Dead Markings
     60 [33344, 33343, 33342, 33341, 33340, ...]

  Dead Transition Instances
     GEN_AUTHERR, ERR_BACK, CHANNEL  A, T0, T11, T4, T1, T2, T3, CHANNEL  C,
     CHANNEL  AGM, CHANNEL  C', PROCESS, T_WRITE, T_READ, FEEDBACK, TT_CH,
     T_CH, SEC_NACR, CH1, CH2

  Live Transition Instances
     None
------------------------------------------------------------------------
```

Figure 13: State space report of protocols in full attack mode

The number of nodes and directed arcs obtained by the strongly connected graph is consistent with the results given by the state space, which shows that all state nodes of the protocol model are reachable and will not lead to infinite state occurrence and behavior iteration. There are 60 dead mark states in the model. There are 21 dead transitions in the model, and dead transitions refer to the transitions that can not be triggered in the model. Usually due to the defects in the design of the model, there are many dead transitions. However, due to the introduction of the attacker model, the 21 dead transitions in this model have a special meaning. The living transition in the model refers to the transition that can be triggered during the operation of the model. Because there is a dead mark state in the model established in this paper, that is, all the changes can not be triggered under the dead mark state, that is to say, the dead mark indicates the termination state of the completion of data transmission. therefore, the report shows that there is no living transition in the model established in this paper. This paper will conduct an in-depth analysis of the state space report to verify whether the protocol function is completed and whether the security is satisfied.

5.3 Behavior Consistency Analysis

This section verifies whether the behavior of the attacker model is consistent with that of the original protocol model when the attack is not enabled, The statistics of the state space reports of the protocol model in various modes are shown in Table 2.

The 0_ATK column only introduces the attacker model, but does not enable any attacks. RP_ATK is listed as the status space report obtained after the enable replay attack. MD_ATK is listed as a status space report obtained after enabling tampering attacks. SP_ATK is listed as a status space report obtained after enabling spoofing attacks. MRS_ATK is listed as the status space report obtained after enabling the above three attacks. In the attacker model, due to the introduction of the improved Delov-Yao attacker model, the number of state space nodes and arcs increases significantly compared with the original model, which is in line with expectations. And the number of arcs and nodes in state space is the same as that of strongly connected arcs and nodes, which shows that all state nodes in the attacker model are reachable, and there is no loop and iterative behavior that leads to the infinite occurrence of states, which further shows that the improved Delov-Yao attacker model is effective.

In 0_ATK mode, the number of dead nodes in the attacker model is the same as that in the original model, which indicates that all requests are successfully authenticated and executed by slave O, which is consistent with the expected behavior of the protocol without enabling any attack parameters. However, the number of dead transitions increases to 14, which indicates that many transitions in the model have not occurred. Furthermore, we use the ListDead Transitions () function to query the state space and find that the dead transition GEN_AUTHERR and ERR_BACK, are consistent with the original model, indicating that because the attack

mode is not turned on, there is no request message of authentication failure in the model, which is consistent with the expected behavior. The other 12 dead transitions reflect that the transition in the attacker model including the red part (replay attack), the blue part (tamper attack) and the purple part (spoofing attack) in Figure 12 failed to occur. Due to the lack of enabling attack parameters, these dead transitions are still consistent with the expected behavior of the protocol. The above analysis shows that when the attack parameters are not enabled, the attacker model does not change the behavior of the protocol, and the identity authentication function of the master station in the protocol specification can still be realized.

6 Attack Analysis and Vulnerability Mining

6.1 Attack Analysis

This section will enable all attacks, including RP_ATK (replay attack), MD_ATK (tamper attack), and SP_ATK (spoofing attack), to form the final attacker model MRS_ATK. For tampering attacks, it is necessary to set the initial token value of the library TAG_COUNT in the NET subpage model to 1, and the tamper attack flag MATTACK on the transition CHANNEL output arc to be "ture" to manipulate the challenge information received from the slave station O (Figure 12). The above settings enable the attacker to obtain enough challenge information issued by the slave station in NACR mode, so as to obtain the latest challenge information, which is used to forge the request information of the master station in AGM mode.

As shown in Table 2, MRS_ATK lists 60 dead nodes and 21 dead transitions, which indicates that unexpected behavior has occurred in the attacker model in full-state attack mode. Furthermore, the above 60 dead nodes are classified and analyzed by using SML query statements, including ListDeadMarkings(), SearchNodes(), Reachable $(M1, m2)$. The ListDeadMarkings() function is used to determine all the dead node sequence numbers of the model. The SearchNodes() function is used to determine whether a state contains an attack sequence. The Reachable() function is used to determine whether there is a path to the protocol security state, that is, the request initiated by the attacker is not authenticated by the slave O. S0-7 is used to classify the state points satisfied under different conditions. The number and meaning of states in each category are shown in Table 3.

In order to further analyze the authentication attributes of the protocol, the dead nodes are classified into expected and unexpected types. The expected dead node refers to the state node when the first two NACR requests are successful and the third AGM attack request fails. The unexpected dead node is the state node other than the expected dead node. The classification basis of the expected dead nodes is that the attacker model in this paper assumes that the AGM mode is used in the third communication, and the attacker expects to use the latest challenge information intercepted to launch three types of man-in-the-middle attacks in the third communication. According to the above description of unexpected dead nodes, it can be seen that the result of S4 query reflects the number of such dead nodes.

Further, the SML statement is used to query the status of the above dead nodes, and the process and results are as follows:

1) Using the above SML function AttackallInstaces query, the result shows that there are 36 unexpected dead nodes;

2) Use the SearchNodes statement to find all the states that contain the attacker request message, and get the status node 173;

3) The reachability from the node containing the attack request status to all unexpected dead nodes is verified by using the Reachable (M1-M2) statement.

The return value of the Reachable (M1-M2) statement is "ture". According to the definition of the authentication attribute of the protocol in section 3, the behavior of the attacker leads to the unexpected termination of the protocol, violating the authentication attribute of the protocol. Moreover, the attacker can still control the slave station to execute the key request message without obtaining the session key and without a valid message authentication code. Therefore, the protocol does not meet the authentication requirements claimed in IEC62351 and IEEE-1815-2012 specifications.

6.2 Attack Path Analysis

The attack results in the full attack state show that the attacker can send valid AGM messages to the slave station O without knowing the session key, and be executed by the slave station, including read and write registers, and even restart the application service. Considering the plaintext transmission of protocol messages, legitimate request and challenge information can be intercepted and tampered with, thus preventing the formation of legitimate message sequences. This results in the following attack threats:

1) The first kind of protocol loophole: in NACR mode, randomly changing the value of message sequence number Ksn and random number will make the protocol lose synchronization and cause unexpected authentication failure. By observing the challenge information of the slave station in continuous NACR mode, the attacker can find that whenever there is a new NACR request, the message sequence number Ksn will be increased by 1. Therefore, the attacker can modify the sequence number in the challenge information from the slave station O to form a new

Table 2: Comparison of the state space of each model

Types	Original model	Attack model				
		0_ATK	RP_ATK	MD_ATK	SP_ATK	MRS_ATK
State Space Nodes	11829	13327	12979	12674	12708	63344
State Space Arcs	129394	135602	177604	145866	124450	911557
SCC Graph Nodes	11829	13327	12979	12674	12708	63344
SCC Graph Arcs	129394	135602	177604	145866	124450	911557
Dead Markings	1	1	1	1	9	60
Dead Transitions	2	14	16	16	19	21

Table 3: SML functions mainly used in the classification of dead nodes

Item()	Description
S0(60)	The number of DeadMarking for the entire model
S1(60)	Contains the number of successful or failed DeadMarking
S2(0)	Contains the number of successful and failed DeadMarking
S3(24)	Number of DeadMarking which authentication failed
S4(36)	Number of DeadMarking which authentication successful
S5(16)	Number of DeadMarking caused by replay attack
S7(18)	Number of DeadMarking caused by tampering attack
S7(26)	Number of DeadMarking caused by spoofing attacks

challenge information and send it to the master station M, so as to induce the master station to send the wrong message authentication code, resulting in the failure of message authentication. Figure 14 shows the MSC model of this attack.

2) The second kind of protocol vulnerability: through the observation of a large number of plaintext challenge information, when an attacker finds challenge information with the same message sequence number Ksn and random number, he can replay the message authentication code tag_old intercepted in previous sessions to the slave O, thus causing the slave station to execute false commands. Prented_RSP is a false response message defined by an attacker based on a bogus request message. Figure 15 shows the MSC model of this attack.

3) The third kind of protocol vulnerability: by observing a large number of plaintext challenge messages, when an attacker finds challenge information with the

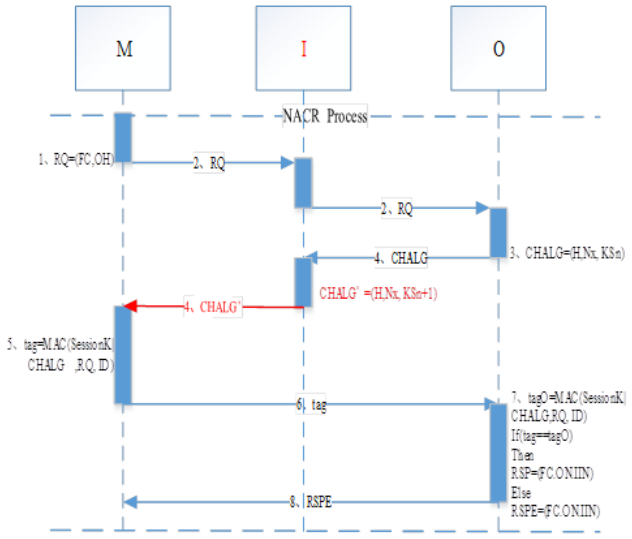


Figure 14: The first type of attack path

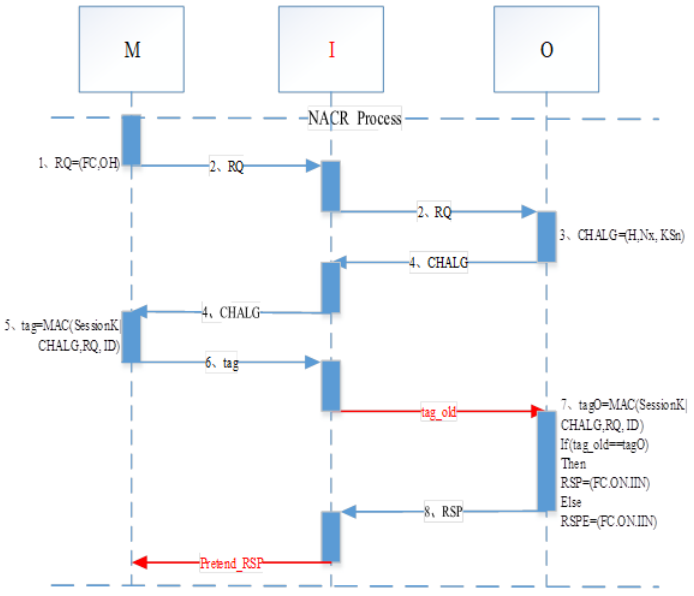


Figure 15: The second type of attack path

same message sequence number Ksn, he can replay the active request message Old_AGMRQ in AGM mode intercepted in previous sessions to the slave station, which causes the slave station to mistakenly assume that the master station will initiate an AGM session later, which in turn causes the slave station to perform unauthorized operations. Through the analysis of the state space of the model in the full attack state, it is concluded that in the attack mode, the transition T_WRITE in the O_EXE_RQ sub-page is triggered, and the slave station O still executes the false request command, resulting in 12 dead transitions in the model. Figure 16 shows the MSC model of this attack.

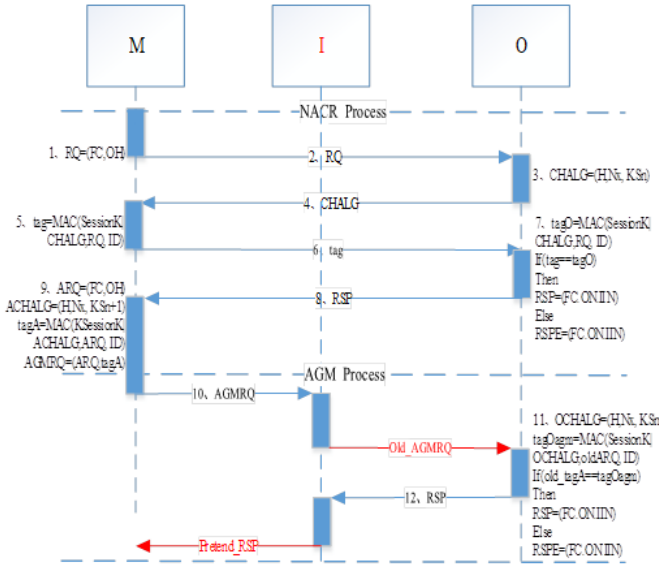


Figure 16: The third type of attack path

7 Conclusion

In This paper, we introduces an improved Delov-Yao attacker model to the DNP3-SA protocol, establishes a parameterized CPN attacker model of the protocol, and verifies the authentication attribute of the protocol and its ability to resist man-in-the-middle attacks. Firstly, the improved Dolov-Yao attacker model is studied, including message splitting and combination and parameterization. The advantage of the improved attacker model is verified by the application example of NS protocol. Secondly, the improved attacker model is applied to the NET subpage of the protocol, and three kinds of attacks such as replay, tampering and spoofing attack are introduced, the state space report of the attacker model is generated, and the functional consistency of the attacker model is analyzed. Finally, the vulnerability mining in the full attack state of the protocol is carried out, and the vulnerability exploitation path is given according to the results of SML query.

The running results of the protocol attacker model proposed in this paper show that the protocol actually does not meet the authentication requirements claimed in the IEC62351 and IEEE-1815-2012 specifications, and can not resist the three attacks listed in the specification: replay, tampering and spoofing.

Acknowledgments

This work is funded by the Shaanxi Provincial Department of Education Scientific Research Project (19JK0040), Science and Technology Project of Shaanxi Province (NO.2020GY-041), Doctoral research project of Baoji College of Arts and Sciences (209040080). The authors would like to thank the anonymous reviewers and the editors for their suggestions.

References

[1] R. Amoah, S. Camtepe, E. Foo, "Formal modelling and analysis of DNP3 secure authentication," *Journal of Network & Computer Applications*, no. 59, pp. 345-360, 2016.

[2] R. Amoah, S. Camtepe, E. Foo, "Securing DNP3 broadcast communications in SCADA systems," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 4, pp. 1474-1485, 2016.

[3] Y. Bai, *Research on Security Protocol Formal Method Based on Colored Petri Nets Model*, Inner Mongolia University, 2013.

[4] J. A. Crain, S. Bratus, "Bolt-on security extensions for industrial control system protocols: A case study of DNP3 SAv5," *Security & Privacy IEEE*, vol. 13, no. 3, pp. 74-79, 2015.

[5] C. Cremers, M. Dehnel-Wild, K. Milner, "Secure authentication in the grid: A formal analysis of DNP3: SAv5," in *Proceedings of European Symposium on Research in Computer Security*, pp. 389-407, 2017.

[6] I. Darwish, O. Igbe, T. Saadawi, "Vulnerability assessment and experimentation of smart grid DNP3," *Journal of Cyber Security*, vol. 5, no. 1, pp. 23-54, 2016.

[7] I. Darwish, T. Saadawi, "Attack detection and mitigation techniques in industrial control system - Smart grid DNP3," in *Proceedings of International Conference on Data Intelligence & Security*, 2018. DOI: 10.1109/ICDIS.2018.00028.

[8] D. Delov, A. Yao, "On the security of public key protocols," *IEEE Transactions on Information Theory*, vol. 29, no. 2, pp. 198-208, 1983.

[9] T. R. de Toledo, N. M. Torrisi, "Encrypted DNP3 traffic classification using supervised machine learning algorithms," *Machine Learning and Knowledge Extraction*, vol. 1, pp. 384-399, 2019.

[10] IEEE, "1815-2012 - IEEE Standard for Electric Power Systems Communications-Distributed Network Protocol (DNP3)," *IEEE*, 2012. DOI: 10.1109/IEEESTD.2012.6327578.

[11] C. Irvene, T. Shekari, D. Formby, and R. Beyah, "If I knew then what I know now: On reevaluating DNP3 security using power substation traffic," in *Proceedings of the Fifth Annual Industrial Control System Security Workshop (ICSS'19)*, pp. 48-59, 2019.

[12] J. Jiang, H. Mao, R. Shao, *et al.*, "Formal verification and improvement of the PKMv3 protocol using CSP," in *Proceedings of IEEE Computer Software & Applications Conference*, 2018. DOI: 10.1109/COMPSAC.2018.10318.

[13] O. S. Kidege, S. P. Maj, "Industrial network security – A critical review," *Modern Applied Science*, vol. 11, no. 6, pp. 24-32, 2017.

[14] S. Kwon, H. Yoo and T. Shon, "RNN-based anomaly detection in DNP3 transport layer," in *Proceedings of IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, 2019. DOI: 10.1109/SmartGridComm.2019.8909701.

[15] D. Lee, H, Kim, K, Kim, P. D. Yoo, "Simulated attack on DNP3 protocol in SCADA system," in *Proceedings of the 31th Symposium on Cryptography and Information Security*, 2014. (`https://caislab.kaist.ac.kr/publication/paper_files/2014/SCIS2014_DS.pdf`)

[16] F. Liu, M. Heiner, D. Gilbert, "Coloured Petri nets for multilevel, multiscale and multidimensional modelling of biological systems," *Briefings in Bioinformatics*, vol. 20, no. 3, pp. 877–886, 2019.

[17] Z. E. Mrabet, N. Kaabouch, H. E. Ghazi, H. E. Ghazi, "Cyber-security in smart grid: Survey and challenges," *Computers & Electrical Engineering*, vol. 67, pp. 469-482, 2018.

[18] T. C. Pramod, N. R. Sunitha, "SCADA: Analysis of attacks on communication protocols," in *Proceedings of International Symposium on Sensor Networks, Systems and Security*, pp. 219-234, 2018.

[19] M. A. Rahman and A. Datta, "Impact of stealthy attacks on optimal power flow: A simulink-driven formal analysis," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 3, pp. 451-464, 2020.

[20] N. R. Rodofile, K. Radke, E. Foo, "Real-time and interactive attacks on DNP3 critical infrastructure using scapy," in *Proceedings of the 13th Australasian Information Security Conference*, 2015. (`https://eprints.qut.edu.au/81587/21/Vol161_AISC2015_paper09.pdf`)

[21] C. Rosborough, "Colin gordon and brian waldron, all about eve: Comparing DNP3 secure authentication with standard security technologies for SCADA communications," in *Proceedings of Power and Energy Automation Conference Spokane*, 2019. (`https://ccaps.umn.edu/documents/CPE-Conferences/MIPSYCON-Papers/2019/AllAboutEve.pdf`)

[22] A. Shahzad, M. Lee, S. Kim, K. Kim, J. Y. Choi, Y. Cho, K. K. Lee, "Design and development of layered security: Future enhancements and directions in transmission," *Sensors*, vol. 16, no. 1, pp. 37-52, 2016.

[23] A. Volkova, M. Niedermeier, R. Basmadjian and H. de Meer, "Security challenges in control network protocols: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 619-639, 2019.

[24] M. Xiao, W. Li, X. Zhong, K. Yang And J. Chen, "Formal analysis and improvement on ultralightweight mutual authentication protocols of RFID," *Chinese Journal of Electronics*, vol. 28, no. 5, pp. 1025-1032, 2019.

[25] X. C. Yin, Z. G. Liu, L. Nkenyereye, B. Ndibanje, "Toward an applied cyber security solution in IoT-based smart grids: An intrusion detection system approach," *Sensors*, vol. 19, no. 22, pp. 4952, 2019.

[26] J. Zhang, L. Yang, W. Cao and Q. Wang, "Formal analysis of 5G EAP-TLS authentication protocol using proverif," *IEEE Access*, vol. 8, pp. 23674-23688, 2020.

Biography

Dr. Ye Lu was born in Shannxi, China in 1986. He received his Bachelor's degree from The Beijing Institute of Technology, China in 2008 and Ph.D degree degree in Lanzhou University of Technology, Lanzhou City, China. He is currently a lecturer at Baoji College of Arts and Sciences, Baoji, China. His research interests are in the areas of Blockchain, Internet of things and protocol security.

Wei-Bin Ou was born in Shannxi, China in 1977. He received his Bachelor's degree and Ph.D degree from Xi'an Technological University, Xi'an City, China. He is currently a lecturer at Baoji College of Arts and Sciences, Baoji, China. His research interests are in the areas of Industrial networked control system and industrial fieldbus.

Identification and Detection of Network Intrusion Data Using the Deep Learning Method

Yonghe Zeng
(Corresponding author: Yonghe Zeng)
Zhangjiajie Institute of Aeronautical Engineering
Zhangjiajie, Hunan 427000, China
Email: yongchici479907@126.com
(Received Feb. 14, 2019; Revised and Accepted Jan. 16, 2021; First Online July 11, 2021)

Abstract

This study mainly analyzed the application of the deep learning method in the identification and detection of network intrusion data. Firstly, the deep learning method was briefly introduced. Then, the automatic encoder (AE) was mainly analyzed, and the adversarial autoencoders (AAE) combined with generative adversarial networks (GAN) were introduced. Finally, an experiment was carried out on the UNSW-NB15 data set. The results showed that the accuracy of AAE was higher than 99% in the binary classification of normal and intrusion data; in the multi-classification of intrusion data, for some types with few data, AAE had a poor performance, but the overall accuracy was above 85%; compared with other methods such as Principal Component Analysis (PCA) and convolutional neural network (CNN), AAE shows the best performance. The results verify the reliability of AAE in intrusion data identification and detection, and AAE can be further applied in network security.

Keywords: Automatic Encoder; Deep Learning; Generative Adversarial Networks; Network Intrusion

1 Introduction

With the further development of the network, the number of network users and the use time of the network are also further expanded, which has brought great changes to people's life, study, and work. But simultaneously, the network environment is more complex, and the amount of data generated further increases [29]. A large number of network attacks have brought great threats to network security [4, 20]. Some common security technologies, such as firewall [23], anti-virus softwar [16], *etc.*, are passive strategies. Facing the constantly updated intrusion means, there is a lack of active response measures. With the development of big data and artificial intelligence [1, 12, 27], many new methods have been successfully applied in the identification and detection of intrusion data [6].

Based on the game theory, Subba *et al.* [24] designed an intrusion detection framework, which combined the Shapley value mechanism and Vickery-Clark-Grooves (VCG) mechanism. Through simulation experiments, they found that the framework had a high detection rate for the intrusion. Vahid *et al.* [26] designed a method combining K-means clustering with multiple classifiers (KCMC). Through experiments, they found that the hybrid method had a better efficiency than the single classifier and the detection rate reached 99.5%.

Ali *et al.* [2] proposed a fast learning network (FLN) method based on particle swarm optimization (PSO), named the PSO-FLN model, and found through an experiment on the KDD99 data set that the model had an excellent performance. Chiba *et al.* [8] detected known attacks with the signature-based detection and detected network anomalies with the back-propagation neural network (BPN) and found through an experiment that the method could reduce the computational cost and improve the detection rate [11, 25]. This paper mainly analyzed the automatic encoder (AE) and its applicability in intrusion detection of intrusion data. This study makes some contributions to enrich network security technology and further improve the efficiency of intrusion detection.

2 Identification and Detection of Network Intrusion Data

There are mainly two methods for the identification and detection of network intrusion data. One is anomaly detection [28]. It is assumed that the network attack behavior is abnormal. Then, the normal behavior profile of the user is established. When identifying and detecting, the current user behavior is compared with the normal behavior profile to judge whether it is normal behavior. This method needs to simplify the feature quantity as much as possible and select the appropriate reference threshold. It can recognize and detect intrusion data in time, but it needs a large amount of computation to calculate

behavior profiles [9].

The other is misuse detection [22]. It digitally signs the abnormal intrusion data through the monitoring of network data and compares it with the digital signature of user behavior to determine whether it is intrusion data. This method relies on the digital signature, and can only identify and detect known attacks, which may cause a missing report to unknown intrusion. Overall, in order to improve the performance of misuse detection, anomaly detection or other technologies need to be added on its basis.

3 Recognition and Detection Method Based on Automatic Encoder

3.1 Deep Learning Method

In the identification and detection of intrusion data, data feature extraction is always a very important problem. Deep learning can select the optimal features from a large number of unordered data by establishing a learning model. It has also been widely used in intrusion data detection [15]. The common methods are as follows.

1) Multi-layer perceptron (MLP) [10]: It is an artificial neural network structure, which can solve the classification problem. It has a simple principle, but the disadvantages are also obvious. It needs a long training time and is easy to fall into the local extremum. In the case of a large network structure, its training effect is poor.

2) Convolutional neural network (CNN) [7]: CNN extracts high-level features through the convolution layer and selects and filters features through the pooling layer to reduce the amount of data processing. It has a good generalization ability and has been widely used in pattern recognition, object detection, *etc.*

3) Recurrent neural network (RNN) [21]: RNN can learn the characteristics of the packet and network flow, respectively, to get a flow feature vector and recognize and detect the data through softmax or other classifiers.

4) Automatic encoder (AE) [13]: AE is a kind of neural network, which is an unsupervised algorithm. It can learn the feature distribution of data through encoder and decoder. By adding different constraints, different AE can be established to obtain different types of features of data.

3.2 Adversarial Autoencoders (AAE)

AE is composed of an encoder and a decoder, which encodes to reduce data dimension and decodes to reconstruct data. For the original high-dimensional data, the reconstruction error between output and input is reduced by adjusting the encoding and decoding parameters. Overall, the input layer and hidden layer of the network are the decoding part, and the hidden layer and output layer are the decoding part. The specific operation process is as follows.

Encoding. For input x, the result generated by the encoder is:

$$\begin{aligned} h &= f(x) \\ &= S_1(W_1 x + b_1), \end{aligned}$$

where S_1, W_1, and b_1 are the activation function, weight, and bias between the input and hidden layers and S_1 is the sigmoid function.

Decoding. The result of decoder output is:

$$g(h) = S_2(W_2 h + b_2),$$

where S_2, W_2, and b_2 are the activation function, weight, and bias between the hidden and output layers and S_2 is the sigmoid function.

Reconstruction Error. $J_{AE} = \sum_{x \in l} L(x, g(f(x)))$, where L refers to the reconstruction error function. The cross-entropy loss function used in this study is:

$$L(x, y) = -\sum_{i=1}^{n} x_i \log y_i + (1 - x_i) \log(1 - y_i).$$

Generative adversarial network (GAN) is derived from zero-sum game [17], which is composed of generator (G) and discriminator (D). For sample $z \in p_z$, the prior distribution p_z is mapped to real distribution $x \in p_{data}$. G outputs samples $x' \in p_G$. In order to make the difference between p_{data} and p_G the smallest, the mapping relationship is learned through adversarial training. In discriminator D, x is classified as a real sample, and x' is classified as a false sample. G attempts to cheat the D network to make D classify the output of G as the real sample. In the process of back propagation, D can find the real features. This adversarial learning process can be written as:

$$\begin{aligned} \min_G \max_D V(D, G) &= E_{x \sim p_{data}(x)}[\log D(x)] \\ &\quad + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \end{aligned}$$

When training D, the goal is to maximize its discrimination ability. When training G, the goal is to generate the data closest to the real data; therefore, the generative network G can be written as:

$$G' = arg \min_G \max_D V(D, G),$$

and the discriminative network D can be written as:

$$D' = arg \max_D V(D, G).$$

AAE is obtained after combining GAN with AE. In the AE part, if the real sample is x, after decoding by $q(z|x)$, the decoding vector z is obtained. New data are obtained after decoding with the decoder. In the GAN part, the D network judges the input coding vector. At that moment, G cheats the D network. If the judgment of the D network is successfully prevented, the output coding is closer to the predefined distribution function. AAE is applied to the detection and recognition of intrusion data. For the input training samples, G and D are initialized, and the parameters are constantly updated in the training process. Through the cheating and game of the two models, the effective feature extraction is realized. Finally, the classification of data is realized using the softmax classifier.

4 Experimental Analysis

4.1 Experimental Data Set and Preprocessing

An experiment was carried out on the UNSW-NB15 data set. The data set includes one kind of normal data and nine kinds of attack data, namely Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. The experiment was carried out on one subset, as shown in Table 1.

Table 1: Experimental data set

Data Type	Training Data/n	Test Data/n
Normal	35983	53122
Fuzzers	4885	16852
Analysis	69	636
Backdoors	83	443
DoS	1452	3399
Exploits	8281	21595
Generic	18830	39754
Reconnaissance	3217	8874
Shellcode	378	1133
Worms	44	130

In the dataset, there are 49-dimensional features [18], as shown in Table 2.

When mapping the character data, for example, for Label items, Normal is labeled as 0, and the other intrusion types as 1. Then, the features are normalized and mapped to the interval of (0,1). The formula is:

$$y = \frac{(x - x_{min})}{x_{max} - x_{min}}$$

where y is the normalized value, x is the original value, and x_{max} and x_{min} are the maximum value and the minimum value, respectively.

4.2 Evaluation Index

In the problem of data classification, the results can be divided into the following four types:

True Positive (TP): Intrusion data are identified as an intrusion;

True Negative (TN): Normal data are identified as normal;

False Positive (FP): Normal data are identified as an intrusion;

False Negative (FN): Intrusion data are identified as normal.

The evaluation indexes of intrusion detection results include:

Accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$;

Precision Rate: $P = \frac{TP}{TP+FP}$;

Recall Rate: $R = \frac{TP}{TP+FN}$;

F_{-1} **(F-score):** $F_{-1} = \frac{2PR}{P+R}$

4.3 Experimental Results

Firstly, binary classification was performed on the intrusion data using the AAE method, i.e.,, only the normal data and intrusion data were identified and detected, and the results are shown in Table 3.

It was seen from Table 3 that the AAE method showed an excellent performance in the binary classification, and the evaluation index was more than 99%, which indicated that the AAE method could distinguish normal data and intrusion data well, i.e., it had a good identification and detection ability.

The performance of AAE in multi-classification was compared, i.e., nine kinds of data in the dataset were identified and detected, and the results are shown in Figure 1.

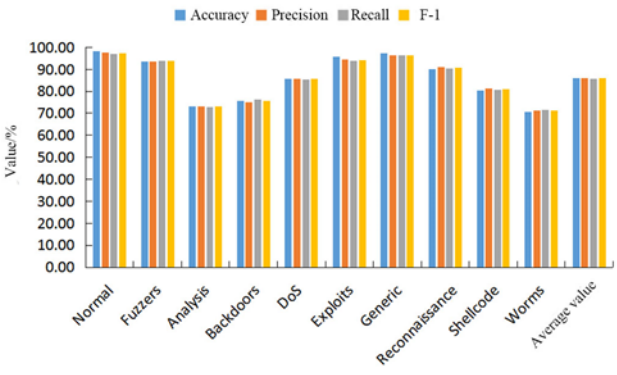


Figure 1: The identification and detection results of AAE

Table 2: Features of UNSW-NB15

Type	No.	Name	Type	No.	Name
Flow Features	1	srcip	Time Features	27	sjit
	2	sport		28	djit
	3	dstip		29	stime
	4	dsport		30	ltime
	5	proto		31	sintpkt
Base Features	6	state		32	dintpkt
	7	dur		33	tcprtt
	8	sbytes		34	synack
	9	dbytes		35	ackdat
	10	sttl	General purpose features	36	is_sm_ips_ports
	11	dttl		37	ct_state_ttl
	12	sloss		38	ct_flw_http_mthd
	13	dloss		39	is_ftp_login
	14	service		40	ct_ftp_cmd
	15	sload	Connection features	41	ct_srv_src
	16	dload		42	ct_srv_dst
	17	spkts		43	ct_dst_ltm
	18	dpkts		44	ct_src_ltm
Content Features	19	swin		45	ct_src_dport_ltm
	20	dwin		46	ct_dst_sport_ltm
	21	stcpb		47	ct_dst_src_ltm
	22	dtcpb	Labeled Features	48	attack_cat
	23	smeansz		49	Label
	24	dmeansz			
	25	trans_depth			
	26	res_bdy_len			

Table 3: Identification and detection results of binary classification/%

	Normal data	Intrusion data	Average value
A	99.87	99.59	99.73
P	99.46	99.34	99.40
R	99.21	99.56	99.39
F-1	99.33	99.45	99.39

It was seen from Figure 1 that there were slight differences in the performance of the algorithm in identifying different types of intrusion. First, the classification performance of the algorithm on Normal was the best, with all the indexes above 95%, followed by Exploits and Generic. The classification performance of the algorithm on Analysis, Backdoors, and Worms was poor, with all the indexes below 80%. It was found from Table 1 that the amount of Analysis, Backdoors, and Worms data was small, which might lead to the incomplete learning of AAE for features; therefore, the classification performance was poor. From the perspective of the average value, in the multi-classification, the indexes of the AAE algorithm were about 85%. In order to further verify the effectiveness of the proposed method, it was compared with other feature selection methods, such as PCA, deep neural network (DNN), and CNN, and the same classifier softmax was used. The results are shown in Figure 2.

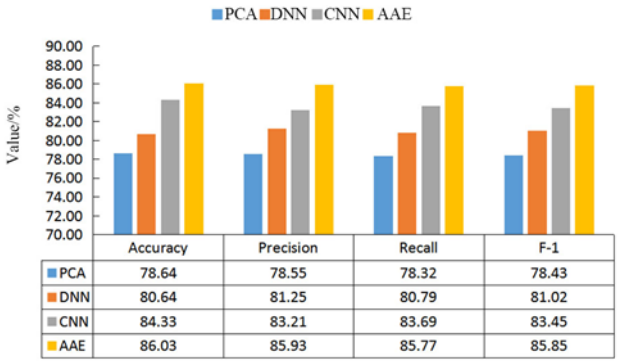


Figure 2: Comparison between AAE and other algorithms

It was seen from Figure 2 that the AAE method showed good performance compared with other algorithms. First of all, the accuracy of AAE was 86.03%, 7.39% higher than PCA, 5.39% higher than DNN, and 1.7% higher than CNN; the precision rate of AAE was 85.93%, 7.38% higher than PCA, 4.68% higher than DNN, and 2.72% higher than CNN. The recall rate of AAE was 85.77%, 7.45% higher than PCA, 4.98% higher than DNN, and 2.08%

higher than CNN. The F-1 value of AAE was 85.85%, 7.42% higher than PCA, 4.83% higher than DNN, and 2.4% higher than CNN. It was found that the performance of the AAE method was the best in identifying and detecting intrusion data.

5 Discussion

The identification and detection of network intrusion data [30] is the classification of normal data and intrusion data. At present, the commonly used methods include support vector machine (SVM) [14], k-nearest neighbor [3], logistic algorithm [5], naive Bayes algorithm [19], *etc.* However, with the emergence of more and more complex intrusion data, identification and detection technology is facing great challenges. The update of attack types and the increase of attack complexity make many traditional methods unable to maintain high performance. Also, manual selection of features not only costs a lot of workforce and time but also has low flexibility in new attacks. With the successful application of the deep learning method in speech recognition, image processing, *etc.*, its application in network intrusion data recognition and detection has attracted more and more attention.

This paper mainly analyzed the AAE algorithm in the deep learning method and its effectiveness in intrusion data identification and detection. It was found from the results that AAE showed a strong performance in the binary classification of intrusion and normal data, and all the indexes were more than 99%, which indicated that it could distinguish normal data and intrusion data well. Then, in the multi-classification, it was seen from Figure 1 that the performance of the algorithm had a relationship with the amount of data in model training. For the data types with large amounts of data, the classification performance of AAE was good; however, for the data types with small amounts of data, such as Analysis and Backdoors, the classification performance of AAE was poor. The reason for the above results was that there were no enough features for learning in model training. It was also found from the comparison with other algorithms that the accuracy and precision rate of AAE were higher than other algorithms, which indicated that the AAE algorithm had good reliability in identifying and detecting intrusion data. Although this paper obtained some results from the identification and detection of intrusion data, there are some shortcomings. In future works, we should

1) Study more deep learning methods;

2) Carry out experiments on more data sets;

3) Further optimize the performance of the AAE algorithm.

6 Conclusion

For the identification and detection of network intrusion data, this paper mainly analyzed the application of the deep learning method, the AAE algorithm, and carried out experiments on the UNSW-NB15 data set. The results showed that:

1) The indexes of AAE were all above 99% in the identification and detection of normal and intrusion data;

2) In the multi-classification of intrusion data, the identification and detection effect was related to the amount of data;

3) Compared with methods such as PCA and DNN, AAE had advantages in accuracy and precision rate.

The experimental results verify the reliability of the AAE method in intrusion data identification and detection. The AAE method can be further promoted and applied in practice.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61802243), the Key R&D program in industry field of Shaanxi Province (2019GY-013) and the basic science research program of Shaanxi Province (2019JQ273,2020JM288), and Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) (SKLNST-2020-1-03). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] D. S. Abdul Minaam and E. Amer, "Survey on machine learning techniques: Concepts and algorithms," *International Journal of Electronics and Information Engineering*, vol. 10, no. 1, pp. 34–44, 2019.

[2] M. H. Ali, B. A. D. A. Mohammed, M. A. B. Ismail, M. F. Zolkipli, "A new intrusion detection system based on fast learning network and particle swarm optimization," *IEEE Access*, vol. PP, no. 99, pp. 1-1, 2018.

[3] Y. Y. Aung, M. M. Min, "An analysis of random forest algorithm based network intrusion detection system," in *18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD'17)*, Kanazawa, pp. 127-132, 2017.

[4] Y. R. Bachupally, X. Yuan, K. Roy, "Network security analysis using big data technology," in *IEEE SoutheastCon*, Norfolk, VA, pp. 1-4, 2016.

[5] E. Besharati, M. Naderan, E. Namjoo, "LR-HIDS: Logistic regression host-based intrusion detection system for cloud environments," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 9, pp. 3669-3692, 2019.

[6] A. L. Buczak, E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2016.

[7] Y. H. Chen, T. Krishna, J. S. Emer, V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127-138, 2017.

[8] Z. Chiba, N. Abghour, K. Moussaid, A. E. Omri, M. Rida, "A cooperative and hybrid network intrusion detection framework in cloud computing based on snort and optimized back propagation neural network," *Procedia Computer Science*, vol. 83, pp. 1200-1206, 2016.

[9] A. Dewanje and K. A. Kumar, "A new malware detection model using emerging machine learning algorithms," *International Journal of Electronics and Information Engineering*, vol. 13, no. 1, pp. 24–32, 2021.

[10] G. Dudek, "Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1057-1060, 2016.

[11] E. N. Ganesh, "Study of VoIP network delay using neural networks," *International Journal of Electronics and Information Engineering*, vol. 12, no. 2, pp. 83–91, 2020.

[12] 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.

[13] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659-5670, 2015.

[14] S. T. Ikram, A. K. Cherukuri, "Improving accuracy of intrusion detection model using PCA and optimized SVM," *Journal of Computing and Information Technology*, vol. 24, no. 2, pp. 133-148, 2016.

[15] M. J. Kang, J. W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *Plos One*, vol. 11, no. 6, pp. e0155781, 2016.

[16] D. W. Kim, P. Yan, J. Zhang, "Detecting fake antivirus software distribution webpages," *Computers & Security*, vol. 49, pp. 95-106, 2015.

[17] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, S. P. Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision (ICCV'17)*, Venice, pp. 2813-2821, 2017.

[18] N. Moustafa, J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS'15)*, pp. 1-6, 2015.

[19] Z. Muda, W. Yassin, M. N. Sulaiman, N. Udzir, "K-Means clustering and naive Bayes classification for intrusion detection," *Journal of IT in Asia*, vol. 4, no. 1, pp. 13-25, 2016.

[20] E. U. Opara and O. J. Dieli, "Enterprise cyber security challenges to medium and large firms: An analysis," *International Journal of Electronics and Information Engineering*, vol. 13, no. 2, pp. 77–85, 2021.

[21] F. J. Ordóñez, D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 115, 2016.

[22] T. J. Parvat, P. Chandra, "A novel approach to deep packet inspection for intrusion detection," *Procedia Computer Science*, vol. 45, pp. 506-513, 2015.

[23] X. Song, "Firewall technology in computer network security in 5G environment," *Journal of Physics Conference Series*, vol. 1544, pp. 012090, 2020.

[24] B. Subba, S. Biswas, S. Karmakar, "A game theory based multi layered intrusion detection framework for wireless sensor networks," *International Journal of Wireless Information Networks*, vol. 25, no. 4, pp. 1-23, 2018.

[25] 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.

[26] S. Vahid, M. Ahmadzadeh, "KCMC: A hybrid learning approach for network intrusion detection using K-means clustering and multiple classifiers," *International Journal of Computer Applications*, vol. 124, no. 9, pp. 18-23, 2015.

[27] G. Yadav, S. Dalal, "Improvisation of network security using encryption technique for big data technology," *International Journal of Computer Applications*, vol. 124, no. 11, pp. 27-30, 2015.

[28] X. Zhai, K. Appiah, S. Ehsan, G. Howells, H. Hu, D. Gu, K. McDonald-Maier, "A method for detecting abnormal program behavior on embedded devices," *IEEE Transactions on Information Forensics & Security*, vol. 10, no. 8, pp. 1692-1704, 2015.

[29] C. Zhang, X. Shen, X. Pei, Y. Yao, "Applying big data analytics into network security: Challenges, techniques and outlooks," in *IEEE International Conference on Smart Cloud (SmartCloud'16)*, New York, NY, pp. 325-329, 2016.

[30] R. Zuech, T. M. Khoshgoftaar, R. Wald, "Intrusion detection and big heterogeneous data: A survey," *Journal of Big Data*, vol. 2, no. 1, pp. 3, 2015.

Biography

Yonghe Zeng, born in 1966, was major in computer and its application in the department of electronic engi-

neering in Shenyang Institute of Aeronautical Engineering (Shenyang Aerospace University now) and graduated in 1989. Since July 1989, he has been engaged in teaching and management work at Zhangjiajie Institute of Aeronautical Engineering. He was appointed as the director of the computer teaching and research section in May 1994. Since 1998, he has served successively as the director of electrical engineering department, the director of electrical engineering department (department-level), the director of computer and automatic control department, and the director of information engineering department. He is now the general Party branch secretary of the department of information engineering. He was promoted to associate professor in September 2004. He has been engaged in the teaching and scientific research of electrician, electronics, automation technology, computer and other related majors for a long time.

Efficient Identity-based Proxy Re-encryption Scheme in Blockchain-assisted Decentralized Storage System

Jiayu He[1,2], Dong Zheng[1,2,4], Rui Guo[1,2,3], Yushuang Chen[1,2], Kemeng Li[1,2], and Xiaoling Tao[4]

(Corresponding author: Jiayu He)

School of Cyberspace Security, Xi'an University of Posts and Telecommunications[1]

National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications, China[2]

Guilin University of Electronic Technology, Guilin, China[3]

Westone Cryptologic Research Center, Beijing, China[4]

Email: hjy_xupt@163.com

(Received Apr. 20, 2020; Revised and Accepted Dec. 10, 2020; First Online Aug. 14, 2021)

Abstract

The rapid development of cloud storage has greatly promoted the industrial productivity and social progress. However, with the era of big data coming, there exists several challenges for cloud storage in terms of the difficulty of maintaining data security, the efficiency of data sharing and the cost of the single point of failure. Proxy re-encryption (PRE) cryptographic primitive is regarded as a promising technique to improve the efficiency and security of data. Blockchain can overcome the flaws caused by single point of failure. These two technologies have attracted much attention recently. Nevertheless, the conventional PRE scheme incurs intricate certificate management and the blockchain is not suitable for storing big data because of the high cost. In order to figure out the above problems, in this paper, we propose a novel identity-based proxy re-encryption (IBPRE) scheme,data owner-manipulative IBPRE (DOM-IBPRE), which is achieved by combining IBPRE, blockchain and the inter planetary file system (IPFS) technology. The scheme can avoid the complexity of certificate management, enhance the security of big data storage and ameliorate the efficiency of big data sharing. In addition, we assess the security performance of the scheme by Chosen-Plaintext Attack (CPA) in the standard model and simulate our scheme with Pairing-Based Cryptography (PBC) library, by comparing with other PRE schemes, our scheme has the better performance due to the reduction of the computation times of exponential and bilinear pairings. Finally, we implement the blockchain scheme under Linux system with the hyperledger test network fabric and develop the interface to client, the test results show that our scheme is practical.

Keywords: Blockchain; Decentralized Storage System; Hyperledger Fabric; Inter Planetary File System; Proxy Re-encryption

1 Introduction

With the emergency advancement of information diversification, cloud storage is becoming increasingly popular in our daily life. Cloud storage enables enterprises or individuals to obtain cloud data at anytime and anywhere, this feature brings wonderful convenience to our lives [31]. However, there are also some challenges in existing cloud storage system. Above all, shared data is required to be encrypted to guarantee data privacy, with the era of big data coming, the difficulty of maintaining data security is getting increasingly and the efficiency of data sharing is getting reducing. In addition, many cloud storage systems are operated by a centralized corporation which has powerful ability to store and supervise data, the corporation can be seen as a third party, it ineluctable inherits the single point of failure flaws. Last but not least, with the upgrade of cloud storage devices and the rise of employee wages, the cost of centralized cloud storage is increasingly higher. Therefore, to better guarantee data privacy and availability, we should achieve a flexible access control over the encrypted data and change data storage from centralized systems to the decentralized systems [18], which not only have the cheaper price than existing centralized storage systems, but also can relieve of our concern about a single point of failure. Fortunately, the emergence of bitcoin has driven the development of blockchain technology, a blockchain can be thought of a decentralized ledger, where the data and transactions are not under the control of any third party, the data that stored on the blockchain cannot be tampered. Therefore, blockchain has strong data stability and reliability.

Inter planetary file system (IPFS) and hyperledger fab-

ric as typical distributed storage applications, they use blockchain as their core structure and attract more attentions in recent years. For the purpose of promoting efficiency and security of data encryption in cloud storage, proxy re-encryption (PRE) scheme was proposed [15], it can provide a flexible and feasible method for storing and sharing data. Nevertheless, the original PRE was proposed in the public key model which incurs intricate certificate supervision. To ease this problem, identity-based PRE (IBPRE) scheme was proposed [10], take advantage of this scheme, the identity of receivers can be seen as public keys, the process of verifying certificates is replaced by knowing the identities of users, which is more convenient in application. However, in practical applications, the master key in system is generated by single public key infrastructure (PKG), if the authenticity of the PKG cannot be trusted or the PKG is attacked maliciously, the master key may be leakage. From what we have discussed above, how to combine with IBPRE, IPFS and blockchain technology to design a flexible and efficient identity-based proxy re-encryption scheme for decentralized cloud storage is a worthy study. The specific contributions of this paper are as follows:

1) We design a frame which unites the decentralized storage system IPFS, the hyperledger fabric and IBPRE technology to achieve secure and efficient control data in decentralized system. The data owner is the sole one who can control their data, besides, a single PKG is replaced by multi trusted authority. The problem that the malicious attack on the PKG results in the leak of the private key of each user in conventional IBPRE scheme is solved.

2) We also analysis the security performance of the scheme by Chosen-Plaintext Attack (CPA) based on a modified decisional bilinear Diffie-Hellman (mDBDH) problem in the standard model. The theoretical analysis indicates that our scheme is correct with security.

3) For the purpose of mitigating the storage burden of blockchain, we only store the index of sharing files in the blockchain system. By designing the smart contract, the data user who has access rights can obtain the index of file and decrypt the file.

4) We simulate our scheme under the Linux system, the simulation of the data owner-manipulative IBPRE (DOMIBPRE) algorithm is ran based on the pairing-based cryptography library (PBC) [14], and the performance was analyzed. The implementation of the blockchain framework is performed through the hyperledger test network fabric, we design the smart contract and store the index of files to test our blockchain network, the test results prove our scheme is practical.

The remainder of paper is arranged as follows, Section 2 introduces of related work, Section 3 reviews the preliminaries, Section 4 details the system model, including scheme model and security model. The system and smart contract description are shown in Section 5. Section 6 discusses the security analysis. The performance analysis is shown in Section 7. The implementation of blockchain framework is described in Section 8. Finally, Section 9 concludes this paper.

2 Related Work

2.1 Blockchain Technology

Blockchain is a decentralized storage database, it is a novel application of computer technology such as cryptography, consensus mechanism, point-to-point transmission, smart contract and other technologies [20]. It records all transaction information which occurring on the node. All processes are highly transparent. These days, electronic cryptocurrency (such as Bitcoin [19], Ethereum [7], Zcash [5], *etc.*) have grown popular, the emergence of these cryptocurrency have promoted the development of blockchain. Taking bitcoin as an example, the transaction processes in blockchain are shown in Figure 1.

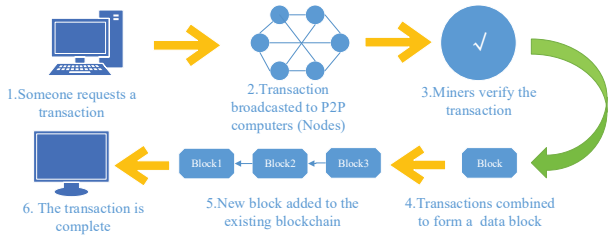


Figure 1: Transaction process of blockchain

The blockchain technology as the supporting technology of cryptocurrency has been attached more attention. With the characteristics of decentralization, transparent, anonymous traceability and non-tampering of information, it has been playing an indispensable role in many territories, such as: decentralized supply chain [1], decentralized identity-based PKI [16], decentralized IoT [17], decentralized scheme [13], decentralized storage [3], *etc.* Many researchers in cloud storage field focus on guaranteeing security of data, an individual data control system based on blockchain [34] has been proposed, this system can better guarantee the data privacy. To address the problem of the difficulty of maintaining data security and the efficiency of data sharing that impede the progress of big data, a blockchain based access control frame for reinforcing the privacy of big data platforms [27, 28] was presented. However, these schemes do not give the data owner the power to control the data efficiently and flexibly. For the purpose of achieving flexible and fine-grained data access, a scheme for cloud computing based attribute-based encryption (ABE) was proposed [22], although the scheme achieved flexible data

access, it is not effectively combined with the blockchain technology, moreover, the computation cost is large in this scheme.

Decentralized storage systems [25, 29, 33] do not rely on centralized service providers, it allows users to upload files in nodes which lease free storage space on the internet. These systems utilize blockchain as their primary structure. As a distributed storage terrace, IPFS [33] uses Filecoin as an incentive mechanism to encourage nodes to contribute retrieval service. After research, it was found that IPFS does not offer a doughty privacy encryption algorithm port for user-uploaded data. In order to solve this problem, the Storj system offers an end-to-end encryption way [29]. This system stores hash value of data on blockchain while offering a means of validating data integrity. Blockchain technology and peer to peer storage network were also employed in Sia platform [25], this platform unpicks the uploaded data into multiple data parts, and encrypts each part of data. Encrypted data are sent through smart contracts to each node that provides the storage service, the users pay siacoin for the storage service, as the storage node, they need termly submit proof of the stored data to prevent the storage node from removing the stored data. Nevertheless, in these systems, the process of data encryption is not fully controlled by the data owner.

2.2 Identity-based Proxy Re-encryption Technology

The first PRE scheme was proposed by Blaze, Bleumer and Strauss in [10], following this founding work, a series of PRE schemes were proposed in classical public key setting, such [9,12,26], these schemes require a certificate to validate the public key before encrypting a plaintext.

For the purpose of avoiding the overhead to notarize public keys certificates, Green and Ateniese [8] proposed the first IBPRE scheme, it avoids the complexity of the certificate management, after that, a number of IPRE schemes [2, 24] have been presented based the thought of identity-based encryption [23].

The above PRE schemes only allow data sharing in a rough level. In other words, the schemes could not control the process of data re-encryption. This issue is addressed in the recent conditional proxy re-encryption (CPRE) schemes [11, 32], they achieve fine-grained data sharing. In order to further improve the CPRE scheme, the conditional identity-based broadcast PRE (CIBPRE) scheme [30] was proposed, it combines the thoughts of CPRE, IPRE and broadcast encryption. In CIBPRE system, a sender can encrypt the data with the identity of receiver and data sharing conditions, the primal authorized receivers can obtain the data with their private keys, the new authorized receiver can also access the data by decrypting the ciphertext with their private keys. It avoids repeated downloading and encryption of data by the sender. These merits make CIBPRE get a practical and secure tool to store data, particularly when there are diverse receivers to share the data. The recent work in [4] proposed a more efficient conditional identity based broadcast proxy re-encryption(CIBPRE) scheme that supports diverse receivers to share the data as time passes. However, these schemes are not used decentralized storage system.

3 Preliminaries

In this part, we retrospect few of the notations and correlative background knowledge that will be needed in our paper. Table 1 shows a few of the notations deployed in the paper.

Table 1: Notations

Notations	Descriptions
MTA	Multi Trust Authority
DO	Data Owner
DU	Data User
EP	Encryption Proxy Server
Dec	Decryption Proxy Server
S	System master key
sk_{id}	The secret key of system user
pk_{id}	The public key of System user
C_{owner}	Data owner encrypted ciphertext
C_{Enc}	Encryption proxy encrypted ciphertext
C_{Dec}	Re-encryption proxy encrypted ciphertext
C'_{Dec}	Decryption Proxy encrypted ciphertext
C_T	Decryption Proxy decrypted ciphertext
tid_{An}	Transaction number
Loc	File Location

3.1 Bilinear Mapping and Computational Assumption

1) Bilinear mappin: Let G_1 and G_T be cyclic multiplicative group of big prime order q, g be a generator of G_1, We define e: $G_1 \times G_1 \rightarrow G_T$ e:is a bilinear pairing map. If e: $G_1 \times G_1 \rightarrow G_T$ has the following characters:

   Bilinear: $e(g^a, g^b) = e(g, g)^{ab}$ for all $a, b \in Z_q^*$.

   Non-degenrate: There exists $g_1, g_2 \in G_1$,such that $e(g_1, g_2) \neq 1$.

   Computable: There is an competent algorithm to compute $e(g_1, g_2)$ for all $g_1, g_2 \in G_1$.

2) Computational assumption: Let $(q, g, G_1, G_T, e) \leftarrow Setup(k)$ for a security parameter k. The modified decisional bilinear Diffie-Hellman (mDBDH) problem is to determine whether $T = e(g, g)^{b/a}$ given a tuple $(g, g^a, g^b, g^c, T) \in G_1^4 \times G_T$, where $a, b \in_R Z_q^*$.

   We set the k to be a security parameter. Generally, we think that mDBDH assumption holds in G_1,G_T,

if for any probabilistic polynomial time (PPT) algorithm A, the following formula holds:

$$\left|Pr[A(g, g^a, g^b, g^c, e(g,g)^{b/a}) = 1 | a, b \leftarrow_R z_q^*]\right| - \left|Pr[A(g, g^a, g^b, g^c, T) = 1 | a, b \leftarrow_R z_q^*]\right| \leq V(K),$$

where $V(.)$ is a negligible function, for all polynomial functions$P(.)$, we have $V(K) < 1/p(k)$.

3.2 Blockchain Technology and Hyperledger Fabric

1) Blockchain technology: Blockchain is a decentralized and distributed ledger jointly maintained by all nodes in the blockchain network. The nodes in the network keep the same ledger, there are no third party and central authority can control the entire network.

A block header and some transactions are included in each block, each block header includes of the connection pointer to the previous block header, a merkle root with a binary tree structure and a timestamp. By this way, blocks are connected together in an orderly fashion by the hash value of the header pointer. The hash algorithm guarantees that the transaction data in each block is immutable. The structure of blockchain is shown in Figure 2.

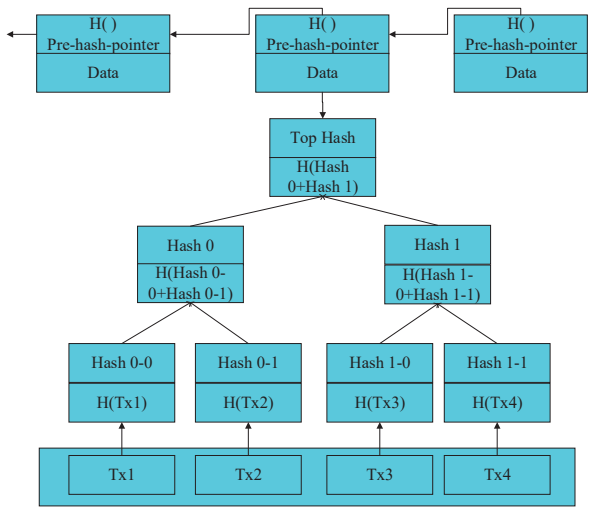


Figure 2: The structure of blockchain

In this paper, all system users are used as peer nodes to constitute a blockchain network, all of nodes are responsible for recording the index of sharing files from IPFS. There are two kinds of data in network, one is the file location, the other is a transaction number. Because the blockchain is not fit for storing big data, so we merely store the index of files in the blockchain, this way not only can alleviate the storage burden on the blockchain, but also can ameliorate the operation efficiency of the blockchain system.

2) Hyperledger fabric: Hyperledger fabric is a consortium blockchain platform and one of the hyperledger projects hosted by the Linux Foundation. It is a popular implementation of the blockchain network framework [6, 21]. As other blockchain technologies, hyperledger fabric contains a ledger, smart contracts and consensus mechanism. It is a system for managing transactions through all participants. The most obvious distinction from other blockchain systems is that fabric is private and licensed, the members of hyperledger fabric network are registered through a trusted membership service provider (MSP). Hyperledger fabric also provides the ability to create channels, permitting a group of participators to build separate transaction ledgers.

Hyperledger fabric node: Node plays a significant role in the hyperledger fabric platform, it is the core of the entire blockchain network, there are five kinds of node in fabric, CA, orderer, endorser, leader and committer. CA node is similar to a certificate authority, providing identity registration services for users. Orderer is responsible for sorting and packaging transactions into blocks on the network. Endorser is responsible for receiving a message request from client proposal, after checking the message, this node simulates the message proposal and uses its secret key to sign the simulation result, so as to show that the transaction is legal and effective, the endorser node returns the endorsement result to the client. Leader, as a representative of all nodes in the organization, it is able to connect to the orderer and broadcast the block received from orderer to all nodes in the organization. Committer is used to verify the integrity and legitimacy of the transaction message structure and to store it in the ledger.

Hyperledger fabric channel: Channel is an important concept in fabric. In essence, it is a private atomic broadcast channel divided and managed by orderer. For the purpose of isolating the information in the channel, the entities outside of the channel cannot access the information in the channel, it achieves the privacy of the transaction. At present, the channel consists of system channel and application channel. The orderer manages the application channel through the system channel, and the user's transaction information is transmitted by the application channel. For the general user, the channel is an application channel. The information in the channel is transparent to the nodes who joined the channel.

Hyperledger fabric chaincode: Smart contract is called chaincode in hyperledger fabric, the chaincode is written by developer, Go and Java language as the development languages. The

chaincode is deployed in the blockchain network node, it can run independently in a secure docker container, using the Go remote procedure calls (GRPC) protocol to correspond with related nodes and managing the data in the ledger.

To illustrate the hyperledger fabric visually, the architecture of fabric is shown in Figure 3, the complete transaction processes of fabric are shown in Figure 4.

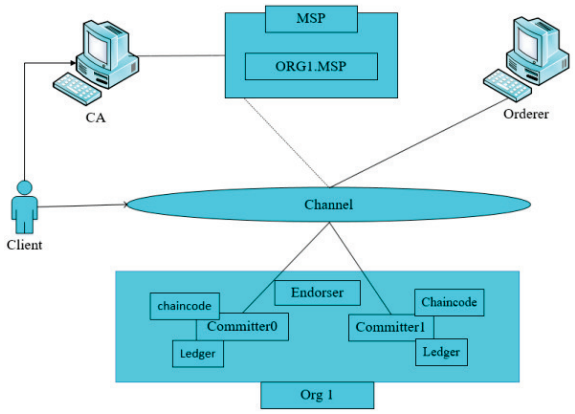


Figure 3: The architecture of fabric

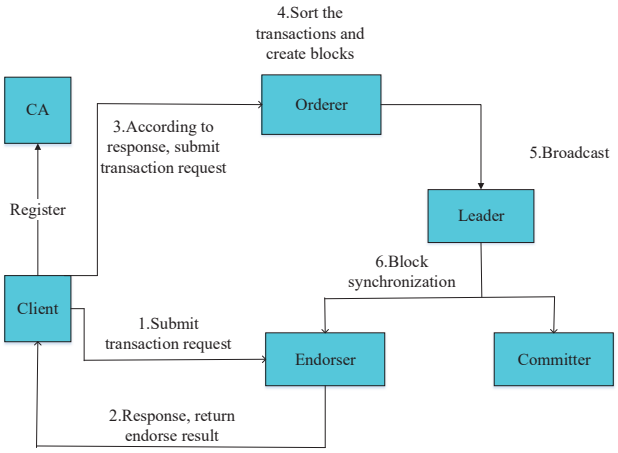


Figure 4: Transaction processes of fabric

3) IPFS: Inter planetary file system (IPFS) is peer-to-peer decentralized file system. According to the underlying protocol, the files which saved on IPFS can be obtained in anywhere. It offers a high throughput file store frame and unites techniques such as distributed hash tables (DHT), self-certifying *etc.* The merit of IPFS over extant storage is that there is no central server, therefore, this system conquers a single point of failure. When we upload the file to the system, we can obtain the hash value about file. This hash value can be seen as a Uniform Resource Locator (URL) in the web. We call this hash value file location. When the data user requires to obtain the file in IPFS, according to hash value, data user downloads the encrypted files from IPFS, and decrypted it. The complete processes of IPFS are shown in Figure 5.

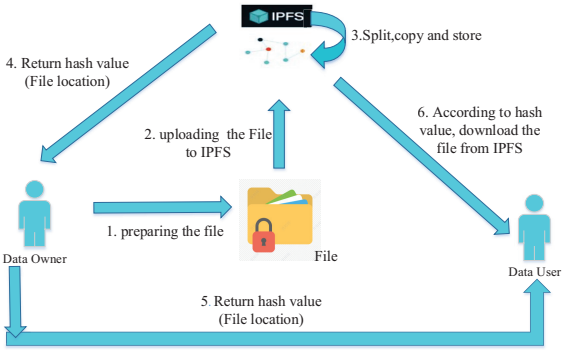


Figure 5: Transaction processes of IPFS

4 System Model

4.1 System Model

The scheme consists of the following eight entities:

Multi Trust Authority (MTA): MTA is an organization in system that cooperates to generate the master key for each user.

Data Owner (DO): DO is a person or organization that has some of files to share. He or She encrypted the file and sent it to IPFS, when received a file request from DU, DO generates a re-encryption key and sends it to the proxy, then returns the index address of the file to DU. They also write smart contracts in the blockchain to control access permissions.

Data Users:(DU): DU is the data client of DO that they are authorized to access some of sharing files. When they require to view some of the file, they send request to DO. With the help of Dec, DU decrypted the file which is shared in IPFS.

Encryption Proxy (EP): EP is a proxy that helps the DO encrypts the file again. This encryption makes the data more secure. On the basis of encryption, some ciphertext conversion work is also done for following decryption.

Inter Planetary File System (IPFS): IPFS is a peer-to-peer distributed file system. They are responsible for splitting the data into blocks and encrypting each block. Then, IPFS returns the encrypted hash value as an index to the DO.

Decryption Proxy (Dec): Dec is a proxy that helps the DO transfer the ciphertext.

They also have a responsibility to help DU decrypt the transformed ciphertext partially.

Smart contact: Smart contact is a self-executing protocol that help people to disseminate, validate, or implement a contract in an automatic manner. In our scheme, smart contract is used to authenticate the identity of user and help them upload the file index.

Blockchain:We choose hyperledger fabric as our blockchain platform. It is one of a consortium blockchain that hosted by the Linux foundation. The file index and the transaction number are stored in this blockchain. The smart contact is used on this blockchain.

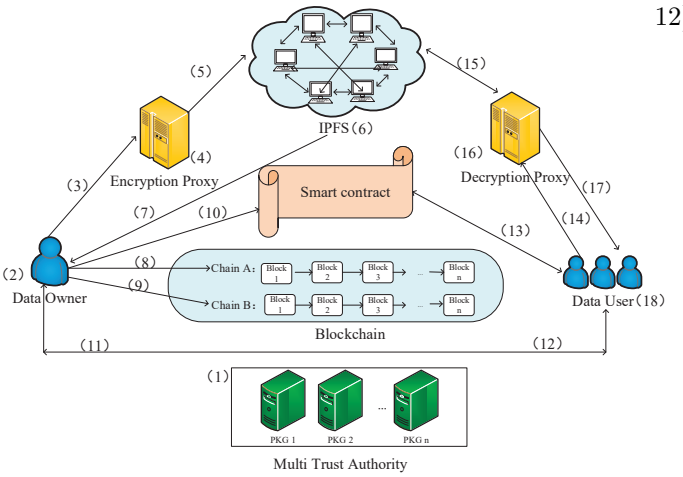


Figure 6: System framework

For clarity, the system frame is described in Figure 6. The corresponding description of each step number in Figure 6. is shown as follows:

1) Users (Including DO, DU, Enc-proxy, Dec-proxy) register in the MTA (Multi Trust Authority) and hyperledger fabric ca with their own ID, and obtain their unique public and private key.

2) DO encrypts the plain-text message m to obtain the ciphertext C_{owner}.

3) DO uploads the C_{owner} to the EP.

4) EP re-encrypts the ciphertext C_{owner} and obtains the ciphertext C_{Enc}.

5) The EP uploads the ciphertext C_{Enc} to IPFS.

6) IPFS divides the ciphertext C_{Enc} into n blocks and calculates the hash of each block. Eg: Divided C_{Enc} into n blocks: calculate $hash1 = hash(Block1)...hashn = hash(Blockn)$, the array $Arr[n] = [hash1, hash2, ..., hashn], hash(file) = hash(Arr[n])$.

7) The IPFS packages hash(file) and Arr[n] as the Loc and returns Loc to DO.

8) DO uploads Loc to chain A and records transaction number tid_{An}.

9) DO uploads the transaction number tid_{An} to chain B.

10) DO writes a smart contract, making chain B is only accessed by the DO, when the DU node queries the transaction content corresponding to the transaction tid_{An}, after confirming the identity of user, the content about tid_{An} can be returned to the DU.

11) The DU sends his/her own ID to the DO and requests to obtain the data of the DO.

12) With the help of CA and MSP, DO verifies the identity of DU, when it passes validation, DO obtains the transaction number tid_{An} from the chain B, and runs the Transkey algorithm to generate TransKey:

    a. Firstly, DO chooses the secret parameter t and computes $C_1 = g^t$.

    b. Secondly, DO chooses the secret parameter k and a hash function H_1.

    c. Thirdly, according to the identity of DU and DO, DO computes Transkey: $< R_1, R_2 >=< C_1^{k/t}, H_1(id_{DO})^{-t} \cdot H_1(id_{DU})^k >$

    d. Finally, DO returns the Transkey to DU through the fabric network. (In our scheme, we supposed that fabric network is a safe channel).

13) After verifying the identity of DU, DU gets the content from the chain A by the transaction number tid_{An}, the content is the file index Loc.

14) DU sends $(Loc, TransKey)$to the Dec.

15) Dec downloads ciphertext $CEnc$ from the IPFS according to Loc.

16) Dec converts C_{Enc} to C_{Dec} according to the Rekey, and C_{Dec} to C'_{Dec}.

17) Dec Returns C'_{Dec} to DU.

18) DU decrypts C'_{Dec} with its own private key sk_{Du}, obtains m.

In our paper, the system model consists of six parts as following:

1) $par \leftarrow setup(k)$: This step completes the system initialization setup. It takes as input security parameter k, the system run the setup algorithm, outputs a series of public parameters.

2) $(pki, ski) \leftarrow keyGen(id)$: Users in this system register in the MTA with his or her own ID, according to each ID, MTA runs algorithm to generate public and private key for each user. As shown in Step (1) of Figure 6.

3) The encryption by Data Owner: In this part, it is made up of the following two sub-algorithms: The encrypt and TranskeyGen algorithm are run by DO.

   a. $Encrypt.owner \leftarrow (ski, pki, m)$: The file encrypt algorithm takes as input the shared file m. DO chooses two secret parameters and computes, it outputs file ciphertest $C_{owner.}$ DO uploads the ciphertext C_{owner} to EP. As shown in Steps (2) and (3) of Figure 6.

   b. $TransKey \leftarrow TranskeyGen(id_{DO}, id_{DU})$: Data Owner chooses a secret parameter, receive ID from DU, then, DO runs the TranskeyGen algorithm to generate TransKey and sends it to DU. As shown in Steps (11) and (12) of Figure 6. The detailed process of TransKey algorithm is shown in the part 5.1.3).

4) The computation by Encryption Proxy: This step consists of two algorithms: RekeyGen and ReEncrypt algorithm. They are all run by EP.

   a. $Rekey \leftarrow RekeyGen(pk_{DO}, pk_{Enc}, pk_{Dec})$: EP received three public key of the Data Owner DO, the delegator EP and the delegatee Dec. EP runs Rekey algorithm to generate Rekey and send it to Dec.

   b. $ReEncrypt.Enc \leftarrow C_{owner}$: EP received the ciphertext C_{owner} from DO, EP runs encrypt algorithm to generate ciphertext C_{Enc}, EP uploads C_{Enc} to IPFS, IPFS distributes the ciphertext and storage, then returns the index address Loc of the file to the DO. As shown in Steps (4) $\sim$ (7) of Figure 6.

5) The computation by Decryption Proxy: In this part, it consists of the following two sub-algorithms, ReEncrypt and Decrypt algorithm, both algorithms run on the Dec. Firstly, DO stores some index information into the blockchain and writes smart contracts, returns some correlative information to DU, DU sends the returned information to Dec. Secondly, according to the index information, Dec downloads the relevant files from IPFS and decrypts them. The purpose of the ReEncrypt process is to convert the ciphertext C_{Enc} into the ciphertext C_{Dec} that DU can decrypt it. In order to reduce the computation of DU, decrypt algorithm is used for partial decryption of ciphertext. This step decrypts C_{Dec} to ciphertext C'_{Dec}, then, Dec sends the ciphertext C'_{Dec} to DU. As shown in Steps (8) $\sim$ (17) of Figure 6.

6) $Decrypt.user \leftarrow (c'_{Dec})$:DU received the (c'_{Dec}), then decrypts it by his or her own secret key sk_{DU}. DU computes a bilinear pair operation and xor to get the plaintext message m. As shown in Step (18) of Figure 6.

4.2 Security Model

We supposed that the encryption proxy and the decryption proxy server are curious but honest in our system model, supposed that $\varepsilon = (Setup, KeyGen, Enc, ReKeyGen, ReEnc, Dec)$ is a DOMIBPRE scheme. In order to describe our game, denoted by $Exp_{\varepsilon,A}^{IND-DOMIBPRE-CPA}$, we devise the oracle between a PPT adversary Adv and a challenger C, that the adversary Adv can query during the game:

1) $G_{pk}(id)$: The oracle that generates the public key. Give as input id, C operates $keyGen(id)$ algorithm to generate (pki, ski), returns pki to Adv and stores in an empty table T_{pk}.

2) $G_{sk}(pki)$:The oracle that generates the private key. Give as input pki by Adv. C searches pki from the table T_{pk} and returns the relevant to Adv.

3) $G_{rk}(pk_{DO}, pk_{Enc}, pk_{Dec})$: The oracle that generates the Rekey. Take three public keys as input, $pk_{DO}, pk_{Enc}, pk_{Dec}$ of the DO, the delegator EP and delegatee Dec. Firstly, C seeks the private key sk_{Enc} of the delegator EP from the T_{pk}. Then, C operates Rekey algorithm, returns Rekey to Adv.

4) $G_{reenc}(pk_{DO}, pk_{Enc}, pk_{Dec}, C_{owner})$: The oracle that Re-encryption. Take three public keys as input, $pk_{DO}, pk_{Enc}, pk_{Dec}$ of the DO, delegator EP and delegatee Dec respectively, an original ciphertext C_{owner} of the delegator EP, C first seeks the re-encryption key Rekey by running Rekey algorithm. Next, C operates Enc algorithm to compute C_{Enc} and sends C_{Enc} to Adv.

5) $G_{Dec}(C_{\cdot Enc}, pk_{Dec})$: If it is CPA, this oracle is refused to Adv. Now, the game $Exp_{\varepsilon,A}^{IND-DOMIBPRE-CPA}$ can be described as below:Game $Exp_{\varepsilon,A}^{IND-DOMIBPRE-CPA}$ (k):

   a. $Par \leftarrow Setup(k)$;

   b. $(pk^*_{DO}, pk^*_{DU}, m_0, m_1, sp) \leftarrow A_1^{G_{pk}, G_{sk}, G_{rk}, G_{reenc}, G_{Dec}}$, where$|m_0| = |m_1|$; (Find step)

   c. $d' \leftarrow_R \{0, 1\}$;

   d. $c^*_{Dec} = Enc(sk^*_{DO}, pk^*_{DU}, m_d)$;(Challenge step)

   e. $d' \leftarrow A_2^{G_{pk}, G_{sk}, G_{rk}, G_{reenc}, G_{Dec}}$, (par, c^*_{Dec}, SP); (Guess step);

   f. Return d';
   In the $Exp_{\varepsilon,A}^{IND-DOMIBPRE-CPA}$, the adversary Adv works in Find and Guess phase, respectively. Furthermore, some of limits are needed to Adv. Including:

      i. Adv is not permitted to make any private key generation queries either on pk^*_{DO}, pk^*_{DU}.

ii. *Adv* is not permitted to participate collusion attack. To be more precise, if *Adv* has made $G_{rk}(pk^*_{DO}, pk^*_{DU}, pk_{Dec})$ or $G_{reenc}(pk^*_{DO}, pk^*_{DU}, pk_{Dec}, c^*_{Dec})$ queries, then *Adv* is not permitted to make a $G_{sk}(pki)$ query, whereas the same, if *Adv* has made $G_{sk}(pki)$ query, *Adv* is not permitted to make $G_{rk}(pk^*_{DO}, pk^*_{DU}, pk_{Dec})$ or $G_{reenc}(pk^*_{DO}, pk^*_{DU}, pk_{Dec}, c^*_{Dec})$ queries.

iii. *Adv* is not permitted to initiate $G_{Dec}(c^*_{Dec}, pk^*_{Dec})$ query.

iv. *Adv* is not permitted to initiate $G_{Dec}(c, pk_{Enc})$ query, if $c = ReEnc(c^*_{Dec}, Rekey^*)$. Now, the advantage of *Adv* in the game $Exp^{IND-DOMIBPRE-CPA}_{\varepsilon,A}$ is defined:

$$\begin{aligned} & Adv^{IND-DOMIBPRE-CPA}_{\varepsilon,A}(k) \\ = \; & |2Pr[Exp^{IND-DOMIBPRE-CPA}_{\varepsilon,A}(k) \\ & = d] - 1| \end{aligned}$$

Definition 1. *We think DOMIBPRE scheme ε is IND-CPA secure if for all PPT algorithms A, we have $Adv^{IND-DOMIBPRE-CPA}_{\varepsilon,A}(k) \leq v(k)$, $v(.)$ is a negligible function.*

Notice 1. *The re-encrypted ciphertext is not offered to the adversary Adv, but we permit Adv to make re-encryption key generation query, therefore, Adv can convert the ciphertext by the re-encryption key. Nevertheless, the below situations are prohibited:*

1) If Adv has accessed to a Rekey from pk^*_{DU} to pk_{Enc}, then Adv is not permitted to make a private key query on pk_{Enc}.

2) If Adv has obtained private key of user pk_{Enc}, then Adv is not permitted to make Rekey query from pk^*_{DU} to pk_{Enc}.

5 System Description

In this section, we describe our concrete scheme design and smart contract design.

5.1 Scheme Description

1) $par \leftarrow setup(k)$: Input a security parameter k, Let e: $G_1 \times G_1 \rightarrow G_T$ is a bilinear map, g is a generator of G_T, let G_1 and G_T be two cyclic multiplicative groups of big prime order q. Selects two functions H_1 and H_2. $H_1\{0,1\}^* \rightarrow G_1$, $H_2 : G_T \rightarrow \{0,1\}^n$, $g_1 = e(g, g)$ selects $s \leftarrow z^*_q$, as the primary secret key (The generation of s is different from the traditional scheme, it is not generated by a single PKG, it is combined with multi PKG, each PKG contributes a portion of the private key, then the system aggregates them to generate the primary private key s, the detailed generation of is shown in Section 2.2 of [13]), output the system parameter $par(G_1, G_T, H_1, H_2, g_1, g, g^s)$.

2) $(pki, ski) \leftarrow keyGen(id)$: According to the ID of each user, MTA generates public key and private key for users.

$$\begin{aligned} sk_i &= H_1(id)^s \\ pk_i &= g^{ski} \end{aligned}$$

3) The encryption by Data Owner: This step consists of the following two sub-algorithms. Encrypt and TranskeyGen algorithms are run by DO.

a. $Encrypt.Owner \leftarrow (ski, pki, m)$:Data Owner chooses $t, \sigma \leftarrow z^*_q$ computes:

$$\begin{aligned} c_1 &= g^t \\ c_2 &= (m \oplus H_2(\sigma)) \cdot g_1^{1/sk_{DO}} \\ c_3 &= \sigma.e(g^s, H_1(id_{DO})) \\ c_4 &= pk_{Enc}^{1/sk_{DO}} \end{aligned}$$

Output: $C_{owner} = (c_1, c_2, c_3, c_4)$.
Data Owner uploads the C_{owner} to Encryption Proxy.

b. $TransKey \leftarrow TranskeyGen(id_{DO}, id_{DU})$:
Data Owner chooses$k \leftarrow z^*_q$, computes: TransKey:

$$< R_1, R_2 > = < C_1^{k/t}, H_1(id_{DO})^{-t} \cdot H_1(id_{DU})^k >$$

Data Owner returns the $TransKey$ to Data User.

4) The computation by Encryption Proxy: This step consists of two algorithms, RekeyGen and re-encrypt algorithms. They are all run by EP.

a. $Rekey \leftarrow RekeyGen(pk_{DO}, pk_{Enc}, pk_{Dec})$ computes:

$$\begin{aligned} Rekey &= (tk1_{DO \rightarrow Enc \rightarrow Dec}, tk2_{DO \rightarrow Enc \rightarrow Dec}) \\ &= ((pk_{DO}^{1/sk_{Enc}}, (pk_{Dec})^{1/sk_{Enc}})) \\ &= (g^{sk_{DO}/sk_{Enc}}, g^{sk_{Dec}/sk_{Enc}}) \end{aligned}$$

b. $Re - Encrypt.Enc \leftarrow C_{owner}$: Encryption Proxy computes:

$$\begin{aligned} c'_i &= c_i \qquad (i = 1, 2, 3) \\ c'_4 &= e(c_4, tk2_{DO \rightarrow Enc \rightarrow Dec}) \\ &= (g^{sk_{DO}/sk_{Enc}}, g^{sk_{Dec}/sk_{Enc}}) \end{aligned}$$

Output: $C_{Enc} = (c'_1, c'_2, c'_3, c'_4)$;
Encryption Proxy uploads the C_{Enc} to IPFS.

5) The decryption by Decryption Proxy: In this part, it consists of the following two sub-algorithms, ReEncrypt and Decrypt algorithms, both algorithms are

run on the Dec. The purpose of the ReEncrypt process is to convert the ciphertext C_{Enc} into a ciphertext C_{Dec} that DU can decrypt it by his or her own secret key. In order to reduce the computation of DU, decrypt algorithm is used for partial decryption of ciphertext. This step decrypts ciphertext C_{Dec} into ciphertext C'_{Dec}. Then, Dec sends the ciphertext C'_{Dec} to DU. The specific encryption processes are shown below:

a. $Decrypt.Dec \leftarrow (C_{Enc}, Transkey)$

$$\begin{aligned} C''_1 &= R_1 = g^k \\ C''_2 &= C'_2 \\ C''_3 &= C'_3.e(g^s, R_2) \\ C''_4 &= C'_4. \end{aligned}$$

Output: $C_{Dec} = (C''_1, C''_2, C''_3, C''_4)$.

b. $Decrypt.Dec \leftarrow C._{Enc}$ Decryption Proxy computes:

$$C_T = \frac{C''_2}{e(g, C_4^{''1/sk_{Dec}})}$$

6) $Decrypt.user \leftarrow C'_{Dec}$,Data User computes:

$$\begin{aligned} \sigma &= \frac{C''_3}{e(C''_1, sk_{DU})} \\ m &= C_T \oplus H_2(\sigma). \end{aligned}$$

Consistency. For the ciphertext $C'_{Dec} = (C''_1, C''_3, C''_T)$, where $C''_1 = g^k$;

$$\begin{aligned} C''_3 &= C'_3.e(g^s, R_2) \\ &= \sigma.e(g^s, H_1(id_{DO})^t).e(g^s, H_1(id_{DO})^{-t} \\ &\qquad \cdot H_1(id_{DU})^k) \\ &= \sigma.e(g^s, H_1(id_{DU})^k) \\ C''_4 &= e(C_4, pk_{Dec}^{1/sk_{Enc}}) \\ &= g^{sk_{Dec}/sk_{DO}} \\ C_T &= \frac{C''_2}{e(g, C_4^{''1/sk_{Dec}})} \\ &= \frac{(m \oplus H_2(\sigma)).g_1^{1/sk_{DO}}}{e(g, g^{1/sk_{DO}})} \\ &= m \oplus H_2(\sigma) \\ \sigma &= \frac{C''_3}{e(C''_1, sk_{DU})} \\ &= \frac{\sigma.e(g^s, H_1(id_{DU})^k}{e(g^k, H_1(id_{DU})^s)} \\ m &= C_T \oplus H_2(\sigma) \end{aligned}$$

5.2 Smart Contract Description

In this section, we present the smart contract algorithm logic employed in our paper. In the hyperledger fabric, smart contract is coded by Go language. There are three algorithms in our design. They are AddIndex, AddTxid and QueryIndex. We will introduce each of the four algorithms in the following:

1) AddIndex(Loc): This function can merely be performed by Data Owner(DO), when DO uploads some files to IPFS, DO can obtain Index Hash from IPFS, we call it Loc. According to smart contract we store the Loc in blockchain and record this transaction number txid.

Algorithm 1 AddIndex

```
Begin
Input: New Loc
Output:txid
if sender is not DO, then
    throw;
end if
New Loc has existed,then
return false;
Existing Index[New loc]⇐true;
return ture;
End
```

2) AddTxid(txid,IDO): This function can only be executed by DO. DO stores the txid in blockchain, and writes their ID as an index into the smart contract, by matching the ID, DO can obtain txid.

Algorithm 2 AddTxid

```
Begin
Input: txid,IDO
Output:bool
if sender is not DO, then
    throw;
end if
Mapping txid to IDO, and add it to extant Txid collection
return ture;
End
```

Algorithm 3 QueryTxid

```
Begin
Input: txid
Output:query Result
if sender is not authorized DU then then
    throw;
end if
Query the transaction content that corresponding to txid
Query Result⇐ Index[Loc]
return Query Result ;
End
```

3) QueryIndex(txid,IDU): DU submits a file sharing request to DO, after passing the identity authentication, DO returns txid to DU, DU uses txid as a condition to retrieve the corresponding data Loc in the blockchain network. According to the Loc, DU downloads the corresponding encrypted file in IPFS.

6 Security Proof

Theorem 1. *We assumed that mDBDH problem is difficult, then, our DOMIBPRE scheme proposed in 5 is IND-CPA security in the standard model.*
Proof: We require to attest that if there is a PPT adversary Adv has an advantage ω to attack DOMIBPRE scheme, and we need construct the other PPT adversary B, with the help of Adv, B can solve the mDBDH problem in G_1, with an advantage. Given a tuple $(g, g^a, g^b, g^c, T) \in G^4 \times G_T$, where $a, b \in z_q^*$ B can output 1 if $T = e(g,g)^{b/a}$ and 0 others.

The interaction between the two adversaries and B is presented as below:

Setup: B inputs a security parameter k to generate the system parameters $Par = (q, g, G, G_T, e, g_1)$ and return par to *Adv*, $g_1 = e(g,g)$.

Find: In this step, B answers the queries from *Adv* as below:

1) On a $G_{pk}(id)$ query: B chooses a random value $x_i \in z_q^*$, At the same time, B throws a weighted coin,$b_i \in$\{0,1\}, satisfying $b_i = 1$ with probability γ and 0 otherwise. If $b_i = 1$, B sets $pki = g^{ski}$, it signifies that the private key of user is ski. If $b_i = 0$, B sets $pki = (g^a)^{ski}$, it signifies that the private key of user i is αski. In this situation, does not know the private key either. Finally, B returns pki to *Adv* and stores (pki, ski, bi) in T_{pk}, where T_{pk} is a table that records public keys in the game and is vacant when initialized.
   For the sake of generality, assumed that *Adv* has made the suitable G_{pk} query before executing the following query:
2) On a $G_{sk}(pki)$ query : B searches the table T_{pk}, if $b_i = 1$, B returns ski to *Adv*, otherwise, a random number in z_q^* is output by B. B terminates.
3) On $G_{rk}(pk_{DO}, pk_{Enc}, pk_{Dec})$ query:
   Firstly, B searches T_{pk} to get $(pk_{DO}, sk_{DO}, b_{DO})$, $(pk_{Enc}, sk_{Enc}, b_{Enc}), (pk_{Dec}, sk_{Dec}, b_{Dec})$ and responds to *Adv* based on the below situation respectively.
   a. If $b_{Enc} = 1$, B calculates re-encryption key $Rekey = ((pk_{DO})^{1/sk_{Enc}}, (pk_{Dec})^{1/sk_{Enc}})$. If $(b_{DO}, b_{Enc}, b_{Dec}) = (0,0,0)$, B calculates re-encryption key $(g^{sk_{DO}/sk_{Enc}}, g^{sk_{Dec}/sk_{Enc}})$.
   b. If $b_{Enc} = 0$, and $(b_{DO}, b_{Dec}) \neq \{0,0\}$, B terminates.
4) On a $G_{reenc}(pk_{DO}, pk_{Enc}, pk_{Dec}, c_{owner})$ query: B makes a query to get Rekey, and runs the algorithm $ReEnc(C_{owner}, Rekey)$. According to the results return from the Rekey, B returns or terminates.

Challenge: *Adv* puts in two messages, $m_0, m_1 \in G_T$. *Adv* targets a sender pk_{DO}^*, and a target receiver pk_{DU}^*, with the below limits:

1) *Adv* does not make a query about the private key generation either on pk_{DO}^*, pk_{DU}^*, in Find step.
2) *Adv* does not initiate any $G_{rk}(pk_{DO}^*, pk_{DU}^*, pk_{Dec})$ query, where the private key of *Dec* is obtained to *Adv* by making an $G_{sk}(pk_{Dec})$query. B randomly chooses a bit $d \in \{0,1\}$, and retrieves T_{pk} to get $(pk_{DO}^*, sk_{DO}^*, b_{DO}^*)$. Then B calculates the ciphertext:

$$\begin{aligned} c_{Dec}^* &= (c_1^*, c_2^*, c_3^*, c_4^*) \\ &= (g^b, (m_d \oplus p).T^{1/sk_{DO}^*}, \\ &\quad \sigma.e(g^{s*}, H_1(id)_{DO}), (c_1^*)^{sk_{DU}^*/sk_{DO}^*}). \end{aligned}$$

   This process is viewed as DO to DU.
   Finally, c_{Dec}^* is transmitted to *Adv* as a challenge ciphertext.

Guess: More queries can be made by in the find phase, with the below limits:

1) *Adv* is not permitted to make any private key generation queries either on pk_{DO}^*, pk_{DU}^*.
2) If *Adv* has made $G_{rk}(pk_{DO}^*, pk_{DU}^*, pk_{Dec}^*)$ or $G_{reenc}(pk_{DO}^*, pk_{DU}^*, pk_{Dec}, c_{Dec}^*)$ queries, then *Adv* is not allowed to make a $G_{sk}(pki)$ query, whereas the same, if *Adv* has made a $G_{sk}(pki)$ query, *Adv* is not permitted to make $G_{rk}(pk_{DO}^*, pk_{DU}^*, pk_{Dec})$ or $G_{reenc}(pk_{DO}^*, pk_{DU}^*, pk_{Dec}, c_{Dec}^*)$ queries.

Decision: At this step, *Adv* outputs his guess $d' \in \{0,1\}$, if $d = d'$, B outputs 1, which manifests that $T = e(g,g)^{b/a}$, otherwise, B outputs 0, which manifesting that T is a random element in G_T.

Probability analysis that B does not abort:

1) On $G_{sk}(pki)$ query, B will not terminate in situation of bi = 1. Supposed *Adv* made qsk private key generation queries during the interaction, then the probability that B does not terminate in this situation is γ^{qsk}.
2) On $G_{rk}(pk_{DO}, pk_{Enc}, pk_{Dec})$query, B will not terminate in situation of *bi =1 or bi =bs=bj=0.* Supposed *Adv* made qrk re-encryption key queries, then the probability that B does not terminate is more γ^{qrk}.
3) On $G_{reenc}(pk_{DO}, pk_{Enc}, pk_{Dec}, C_{owner})$query, B will not terminate if there is a re-encryption key returned for this query. The probability analysis is as identical as the situation in $G_{rk}(pk_{DO}, pk_{Enc}, pk_{Dec})$query. Supposed that *Adv* can make qre re-encryption queries, then the probability that B does not terminate is more than γ^{qre}.

4) When the guess bit d_0 is output by Adv. B will not terminate if $b^*_{DU} = 0$ The probability in this situation is $1-\gamma$. It is distinct that if B does not terminate, the view of Adv during the interaction is as identical as the one in the actual attack. The total probability that B does not terminate is $f(\gamma) = (1-\gamma)\gamma^{qsk+qrk+qre}$. It is effortless to show that the function $f(\gamma)$ has a maximum value:

$$\frac{1}{qsk+qrk+qre} \times (1-\frac{1}{qsk+qrk+qre+1})^{qsk+qrk+qre+1}$$

at the $\gamma_{opt} = 1-1/(qsk+qrk+qre)$ using γ_{opt},the probability that B does not terminate is at least $\frac{1}{e(qsk+qrk+qre+1)}$ for large value of $qsk+qrk+qre$. This demonstrates that B has an advantage of at least $\varepsilon/e(qsk+qrk+qre+1)$.

7 Performance Analysis

In this section, some crucial features will be discussed. In order to assess the costs of this scheme in algorithm, we simulated our scheme with Pairing-Based Cryptography (PBC) library in C language (pbc-0.5.14). The elliptic curve parameter is chosen in Type-A, and the order of group is 160 bits. Using a computer with 64-bit windows 10 operation system, 2.60GHz Intel Core 8850H CPU, with 16 GB RAM. The computation cost of primitive cryptography operations shown in Table 2, E is represented the exponentiation operation in G or G_T, P is represented the pairing operation in G_T.

Table 2: Time cost of cryptography operations

Operation Names	T_E	T_p
$Times$	6.648	9.461

The performance comparison between our scheme and other related schemes is carried out results are shown in Table 3, from Table 3 we can see that [2, 28] and [32] did not use outsourcing computing, which brings a lot of computing overhead to users, in addition, these schemes did not realize the significance of multi authority. The scheme [31] deployed outsourced decryption and multi authority, but it lacks of completeness and flexibility. By comparing with the above schemes, it can be concluded that our scheme satisfies all properties.

In addition, we discussed the computation cost, we had a comparative summary of the computation costs in Table 4. From Table 4 we can see that our scheme reduced the computation of exponential and bilinear pairings, it is more efficient than other schemes (including ReKeyGen, Enc, ReEnc, and Dec). To make the comparison more intuitive, the number of computation is tuned from 0 to 100 in steps of 10 to record the running time of different schemes, as shown in Figure 7. We can see that in any calculation number, the running time of our scheme is improved compared to others. We also plotted the comparison of operation times in the bar chart, as shown in Figure 8, it can be seen that compared with [2,11,12,32], the operation times of our scheme has been reduced.

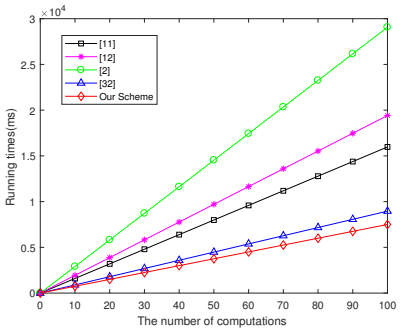


Figure 7: Comparison of time cost

Finally, by experimenting and comparing with other schemes, we illustrated that our scheme is more suitable for big data storage from the following three aspects:

1) **Storage cost:** According to reference [28], we get the cost of storing data on Ethereum, as show in Table 6: Obviously, as the amount of data stored increases, there is a huge storage cost, which is not suitable for big data storage. IPFS is a free storage platform for users, we also set up an IPFS network and test the time cost of IPFS, as shown in Table 5, its time cost can be accepted in practical. Therefore, IPFS is more suitable for storage big data.

2) **Trading efficiency:** Ethereum is a private blockchain. It is a completely decentralized organization and open organization, the consensus mechanism consumes a lot of energy, which affects its trading efficiency. Hyperledger fabric is a consortium blockchain, it is jointly managed by several institutions. We chose the hyperledger fabric as our blockchain network, we compared and analyzed the properties of Hyperledger Fabric and Ethereum, as shown in Table 7, hyperledger fabric has superior performance in TPS. Therefore, our scheme can enhance the efficiency of big data sharing.

3) **Data security:** Although IPFS provides the function of data encryption, the data owner was not involved in the whole process of data encryption, the data security can not be controlled by the data owner, in our scheme, before IPFS encrypted data, the data owner encrypts data by means of outsourcing, then they upload the encrypted data to IPFS system. Therefore, our scheme enhance the security of big data storage.

Table 3: Comparison of features

Schemes	Types of authority	Outsourced Encryption	Outsourced Decryption
[30]	Multiple	NO	YES
[23]	Single	NO	NO
[18]	Single	NO	NO
[4]	Single	YES	YES
Ours	Multiple	YES	YES

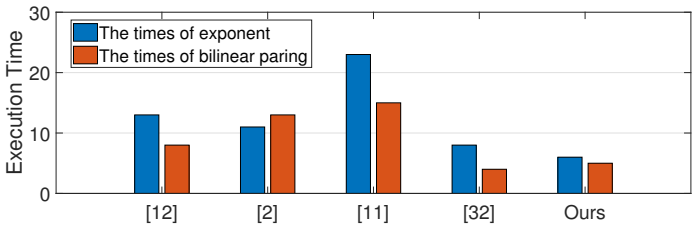


Figure 8: Comparison of computation time

Table 4: Comparison of computation cost

Schemes	[12]	[2]	[11]	[32]	Ours
$ReKeyGen$	1E	6E+1P	9E+1P	2E	2E
Enc	6E+1P	5E+1P	5E+1P	3E	3E+1P
$ReEnc$	4E+2P	9P	6E+7P	3P+E	1E+2P
Dec	2E+5P	2P	3E+6P	2E+1P	2P
$Runtimes(ms)$	159.772	194.141	290.679	89.588	86.133

Table 5: Time cost in IPFS

Names	Task Posting (Uploading)		Task Posting (Downloading)	
	Size(kb)	Time(ms)	Size(kb)	Time(ms)
Task_1	1158.39	495.41	1158.39	4.3265
Task_2	2357.45	607.56	2357.45	9.6875
Task_3	3562.81	736.82	3562.81	15.1238
Task_4	4823.76	912.49	4823.76	21.2926

8 Implementation of Blockchain Architecture

We simulated our blockchain scheme in personal computer, CPU 2.60 GHz, RAM 16GB. The operating system is ubuntu 16.04. We chose the hyperledger official test network fabric1.4.1 as our development platform and the solo as our consensus algorithm. Meanwhile, we build the blockchain network that consists of two organizations, four nodes and two channels. To illustrate our fabric blockchain system more clearly, the architecture of our scheme is shown in Figure 9.

As shown above, the member of nodes in Org 1 consist of data user, all data users join the channel A to get

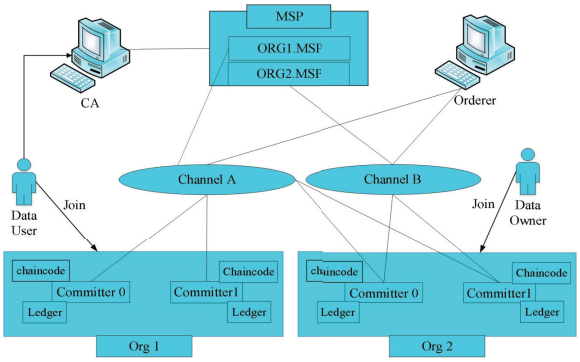


Figure 9: The architecture of our fabric network

Table 6: The cost of storing data

File Size(bytes)	Gas Used	Actual Cost(ether)	USD
208	28344	0.000056688	0.0236
1088	94984	0.000189968	0.0791
710	69280	0.00013856	0.0577

Table 7: The cost of storing data

Names	Type of blockchain	Consensus mechanism	participator	TPS
Ethereum	private chain	Pow/Pos/Dpos	Everyone	3-20
HyperledgerFabric	consortium chain	Solo/Kafka/Raft	Pre-Specified	1000-100000

the index of sharing files. At the same time, the member of nodes in Org 2 consists of data owners, they join the channel A and B simultaneously. Based on the smart contract, firstly, the index of the files which stored in IPFS were written in channel A by data owner, secondly, data owner records the transaction number tid_{An} which generated by the previous step to channel B. The data in channel B can only be accessed by data owner. When data user initiates a request to files, data owner returns the transaction number tid_{An} to data user after verifying his/her identity. According to tid_{An}, data user gets index of sharing files from the channel A.

Based on the hyperledger fabric blockchain network, we implement the aforesaid blockchain network. In order to make the operation of the client conveniently, we developed the software development kit (SDK) with go language. The interface and functions of the client are shown in the Figure 10.

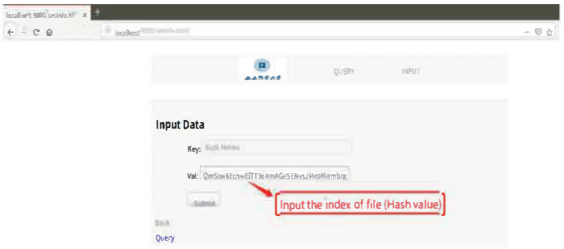


Figure 10: The process of uploading the index

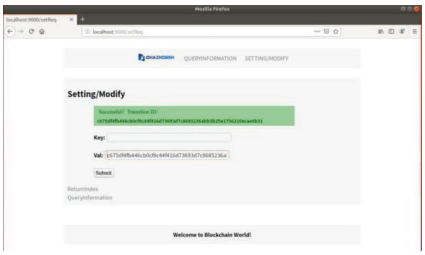


Figure 11: The feedback of storage in fabric

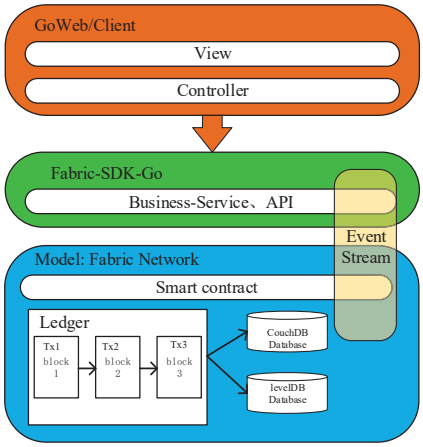


Figure 12: The architecture of MVC

As shown in Figure 10, data owner uploads the file index (hash string) into the blockchain system, when the data owner meets the conditions of smart contact, the system will return a transaction ID to client interface. As shown in Figure 11, when the data owner receives the transaction ID, the storage is successful. The experimental results prove our scheme is user-friendly. To explain clearly how smart contract help user upload the file index, we show the architecture of our MVC.

As shown in Figure 12, MVC architecture pattern consists of model, view and controller.

How to help DO upload the file index? When users upload file index from the view, they click the submission button, the controller will accept a requests from the client, it connects and invokes the corresponding API to access the smart contract through SDK. (In the step of uploading data, the smart contract which be invoked has the ability to write data in blockchain). When the requests meet the condition of smart contract, the system automatically executes the smart contract to write data in blockchain and sends the transaction ID to respond client. When the ID is returned, the storage is successful.

9 Conclusion

In this paper, a novel IBPRE scheme, DOM-IBPRE scheme is proposed and demonstrated, which is achieved by combining IBPRE, blockchain and the inter planetary file system (IPFS) technology. The scheme can avoid the complexity of certificate management, enhance the security of big data storage and ameliorate the efficiency of big data sharing. In addition, a single PKG is replaced by multi trusted authority in our scheme, the problem that the malicious attack on the PKG results in the leak of the private key of each user in conventional IBPRE scheme is solved. At the same time,we also assessed the security performance of the scheme by Chosen-Plaintext Attack (CPA) based on a modified DBDH problem in the standard model and simulated our scheme with Pairing-Based Cryptography (PBC) library, by comparing with other PRE schemes, our scheme has the better performance in computation. Finally, we implemented our blockchain architecture under linux system with the hyperledger official test network fabric, based on the smart contract, the index of sharing file is stored in blockchain, all authorized users in this system can store and read data in blockchain through the web. The test results validate that our scheme is practical.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB0802000, the Natural Science Foundation of China under Grant 61802303, 61772418 and 61602378, the Key Research and Development Program of Shaanxi under Grant 2019KW-053, the Innovation Ability Support Program in Shaanxi Province of China under Grant 2017KJXX-47, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2019JQ-866, 2018JZ6001 and 2016JM6033, the Research Program of Education Bureau of Shaanxi Province under Grant 19JK0803, the New Star Team of Xi?an University of Posts and Telecommunications under Grant 2016-02, the Fundamental Research Funds for the Central Universities under Grant GK201903005, and Guangxi Cooperative Innovation Center of Cloud Computing and Big Data under Grant YD1903.

References

[1] A. Banerjee, "Blockchain technology: Supply chain insights from ERP," in *Advances in Computers*, vol. 111, pp. 69–98, 2018.

[2] J. Bankar and J. Raghatwan, "Identity based proxy re-encryption using forward security in cloud framework," in *International Conference on Computing, Communication, Control and Automation (ICCUBEA'17)*, pp. 1–5, 2017.

[3] J. Benet and N. Greco, "Filecoin: A decentralized storage network," *Protocol Labs*, pp. 1–36, 2018. (`https://research.protocol.ai/publications/filecoin-a-decentralized-storage-network/`)

[4] G. Chunpeng, Z. Liu, and J. Xia, "Revocable identity-based broadcast proxy re-encryption for data sharing in clouds," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, 2019.

[5] D. Hopwood, S. Bowe, T. Hornby, N. Wilcox, *Zcash Protocol Specification*, Ver. 2021.2.7, NU5 Proposal, June 28, 2021. (`https://raw.githubusercontent.com/zcash/zips/master/protocol/protocol.pdf`)

[6] E. Androulaki, B. Artem, and V. Bortnikov, "Hyperledger fabric: A distributed operating system for permissioned blockchains," in *Proceedings of the Thirteenth EuroSys Conference*, pp. 1–15, 2018.

[7] W. G. Ethereum, "A secure decentralised generalised transaction ledger," *Bitcoin*, vol. 151, pp. 1–32, 2014. (`https://gavwood.com/paper.pdf`)

[8] M. Green and G. Ateniese, "Identity-based proxy re-encryption," in *International Conference on Applied Cryptography and Network Security*, pp. 288–306, 2007.

[9] H. Guo, Z. Zhang, and J. Xu, "Accountable proxy re-encryption for secure data sharing," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 1, pp. 145–159, 2018.

[10] J. Bankar and J. Raghatwan, "Identity based proxy re-encryption using forward security in cloud framework," in *International Conference on Computing, Communication, Control and Automation (ICCUBEA'17)*, pp. 1–5, 2017.

[11] K. Liang, W. Susilo, and J. K. Liu, "Efficient and fully CCA secure conditional proxy re-encryption from hierarchical identity-based encryption," *The Computer Journal*, vol. 58, no. 10, pp. 2778–2792, 2015.

[12] B. Libert and D. Vergnaud, "Unidirectional chosen-ciphertext secure proxy re-encryption," in *International Workshop on Public Key Cryptography*, pp. 360–379, 2008.

[13] Y. Cheng, W. Luo, S. Wen, "Blockchain-based electronic health record sharing scheme," *Journal of Computer Applications*, vol. 40, no. 1, pp. 157–161, 2020.

[14] B Lynn, *PBC Library - Pairing-Based Cryptography Library*, June 29, 2021. (`https://github.com/blynn/pbc`)

[15] M. Blaze, G. Bleumer, and M. Strauss, "Divertible protocols and atomic proxy cryptography," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 127–144, 1998.

[16] L. Mand, N. Kannouf, and Y. Chahid, "Blockchain-based pki for content-centric networking," in *The Proceedings of the Third International Conference on Smart City Applications*, pp. 656–667, 2018.

[17] A. Manzoor, M. Liyanage, and A. Braeke, "Blockchain based proxy re-encryption scheme

for secure IoT data sharing," in *IEEE International Conference on Blockchain and Cryptocurrency (ICBC'19)*, pp. 99–103, 2019.

[18] Y. N. Li, X. Feng, and J. Xie, "A decentralized and secure blockchain platform for open fair data trading," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 7, p. e5578, 2020.

[19] S. Nakamoto, "A peer-to-peer electronic cash system," *Ethereum Project Yellow Paper*, 2008. (`https://bitcoin.org/bitcoin.pdf`)

[20] M. Pilkington, "Blockchain technology: Principles and applications," in *Research Handbook on Digital Transformations*, 2016. DOI:10.4337/9781784717766.00019.

[21] S. Ranjan, A. Negi, and H. Jain, "Network system design using hyperledger fabric: Permissioned blockchain framework," in *The Twelfth International Conference on Contemporary Computing (IC3'19)*, pp. 1–6, 2019.

[22] S. Rezaei, M. A. Doostari, and M. Bayat, "A lightweight and efficient data sharing scheme for cloud computing," *International Journal of Electronics and Information Engineering*, vol. 9, no. 2, pp. 115–131, 2018.

[23] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 457–473, 2005.

[24] J. Shao, G. Wei, and Y. Ling, "Identity-based conditional proxy re-encryption," in *IEEE International Conference on Communications (ICC'11)*, pp. 1–5, 2011.

[25] D. Vorick and L. Champine, *Sia: Simple Decentralized Storage*, 2014. (`https://sia.tech/sia.pdf`)

[26] Q. Wang, W. Li, and Z. Qin, "Proxy re-encryption in access control framework of information-centric networks," *IEEE Access*, vol. 7, pp. 48417–48429, 2019.

[27] S. Wang, X. Wang, and Y. Zhang, "A secure cloud storage framework with access control based on blockchain," *Ieee Access*, vol. 7, pp. 112713–112725, 2019.

[28] S. Wang, Y. Zhang, and Y. Zhang, "A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems," *IEEE Access*, vol. 6, pp. 38437–38450, 2018.

[29] S. Wilkinson, T. Boshevski, and J. Brandoff, *Storj a Peer-to-Peer Cloud Storage Network*, 2014. (`https://www.storj.io/storj2014.pdf`)

[30] P. Xu, T. Jiao, and Q. Wu, "Conditional identity-based broadcast proxy re-encryption and its application to cloud email," *IEEE Transactions on Computers*, vol. 65, no. 1, pp. 66–79, 2015.

[31] K. Yang and X. Jia, "Expressive, efficient, and revocable data access control for multi-authority cloud storage," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 7, pp. 1735–1744, 2013.

[32] P. Zeng and K. K. R. Choo, "A new kind of conditional proxy re-encryption for secure cloud storage," *IEEE Access*, vol. 6, pp. 70017–70024, 2018.

[33] Q. Zheng, Y. Li, and P. Chen, "An innovative ipfs-based storage model for blockchain," in *IEEE/WIC/ACM International Conference on Web Intelligence (WI'18)*, pp. 704–708, 2018.

[34] G. Zyskind and O. Nathan, "Decentralizing privacy: Using blockchain to protect personal data," in *IEEE Security and Privacy Workshops*, pp. 180–184, 2015.

Biography

Jiayu He received the B.S. degree from the Xi'an University of Posts and Telecommunications in 2018. He is currently pursuing the M.S. degree with the National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications, China. His research interests include blockchain technology and security in cloud storage

Dong Zheng received the Ph.D. degree from Xidian University in 1999. He joined the School of Information Security Engineering, Shanghai JiaoTong University. He is currently a Professor with the Xi'an University of Posts and Telecommunications, China. His research interests include information theory, cryptography, and information security. He is a Senior Member of the Chinese Association for Cryptologic Research and a member of the Chinese Communication Society.

Rui Guo received the Ph.D. degree from the State Key Laboratory of Networking and Switch Technology, Beijing University of Posts and Telecommunications, China, in 2014. He is currently a Lecturer with the National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications. His present research interests include attribute-based cryptograph, cloud computing, and blockchain technology.

Yushuang Chen received the B.S. degree from the Xi'an University of Posts and Telecommunications in 2018. She is currently pursuing the M.S. degree with the National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications, China. Her research interests include blockchain, information security, and the Internet of Things (IoT).

Kemeng Li received the bachelors degree from Xi'an University of Posts and Telecommunications in 2018. He is currently pursuing his master degree at National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications. His research interests include blockchain technology and security in cloud storage.

Xiaoling Tao received the M.S. degree in computer application technology from Guilin University of Electronic Technology. She is a professor at the school of computer science and information security, Guilin University of Electronic Technology. Her research interests include cloud computing and security and network security.

A Coercion-Resistant E-Voting System Based on Blockchain Technology

Kaili Ye[1,2], Dong Zheng[1,2], Rui Guo[1,2,3], Jiayu He[1,2], Yushuang Chen[1,2], and Xiaoling Tao[3]
(Corresponding author: Kaili Ye)
School of Cyberspace Security, Xi'an University of Posts and Telecommunicationsr[1]
710121, Xi'an, China
National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications[2]
Guilin University of Electronic Technology, Guilin, China[3]
Email: kailiye@stu.xupt.edu.cn
(Received Apr. 21, 2020; Revised and Accepted Dec. 10, 2020; First Online Aug. 14, 2021)

Abstract

People widely utilize voting to express their opinions in a democratic society. In recent years, based on the Internet and information technology, the electronic voting (e-voting) system has replaced the traditional voting system because of its convenience and flexibility. However, many existing e-voting systems are suffered from the excessive power of counting agencies and tampering, which causes voting to no longer be transparent and fair. In addition, other threats, including intimidation and vote bribe, destroy the security of the e-voting system. Therefore, to guarantee the security and fairness of the e-voting system, designing a coercion-resistant e-voting system is crucial. This paper proposes an e-voting system based on blockchain technology, which offers fairness and transparency in the voting procedure. And our proposed system achieves coercion-resistance by employing a receiver-deniable encryption scheme. In particular, we design a time-release encryption algorithm to make the counting operation only take place after a specified time, which can guarantee the fairness of the voting process. We provide a detailed security analysis to prove other safety requirements that our system contains, such as verifiability and correctness. In addition, it can be seen that our proposed system is more suitable for small-scale elections from the performance comparison.

Keywords: Blockchain; Coercion-Resistant; Electronic Voting System; Receiver-Deniable; Time-Release Encryption

1 Introduction

Voting is seen as an essential way for people to exercise their democratic rights, and almost all of governments invest a big budget for designing a stronger and more trustworthy voting system. As Internet and information technology continue to develop, electronic voting (e-voting) system has become a convenient tool instead of the traditional paper-voting and handshow for the advantage of efficiency and flexibility [21].

The design and operation of e-voting system is a complex subject, which must be based on a secure electronic voting protocol. In 1981, Chaum [6] first presented an electronic voting scheme according to public key cryptography, which changed the way people vote. After that, plenty of electronic voting protocols were proposed with the growth of cryptographic technology. For a robust e-voting protocol, it should be designed to fulfill a series of security requirements such as transparency, accuracy, verifiability, integrity, privacy and availability, and so on. However, there are two other threats in the electronic voting process: Voter-coercion and vote-buying, which are not considered in the most of electronic voting protocols. Vote-buying means that briber buys a special ballot from a voter, and voter need to prove this special ballot to the briber. Voter-coercion refers to that adversary choose voters to coerce and require them to vote according to adversary's intentions. Many electronic voting protocols are designed without considering these two problems, which is crucial to construct a fair and democratic e-voting system.

To avoid the above two threats, Benaloh and Tuinstra proposed the notion of receipt-free in 1994 [3]. Receipt-free means that a voter cannot prove to any adversary that he voted in a special manner, even if the voter want to do so [30]. Although the receipt can be a evidence that voters have cast their ballots, it may be utilized by adversary to coerce voters. This property ensures that voter cannot be coerced by requiring to vote for one candidate, and adversary cannot determine voter's behavior about voting, so as to protect against vote-buying and voter-coercion. Coercion-resistance is a more extensive and stronger notion than receipt-free. While satisfying the demand of receipt-free, a coercion-resistant voting system

can also resist randomization attack, forced-abstention attack and simulation attack [22], which are real existential threats in the real world.

Scholars have presented a lot of coercion-resistant e-voting schemes. However, many of these schemes were implemented under strong physical assumptions and constraints, which is a disadvantage for the application of e-voting system. In 2010, Meng and Wang first proposed a coercion-resistant e-voting protocol without physical assumptions and constraints [27]. The protocol was implemented using deniable encryption, which enables the sender (or receiver or both parties) of a message to deny the true parameters used in the encryption process and give a fake parameter, and keep the coercer from discovering. Deniable encryption provides a new direction for designing a coercion-resistant e-voting scheme without physical constraint.

The adaptive issue in e-voting protocols, which is that some authorities maybe know the counting results before the polls closed, cannot be ignored. In traditional e-voting schemes, there may be situations where the last voter already knows the final voting result of other voters before casting. Further, This voter might be affected by the result. The timed-release encryption refers to that an encrypted message only can be decrypted after an appointed time. In recent years, many proposed e-voting protocols used timed-release encryption to solve the adaptive issues in the voting process.

In most e-voting protocols with verifiability, it usually assumes that there is a public bulletin board to provide a view to all voters [26]. The bulletin board is a storage system where all the public information is published for anyone to see, review and audit, which has some similarities to the blockchain technology. The blockchain technique is a distributed data structure that is replicated among the nodes in a network [7], and the core advantages of it is tamper-proof and transparency. Transactions and information will be shared between different nodes in the blockchain network [1]. In a blockchain, the nodes have the ability to operate peer-to-peer transactions without mutual trust by means of data signature, timestamp, distributed consensus and economic incentives and so on, which prevents the transaction data from being tempered and revealed. Therefore, blockchain technology is suitable for constructing an e-voting system. This paper achieves integrity, fairness and privacy preservation in the proposed e-voting system by employing the blockchain technique and smart contracts.

1.1 Our Contribution

In this paper, based on the blockchain technology, we design a coercion-resistant e-voting system with receiver-deniable encryption. The advantages of our e-voting system are as follows:

1) We propose a voting scheme on the basis of the property of coercion-resistant of JCJ scheme [22]. Concretely, we adopt a receiver-deniable encryption to issue voter's credential. According to this method, the voter is distributed two credentials (valid one and false one), and provides the coercer with false credential when it is coerced.

2) Our e-voting system is constructed based on the blockchain technology. Each participant shares all the information of the system and can verify the authenticity and integrity of the information recorded on the blockchain. Therefore, our proposed system can ensure the fairness and verifiability of the voting process and counting result.

3) We design a time-release encryption algorithm based on M-ElGamal encryption to run the counting operation after a specified time. It can avoid election fraud which means all participants involved do not have access to the results in advance.

4) We provide a specific security analysis similar to JCJ model to prove the proposed e-voting system is coercion-resistant under the Decisional Diffie-Hellman (DDH) assumption. Coerced voters offer the adversary an invalid credential, which will be used to vote according to adversary's instruction, then cast their ballots with valid credential in a privacy time. In the meantime, these operations will not be detected by the adversary.

1.2 Realted Work

In 1981, Chaum published the first paper on encrypted e-voting scheme [6], he took advantage of an anonymous channel and public key cryptosystem to encrypt votes. However, the election has to be restarted if there is a failure of single voter. Subsequently, DeMillo proposed a protocol, which calls for all voters to participate in the voting and ballots should be cast after being encrypted [14]. In 1985, Cohen and Fisher proposed an e-voting protocol based on homomorphic cryptography for secure voting [9]. However, it requires that all voting stages must be operated at the same time. In 1992, Fujioka *et al.* proposed an e-voting scheme that suitable for large-scale elections [17], which took blind signature to endorse voters' messages in the process of voting and sent the results to the administrator. This protocol satisfied the properties of privacy and fairness. Based on this excellent protocol, a number of electronic voting software have been implemented in the real-world application scenarios, such as the MIT's EVOX system [15] and Sensus of Washington University [12].

As e-voting system develops, a series of crimes associated with this emerged, such as vote buying, coercing and bribing voters. In order to overcome these challenges, it has put forward some new requirements and properties on e-voting system including receipt-free, coercion-resistant, and so forth. The concept of receipt-free was first proposed by Benaloh in 1994 [2], which meant that voters have the incapability of constructing a receipt as a evidence that they cast a specific vote, even if voters

wished to. However, Hirt [19] believed that Benaloh's protocol would be invalid if there existed more than one counting organization. In 1995, Sako and Kilian proposed the first receipt-free protocol based on mix network [32], which was deployed on the premise that there was an untappable private channel between Vote Center and voters. But it is not suited to large-scale elections. A year later, Okamto adopted a non-anonymous channel, a private channel, and a bulletin board to construct a receipt-free voting scheme that apply to large scale elections [29]. But in 1998, Okamto pointed out [30] there were receipts for this voting scheme. Therefore, Okamto modified the scheme [29] and proved revised schemes can fulfill receipt-free by using an untappable channel or a voting booth.

Generally, a coercion-resistant voting system is one that the voter enables to cheat the coercer that he has voted according to the coercer's command, while the voter has in fact voted as his own intentions. However, once the coercer has the power to know whether or not voters voted as ordered, the coercer is able to influence the election process or blackmailing voters. Previous researches of coercion resistance had been restricted to the receipt-free. In 2002, Juels, Catalano and Jakobsson(JCJ) first introduced the strong concept of coercion-resistant in electronic voting protocol and proposed a voting scheme [22]. They also noted that a coercion-resistant voting system is not only receipt-free, but also can resist some attacks, for example the simulation attack. In 2008, based on JCJ scheme [22], Clarkson *et al.* proposed a coercion-resistant e-voting system, called Civitas [8]. It is the first electronic voting system that is coercion-resistant, universally and voter verifiable, and suitable for remote voting.

In 1996, Canett presented a concept which is called deniable encryption [5]. In 2010, Meng and Wang [28] proposed an receiver deniable encryption scheme on the base of the thought of Klonowski *et al.* [24] and a cryptosystem [4]. On this basis, they proposed a coercion-resistant remote Internet voting protocol in the same year [27]. Compared with other coercion-resistant voting scheme, this one was implemented without physical assumptions.

There are two main approaches to implement time-release encryption: One is on the base of time-lock puzzles, the other is to add a time-server. The former refers to a computational problem which cannot be solved until a computer run the puzzles continuously for a definite amount of time. A time-server is a trusted agent who will never reveal information until a specified date. The time server provides a time reference for users by periodically releasing time-trapdoor information which is necessary for decrypting the ciphertext at the release-time. The scheme of realizing the time-release by using a time-server was first proposed by Crescenzo [13] in 1999. However, the anonymity of the scheme cannot be guaranteed. In this paper, we design a time-release encryption algorithm by introducing a time-server to make the counting operation only can take place after a specified time, which guarantee the fairness of voting process. Meanwhile, our proposed scheme can get rid of the trust dependence on the time-server.

In recent years, with the extensive application of blockchain technology, it has become a new trend to design blockchain-aided electronic voting system. Shahzad *et al.* proposed a scheme to adjust the block generation and sealing, and then achieve the trustworthiness and fairness of the voting by altering the hash function in the blockchain [33]. McCorry *et al.* [26] presented a decentralized and self-tallying e-voting protocol on the blockchain with maximum privacy for the voter. They arranged a smart contract on Ethereum for the board of elections, which relied on no trusted authority to tally all the votes and preserve voter privacy. Yang *et al.* constructed a e-voting system based on the Ethereum blockchain [35], which adopted the homomorphic encryption to ensure vote's confidentiality. Depending on IoT equipment, Li *et al.* proposed a self-tallying voting scheme based on blockchain [25], which used time-locked encryption to solve adaptive issue in self-tallying voting system. However, the protocol cannot support multi-candidate voting.

Although these blockchain-enabled e-voting schemes satisfy some basic security requirements, few of them take stress coercion-resistant into consideration. Hardwick *et al.* made use of the blockchain as a transparent ballot box to achieve a voting system of open, fair, and independently verifiable [18]. The system has the property of forgiveness, which is able be regarded as a feeblish version of coercion-resistant. In other words, this system does not really achieve coercion-resistance. Yu *et al.* put forward a platform-independent secure and verifiable voting scheme on the blockchain [36]. This proposal analyzed the property of coercion-resistance and drew the conclusion that their e-voting system was receipt-free and even free from double voting and randomization attack. Unfortunately, their scheme is incapable of resisting the forced-abstention attack. Thus, it can be seen that there are few electronic voting schemes that can truly realize coercion-resistance. To solve this problem, this paper provides a blockchain-based coercion-resistant e-voting system.

1.3 Organization

The rest of this article is organized as follows: Section 2 descries the computational assumptions and cryptographic techniques used in our proposed system. Section 3 introduces the system model and the security model of our proposed system. Section 4 presents the proposed blockchain-based coercion-resistant e-voting system. The security proof and performance analysis are given in Sections 5 and 6, respectively. In the end, we conclude the paper in Section 7.

2 Preliminaries

This part introduces some assumptions and preliminaries related to our construction. We modify a deniable encryption scheme to satisfy our setting. Specifically, we employ a revised version of the ElGamal scheme.

2.1 Computational Assumption

(1) **Partial Discrete Logarithm Problem (PDLP) Assumption.**
Let $N = pq$ be a a safe-prime modulus where p and q are primes of $p = 2p' + 1$ and $q = 2q' + 1$, and p' and q' are prime numbers. $G = QR_{N^2}$ is a cyclical group of $Z^*_{N^2}$, which the order of group G is denoted by $ord(G) = \lambda(N^2)/2 = pp'qq' = N\lambda(N)/2$, with $\lambda(N) = 2p'q'$. The element with order N is denoted as $\alpha = (1 + KN)$. Paillier introduced the partial discrete logarithm problem (PDLP) [31] as a new computational problem. The problem is stated as follows.

Definition 1. *Let g denote an element of maximal order in QR_{N^2}. Given g and $h = g^a \bmod N^2$ for some $a \in [1, ordQR_{N^2}]$. And $ordQR_{N^2}$ denotes the order of QR_{N^2}. The PDLP over QR_{N^2} is to find a.*

The author in [4] assumed that this problem is difficult if people can not know the factorization of the modulus N, as follows.

Assumption. For each probabilistic polynomial time algorithm $\mathcal{A}$, there is a negligible function $negl()$ such that for large enough ℓ,

$$\mathbf{Pr}\left[\begin{array}{c|c} \mathcal{A}(N,g,h) & p,q \leftarrow SP(\ell/2); N = p \times q; \\ = a \bmod N & g \leftarrow QR_{N^2}; a \leftarrow [1, ordQR_{N^2}]; \\ & h = g^a \bmod N^2 \end{array}\right] = negl(\ell)$$

Pallier also proved that this problem can be efficiently solved when know the factorization of the modulus [31].

(2) **Decisional Diffie-Hellman (DDH) Assumption.**
Let $G =< g >$ be a cyclic group of order p, where p is a prime number. Decisional Diffie-Hellman (DDH) problem [23] states that given random g^a and g^b for $a, b \in Z_p$, draw a distinction between g^{ab} and a random value g^c. Decisional Diffie-Hellman assumption states that the advantage $Adv^{DDH}_{\mathcal{A}}(\lambda)$ is negligible in security parameter λ for any probabilistic polynomial time algorithm $\mathcal{A}$, as shown in follows.

$$\begin{aligned} Adv^{DDH}_{\mathcal{A}}(\lambda) = &|\mathbf{Pr}[C(g, g^a, g^b, g^{ab} = 1)] \\ &- \mathbf{Pr}[C(g, g^a, g^b, g^c) = 1]| \end{aligned}$$

2.2 Modified ElGamal Encryption

In this paper, a modified version of the ElGamal scheme is adopted, which simplified the Cramer-Shoup cryptosystem [10].

Key Generation. Suppose there is a multiplicative cyclic group G of prime order p, and g_1, g_2 are generators of G. Choose two random integer $x_1, x_2 \in Z_p$ as the private key. Compute $y = g_1^{x_1} g_2^{x_2} \bmod p$. The corresponding public key is (G, p, g_1, g_2, y).

Encryption. To encrypt a message $m \in G$: Choose a integer $r \in Z_p$, the ciphertext (A, B, C) can be computed as follows:

$$A = g_1^r \bmod p, B = g_2^r \bmod p, C = my^r \bmod p.$$

Decryption. The message m can be obtained by computing $m = C/(A^{x_1} B^{x_2}) \bmod p$.

2.3 BCP Cryptosystem

Bresson, Catalano, and Pointcheval (BCP) introduced a public-key cryptosystem with a double trapdoor decryption mechanism [4]. The idea of this scheme is based on [11]. It offers two different decryption methods, which can be used within deniable encryption. It is also an additively homomorphic variant of the ElGamal cryptosystem [16].

Key Generation. Let p and q be prime number and $p = 2p' + 1$ and $q = 2q' + 1$, which p' and q' are also prime numbers. There is $N = pq$. $G = QR_{N^2}$ is the cyclic group of quadratic residues modulo N^2. Let h and g denote elements of maximal order in G. The message space is Z_N. Select a random element $\alpha \in Z^*_{N^2}$ and a random value $a \in [1, ord(G)]$. Set $g = \alpha^2 \bmod N^2$ and $h = g^a \bmod N^2$. We have $ord(G) = \lambda(N^2)/2 = pp'qq' = N\lambda(N)/2$, with $\lambda(N) = 2p'q'$. $N\lambda(N)/2$ is the maximal order of an element in G. The private key is a, and the corresponding public key is (N, g, h).

Encryption procedure. Let $m \in Z_N$ represent the message to be encrypted, and then choose a random element r uniformly in Z_{N^2}. Finally, the ciphertext (A, B) can be calculated as follows:

$$A = g^r \bmod N^2, B = h^r(1 + mN) \bmod N^2.$$

Decryption procedure. The message m can be computed when knowing a:

$$m = \frac{B/A^a - 1}{N} \bmod N^2.$$

Commitment scheme. The sender chooses $r' \in_R Z_{N\lambda(N)/2}$ and then commit to a message $m' \in Z_N$:

$$\mathcal{C}(r', m') = h^{r'}(1 + m'N) \bmod N^2.$$

3 System Model and Security Model

This part first puts forward the system model of our coercion-resistant e-voting system which is based on blockchain, depicted in Figure 1. Then we introduce the security model of our proposed system. In the end, we list the symbols involved in the scheme.

3.1 Entities

Supervisor: A supervisor is required to initialize and publish details of voting, upload the candidate list and the list of registered voters to the block, and deploy the smart contract. Supervisor are selected by all nodes from the ordinary users in the blockchain before a vote.

Voter: A voter is a person who have a right to cast a ballot for one candidate. They should register in the registration authority first. Set there are n voters, and a certain voter is denoted as voter V_i.

Candidate: The person who will be chosen to be the subject of voter's ballots before the operation of system. Let $\vec{C}$ denote the shortlist which is produced by the registration authority.

Registration Authority (RA): A registration authority is deemed necessary for providing assurance that only eligible voters are able to vote. To ensure the validity of user's identity, they must register in RA and get an Identity Certificate. In our voting scheme,we suppose RA is fully credible.

Smart Contract: Smart contract is a protocol with specific rules in blockchain. It cannot be modified once the system starts running. The functions of smart contract include:

1) Verify the validity of the ballots.
2) Verify the validity of the voter's credential.
3) Publish the voter roll, candidate roll and the ballots to the blockchain.
4) Tally and publish the final result.

Time server (TS): An third party that can verify whether the current time is release-time of ballots, if so, it will sends the information needed for the decryption of ballots to smart contract.

3.2 Functions

We define a voting scheme $\mathcal{VS}$ consists of following stages:

Setup: Generating key pairs of RA, voters, Smart contract and TS.

Registration: RA checks the identity and eligibility of would-be voters. If the verification is successful, RA provides them a credential. Then RA will publish a voter roll $\vec{L}$ and a candidate roll $\vec{C}$ to blockchain.

Voting: Registered voters choose a candidate from candidate slate $\vec{C}$ and cast ballots using their credentials and private keys.

Verification: Smart contract verifies the validity of proofs and credentials. TS verifies the timestamp freshness limitation.

Tallying: After the voting deadline, smart contract computes the voting results and publish it to blockchain.

3.3 Security Model

For the security of an e-voting system, the e-voting protocol must be designed to satisfy following security requirement.

Privacy: All ballots submitted by voters have to be kept in a safe and secret environment. No one can reveal any information of the ballots.

Fairness: This property means that in the process of voting, no one can obtain a early results before the end of the voting, which prevents the remaining voters from being influenced in their vote.

Eligibility: Eligibility states that only authorized voters can be allowed to cast their ballots.

Coercion Resistance: This property can be considered an extension of the privacy. There exists an assumption that the coercer is likely to interact with voters. Even the coercer may claim voter's credential to vote, or may instruct voters to vote for a specified candidate. In a coercion-resistant e-voting system, coerced voters can deceive the adversary by providing a fake credential. They pretend to submit to the coercion and vote for adversary's choice using the fake credential, but cast their ballots with valid credential in a privacy time.

We assume the registration authority RA is fully trustworthy in this work. The adversary can coerce a voter either prior or during the registration stage. Voter can evade coercer by giving a fake credential to adversary and cast their own ballots at a private time. So we assume the adversary cannot distinguish a fake credential from valid one. We also assume there is an anonymous channel. Voter's ballots are submitted through anonymous channels, so that adversary cannot know when the coerced voter cast their ballot.

The definition of coercion resistance centers on a game between an adversary $\mathcal{A}$ and a voter that the target of coercion attack. Let k_1, k_2, k_3 be the security parameters. n_V denotes the total number of eligible voters, n_C denotes the number of candidates, and n_A denotes the

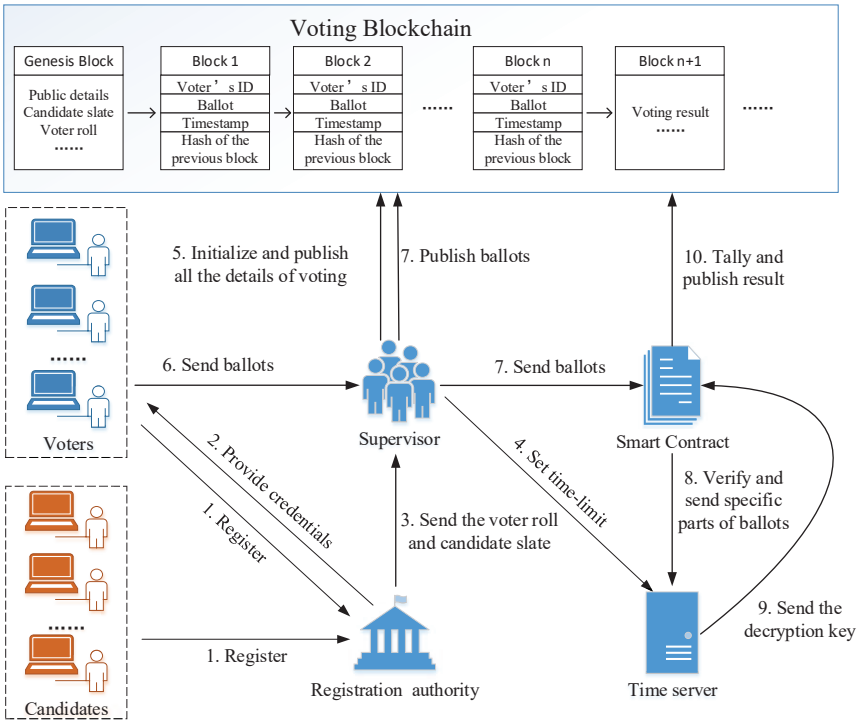


Figure 1: Voting system

number of voters which may be completely corrupted by the adversary. We use D_{n,n_C} to characterize the probability distribution of the voting pattern of honest voters. Let ϕ denote a null ballot, and λ represent a ballot that contains an invalid credential. BC denotes blockchain. A coin b is flipped: If $b = 1$, voter gives $\mathcal{A}$ a valid credential and follows the instruction; If $b = 0$, voter gives $\mathcal{A}$ a fake credential and casts a ballot of choice β in a private time. The goal of $\mathcal{A}$ is to guess whether the value of b is 0 or 1. After this, $\mathcal{A}$ can determine what kind of the behavior the voter has adopted during the period of implementation of $\mathcal{VS}$.

We present experiment *c-resist* according to the experiment in JCJ protocol [22]. This experiment defines a game between coercer $\mathcal{A}$ and voter who is target of coercion, as shown in Experiment 1. We consider that adversary $\mathcal{A}$ may corrupts voters prior to the registration phase. Voters get a key pair (sk_i, pk_i) after registering with the identity information i. Then $\mathcal{A}$ choose a target voter j and his choice β to coerce. If $b = 0$, then the voter provides the adversary with a fake private key sk^* (generated by RA according to voter's key pair) and cast a ballot of his own choice β to blochchain. If $b = 1$, then the voter provide the adversary with a valid private key, i.e. voter submits to the coercion. Honest voters cast their ballots to blockchain according to distribution D_{n,n_C}. Then the smart contract tallies the voting result and get a final result $\vec{X}$ and a proof P that the count is correct. Adversary $\mathcal{A}$ guess the value of b according to $\vec{X}$ and P. This experiment describes the possibilities of an adversary $\mathcal{A}$ in a "real world". And the success probability of the adversary $\mathcal{A}$ can be defined as $\mathbf{Succ}^{c\text{-}resist}_{\mathcal{VS},\mathcal{A}}(k_1, k_2, k_3, n_V, n_C, n_A) = \mathbf{Pr}[\mathbf{EXP}^{c\text{-}resist}_{\mathcal{VS},\mathcal{A}}(k_1, k_2, k_3, n_V, n_C, n_A) = 1]$.

Experiment 1 $\mathbf{EXP}^{c\text{-}resist}_{\mathcal{VS},\mathcal{A}}(k_1, k_2, k_3, n_V, n_C, n_A)$

$V \leftarrow \mathcal{A}$(voter names, "control voters");
Initialize the *observations storage*.
$\{(sk_i, pk_i) \leftarrow register(SK_R, i, k_1)\}_{i=1}^{n_V}$;
$(j, \beta) \leftarrow \mathcal{A}(\{sk_i\}_{i \in V}$, "set target voter and vote");
if $|V| \neq n_A$ or $j \notin \{1, 2, \ldots, n_V\} - V$ **then**
  output '0'
end if
$b \in_U \{0, 1\}$;
if $b = 0$ **then**
  $sk^* \leftarrow fake(sk_j, pk_j, SK_R)$;
  $BC \Leftarrow vote(sk_j, PK_S, n_C, \beta, k_2)$;
else
  $sk^* \leftarrow sk_j$;
  $BC \Leftarrow vote(\{sk_i\}_{i \neq j}, PK_S, n_C, D_{n,n_C})$;
  $BC \Leftarrow \mathcal{A}(sk^*, BC$, "cast ballots");
  $(\vec{X}, P) \leftarrow tally(SK_S, BC, n_C, \{pk_i\}_{i=1}^{n_V}, k_3)$;
  $b' \leftarrow \mathcal{A}(\vec{X}, P$, "guess b");
end if
if $b' = b$ **then**
  output '1';
else
  output '0';
end if

The adversary $\mathcal{A}$ in the above experiment is very powerful, he can completely coerce the targeted vote. So second adversary $\mathcal{A}'$ is employed to describe the success of $\mathcal{A}$ by a comparison. In other words, to describe a security

requirement that we would like to achieve in $\mathcal{VS}$.

In the *c-resist-ideal* experiment, we assume that voters are always want to evade coercion. Further, the voter always provides a fake key. In other words, $\mathcal{A}'$ will learn noting about the private key. We use two new functions in the ideal experiment: *revote* and *ideal-tally*. Whatever the value of b, voter always provides the adversary with a fake private key, and smart contract tallies the ballots posted to BC according to the pre-defined rules. Specially, only tally the ballots with valid credential. See Experiment 2 for experiment $\mathbf{EXP}^{c\text{-}resist\text{-}ideal}_{\mathcal{VS},\mathcal{A}'}$.

Experiment 2 $\mathbf{EXP}^{c\text{-}resist\text{-}ideal}_{\mathcal{VS},\mathcal{A}'}(k_1,k_2,k_3,n_V,n_C,n_A)$

$V \leftarrow \mathcal{A}'$(voter names, "control voters");
$\{(sk_i,pk_i) \leftarrow register(SK_R,i,k_1)\}_{i=1}^{n_V}$;
$(j,\beta) \leftarrow \mathcal{A}'(\{sk_i\}_{i\in V}$, "set target voter and vote");
if $|V| \neq n_A$ or $j \notin \{1,2,\ldots,n_V\} - V$ **then**
  output '0'
end if
$b \in_U \{0,1\}$;
if $b = 0$ **then**
  $sk^* \leftarrow fake(sk_j,pk_j,SK_R)$;
else
  $sk^* \leftarrow fake(sk_j,pk_j,SK_R)$;
end if
$BC \Leftarrow vote(\{sk_i\}_{i\neq j},PK_S,n_C,D_{n,n_C})$;
$BC \Leftarrow \mathcal{A}'(sk^*,BC$, "cast ballots");
$BC \Leftarrow revote(sk_j,PK_S,n_C,\beta,k_2)$;
$(\vec{X},P) \leftarrow ideal\text{-}tally(SK_S,BC,n_C,\{pk_i\}_{i=1}^{n_V},k_3)$;
$b' \leftarrow \mathcal{A}'(\vec{X},P$, "guess b");
if $b' = b$ **then**
  output '1';
else
  output '0';
end if

Correctness: We define a notion of correctness follows that of Juels *et al.* [22]. We model voters as posting a series of ballots. We call a tallying is correct when, regardless of the behavior of corrupted voters, only the ballot that closest to the deadline of each valid credential can be counted. In the experiment characterizing correctness, we assume not only can the adversary choose a set V of voters to corrupt, but also can select the candidate slate n_C and ballots that will be submitted by honest voters. The goal of the adversary is to make more than $|V|$ ballots to be counted into the end result, or to alter the ballot of at least one honest voter, or to delete the ballot of at least one honest voter.

See the $\mathbf{EXP}^{corr}_{\mathcal{VS},\mathcal{A}}$ in Experiment 3, our voting protocol is correct if $\mathbf{Succ}^{corr}_{\mathcal{VS},\mathcal{A}}(k_1,k_2,k_3,n_V,n_C,n_A)$ is negligible in k_1, k_2, k_3 for any adversary $\mathcal{A}$.

Verifiability: This property is that any participants has the ability to check the correctness of the tally X, and find whether smart contract has misbehavior while running the function *tally*. See Experiment 4 for experiment $\mathbf{Exp}^{ver}_{\mathcal{VS},\mathcal{A}}$.

Experiment 3 $\mathbf{Exp}^{corr}_{\mathcal{VS},\mathcal{A}}(k_1,k_2,k_3,n_V,n_C,n_A)$

$\{(sk_i,pk_i) \leftarrow register(SK_R,i,k_1)\}_{i=1}^{n_V}$;
/*voters are registered*/
$V \leftarrow \mathcal{A}(\{pk_i\}_{i=1}^{n_V}$ "choose controlled voter set");
/*$\mathcal{A}$ corrupts voters*/
$\{\beta_i\}_{i\notin V} \leftarrow \mathcal{A}$("choose votes from honest voters");
/*$\mathcal{A}$ voters from honest voters*/
$BC \Leftarrow \{vote(sk_i,PK_S,n_C,\beta_i)\}_{i\notin V}$;
/*honest voters cast ballots*/
$(\vec{X},P) \leftarrow tally(SK_S,BC,n_C,\{pk_i\}_{i=1}^{n_V},k_2)$;
/*honest ballots are counted*/
$BC \Leftarrow \mathcal{A}(BC$, "cast ballots");
/*$\mathcal{A}$ post ballots to BC*/
$(\vec{X}',P') \leftarrow tally(SK_S,BC,n_C,\{pk_i\}_{i=1}^{n_V},k_3)$;
/*all ballots are counted*/
if $verify(PK_S,BC,n_C,\vec{X}',P')$='1' and
$(\{\beta_i\} \not\subset \langle\vec{X}'\rangle)or|\langle\vec{X}'\rangle| - |\langle\vec{X}\rangle| > |V|$ **then**
  output '1';
else
  output '0';
end if

Experiment 4 $\mathbf{Exp}^{ver}_{\mathcal{VS},\mathcal{A}}(k_1,k_2,k_3,n_V,n_C,n_A)$

$\{(sk_i,pk_i) \leftarrow register(SK_R,i,k_1)\}_{i=1}^{n_V}$;
/*voters are registered*/
$(BC,\vec{X},P) \Leftarrow \mathcal{A}(SK_S,\{(sk_i,pk_i)\}_{i=1}^{n_V},k_2$, "forge election"); /*$\mathcal{A}$ concocts full election*/
$(\vec{X}',P') \leftarrow tally(SK_S,BC,n_C,\{pk_i\}_{i=1}^{n_V},k_3)$;
/*$tally$ is taken on BC*/
if $\vec{X} \neq \vec{X}'$ /*does $\mathcal{A}$'s differ from BC*/
and $verify(PK_S,BC,n_C,\vec{X},P')$='1' **then**
  output '1';
else
  output '0';
end if

We say that our e-voting system satisfies verifiability if quantity $\mathbf{Succ}^{ver}_{\mathcal{VS},\mathcal{A}}(k_1,k_2,k_3,n_V)$ is negligible for all n_V and the adversary $\mathcal{A}$ with polynomial running time. The symbols used to explain our scheme are shown in Table 1.

4 Concrete Scheme

4.1 The Time-Release Encryption Scheme

We design a time-release encryption scheme based on M-ElGamal encryption, which enables the sender to encrypt the message so that it cannot be decrypted before the

Table 1: List of the symbols used in our scheme

Notation	Meaning
V_i	The voter i.
σ_i	The credential of voter$i$.
σ_i'	The fake credential.
$\vec{L}$	The list of voter's credential.
C_j	The candidate choice of voter i.
C_j'	The choice of the coercer.
$\vec{C}$	The list of candidates.
T	The designated condition of time.
$H(.)$	A one-way hash function.
(Pk_S, Sk_S)	The key pair of smart contract.
(Pk_T, Sk_T)	The key pair of Time-Server.
(Pk_R, Sk_R)	The key pair of Registration Authority.
$Enc_{Pk}(m)$	Encrypt the message m with public key.
$Sig_{Sk}(m)$	Digital signature of message m.

appointed time. We assume there are three parties in this protocol: Sender, agent and time-server. The sender encrypts a message using his public key and gives the agent a delegation key, which enables the agent to decrypt the ciphertext under the control of the time server. The process of this protocol is as follows:

1) We assume the public key of sender is (g_1, g_2, h), and the corresponding private key is $s_1, s_2 \in Z_p$ such that $h = g_1^{s_1} g_2^{s_2}$. g_1, g_2 are the generators of G. Let (Pk_T, Sk_T) denote the public key and private key of time-server. There is a time condition T generated by the sender on the decryption. Only when the proxy fulfills the condition can he decrypt the ciphertext. The sender encrypts the message m according to the M-ElGamal encryption scheme using the public key, then gets the ciphertext $c = (A, B, C) = (g_1^r, g_2^r, mh^r)$, where $r \in Z_p$ is a random chosen by sender. Then he choose ξ randomly and computes the delegation key $\varphi =< u_T, u_{V_1}, u_{V_2} >=< Enc_{Pk_T}(\xi), s_1 - H(T, \xi), s_2 - H(T, \xi) >$, then sends (c, φ, T) to proxy.

2) The proxy sends (u_T, A, B, T) to the time-server after receiving (c, φ, T). Then the time-server will verify whether the proxy fulfill the condition T. If condition are met, the time-server computes $M_T = A^{H(T, Dec(Sk_T, u_T))}$ and $M_{T'} = B^{H(T, Dec(Sk_T, u_T))}$ needed for decryption and sends them to the proxy.

3) Finally, the proxy can get m according to $m = CA^{-u_{V_1}} B^{-u_{V_2}} M_T^{-1} (M_T^{'})^{-1}$.

In our time-release scheme, the private key of sender cannot be disclosed to the proxy or the time-server. Proxy can decrypt the ciphertext without the private key of sender. And time-server cannot decrypt the ciphertext.

4.2 The E-voting Scheme

We provide the schematic diagram of our e-voting scheme, as Figure 2 shows.

4.2.1 Initialization

Voter chooses a random element $\alpha \in Z_{N^2}^*$ and sets $g = \alpha^2 \bmod N^2$, publishes (N, g) publicly. RA can get (N, g) and choose a random number $x_1 \in [1, ord(G)]$, then compute $h = g^{x_1} \bmod N^2$, publish (h) publicly.

The first public key of voter is given by the triplet (N, g, h), and the corresponding private key is (p, q). The public key of RA is the triplet (N, g, h), and private key is x_1 based on BCP cryptosystem. Voter can get x_1 from $N = p \times q$ and $h = g^{x_1} \bmod N^2$ because of PDLP assumption, then he randomly chooses $x_2 \in Z_p$ and generates $g_1, g_2 \in G$, output the second public key (g_1, g_2, y) and corresponding secret key (x_1, x_2) based on M-ElGamal cryptosystem, wherein $y = g_1^{x_1} g_2^{x_2}$.

The key rules in the smart contract are set by RA, and this part of the rules is not visible to the participants in the voting system. The key pairs of smart contract and TS are expressed as (Pk_S, Sk_S) and (Pk_T, Sk_T) respectively.

RA generates a candidate slate $\vec{C}$ and send it to supervisor, complete with a proof of this slate. Then supervisor check the proof and upload it to the blockchain. Furthermore, supervisor initializes the voting system and publishes the details of voting publicly, which include a time limit T on the ballots.

4.2.2 Registration

In this work, we modify Meng *et al.*'s [28] scheme to issue the certificate, which can achieve deniable encryption. Specially, we adopt a modified version of the basic ElGamal scheme to satisfy our setting. And the modified version is semantically secure under the decisional Diffie-Hellman assumption. This process can be finished before voting begins. The specific implementation steps are as follows:

1) The participant in blockchain who wants to be legitimate voter sends a registration request to RA, the request contains identity information and a security parameter.

2) RA verifies voter's identity information, and chooses a random number $\lambda_i \in< g >$ and generate a string $\sigma_i \in Z_N$ according to voter's identity information as the credential of V_i. Then RA compute commitment $\mathcal{C} = \mathcal{C}(\sigma, \lambda) = h^{\lambda_i}(1 + \sigma_i N) \bmod N^2$ to the credential according to the commitment scheme proposed by Bresson *et al.* [4]. Generating a fake credential $\sigma_i^{'} \in Z_N$, RA can calculate $\mu_i = \lambda_i + (\sigma_i - \sigma_i^{'}) d\lambda(N) \bmod N\lambda(N)/2$ according to $\mathcal{C} = \mathcal{C}(\sigma, \lambda) = \mathcal{C}(\sigma^{'}, \mu)$. The d here is the inverse of k which is such that $h^{\lambda(N)} = (1 + kN) \bmod N^2$. Then RA computes

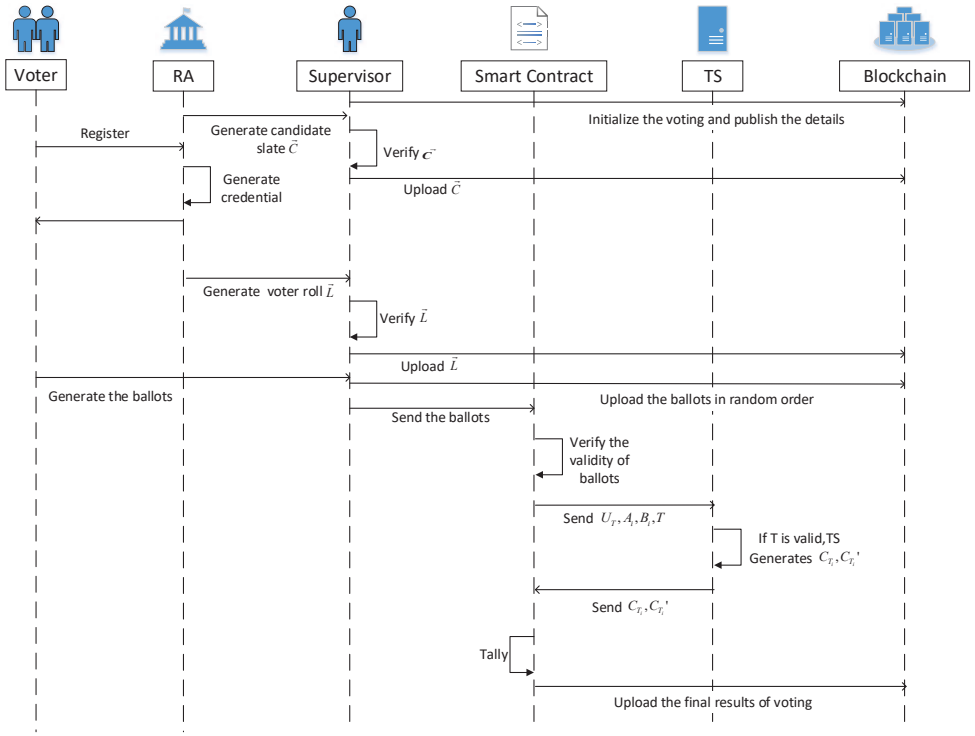


Figure 2: The schematic diagram of protocol timing

$(A^{'} = g_1^{H(\lambda_i)} \cdot \mu_i, B^{'} = g_2^{H(\lambda_i)}, C^{'} = (y^{H(\lambda_i)} \cdot \mu_i^{x_1}) \cdot \lambda_i)$ using M-ElGamal cryptosystem. Finally, sending the ciphertext $(A^{'}, B^{'}, C^{'})$ and the commitment $\mathcal{C} = \mathcal{C}(\sigma, \lambda)$ to V_i.

3) After receiving the ciphertext and commitment sent by RA, V_i computes λ_i and μ_i as follows,

$$\begin{aligned}\lambda_i =& C^{'}(A^{'})^{-x_1}(B^{'})^{-x_2} \\ =& (y^{H(\lambda_i)} \cdot \mu_i^{x_1}) \cdot \lambda_i \cdot (g_1^{H(\lambda_i)} \\ & \cdot \mu_i)^{-x_1} \cdot (g_2^{H(\lambda_i)})^{-x_2} \\ \mu_i =& A^{'} \cdot g_1^{-H(\lambda_i)}\end{aligned}$$

Then he can recover the valid credential σ_i and the fake credential $\sigma_i^{'}$ according to

$$\sigma_i = \frac{B^{'}/(g^{\lambda_i} \bmod N^2)^a - 1 \bmod N^2}{N}$$

$$\sigma_i^{'} = \frac{B^{'}/(g^{\mu_i} \bmod N^2)^a - 1 \bmod N^2}{N}$$

4) RA send the voter roll $\vec{L} = \{Enc_{Pk_S}(\sigma_i), Sig_{Sk_R}(\sigma_i)\}$ to the supervisor, where $Enc_{Pk_S}(\sigma_i)$ refers to the ciphertext of σ_i encrypted by RA using Pk_S (the public key of the smart contract), $Sig_{Sk_R}(\sigma_i)$ refers to the signature of the identity certificate σ_i by RA.

5) The supervisor verifies whether RA's signature is valid or not. If so, uploading $\vec{L}$ to the blockchain.

In this phase, RA adopts deniable encryption scheme to send the certificates to voters. Voters can deny the real parameter used in the encryption procedure to others. If voters are coerced to reveal parameters, voters can reveal the false parameter to coercer without being discovered.

4.2.3 Voting

Voter create ballots and submit them to the smart contract through an anonymous channel. In this process, there are two cases:

1) **The voter V_i is not coerced.** V_i will choose a candidate C_j which he/she want to vote and encrypt the candidate choice with the public key $y = g_1^{x_1} g_2^{x_2}$ based on the M-ElGamal cryptosystem, then he can get the ciphertext $(A_i, B_i, C_i) = (g_1^{r_i}, g_2^{r_i}, C_j y^{r_i})$, r_i is chosen at random by V_i. The contents of a valid ballot is as follows:

$$\begin{aligned}Ballot =& \{(A_i, B_i, C_i), Sig_{Sk_{V_i}}((A_i, B_i, C_i)), \\ & Enc_{Pk_S}(\sigma_i), Sig_{Sk_R}(\sigma_i), \phi, T\}\end{aligned}$$

Here, $Sig_{Sk_{V_i}}((A_i, B_i, C_i))$ is a signature of V_i on the ciphertext (A_i, B_i, C_i). V_i chooses ξ and computes the key ϕ that needed for decrypting the encrypted ballot according to

$$\begin{aligned}\phi =& < U_T, U_{V_1}, U_{V_2} > \\ =& < Enc_{Pk_T}(\xi), x_1 - H(T, \xi), x_2 - H(T, \xi) >\end{aligned}$$

2) **The voter V_i is coerced.** V_i will use the fake credential $\sigma_i^{'}$ to generate a ballot. The contents of the

ballot is as follows:

$$Ballot = \{(A_i, B_i, C_i^{'}), Sig_{Sk_{V_i}}((A_i, B_i, C_i^{'})), Enc_{Pk_S}(\sigma_i^{'}), Sig_{Sk_R}(\sigma_i), \phi, T\}$$

Here, $(A_i, B_i, C_i^{'}) = (g_1^{r_i}, g_2^{r_i}, C_j^{'} y^{r_i})$. And the $C_j^{'}$ is a choice of the coercer. V_i can cast their valid ballots in their own private moments using the true credential.

4.2.4 Verification

The specific implementation steps of this process are as follows:

1) Smart contract verifies the validity of the signature of V_i and RA by using their public key.

2) Smart contract decrypts the ciphertext of σ_i, then verify the validity of σ_i by checking whether it is on the voter roll $\vec{L}$.

3) If σ_i and signatures are valid, smart contract will send U_T, A_i, B_i, T to TS. Note that U_{V_1}, U_{V_2} and C_i are not sent to TS. Then smart contract numbers each ballots at random and publishes (A_i, B_i, C_i), $Sig_{Sk_{V_i}}((A_i, B_i, C_i))$, $Sig_{Sk_R}(\sigma_i)$, ϕ, T, N_i to blockchain, wherein N_i is the serial-number of ballots. $Enc_{Pk_S}(\sigma_i)$ cannot be published to others except smart contract.

4) TS checks the validity of T after receiving it. If it is valid, TS will compute $C_{Ti} = A_i^{H(T, Dec(Sk_T, U_T))}$ and $C_{Ti}^{'} = B_i^{H(T, Dec(Sk_T, U_T))}$, then send them to the smart contract.

4.2.5 Tallying

After receiving C_{Ti} and $C_{Ti}^{'}$, smart contract can obtain the candidate choice of V_i by computing as below,

$$\begin{aligned} C_j &= C_i A_i^{-U_{V_1}} B_i^{-U_{V_2}} C_{Ti}^{-1} (C_{Ti}^{'})^{-1} \\ &= C_i A_i^{-U_{V_1}} B_i^{-U_{V_2}} A_i^{-H(T,\xi)} B_i^{-H(T,\xi)} \\ &= C_i A_i^{-x_1} B_i^{-x_2}. \end{aligned}$$

Our e-voting system allows voters to re-vote, so coerced voters are able to cast their ballots again after they have cast according to the adversary's demand. Because of the special counting rules, ballots cast according to the adversary's demand will not be counted. The counting rules in the smart contract:

1) Only count valid ballots, that is, the ballots which credential can be found on the voter roll;

2) For multiple valid ballots, only the one closest to the voting deadline can be counted. Finally, smart contract counts ballots and upload the results of election to the blockchain.

5 Security Analysis

Theorem 1. *Under the assumption that PDLP is hard, the above commitment is a perfectly hiding trapdoor commitment scheme.*

Proof. For any λ of $\mathcal{C}(\lambda, \sigma)$, when σ is evenly distributed in $Z_{N\lambda(N)/2}$, then $\mathcal{C}(\lambda, \sigma)$ is also evenly distributed in G. Because any $1 + \sigma N$ is in G, and h^{λ} is uniformly distributed in G, where h is generating element.

Provide $\mathcal{C} = \mathcal{C}(\lambda, \sigma)$ and the corresponding (λ, σ). One can find a collision for any $\sigma^{'}$ if he or she knows the factorization of the modulus. Let $d \in Z_N^*$ denote the inverse of k in $h^{\lambda(N)} = (1 + kN) \bmod N^2$. Thus, we can get,

$$\begin{aligned} \mathcal{C}(\lambda, \sigma) &= h^{\lambda}(1 + \sigma N) \bmod N^2 \\ &= h^{\lambda}(1 + kd\sigma N) \bmod N^2 \\ &= h^{\lambda + d\sigma\lambda(N)} \bmod N^2. \end{aligned}$$

And we also can get the μ as follows,

$$\mu = \lambda + (\sigma - \sigma^{'}) d\lambda(N) \bmod N\lambda(N)/2.$$

Given an input (N, h), we assume that exists an algorithm A that can find two couples (λ, σ) and $(\mu, \sigma^{'})$ such that $\mathcal{C} = \mathcal{C}(\sigma, \lambda) = \mathcal{C}(\sigma^{'}, \mu)$. At the same time, because here is $\sigma = \sigma^{'}$ if $\lambda = \mu$, we can assume that $\lambda \neq \mu$. Then one can get the equation as follows:

$$h^{\lambda}(1 + \sigma N) = h^{\mu}(1 + \sigma^{'} N) \bmod N^2$$

And thus, letting $\Delta_{\lambda} = \lambda - \mu$ and $\Delta_{\sigma} = \sigma^{'} - \sigma$,

$$h^{\Delta_{\lambda}}(1 + \Delta_{\sigma} N) \bmod N^2.$$

Since that h has the order $\lambda(N)N/2$ and $(1 + \Delta_{\sigma} N)$ has the order (at most) N, Δ_{λ} is a multiple of $\lambda(N)/2$. This is enough to factor according to the paper [4].

Theorem 2 (Coercion-resistant)**.** *Our e-voting scheme is Coercion-resistant under the DDH assumption.*

Proof. We prove the proposed e-voting system is coercion-resistant by closely following the JCJ techniques [22]. The formal definition of coercion resistance is described through a type of games between the coercer $\mathcal{A}$ and a targeted voter that will be coerced. we construct a polynomial-time algorithm $\mathcal{S}$, and assume it takes a set of ballots $W \in D_{n,n_C}$ of honest voters and simulates the $\mathcal{VS}$ in the experiment *c-resist*.

Under the DDH assumption, We prove that our proposed e-voting system is coercion-resistant when $\mathcal{A}$ cannot distinguish between the random ciphertexts provided by $\mathcal{S}$ and the ciphertexts that would be processed in a true execution of $\mathcal{VS}$.

The simulator will receive a quadruple (g_1, g_2, h_1, h_2) from the start. Whether or not the quadruple is a Diffie-Hellman (DH) one does depends on a hidden bit $\boldsymbol{d}$. If $\boldsymbol{d} = 1$, then the quadruple is a DH one, otherwise it is a random one. The goal of the simulator is to guess the value of $\boldsymbol{d}$. The simulation steps are as follows.

Setup: $\mathcal{S}$ chooses elements $x_1, x_2 \in_U Z_q$ uniformly and randomly, and computes $h = g_1^{x_1} g_2^{x_2} \bmod p$. Then the public key (g_1, g_2, h) can be published by $\mathcal{S}$.

Registration: $\mathcal{S}$ simulates the registrar RA. Every voter is assigned a credential σ_i. $\mathcal{S}$ encrypts the credential with the public key, then publishes the voter roll $\vec{L}$ and the candidate slate $\vec{C} = \{C_i\}_{i=1}^{n_C}$.

Corruption: The adversary selects n_A voters to corrupt, and a voter j with her honest choice β to coerce.

Coin Flip: A coin $b \in \{0, 1\}$ is flipped.

Credential Release: If $b = 0$, $\mathcal{S}$ gives $\mathcal{A}$ a fake credential $\sigma_j^{'}$ of coerced voter j; If $b = 1$, $\mathcal{A}$ is given the valid credential σ_j.

Honest voter simulation: Each ballot cast by honest voters contains two ciphertexts $(\alpha_{i,1}, \alpha_{i,1}^{'}, \beta_{i,1})$, (U_T, U_{V_1}, U_{V_2}). $\mathcal{S}$ posts these ballots and also simulates some proofs, like the signature of credential. Let $\vec{A}$ denote the list of these ballots, and $\vec{B}$ denote the list of the corresponding plaintext choices for each ballot in W.

The above ciphertexts is computed as follows,

$$Enc(C_j) = (\alpha_{i,1} = h_1^{r_i}, \alpha_{i,1}^{'} = h_2^{r_i}, \beta_{i,1} = h_1^{r_i x_1} h_2^{r_i x_2} C_j),$$

$$\begin{aligned}\phi &= < U_T, U_{V_1}, U_{V_2} > \\ &= < Enc_{Pk_T}(\xi), x_1 - H(T, \xi), x_2 - H(T, \xi) > .\end{aligned}$$

r_i and ϕ are random elements chosen by $\mathcal{S}$. The ballot has the format $(Enc(C_j), Enc_{Pk_S}(\sigma_i), proofs, \phi, T)$.

Adversarial ballots: A set of ballots $\vec{A^*}$ and corresponding proofs are posted by adversary $\mathcal{A}$.

Verification of ballots posted by the adversary: $\mathcal{S}$ simulates the smart contract to check the proofs in $\vec{A^*}$. Let $\vec{A_1}$ denote the list of ballots which have valid proofs.

Tallying: $\mathcal{S}$ simulates the time server (TS), generates the decryption key of ciphertext $Enc(C_j)$ and decrypts $Enc(C_j)$ and $Enc_{Pk_S}(\sigma_i)$. A credential may correspond to several ballots. Only the one closest to the deadline will be count. Let $\vec{B_1}$ be the list of associated ballots after decrypting, and $\vec{B_2}$ be the list of credentials. $\vec{E}$ denotes the union of $\vec{A_1}$ and $\vec{B_1}$. $\mathcal{S}$ simulates tallying in a straightforward way.

Guess: $\mathcal{A}$ decides the value of $b^{'}$.

Output: $\mathcal{S}$ output 1 if $b^{'} = b$, else output 0.

Apart from the data the adversary $\mathcal{A}$ produced, he can only see encrypted choices and associated signatures of choices and credentials. We denote the view of the adversary $\mathcal{A}$ is $\mathcal{V}$. Simulator will input $b^{'}$ as the guess for the DDH challenge after $\mathcal{A}$ outputs $b^{'}$.

When $b^{'}$ is a Diffie-Hellman triplet ($\boldsymbol{d} = 1$), then assuming $(g_1, g_2, h_1, h_2) = (g, g^a, g^b, g^{ab})$, and any ciphertext of the message m with a following form is actually a valid one.

$$\begin{aligned}(\alpha_{i,1}, \alpha_{i,1}^{'}, \beta_{i,1}) =& (h_1^{r_i}, h_2^{r_i}, h_1^{r_i x_1} h_{12}^{r_i x_2} m) \\ =& (g^{br_i}, g^{abr_i}, g^{br_i x_1} g^{abr_i x_2} m) \\ =& (g_1^{br_i}, g_2^{br_i}, h^{br_i} m)\end{aligned}$$

This means that

$$\begin{aligned}\mathbf{Pr}[S = 1 | \boldsymbol{d} = 1] =& \mathbf{Pr}[\mathbf{Exp}_{\mathcal{VS},\mathcal{A}}^{c\text{-}resist}(\mathcal{V}) = 1] \\ =& \mathbf{Succ}_{\mathcal{VS},\mathcal{A}}^{c\text{-}resist}(\mathcal{V})\end{aligned}$$

If $b^{'}$ is not a Diffie-Hellman triplet ($\boldsymbol{d} = 0$), then assuming $(g_1, g_2, h_1, h_2) = (g, g^a, g^b, g^c)$. This means that the ciphertext generated by simulator $\mathcal{S}$ reveal no information about the ballots cast by honest voters(in a strong information theoretic sense). Let $c^{'} = c/a$ and $c^{''} = c^{'}/b$.

$$\begin{aligned}(\alpha_{i,1}, \alpha_{i,1}^{'}, \beta_{i,1}) =& (h_1^{r_i}, h_2^{r_i}, h_1^{r_i x_1} h_{12}^{r_i x_2} m) \\ =& (g^{br_i}, g^{cr_i}, g^{br_i x_1} g^{cr_i x_2} m) \\ =& (g_1^{br_i}, g_2^{c^{'} r_i}, (g_1^{bx_1} g_2^{c^{'} x_2})^{r_i} m) \\ =& (g_1^{br_i}, g_2^{c^{''} br_i}, h^{br_i} g_2^{(c^{''} - 1) bx_2 r_i} m)\end{aligned}$$

This means that

$$\begin{aligned}\mathbf{Pr}[S = 1 | \boldsymbol{d} = 0] =& \mathbf{Pr}[\mathbf{Exp}_{\mathcal{VS},\mathcal{A}'}^{c\text{-}resist\text{-}ideal}(\mathcal{V}) = 1] \\ =& \mathbf{Succ}_{\mathcal{VS},\mathcal{A}'}^{c\text{-}resist\text{-}ideal}(\mathcal{V})\end{aligned}$$

Finally,

$$\begin{aligned}\mathbf{Adv}_{\mathcal{S}}^{DDH} =& \mathbf{Pr}[S = 1 | \boldsymbol{d} = 1] - \mathbf{Pr}[S = 1 | \boldsymbol{d} = 0] \\ =& \mathbf{Adv}_{\mathcal{VS},\mathcal{A}}^{c\text{-}resist}\end{aligned}$$

This quantity is negligible if the DDH assumption holds.

After sending the certificates using the deniable encryption scheme in registration phase, voter will get two certificates (valid one and false one). According to above result, coercer can not distinguish the ballots using these two different certificates under the DDH assumption. Voters can avoid coercion by giving the false certificate to coercer. So our e-voting system is coercion-resistant.

Theorem 3 (Correctness)**.** *Our proposed e-voting system has the property of correctness.*

Proof. According to the experiment $\mathbf{Exp}_{\mathcal{VS},\mathcal{A}}^{corr}$, we prove the scheme satisfies correctness by inferring number n on the blockchain in step 8 of experiment $\mathbf{Exp}_{\mathcal{VS},\mathcal{A}}^{corr}$.

n_V is the number of ballots that cast by honest voters, and n_A is the number of ballots posted on the blockchain by adversary. Because of the feature of tamper-proof of blockchain, the ballots can only be addition.

Now assume there are $n = n_V + n_A$ ballots on the blockchain and the probability of output '1' is negligible

in the security parameters for the adversary.

If the added vote is from an uncontrolled voter. Smart contract verifies whether the credential in the ballot is valid, if valid, then this ballot will replace previous votes of the same credential.

If the added vote is from an intimidated voter. According to the mechanisms of voting, only the ballot that have valid credential can be count in the tally phase. Our scheme ensured that the adversary could not get a valid credential, so the controlled voter can give the adversary an invalid credential and vote own ballot at a privacy time. The ballot post by adversary will not be count because of invalid credential.

No matter which of the above cases, always n ballots can be counted in the tallying phase. So our e-voting scheme is correct. □

Theorem 4 (Privacy)**.** *Our proposed e-voting system has the property of privacy.*

Proof. Ballots published on blockchain are encrypted with M-ElGamal cryptosystem, only voters and smart contracts can know the contents of the ballots. That is to say, the contents of the ballots can not be revealed to anyone else because of the semantic security of M-ElGamal cryptosystem. The only link between the real identity of the voter and one's ballots is an credential issued by RA. And the voter roll published on the blockchain was encrypted by the public key of smart contract, so only the smart contract can see the contents of the voter roll (randomly sorted encrypted credentials) in the verification phase. Smart contract is a piece of code that written before the system's start-up, so it can not reveal any information to anyone else. □

Theorem 5 (Fairness)**.** *Our proposed e-voting system has the property of fairness.*

Proof. Our protocol guarantees that the ballots can not be opened in other phases excepts for the tallying phase, so that other voters will not be affected. This is achieved by encrypting the ballots with the voter's own private key, and only after the verification of time server, can smart contract open the ballots using the delegation keys sent by the time server in the tallying phase. □

Theorem 6 (Verifiability)**.** *Our proposed e-voting system has the property of verifiability.*

Proof. This property is typically described through the individual verifiability and the universal verifiability. Because of the feature of blockchain, each voter has the ability to check and count the ballots' data to check the correctness of counting result. This enables individual verifiability. In the respect of universal verifiability, since the blockchain is public, everyone in blockchain can get a copy of transaction data. Further, everyone can check the validity of signature attached to the ballots, tally the ballots and make a comparison with the final results. For all these reasons, our proposed system is verifiable. □

Theorem 7 (Eligibility)**.** *Our proposed e-voting system has the property of eligibility.*

Proof. Eligibility means that only authorized voters are allowed to vote. In the registration stage, voters need to send a registration request (including the voter's identity information and security parameters, etc.) to RA. RA verifies and generates the voter's identity credential, then returns it to the voter. The identity credential can be regarded as an authorization of the RA. In the voting phase, voter generates the ballot containing identity credential and sends them to the smart contract. The smart contract decrypts the ballot and compares it with the voter's list. If the identity credential in the ballot is included in the voter's list, the voter has been authorized. If the identity credential in the ballot is not included in the voter's list, the voter is not eligible to vote, and the ballot will not be counted in the final voting result. Therefore, our scheme meets eligibility. □

6 Performance Analysis

The performance analysis of our voting scheme is on the strength of the computation time required for every phase. There are five phases: Initialization, registration, voting, verification and tallying. In our scheme, computation operation of each stage involves in modular operations like addition, multiplication, inverse, exponent and general hash function. For ease of description, the corresponding definition is shown in Table 2. We calculate the computation overhead of each phase, as shown in Table 3.

Table 2: The calculating operation related to the scheme

T_{mul}	The time needed for a modular multiplication
T_{add}	The time needed for a modular addition
T_h	The execution time of general hash function
T_{esp}	The time needed for a modular exponentiation
T_{inv}	The time needed for a modular reverse algorithm

Table 3: Total execution time in various phase

Phase	**Total execution time**
Initialization	$3T_{esp} + T_{mul}$
Registration	$13T_{mul} + 5T_{add} + 11T_{esp} + T_h + 3T_{inv}$
Voting	$3T_{esp} + 2T_h + 2T_{add}$
Verification	$T_h + 2T_{esp}$
Tallying	$4T_{mul} + 2T_{esp}$

Initialization: During the initializing phase, the total execution time is $3T_{esp} + T_{mul}$.

Registration: This phase requires more calculating operations. The computation overhead of this phase is $13T_{mul} + 5T_{add} + 11T_{esp} + T_h + 3T_{inv}$.

Voting: In the voting phase, voters need to encrypt their ballots with M-ElGamal encryption algorithm, and the computation overhead involves $3T_{esp} + 2T_h + 2T_{add}$.

Verification: In our scheme, time-server need to verify the validity of voting time, then generate the decryption key to smart contract. The computation overhead of this phase involves $T_h + 2T_{esp}$.

Tallying: In the tallying phase, the counting operation of smart contract needs $4T_{mul} + 2T_{esp}$.

We use a high-performance implementation of libgmp via the gmpy2 python module to test our scheme, on a desktop with the following specifications: 3.00 GHz hexa-core Intel Core i5 with 9MB shared L3 cache and with 8GB of 2264 MHz DDR4 on-board memory. Table 4 gives the execution time of each operation.

Table 4: The execution time of each calculating operation

T_{mul}	T_{add}	T_h	T_{esp}	T_{inv}
$5.95ms$	$< 0.01ms$	$0.1ms$	$7.63ms$	$< 0.01ms$

In our scheme, the registration of voters and candidates can be completed before a voting. Hence, it does not increase the total running time of the system. Figures 3 to 5 indicate the relation between running time and the number of voters in the other three stages.

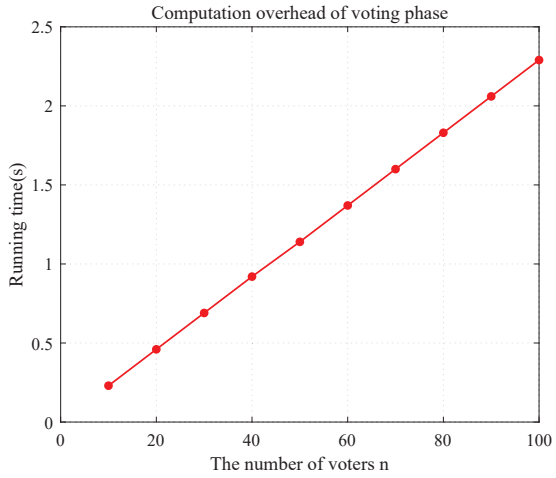


Figure 3: Running time of voting phase

We can see from figures above (Figures 3, 4, 5) that the relationship between running time of each phase and the number of voters is linearly increasing. We also compare the computation overhead of our scheme with [25], [20], [34] in various values of n in Figure 6. In [25], the most complicated computation is zero-knowledge proofs, and the running time of all algorithms for 12 voters is approximately $40ms$. In [20], the computation overhead consists $7n+4T_{mul}$, $3n-1T_{add}$, $2nT_h$ and one modular operation. The running time in [34] involves

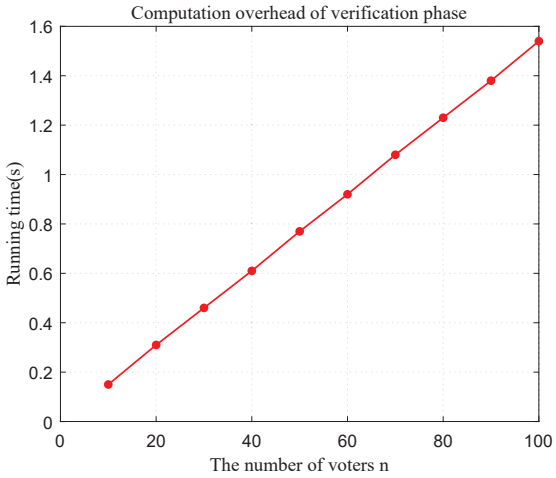


Figure 4: Running time of verification phase

$5T_{mul}$, one T_h and one T_{add}. The most complicated computation is ring signature, and the time needed increases linearly with the ring size. We can see from Figure 6 that the running time of our scheme is smaller than [34] and similar to [20], and longer than [25]. Finally, Figure 6 also shows that our scheme is more suitable for small-scale voting.

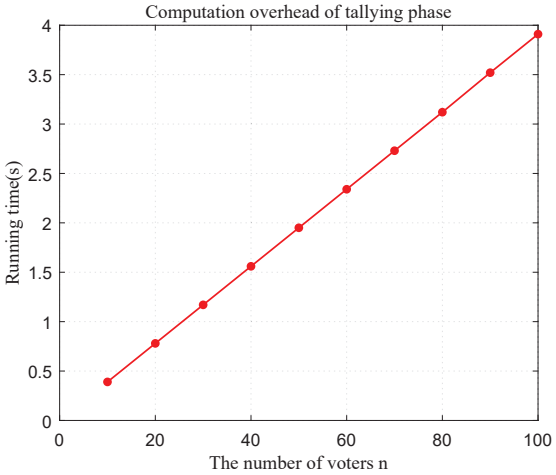


Figure 5: Running time of tallying phase

We also compare the requirement of our scheme and the schemes in [25], [20], [34], as shown in Table 5. We demonstrate in Section 5 that our scheme satisfies the security requirements listed in Table 5, which are all basic properties of voting system except for coercion-resistant. As can be seen form the Table 5, the schemes in [25], [20], [34] do not meet all these basic requirements. And for the property of coercion-resistant, it is only realized in our scheme and the scheme in [20]. The result shows that our scheme can meet more security demands.

Table 5: Comparison of E-voting schemes in requirement

Sceheme	Security requirement					
	Coercion-resistant	Eligibility	Correctness	Verifiability	Fairness	Privacy
[24]	×	×	×	✓	✓	×
[20]	✓	✓	×	×	✓	×
[33]	×	×	✓	✓	×	✓
Ours	✓	✓	✓	✓	✓	✓

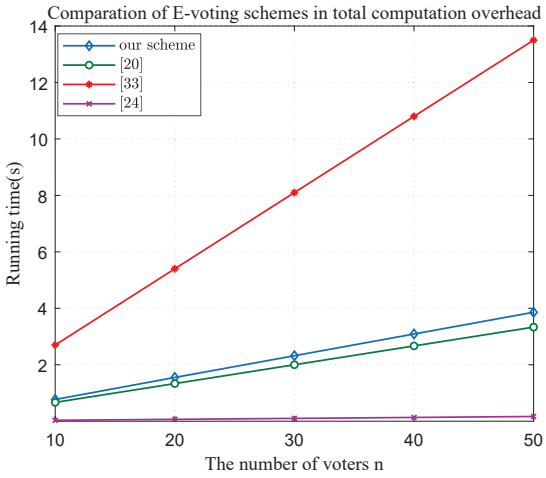


Figure 6: Comparison of E-voting systems in execution time

7 Conclusions

In this work, an e-voting system, which has the characteristic of coercion-resistance, is proposed. We utilize receiver-deniable encryption in the registration phase to realize coercion-resistance. And we adopt blockchain technology to ensure that the process and results of our electronic voting system are fair and transparent. We also design a time-release encryption algorithm in the tallying phase to make the smart contract decrypt ballots and tally after the appointed time. It can guarantee the fairness of our e-voting system further and avoid election fraud.

Our scheme is proved coercion-resistant by simulating the election operations. And we also give detailed safety analysis of other security requirements. Finally, we discuss that the proposed scheme has a better efficiency in a small-scale voting.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB0802000, the Natural Science Foundation of China under Grant 61802303, 61772418 and 61602378, the Key Research and Development Program of Shaanxi under Grant 2019KW-053, the Innovation Ability Support Program in Shaanxi Province of China under Grant 2017KJXX-47, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2019JQ-866, 2018JZ6001 and 2016JM6033, the Research Program of Education Bureau of Shaanxi Province under Grant 19JK0803, the New Star Team of Xi'an University of Posts and Telecommunications under Grant 2016-02, the Fundamental Research Funds for the Central Universities under Grant GK201903005, and Guangxi Cooperative Innovation Center of Cloud Computing and Big Data under Grant YD1903.

References

[1] R. C. Agidi, "Artificial intelligence in nigeria financial sector," *International Journal of Electronics and Information Engineering*, vol. 11, no. 1, pp. 40–47, 2019.

[2] J. Benaloh and D. Tuinstra, "Receipt-free secret-ballot elections," in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 544–553, 1994.

[3] J. Benaloh and D. Tuinstra, "A robust-fragilewatermarking scheme for image authentication," in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 544–553, 1994.

[4] E. Bresson, D. Catalano, and D. Pointcheval, "A simple public-key cryptosystem with a double trapdoor decryption mechanism and its applications," in *International Conference on the Theory and Application of Cryptology and Information Security*, pp. 37–54, 2003.

[5] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky, "Deniable encryption," in *Annual International Cryptology Conference*, pp. 90–104, 1997.

[6] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[7] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *Ieee Access*, vol. 4, pp. 2292–2303, 2016.

[8] M. R. Clarkson, S. Chong, and A. C. Myers, "Civitas: Toward a secure voting system," in *IEEE Symposium on Security and Privacy (SP'08)*, pp. 354–368, 2008.

[9] J. D. Cohen and M. J. Fischer, "A robust and verifiable cryptographically secure election scheme," in *Proceedings of the 26th Annual Symposium on Foundations of Computer Science*, pp. 372-382, 1985.

[10] R. Cramer and V. Shoup, "A practical public key cryptosystem provably secure against adaptive chosen ciphertext attack," in *Annual International Cryptology Conference*, pp. 13–25, 1998.

[11] R. Cramer and V. Shoup, "Universal hash proofs and a paradigm for adaptive chosen ciphertext secure public-key encryption," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 45–64, 2002.

[12] L. F. Cranor and R. K. Cytron, "Sensus: A security-conscious electronic polling system for the internet," in *Proceedings of the Thirtieth Hawaii International Conference on System Sciences*, vol. 3, pp. 561–570, 1997.

[13] G. D. Crescenzo, R. Ostrovsky, and S. Rajagopalan, "Conditional oblivious transfer and timed-release encryption," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 74–89, 1999.

[14] R. A. DeMillo, N. A. Lynch, and M. J. Merritt, "Cryptographic protocols," in *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing*, pp. 383–400, 1982.

[15] B. W. DuRette, *Multiple Administrators for Electronic Voting*, Bachelor's Thesis, MIT, 1999.

[16] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 469–472, 1985.

[17] A. Fujioka, T. Okamoto, and K. Ohta, "A practical secret voting scheme for large scale elections," in *International Workshop on the Theory and Application of Cryptographic Techniques*, pp. 244–251, 1992.

[18] F. S. Hardwick, A. Gioulis, R. N. Akram, and K. Markantonakis, "E-voting with blockchain: An e-voting protocol with decentralisation and voter privacy," in *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 1561–1567, 2018.

[19] M. Hirt and K. Sako, "Efficient receipt-free voting based on homomorphic encryption," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 539–556, 2000.

[20] T. Hsiao, Z. Wu, C. Liu, and Y. Chung, "Electronic voting systems for defending free will and resisting bribery and coercion based on ring anonymous signcryption scheme," *Advances in Mechanical Engineering*, vol. 9, no. 1, pp. 1687814016687194, 2017.

[21] C. T. Li, M. S. Hwang, Y. C. Lai, "A verifiable electronic voting scheme over the internet," in *Sixth International Conference on Information Technology: New Generations*, pp. 449-454, 2009.

[22] A. Juels, D. Catalano, and M. Jakobsson, "Coercion-resistant electronic elections," in *Towards Trustworthy Elections*, pp. 37–63, 2010.

[23] J. Katz, "Digital signatures," *Security and Cryptology*, 2010. ISBN: 978-0-387-27712-7.

[24] M. Klonowski, P. Kubiak, and M. Kutyłowski, "Practical deniable encryption," in *International Conference on Current Trends in Theory and Practice of Computer Science*, pp. 599–609, 2008.

[25] Y. Li, W. Susilo, G. Yang, Y. Yu, D. Liu, and M. Guizani, "A blockchain-based self-tallying voting scheme in decentralized IoT," *Cryptography and Security*, 2019. arXiv:1902.03710.

[26] P. McCorry, S. F. Shahandashti, and F. Hao, "A smart contract for boardroom voting with maximum voter privacy," in *International Conference on Financial Cryptography and Data Security*, pp. 357–375, 2017.

[27] B. Meng, Z. Li, and J. Qin, "A receipt-free coercion-resistant remote internet voting protocol without physical assumptions through deniable encryption and trapdoor commitment scheme," *Journal of Software*, vol. 5, no. 9, pp. 942–949, 2010.

[28] B. Meng and J. Wang, "An efficient receiver deniable encryption scheme and its applications," *Journal of Networks*, vol. 5, no. 6, pp. 683–690, 2010.

[29] T. Okamoto, "An electronic voting scheme," in *Advanced IT Tools*, pp. 21–30, 1996.

[30] T. Okamoto, "Receipt-free electronic voting schemes for large scale elections," in *International Workshop on Security Protocols*, pp. 25–35, 1997.

[31] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 223–238, 1999.

[32] K. Sako and J. Kilian, "Receipt-free mix-type voting scheme," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 393–403, 1995.

[33] B. Shahzad and J. Crowcroft, "Trustworthy electronic voting using adjusted blockchain technology," *IEEE Access*, vol. 7, pp. 24477–24488, 2019.

[34] W. W. Jr, "An efficient and effective decentralized anonymous voting system," *arXiv preprint arXiv:1804.06674*, 2018. (`https://arxiv.org/ftp/arxiv/papers/1804/1804.06674.pdf`)

[35] X. Yang, X. Yi, S. Nepal, and F. Han, "Decentralized voting: A self-tallying voting system using a smart contract on the ethereum blockchain," in *International Conference on Web Information Systems Engineering*, pp. 18–35, 2018.

[36] B. Yu, J. K. Liu, A. Sakzad, S. Nepal, R. Steinfeld, P. Rimba, and M. H. Au, "Platform-independent secure blockchain-based voting system," in *International Conference on Information Security*, pp. 369–386, 2018.

Biography

Kaili Ye received the B.S. degree from the Xi'an University of Posts and Telecommunications in 2018. She is currently pursuing the M.S. degree with the National Engineering Laboratory for Wireless Security, Xi'an

University of Posts and Telecommunications, China. Her research interests include blockchain technology and information security.

Dong Zheng received the Ph.D. degree from Xidian University in 1999. He joined the School of Information Security Engineering, Shanghai JiaoTong University. He is currently a Professor with the Xi'an University of Posts and Telecommunications, China. His research interests include information theory, cryptography, and information security. He is a Senior Member of the Chinese Association for Cryptologic Research and a member of the Chinese Communication Society.

Rui Guo received the Ph.D. degree from the State Key Laboratory of Networking and Switch Technology, Beijing University of Posts and Telecommunications, China, in 2014. He is currently a Lecturer with the National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications. His present research interests include attribute-based cryptograph, cloud computing, and blockchain technology.

Jiayu He received the bachelors degree from Xi'an University of Posts and Telecommunications in 2018. He is currently pursuing his master degree at National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications. His research interests include blockchain technology and security in cloud storage.

Yushuang Chen received the B.S. degree from the Xi'an University of Posts and Telecommunications in 2018. She is currently pursuing the M.S. degree with the National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications, China. Her research interests include blockchain, information security, and the Internet of Things (IoT).

Xiaoling Tao received the M.S. degree in computer application technology from Guilin University of Electronic Technology. She is a professor at the school of computer science and information security, Guilin University of Electronic Technology. Her research interests include cloud computing and security and network security.

Research on Privacy Security Steady State Evaluation Model of Mobile Application Based on Information Entropy and Markov Theory

Ming Yang[1], Li Jia[1], Tilei Gao[1], Tao Zhang[1], and Wanyu Xie[2]
(Corresponding author: Tilei Gao)
School of Information, Yunnan University of Finance and Economics[1]
Kunming 650221, China
Personnel Department, Kunming Metallurgy College[2]
Email: gtllei@ynufe.edu.cn
(Received May 25, 2020; Revised and Accepted Dec. 10, 2020; First Online Aug. 14, 2021)

Abstract

In the process of using mobile applications, if users want to use customized personalized services, they need to provide personal privacy information to businesses, which constitutes a potential threat to their privacy security. In order to help users choose secure mobile applications reasonably, and ensure the security of users during use, this paper combines information entropy and Markov theory to study the measurement of user privacy security of mobile applications, and proposes an effective steady-state evaluation model of user privacy security. Finally, the model is put into a specific case for analysis. The analysis results show that the method can effectively evaluate the privacy security of users, and compared with the previous methods, the method is more real, objective and simple.

Keywords: Information Entropy; Markov Theory; Mobile Application Security; Privacy Security; Risk Evaluation

1 Introduction

With the popularity of smart phones, more and more mobile applications appear in front of people, providing users with a variety of accurate services. However, while enjoying the convenient services brought by these mobile applications, users' privacy security is also threatened. These leakage scenarios include wireless network connection, mobile payment, public equipment use, fingerprint recognition, face recognition, *etc.* [5]. These privacy leakage scenarios include wireless network connection, mobile payment, public equipment use, fingerprint recognition, face recognition, *etc.* In these scenarios, the privacy information of users is continuously collected, transmitted and stored, and is faced with the risk of being leaked, abused and stolen. These risks not only come from individuals, but also from service providers and terminal equipment. There are various ways to steal data, such as internal theft, external hacker intrusion and employee negligence. According to the information collected by the Identity Theft Resource Center and the U.S. Department of health and human services, more than 137 million records were leaked in 2019 [32], and the information leaked due to the privacy security problems of mobile applications accounted for a large proportion.

At present, most of the research on user privacy security is focused on cloud services and big data applications. It is known that, compared with the traditional information system security risk factors, the risk hierarchy of users privacy information in mobile applications is more complex, including traditional information system security risk, user behavior risk [11], third party application risk [6] and unique risk of mobile application service [10]. Therefore, considering the importance of security risk assessment to the ecosystem and sustainable development of mobile e-commerce platform [28], this paper proposes a risk evaluation model for privacy security of mobile applications based on the characteristics of mobile applications and the security model of information system [26].The evaluation model can provide users with practical and intuitive risk evaluation results and protect their privacy security in mobile commerce.

The organizational structure of this paper is as follows: Section 1 - Introduction: The background, content and significance of the research are presented. Section 2 - Related researches: This chapter discusses the privacy security risk attributes and evaluation methods in mobile applications, and puts forward the main problems to be solved in the current research. Section 3 - Research on privacy security measurement and evaluation of mobile applications Based on Information Entropy and Markov Theory: According to the characteristics of mobile application users' privacy security, this chapter puts information entropy and Markov into the research of user

privacy security, proposes a privacy security measurement method based on information entropy, and proposes a dynamic evaluation method of user privacy security based on Markov theory. Section 4 - Steady state evaluation model of mobile application privacy security based on information entropy and Markov: In this chapter, a user privacy security risk evaluation model of mobile application based on information entropy and Markov is established, and the evaluation steps of the model are introduced in detail. Section 5 - Case analysis: In this chapter, the proposed model is put into three different types of mobile applications to carry out case studies. Section 6 - Conclusion: This chapter summarizes the research work of this paper and points out the future research direction.

2 Relevant Researches

Recent researches on users' privacy security of mobile application are studied and summarized for this paper, as follows:

2.1 Research on Users Privacy Security Risk Attribute of Mobile Application

The privacy security research of mobile application is different from the traditional security research. It includes the influence of various risk factors, such as the technical defects of privacy protection, the user's own security weak consciousness, the application environment risk, the terminal equipment risk factor, the operator's management risk, and the privacy risk caused by laws and regulations.

1) Technology risk. Reference [23] studies the privacy security of mobile applications, emphasizing the importance of user authentication technology and protocol. Reference [18] pointed out that the privacy security of users in mobile applications is affected by many technical factors, including data encryption, intrusion detection, identity management, security awareness, *etc.* Literature [1] pointed out that whether to adopt anonymization technology will directly affect the privacy and security of mobile applications. Reference [4] analyzes the security problems of various levels of network physical system (CPS) architecture, and points out that to improve its security, attention should be paid to the influence of related technologies, such as access control, data encryption, attack detection, user authentication and authorization, *etc.* Reference [12] proposed a secure routing protocol with node self-sufficiency resistance, which improves the protection of mobile application privacy.

2) User vulnerability risk. Ampong [3] mentioned that in the process of using mobile applications, users' privacy awareness, privacy concerns and privacy intrusion experience will directly affect their personal privacy security. Literature [7, 34] thinks that location information is extremely sensitive in mobile commerce, and the exposure of location information may cause the risk of mobile application information abuse. Reference [33] analyzes the characteristics of user privacy security in the big data environment, and points out that some common user behaviors will directly cause the disclosure of personal privacy information, such as privacy Association setting, spatial location sharing, information behavior negligence and simple password setting.

3) Application scenario risk. In some mobile applications, there are usually mobile advertisements, which are intrusive to users' privacy, and even forcibly obtain users' personal location information [25]. In addition, the authorization of some application rights will also affect the privacy and security of users. Reference [9] mentioned that users will be forced to agree to open some application permissions before using some mobile commerce applications, resulting in users having no autonomy in whether to share their own information.

4) Mobile terminal device risk. References [13,16,17,19] respectively discussed various privacy risks existing in mobile terminals, including the protection of sensitive data by the device, the location tracking function of the device, the management of the authority of the device, and the malicious monitoring function of the device itself.

5) Management and legal risk. As an important role in the process of mobile application interaction, the improper management of service providers and the restrictions of laws will affect the privacy and security of users. Reference [14] mentioned that establishing a standard privacy policy for mobile applications is the key to solve their privacy security issues. Literature [20] pointed out that the common management risks include imperfect disclosure standards of privacy information, lack of supervision and punishment system, malicious disclosure of internal personnel, *etc.* Literature [27] analyzes the consumption behavior characteristics of mobile application users, and establishes corresponding risk evaluation indicators, which include privacy management mechanism, platform privacy protection investment, information-sharing risk, third-party information collection, privacy law differences and other related factors.

2.2 Research on Privacy Security Risk Ealuation Method

At present, the researches on privacy security evaluation methods for mobile applications are rare. Literature [8, 21, 22, 31] proposed some effective measurement methods for the security of cloud services using

the method of information entropy, but did not specifically evaluate the privacy security. In Reference [24], a privacy-considered information security evaluation model was built with the risk recommendation system based on the identifiability, context of use, quantity, sensitivity, And freshness of the personal identity information data. Lo [15] proposed a user privacy analysis framework called LRPdroid for Android platform, which realized information leakage detection, user privacy leakage assessment and privacy risk assessment of applications installed on mobile devices based on Android. According to the characteristics of mobile application permission, Reference [30] proposed a mobile application risk evaluation strategy based on mobile application permission characteristics. Reference [29] also proposed a risk evaluation method for mobile applications in Android environment based on its permission characteristics.

Summing up the above methods, it can be found that these assessment methods are not targeted for the evaluation of user privacy security, and most of the evaluation objects are only limited to one kind of risk, without considering the interaction between various risk classes, and also not combining with the actual risk environment for dynamic assessment.

Therefore, in order to solve the above problems, this paper proposes a special evaluation method for mobile application privacy security, and establishes the corresponding model to realize the multi-level and multi angle evaluation of mobile application privacy security.

3 Research on Privacy Security Measurement and Evaluation of Mobile Applications based on Information Entropy and Markov Theory

In order to realize the dynamic evaluation of privacy and security of mobile applications, two following research contents have been carried out:

3.1 Research on Privacy Security Measurement Method of Mobile Application based on Information Entropy

Measurement is the premise of evaluation, an objective and accurate measurement result will directly affect the evaluation results. However, the privacy security of mobile applications is an abstract concept, and only be measured by specific methods. Therefore, in order to effectively measure the privacy and security of mobile applications, this paper proposes to use "the uncertainty of risk" to quantitatively describe "the level of privacy security of mobile applications" from the opposite perspective. As shown in Figure 1.

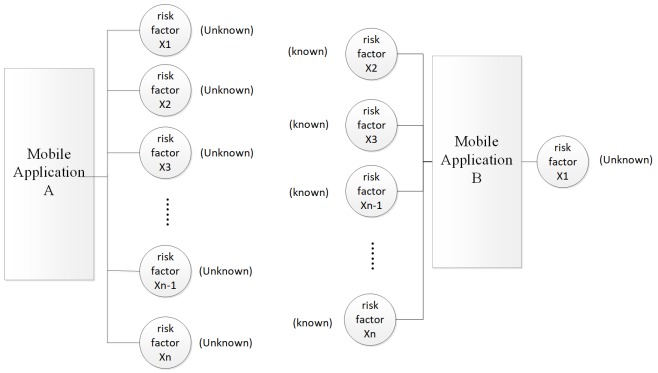


Figure 1: Two extreme mobile application risk environments

According to the theory of information entropy, it is assumed that both A and B contain some unknown risk factors $X_i, X = \{X_1, X_2, \ldots, X_n\}$.

1) The entropy of application A is maximal, and its risk is the highest. There are n unknown risks in application A, and the probability of the risk occurrence $P(X_1) = P(X_2) =, \ldots, = P(X_n)$, the entropy value $H(X) = \log_2 n$ will reach its maximum according to the information entropy equation. At this time, the system contains many risk factors, and its controllable degree will reach the lowest.

2) The entropy of application A is minimal, and its risk is the lowest. There is only one unknown risk in the application B. According to the equation of information entropy, its entropy $H(X)$ will reach the lowest. At this time, the uncertainty of risk factors contained in the system is the lowest, and the risk is completely controllable, that is to say, the security of the system reaches the highest.

However, in practice, a complex mobile application is bound to be affected by a number of different risk factors, and the probability of these factors is different too, so that its entropy is bound to be between the maximum value and the minimum value. Therefore, according to the theory of information entropy, the level of mobile application security in the actual situation can be described by the size of entropy.

3.2 State Description of Mobile Application Risk Environment Based on Markov

On the basis of information entropy measurement, this paper will further combine Markov theory to describe the risk environment state of mobile application, and express it with the form of mathematical matrix, so that to provide support for the follow-up steady-state evaluation research [2].

It is assumed that there are n risk factors X_i in a mobile application environment, so its complex risk environment can be described as the following matrix:

$$R = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{51} & X_{52} & \cdots & X_{nn} \end{bmatrix} \quad (1)$$

Matrix R is the privacy security risk matrix of mobile applications, in which:

1) Element X on diagonal X_{ii} represent the separate occurrence of risk factors X_i;

2) Non diagonal element X_{ij} represent simultaneous occurrence of risk factors X_i and X_j.

It is known that in the actual use of mobile applications, their risk factors have different probability of occurrence, which leads to they have a variety of possible privacy security states. Therefore, in order to objectively evaluate the privacy security of mobile applications, it is necessary to effectively restore their random risk state. Using $P(X_{ij})$ represents the probability of different risks, and substituting it into matrix (1), the risk state transition matrix as follows will be obtained:

$$(R) = \begin{bmatrix} P(X_{11}) & P(X_{12}) & \cdots & P(X_{1n}) \\ P(X_{21}) & P(X_{22}) & & P(X_{2n}) \\ \vdots & & \ddots & \vdots \\ P(X_{n1}) & P(X_{n2}) & \cdots & P(X_{nn}) \end{bmatrix} \quad (2)$$

3.3 Research on Steady State Evaluation Method Based on Information Entropy and Markov

After proposing the privacy security measurement method and risk state description matrix of mobile application, this paper will establish a special privacy risk attribute model to realize a multi-level and multi-dimensional evaluation of the whole mobile application. As shown in Figure 2:

In this paper, the privacy security attribute model of mobile applications is divided into three levels: target layer A, risk class layer β and risk factor layer X. In order to objectively reflect the real risk environment, this paper uses cross lines to describe the relationship between layer 2 and layer 3, which is used to reflect the influence of different risk factor X_i on different dimension β.

As shown in the following example, suppose a mobile commerce contains two types of risk β_1 and β_2. The risk factors involved are shown in Table 1:

As shown in Table 1, risk classes A and B contain a common risk factor X_2, so their transition state matrix

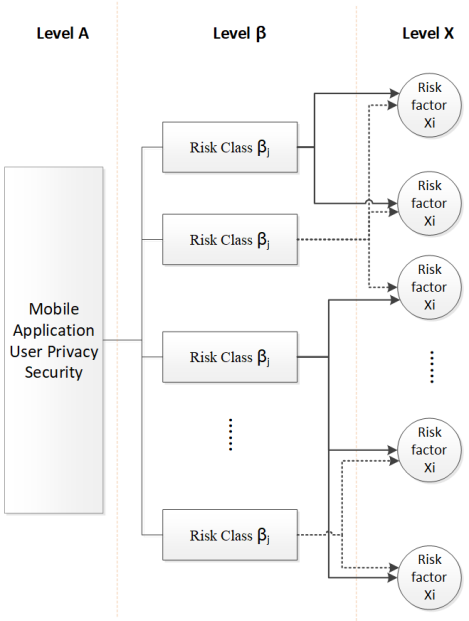


Figure 2: The abstract privacy security risk attribute model of mobile application

Table 1: Risk classes β_1 and β_2 and the risk factors they contain

Risk Class	Risk factors included Risk Class
β_1	X_1, X_2
β_2	X_2, X_3, X_4

$P(R)$ can be calculated as follows:

$$P(R) = \begin{bmatrix} P(\beta_{11}) & P(\beta_{12}) \\ P(\beta_{21}) & P(\beta_{22}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sum_{i=1}^{3} P(X_i)} P(X_1) & \frac{1}{\sum_{i=1}^{3} P(x_i)} P(X_2) \\ \frac{1}{\sum_{i=3}^{4} P(\alpha_i)} P(X_2) & \frac{1}{\sum_{i=3}^{4} P(\alpha_i)} \{P(X_3) + P(X_4)\} \end{bmatrix} \quad (3)$$

According to the above calculation method, it is assumed that a mobile application has m risk classes β_i in the long-term use process, and the steady-state probability of each class is $\widehat{P}(\beta_i)$, which represents the convergence probability of a risk class in the long-term use process. $\widehat{P}(\beta_i)$ can objectively reflect the stable probability of a certain risk class in the long-term use process. It is the result of dynamic evaluation of different risk states through Markov matrix. The relationship between $\widehat{P}(\beta_i)$ and $P(R)$ is as follows:

$$\begin{cases} \widehat{P}(\beta_1) = P(X_{11})\widehat{P}(\beta_1) + P(X_{12})\widehat{P}(\beta_2) + \ldots + P(X_{1m})\widehat{P}(\beta_m) \\ \widehat{P}(\beta_2) = P(X_{21})\widehat{P}(\beta_1) + P(X_{22})\widehat{P}(\beta_2) + \ldots + P(X_{2m})\widehat{P}(\beta_m) \\ \widehat{P}(\beta_3) = P(X_{31})\widehat{P}(\beta_1) + P(X_{32})\widehat{P}(\beta_2) + \ldots + P(X_{3m})\widehat{P}(\beta_m) \\ \vdots \\ \widehat{P}(\beta_m) = P(X_{n1})\widehat{P}(\beta_1) + P(X_{n2})\widehat{P}(\beta_2) + \ldots + P(X_{nm})\widehat{P}(\beta_m) \\ 1 = \widehat{P}(\beta_1) + \widehat{P}(\beta_2) + \ldots + \widehat{P}(\beta_m) \end{cases} \quad (4)$$

Solving the Equation (4), it can obtain the

steady-state probability of various risks $\hat{P}(\beta_i) = \left\{\widehat{P}(\beta_1), \widehat{P}(\beta_2), \ldots, \widehat{P}(\beta_m)\right\}, \sum_{i=1}^{m} \widehat{P}(\beta_i) = 1$.

Further, substituting $\widehat{P}(\beta_i)$ into the information entropy equation and using the reciprocal form, it can quantitatively describe the privacy security of the entire mobile application, the equation is as follows:

$$E = 1/H = 1/-\sum_{i=1}^{m} \widehat{P}(\beta_i) \log_2 \widehat{P}(\beta_i) \tag{5}$$

E represents the security evaluation result of the whole mobile application. The higher the value is, the higher the privacy security of the mobile application is. Similarly, if the probability of occurrence of risk factors contained in class β_i is normalized, the security evaluation result $E(\beta_i)$ of the risk class can be obtained. The higher the value, the higher the privacy security of this risk class is.

As mentioned above, around the hierarchical structure shown in Figure 2, this paper proposes a bottom-up privacy security risk evaluation method for mobile applications based on information entropy and Markov.

4 Steady State Evaluation Model of Mobile Application Privacy Security Based on Information Entropy and Markov

In this chapter, the proposed hierarchical structure in chapter 3 will be embodied, and the calculation steps of the proposed method will be described in detail, so as to build a privacy security steady state evaluation model of mobile application based on information entropy and Markov.

4.1 Risk Attribute Model for Privacy Disclosure of Mobile Commerce Users

In this regard, this paper selects a total of 24 mobile application privacy risk factors, and divides these factors into 5 classes, namely technical risk class β_1. Application risk class β_2. Management and legal risk class β_3. User risk class β_4. Mobile terminal equipment risk class β_5. The risk attribute model established according to the above division is shown in Figure 3.

4.2 The Calculation Steps of Privacy Security Evaluation Method

According to the evaluation method proposed in chapter 3, the calculation is carried out from bottom to top, which is divided into five steps.

Step 1: According to the evaluation standard of risk level shown in Table 2, the occurrence probability grade of each risk factor in the third layer is evaluated, and the probability of single risk factor is obtained by normalization.

Table 2: The level of probability of risk factors occurrence

level	Definition and description
[8, 10)	This factor has a great risk and a direct threat to the user's privacy
[6, 8)	This risk has a high probability of occurrence and exists in most mobile business environments
[4, 6)	This risk is a common risk, which exists in some mobile commerce
[2, 4)	This risk exists and only occurs when special conditions are met
(0, 2)	This factor has high security and hardly causes user privacy risk

Step 2: According to the membership relationship shown in Figure 3, the state transition matrix $P(R)$ is calculated.

Step 3: Solve Equation (4) and calculate the steady-state probability $\widehat{P}(\beta_i)$ of each risk class.

Step 4: According to Equation (5), the privacy security evaluation result of the whole mobile application is calculated.

Step 5: The evaluation results $E(\beta_i)$ of different risk categories are calculated respectively, and the calculation equation is as follows:

$$E(\beta_i) = \frac{\log_2 m}{-\sum_{j=1}^{m} P(X_{j,\beta_i}) \log_2 P(X_{j,\beta_i})} \tag{6}$$

In Equation (6), m is the total number of risk factors included in each risk class, $P(X_j, \beta_i)$ represents the influence weight of risk factor X_j on risk class β_i, which is the result of normalization treatment.

5 Case Study

5.1 Evaluation Process

In order to verify the proposed evaluation model, this paper selects three mobile application products with long operation time in the market. Among them, product A is a mobile application for financial business; product B is a mobile application for catering delivery; product C is a mobile application for map navigation. For these 3 applications, this paper carried out a detailed evaluation process, as follows:

Step 1: According to the definition of Table 2, a total of 10 experts were invited to evaluate the underlying risk factors of three different applications by AHP

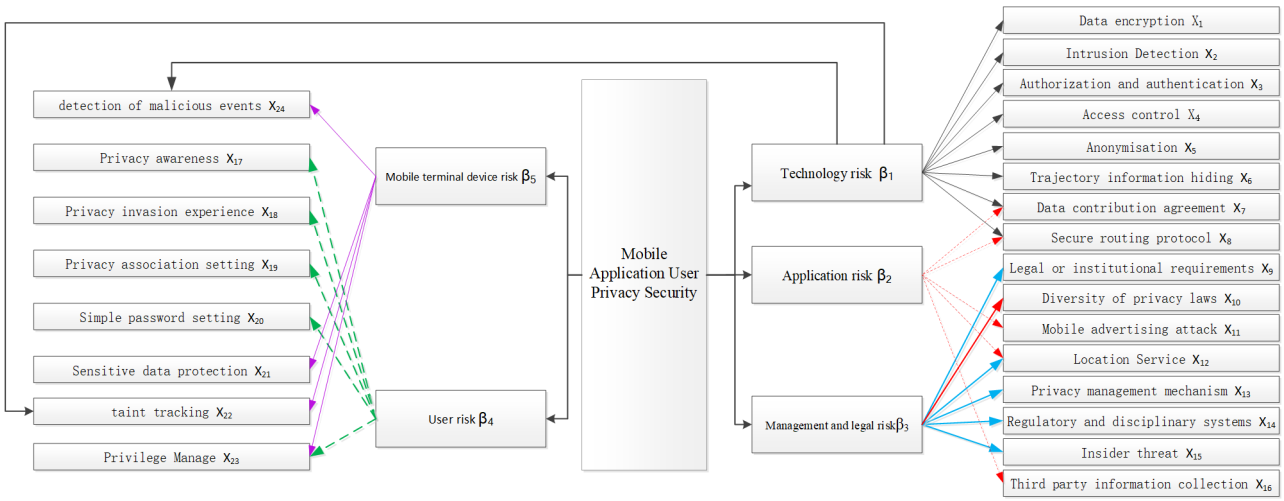


Figure 3: The privacy security risk attribute model of mobile application

method. Finally, the level of each risk factor is calculated, and the probability is obtained by further normalization. The results are shown in Table 3.

Step2: According to Equation (3) and the membership relationship shown in Figure 3, the privacy security risk state transition matrix $P^A(R), P^B(R), P^C(R)$ of the 3 applications are calculated as follows:

$$P^A(R) = \begin{bmatrix} 0.595 & 0.214 & 0.000 & 0.000 & 0.190 \\ 0.450 & 0.450 & 0.100 & 0.000 & 0.000 \\ 0.000 & 0.133 & 0.867 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.794 & 0.206 \\ 0.250 & 0.000 & 0.000 & 0.350 & 0.400 \end{bmatrix}$$

$$P^B(R) = \begin{bmatrix} 0.452 & 0.290 & 0.000 & 0.000 & 0.258 \\ 0.300 & 0.467 & 0.233 & 0.000 & 0.000 \\ 0.000 & 0.389 & 0.511 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.838 & 0.162 \\ 0.263 & 0.000 & 0.000 & 0.316 & 0.421 \end{bmatrix}$$

$$P^C(R) = \begin{bmatrix} 0.548 & 0.262 & 0.000 & 0.000 & 0.190 \\ 0.355 & 0.355 & 0.290 & 0.000 & 0.000 \\ 0.000 & 0.346 & 0.654 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.784 & 0.216 \\ 0.158 & 0.000 & 0.000 & 0.421 & 0.421 \end{bmatrix}$$

Step 3: The data of $P^A(R), P^B(R), P^C(R)$ are respectively substituted into the Equation (4), and the steady-state probability of each risk is calculated, as shown in Table 4.

Step 4: Substitute the results of table4 into Equation (5) to obtain the privacy security evaluation results of the 3 applications, the results are shown in Figure 4.

Step 5: According to Equation (6), the privacy security evaluation results of various risks are calculated respectively.

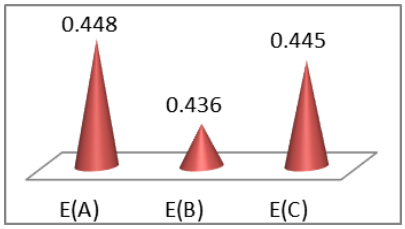


Figure 4: Comparison of evaluation results of three applications

Given that each risk class and its risk factors are as shown in Table 5, the result calculated by Equation (6) is shown in Figure 5.

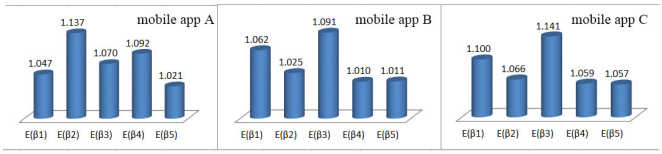


Figure 5: Comparison of entropy values of each risk class in three applications

5.2 Analysis of Evaluation Results

1) Analysis of top-level evaluation results. The comparison of Figure 4 shows that $E(A) > E(C) > E(B)$, which indicates that the financial business application A has the highest privacy security, and the catering delivery application B has the lowest privacy security. But from the data size difference comparison, it can be found that the privacy security performance of the 3 applications is not much different, indicating that the privacy security performance of the 3 applications is similar.

Table 3: Scoring results of probability of occurrence of underlying risk factors

Mobile App	***risk factor***	**level**	$\boldsymbol{P}(\boldsymbol{x}_i)$	***risk factor***	**level**	$\boldsymbol{P}(\boldsymbol{x}_i)$	***risk factor***	**level**	$\boldsymbol{P}(\boldsymbol{x}_i)$
A	X_1	2	1.905%	X_9	2	1.905%	X_{17}	9	8.571%
	X_2	8	7.619%	X_{10}	1	0.952%	X_{18}	8	7.619%
	X_3	7	6.667%	X_{11}	1	0.952%	X_{19}	9	8.571%
	X_4	4	3.810%	X_{12}	2	1.905%	X_{20}	1	0.952%
	X_5	2	1.905%	X_{13}	5	4.762%	X_{21}	5	4.762%
	X_6	2	1.905%	X_{14}	3	2.857%	X_{22}	4	3.810%
	X_7	5	4.762%	X_{15}	2	1.905%	X_{23}	7	6.667%
	X_8	4	3.810%	X_{16}	8	7.619%	X_{24}	4	3.810%
B	X_1	1	0.840%	X_9	2	1.681%	X_{17}	9	7.563%
	X_2	1	0.840%	X_{10}	2	1.681%	X_{18}	9	7.563%
	X_3	3	2.521%	X_{11}	3	2.521%	X_{19}	9	7.563%
	X_4	7	5.882%	X_{12}	9	7.563%	X_{20}	2	1.681%
	X_5	2	1.681%	X_{13}	8	6.723%	X_{21}	3	2.521%
	X_6	9	7.563%	X_{14}	3	2.521%	X_{22}	4	3.361%
	X_7	8	6.723%	X_{15}	2	1.681%	X_{23}	8	6.723%
	X_8	3	2.521%	X_{16}	8	6.723%	X_{24}	4	3.361%
C	X_1	1	0.901%	X_9	2	1.802%	X_{17}	9	8.108%
	X_2	1	0.901%	X_{10}	2	1.802%	X_{18}	9	8.108%
	X_3	3	2.703%	X_{11}	2	1.802%	X_{19}	9	8.108%
	X_4	2	1.802%	X_{12}	9	8.108%	X_{20}	2	1.802%
	X_5	1	0.901%	X_{13}	8	7.207%	X_{21}	3	2.703%
	X_6	9	8.108%	X_{14}	3	2.703%	X_{22}	4	3.604%
	X_7	8	7.207%	X_{15}	2	1.802%	X_{23}	8	7.207%
	X_8	3	2.703%	X_{16}	7	6.306%	X_{24}	4	3.604%

Table 4: The steady-state probability of each risk

Mobile App	$\widehat{\boldsymbol{P}}(\boldsymbol{\beta}_1)$	$\widehat{\boldsymbol{P}}(\boldsymbol{\beta}_2)$	$\widehat{\boldsymbol{P}}(\boldsymbol{\beta}_3)$	$\widehat{\boldsymbol{P}}(\boldsymbol{\beta}_4)$	$\widehat{\boldsymbol{P}}(\boldsymbol{\beta}_5)$
A	0.095	0.145	0.242	0.297	0.234
B	0.141	0.179	0.198	0.267	0.222
C	0.115	0.156	0.178	0.297	0.261

Table 5: The level of probability of risk factors occurrence

Risk class $\boldsymbol{\beta_i}$	**risk factor $\boldsymbol{x_j}$ contained in $\boldsymbol{\beta_i}$**
β_1	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{22}, X_{24}$
β_2	$X_7, X_8, X_{11}, X_{12}, X_{16}$
β_3	$X_9, X_{10}, X_{12}, X_{13}, X_{14}, X_{15}$
β_4	$X_{17}, X_{18}, X_{19}, X_{20}, X_{23}$
β_5	$X_{21}, X_{22}, X_{23}, X_{24}$

2) Comparative analysis of steady-state probability results of risk classes. It can be seen from Table 4 that among the three applications, the steady-state probability $\widehat{P}(\beta_4)$ is the largest, and $\widehat{P}(\beta_1)$ is the lowest. This result shows that the user risk β_4 is the most likely to cause privacy security risk in the long-term use of these 3 applications; on the contrary, technical risk β_1 is the least likely to occur.

3) Comparative analysis of privacy security evaluation result of each risk class. It can be seen from Figure 5 that in application B and Application C, the values of $E(\beta_4)$ and $E(\beta_5)$ are relatively low, indicating that for these two applications, the most difficult to control are the user's own risk β_4 and the terminal equipment risk β_5. These two risks are the main reasons for the privacy security of such applications. On the contrary, the value of $E(\beta_3)$ is the highest, which indicates that this kind of risk is basically controllable and is not easy to cause privacy security problems.

However, in application A, the values of $E(\beta_2)$ and $E(\beta_4)$ are higher, which indicates that financial applications have strict application standards and user behavior control. Compared with other types of applications, the application environment risk β_2 and user risk β_4 are easier to control in financial applications. The most important problem that leads to the privacy security of financial applications is focused on β_3 and β_5.

4) Analysis of risk factors layers. Through the above comparison, combined with table 5, it can be found that the user privacy risk β_4 and β_5 have the highest probability of occurrence in application A and application B. Observe the contained factors of risk class β_4 and β_5, it can be found that the security problems

of the most mobile applications are mainly caused by the factors $\{x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}\}$.

In the financial application A, risk class β_2 has the greater probability of occurrence. Observe the contained factors of risk class β_2, it can be found that the security problems of the financial application are mainly caused by the factors $\{x_7, x_8, x_9, x_{10}\}$

5.3 Countermeasures and Suggestions

To sum up, technical risk is not the main reason leading to the privacy security of mobile applications at this stage. To fundamentally improve the privacy security of applications, we need to shift the focus to the control of users' own risks and mobile terminal equipment risks. For the application providers, it is necessary to strengthen the management and control of user risk, remind users of the existing privacy security risks, standardize the user's operation behavior, and provide users with some suggestions for self-security protection; on the other hand, users themselves need to strengthen their own privacy security awareness and do their own security precautions. For financial applications, the application providers should further clarify the confidentiality agreement with users, reduce the access to users' use rights, and clarify the ownership of relevant responsibilities, and ensure the information security of users through laws and regulations.

5.4 Comparison with other Evaluation Methods

In order to explain the characteristics of the proposed method more intuitively, this paper compares the proposed method with AHP and CMM / CMMI. Among them, AHP (analytic hierarchy process) is a qualitative and quantitative evaluation method, which is easy to operate and has certain objectivity, but it is not suitable for the system with random state; CMM / CMMI is a set of evaluation method based on software process, and its evaluation results have long-term reference value for the management and improvement of the whole software process, but the establishment of its evaluation model needs a lot of manpower and financial resources.

Finally, this paper compares the 3 methods from four aspects which are usability, objectivity, decision support and cost. The results are shown in Table 6.

Table 6: Comparison with AHP and CMM/CMMI

	Usability	Objectivity	Decision support	Cost
Our Method	High	High	Modest	Modest
AHP	Modest	Modest	Low	Low
CMM/CMMI	Low	Low	High	High

In Table 6, usability refers to the ease of use of the method. Objectivity refers to the evaluation objective degree of the method. Decision support refers to the support degree of the method to the decision. Cost refers to the cost of adopting this method.

Through the comparison, it can be found that the method proposed in this paper has advantages in usability and objectivity. It is a simple, easy-to-use and moderate cost evaluation method, which can directly and truly reflect the privacy security degree of the evaluation object. Its evaluation results can provide practical data support for the privacy security protection of mobile applications.

6 Conclusions

In this paper, the privacy security attribute model of mobile applications is established, and the privacy security measurement is carried out based on information entropy, and the privacy risk environment is described based on Markov theory. A steady-state evaluation model of mobile application user privacy security based on information entropy and Markov is proposed. Finally, through the case study, the evaluation results show that the model can achieve multi-level and multi-dimensional analysis of mobile application privacy security, so as to provide the basis for privacy protection of mobile applications. The model is simple, easy to use, and has good objectivity, which is of great significance to the research on privacy security of mobile application. In the future research, the attribute model of risk is a content that needs to be further studied. Only by constantly improving the attribute model of risk can we provide more accurate reference for decision-making. On the other hand, it is necessary to strengthen the dynamic evaluation of security, and introduce the concept of time to dynamically describe the privacy security of mobile applications, so as to provide more realistic security evaluation results for decision makers.

Acknowledgments

This research is supported by National Natural Science Foundation Project (No. 71462036), the Scientific Research Foundation of Yunnan Education Department (No. 2020J0377, No. 2020J0392) and School-level Project of Yunnan University of Finance and Economics (No. 2017D29).

The authors would like to thank the anonymous reviewers and the editors for their suggestions.

References

[1] E. Aghasian, S. Garg, and J. Montgomery, "A privacy-enhanced friending approach for users on multiple online social networks," *Computers*, vol. 7, no. 3, pp. 42, 2018.

[2] M. H. R. Al-Shaikhly, H. M. El-Bakry, and A. A. Saleh, "Cloud security using Markov chain and genetic algorithm," *International Journal of Electronics and Information Engineering*, vol. 8, no. 2, pp. 96–106, 2018.

[3] G. O. A. Ampong, A. Mensah, A. S. Y. Adu, J. A. Addae, O. K. Omoregie, and K. S. Ofori, "Examining self-disclosure on social networking sites: A flow theory and privacy perspective," *Behavioral Sciences*, vol. 8, no. 6, pp. 58, 2018.

[4] Y. Ashibani and Q. H. Mahmoud, "Cyber physical systems security: Analysis, challenges and solutions," *Computers & Security*, vol. 68, pp. 81–97, 2017.

[5] B. Chen and Z. Wu, "Research on personal privacy protection in big data environment," *Journal of Jilin Normal University of engineering and technology*, vol. 35, no. 8, pp. 68–70, 2019.

[6] T. T. El-Khazendar and T. S. M. Barhoom, "Protect local data on personal devices: Third party application," 2015.

[7] M. Fodor and A. Brem, "Do privacy concerns matter for millennials? Results from an empirical analysis of location-based services adoption in germany," *Computers in Human Behavior*, vol. 53, pp. 344–353, 2015.

[8] T. Gao, T. Li, R. Jiang, M. Yang, and R. Zhu, "Research on cloud service security measurement based on information entropy.," *International Journal Network Security*, vol. 21, no. 6, pp. 1003–1013, 2019.

[9] A. Gutierrez, S. O'Leary, N. P. Rana, Y. K. Dwivedi, and T. Calle, "Using privacy calculus theory to explore entrepreneurial directions in mobile location-based advertising: Identifying intrusiveness as the critical risk factor," *Computers in Human Behavior*, vol. 95, pp. 295–306, 2019.

[10] B. N. Jagdale and J. W. Bakal, "Controlled broadcast protocol for location privacy in mobile applications," *Procedia Computer Science*, vol. 78, pp. 782–789, 2016.

[11] M. J. Keith, S. C. Thompson, J. Hale, P. B. Lowry, and C. Greer, "Information disclosure on mobile devices: Re-examining privacy calculus with actual user behavior," *International Journal of Human-Computer Studies*, vol. 71, no. 12, pp. 1163–1173, 2013.

[12] C. T. Li, C. C. Yang, and M. S. Hwang, "A secure routing protocol with node selfishness resistance in manets," *International Journal of Mobile Communications*, vol. 10, no. 1, pp. 103–118, 2012.

[13] H. LI, B. Wang, W. Zhang, Q. Tang, and Y. Zhang, "X-Decaf: Detection of cache file leaks in android social apps," *Journal of Electronics & Information Technology*, vol. 39, no. 1, pp. 10, 2017.

[14] Z. Li, Y. Tian, W. Zhang, and Y. Liu, "Research on china mobile application privacy policy," *Cyberspace Security*, vol. 11, no. 6, pp. 11, 2020.

[15] N. W. Lo, K. H. Yeh, and C. Y. Fan, "Leakage detection and risk assessment on privacy for android applications: Lrpdroid," *IEEE Systems Journal*, vol. 10, no. 4, pp. 1361–1369, 2014.

[16] Y. Nan, Z. Yang, M. Yang, S. Zhou, Y. Zhang, G. Gu, X. Wang, and L. Sun, "Identifying user-input privacy in mobile applications at a large scale," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 3, pp. 647–661, 2016.

[17] A. Ruiz-Heras, P. García-Teodoro, and L. Sánchez-Casado, "Adroid: Anomaly-based detection of malicious events in android platforms," *International Journal of Information Security*, vol. 16, no. 4, pp. 371–384, 2017.

[18] F. Shirazi and A. Iqbal, "Community clouds within m-commerce: A privacy by design perspective," *Journal of Cloud Computing*, vol. 6, no. 1, pp. 22, 2017.

[19] Y. A. Tan, Y. Xue, C. Liang, J. Zheng, Q. Zhang, J. Zheng, and Y. Li, "A root privilege management scheme with revocable authorization for android devices," *Journal of Network and Computer Applications*, vol. 107, pp. 69–82, 2018.

[20] B. Tian, Y. Zheng, P. Liu, and C. Li, "The evaluation index and empirical study on risk of privacy informationdisclosure of mobile app users," *Library and Information Service*, vol. 62, no. 19, pp. 101–110, 2018.

[21] G. Tilei, L. Tong, Y. Ming, and J. Rong, "Research on a trustworthiness measurement method of cloud service construction processes based on information entropy," *Entropy*, vol. 21, no. 5, pp. 462, 2019.

[22] J. Wang, J. Liu, and H. Zhang, "Access control based resource allocation in cloud computing environment.," *International Journal Network Security*, vol. 19, no. 2, pp. 236–243, 2017.

[23] C. H. Wei, M. S. Hwang, and A. Y. H. Chin, "An improved authentication protocol for mobile agent device in rfid environment," *International Journal of Mobile Communications*, vol. 10, no. 5, pp. 508–520, 2012.

[24] Y. C. Wei, W. C. Wu, G. H. Lai, and Y. C. Chu, "pISRA: Privacy considered information security risk assessment model," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 1468–1481, 2020.

[25] V. M. Wottrich, E. A. van Reijmersdal, and E. G. Smit, "The privacy trade-off for mobile app downloads: The roles of app value, intrusiveness, and privacy concerns," *Decision Support Systems*, vol. 106, pp. 44–52, 2018.

[26] Y. Wu, G. Feng, N. Wang, and H. Liang, "Game of information security investment: Impact of attack types and network vulnerability," *Expert Systems with Applications*, vol. 42, no. 15-16, pp. 6132–6146, 2015.

[27] M. Xiang, X. Wang, R. Jia, and L. Wang, "The evaluation index and empirical study on risk of privacy informationdisclosure of mobile app users," *Library and Information Service*, vol. 64, no. 18, pp. 34–44, 2018.

[28] Y. Z. Xu, J. L. Zhang, Y. Hua, and L. Y. Wang, "Dynamic credit risk evaluation method for e-commerce sellers based on a hybrid artificial intelligence model," *Sustainability*, vol. 11, no. 19, pp. 5521, 2019.

[29] Y. Xu and Y. Ji, "Android application risk assessment method based on permission," *Computer Applications and Software*, vol. 37, no. 4, pp. 69–74+100, 2020.
[30] Y. Xu and Y. Ji, "An android application risk assessment strategy based on permission characteristics," *Computer Applications and Software*, vol. 37, no. 4, pp. 69–74,100, 2020.
[31] M. Yang, R. Jiang, T. Gao, W. Xie, and J. Wang, "Research on cloud computing security risk assessment based on information entropy and markov chain," *International Journal Network Security*, vol. 20, no. 4, pp. 664–673, 2018.
[32] L. Ying, "Research on personal privacy protection in big data environment," *Computers and Networks*, vol. 46, no. 2, pp. 68–70, 2020.
[33] G. Zhu, M. N. Feng, Y. Chen, and J. Y. Yang, "Research on fuzzy evaluation of privacy risk for social network in big data environment," *Information Science*, vol. 34, no. 9, pp. 19, 2016.
[34] H. Zhu, C. X. J. Ou, W. J. A. M. van den Heuvel, and H. Liu, "Privacy calculus and its utility for personalization services in e-commerce: An analysis of consumer decision-making," *Information & Management*, vol. 54, no. 4, pp. 427–437, 2017.

Biography

Ming Yang is an associate professor at the school of information, Yunnan University of Finance and Economics, China. He received his Ph.D. in system analysis and integration from the school of software at Yunnan University. His main research interests include information management and data mining.

Li Jia is an associate professor at the school of information, Yunnan University of Finance and Economics, China. His main research fields are network communication and security control, data mining technology.

Tilei Gao is a lecturer at the school of information, Yunnan University of Finance and Economics. He is also a Ph.D candidate in system analysis and integration at the school of software at Yunnan University. His main research interests include software engineering and information management.

Tao Zhang received the M.A. degrees in Information Management and System from School of Information, Yunnan University of Finance and Economics, China, in 2012. At present, he is a lecturer at the School of Information, Yunnan University of Finance and Economics. Also he is a Ph.D. candidate. His research interests include Information Security, Information Management and Information System, Recommendation system.

Wanyu Xie is a lecturer at Kunming Metallurgy College. She received Her master's degree in computer science and technology from the school of information science and engineering at Yunnan University. Her main research interests is information management.

Mining Frequent Sequential Patterns with Local Differential Privacy

Huihua Xia, Wenchao Huang, Yan Xiong, and Fuyou Miao
(Corresponding author: Wenchao Huang)

School of Computer Science and Technology, University of Science and Technology of China
West Campus of USTC, Huang Shan Road, Hefei, Anhui Province, China
Email: huangwc@ustc.edu.cn

(Received May 31, 2020; Revised and Accepted Feb. 11, 2021; First Online Aug. 14, 2021)

Abstract

A massive amount of sequential data are being collected and analyzed. Mining frequent sequential patterns among these data can bring benefits for various real-life applications. However, privacy has been a major concern during the analysis of sequential data. We propose LDPSPM for mining frequent sequential patterns while satisfying local differential privacy. Specifically, each user randomly responds with a set of sanitized sequential patterns. Then the collector can estimate frequencies of the sequential patterns from the sanitized data. Moreover, we propose P-LDPSPM, which further improves the performance by filtering out nonsignificant items. Theoretical analysis and extensive experiments confirm the effectiveness and efficiency of our mechanism.

Keywords: Local Differential Privacy; Privacy Preserving; Sequential Pattern Mining

1 Introduction

With the development of web, mobile and communication technologies, massive sequential data are being recorded and collected [1, 12]. Specifically, sequential data are sequences of items with time correlations, e.g., mobile applications launching sequence, webpage click stream and user moving trajectory. Mining such sequential data can bring benefits for many real-life applications such as travel suggestion, city planning, advertising and *etc.* For example, the server can make travel suggestions for users after mining a trajectory database. In this paper, we focus on the task of sequential pattern mining, *i.e.*, finding interesting sub-sequences in a sequence database.

Privacy is a major concern in the current era of cloud computing [13, 21, 22, 28, 29]. Sequential data contains much privacy information of users. For example, user preference may be inferred from the web browsing history [40]. Personal trajectory data can be used to infer sensitive information such as home address, health status and religious faith [11]. Instead of directly collecting and publishing these sequential data, privacy-preserving mechanisms should be considered during the mining of sequential data.

Some researchers [4, 6, 7, 10, 20, 26, 31, 41] have studied the privacy-preserving mechanisms when processing sequential data. Most of them consider a trusted centralized server. All the users' sequential data are collected and stored on the server. The server takes responsibility for the privacy of the users and carries out privacy-preserving mechanisms. Some of the works [6, 7, 9, 20, 31] study the problem of privacy-preserving data publishing, *i.e.*, publish the entire sequential database after sanitizing the database. The other works [4, 10, 41] publish statistical information (*e.g.*, frequent sequential patterns) of the database other than the entire database. However, the assumption of a centralized server is not practical. The data security can not be guaranteed since the server might be corrupted by attackers.

Recently, local differential privacy (LDP) [3, 15, 17] has become the de facto notion for privacy-preserving while avoiding the assumption of a trusted centralized server. The users take full control of their data and the server never collects the exact value of any personal data. Before sending the data to the server, the users sanitize their data locally with privacy-preserving techniques while ensuring that specific statistical information can be derived from the sanitized data. For example, Google embedded their LDP solutions called RAPPOR [17] into the Chrome browser, which enables Chrome to collect information such as the default homepage of the browser. Samsung [32] developed an LDP system for collecting user data from smart devices, which can deal with not only categorical but also numerical data.

Though solutions of achieving LDP for different tasks have been studied for years, to the best of our knowledge, LDP mechanisms for sequential pattern mining have not been studied yet. Most of the existing works assume the data type as single-valued data [17, 37] or set-valued data [34, 36, 38]. Usually the data of each user is encoded into a binary vector [17], where each bit in the vector de-

notes whether a value is held by the user. Such methods of encoding can not be applied to sequential data since the sequential characteristics between values can not be reflected by a binary vector. Another challenge of designing an LDP protocol for sequential pattern mining comes from the size of the domain. Assume that there are d items in total, then the number of sequences of length-m (suppose there are no duplicate items in a sequence) composed of these items is A_d^m, which is explosive when d is large.

In this paper, we solve the problem of mining the top-k sequential patterns among a set of sequences while satisfying LDP. We propose a novel and efficient mechanism called LDPSPM for mining sequential patterns with LDP. We adopt the idea of exponential mechanism from the traditional differential privacy. Instead of encoding the user data, each user randomly responds with a set of sequential patterns, and the size of the set is proved to be optimized. Moreover, to handle the problem of the large domain, we design a pruning step, which narrows down the size of the domain significantly and improves the effectiveness of our mechanism. Theoretical analysis and extensive experimental results on both synthetic and real-world datasets show the effectiveness of our mechanism and the improvement of the pruning step.

The contributions we made in this paper are as follows:

- We propose a novel and efficient LDP mechanism for mining sequential patterns. We analyze the error bound of the results theoretically and optimize the parameters based on the theoretical analysis.
- We design a pruning step for our protocol, which significantly narrows down the item domain and improves the accuracy of the results.
- We evaluate our protocols on both synthetic datasets and real-world datasets. The results show the utility and effectiveness of our protocol.

The rest of this paper is organized as follows. Section 2 reviews the related works on sequential pattern mining and local differential privacy. Section 3 gives the formal definition of LDP and reviews several existing LDP protocols. Section 4 formalizes the problem of LDP sequential pattern mining. Section 5 proposes the solution and the utility analysis of the LDPSPM method for sequential pattern mining. Section 6 improves LDPSPM with a pruning step. Section 7 gives experimental results of our mechanisms. At last, Section 8 concludes the work.

2 Related Work

In this section, we review the related works on both sequential pattern mining and local differential privacy.

2.1 Sequential Pattern Mining

The task of sequential pattern mining is to find all frequent sub-sequences in a sequence database. Much effort has been made to improve the efficiency of sequential pattern mining. AprioriAll [1] is the first algorithm for sequential pattern mining, which is inspired by the Apriori algorithm for frequent itemset mining. GSP [35] is proposed as an improved version of AprioriAll based on the breadth-first search. Several depth-first search algorithms are proposed which exclude irrelevant patterns and show more efficiency than GSP, such as Spade [43], PrefixSpan [33], Spam [2], Lapin [42], and CM-Spam [19].

The task of privacy-preserving sequential pattern mining has been studied for years. Some of the works [6, 7, 20, 31] focus on publishing a sanitized sequence database which can be used for mining sequential patterns. Other approaches aim to mine the sequential patterns directly on the original database while preserving privacy. Most of them are based on the concept of differential privacy [16]. Bonomi *et al.* [4] propose a two-phase differentially private protocol for mining frequent consecutive-item sequences, which utilizes a prefix tree to find candidate sequences, and then leverages a database transformation technique to refine the support of the candidate sequences. Xiang *et al.* [10] propose DP-MFSM for finding maximal frequent sequences based on the idea of candidate pruning. Xu *et al.* [41] estimate the sequences that are potentially frequent based on sample databases, and then reduce the number of candidate sequences. Li *et al.* [27] solve the problem of time-constrained sequential pattern mining under differential privacy, where the transition time between adjacent items in frequent sequential patterns is constrained. Le *et al.* [25] mine frequent sequential patterns in electronic medical record systems while considering time interval. They restrain the added noise by adding noise only to a set of candidate closed sequences. However, all these approaches assume a centralized server, which collects the exact data of the users and carries out the privacy-preserving mechanisms. If the centralized server is corrupted, the privacy of the users can no longer be preserved.

2.2 Local Differential Privacy

Local differential privacy (LDP) is proposed to avoid the use of the centralized server. LDP protects the privacy of users in a local manner that each user sanitizes his data before sending it to the server.

Most of the existing LDP solutions focus on frequency estimation over categorical data. Duchi *et al.* [15] propose the LDP mechanism for histogram estimation and theoretically analyze the optimality of utility. Google deploys its LDP algorithm called RAPPOR [17] in practice, which is the first LDP solution in real-world applications. RAPPOR encodes user's data by a Bloom filter and then applies the randomized response [39] to perturb it. Bassily *et al.* [3] propose a protocol that uses random matrix projection and reduces the communication cost of each user greatly than RAPPOR. Kairouz *et al.* [23] extend RAPPOR to support categorical attributes with arbitrary number of possible values. Wang *et al.* [37] propose a

general framework and compare several LDP protocols theoretically, and further optimize the parameters of the protocols to provide better utility of the results.

Other works take frequency estimation as a primitive to solve different tasks. Chen *et al.* [8] propose a mechanism to learn the spatial distribution of users under LDP. Kim *et al.* [24] finish the task of collecting indoor positioning data under LDP. Cormode *et al.* [14] solve the problem of answering range counting queries under LDP. Several works [34, 36, 38] deal with set-valued data and solve the task of frequent itemset mining. Qin *et al.* [34] propose a two-phase heavy hitter estimation mechanism to estimate the top frequencies of items from set-valued data. They prune the item domain first and thus improve the estimation accuracy. Wang *et al.* [38] extend the work of [34], which improves the effectiveness of the results and enables the frequency estimation of itemsets. Wang *et al.* [36] propose PrivSet for frequency estimation of single items and set cardinality estimation, which privatizes items in set-valued data as a whole, and takes full utilization of the privacy budget.

However, to the best of our knowledge, none of the existing LDP mechanisms consider the data type of sequential data and solve the task of frequent sequential data mining.

3 Background

3.1 Local Differential Privacy

Local differential privacy is a notion of privacy for data collection originated from differential privacy. A collector collects data from users in the local setting. The users perturb their data through a randomized mechanism before sending their data to the collector. In this paper, we consider the situation that each user has a sequence s. The formal definition of local differential privacy is as follows:

Definition 1. (Local Differential Privacy) *A randomized mechanism K satisfies ϵ-local differential privacy (ϵ-LDP), where $\epsilon > 0$, if and only if for any two sequences s_1 and s_2, we have*

$$\forall y \in Range(K) : \Pr[K(s_1) = y] \leq e^{\epsilon} \cdot \Pr[K(s_2) = y]$$

where $Range(K)$ denotes the output domain of K.

Similar to the differential privacy in the centralized setting, there is a property of sequential composition [30] for ϵ-LDP.

Theorem 1. *Given a set of randomized mechanism K_i, each of which satisfies ϵ_i-LDP, then the whole process of sequentially executing K_i satisfies $(\sum \epsilon_i)$-LDP.*

Given the property of sequential composition, each user can partition the privacy budget ϵ into several portions and adopt several randomized mechanisms, while the whole process satisfies ϵ-LDP.

3.2 Existing Protocols

Next, we overview several existing LDP solutions. All these protocols assume that the users hold categorical data (a single value or a set of values) and aim to estimate the frequencies of the items.

RAPPOR. RAPPOR [18] is designed based on the idea of random response [17].

Assume there are n users, and each user u_i possesses exactly one item v_i (an integer). The items come from a domain containing d items, and a collector wants to estimate the frequency of each item. For a user with an item v_i, he first encodes v_i into a length-d binary vector B_i such that $B_i[v] = 1$ and the other bits are 0. Then the user applies a random response on B_i and gets a new binary vector B_i', such that:

$$\Pr[B_i'[j] = 1] = \begin{cases} p = \frac{e^{\epsilon/2}}{1+e^{\epsilon/2}}, & \text{if } B_i[j] = 1 \\ q = \frac{1}{1+e^{\epsilon/2}}, & \text{if } B_i[j] = 0 \end{cases} \quad (1)$$

Such mechanism satisfies ϵ-LDP for each user. The binary vector B_i' is sent to the collector.

Upon getting all the n users' responses, the collector can estimate the frequency of an item v as follows:

$$f(v) = \frac{\sum_i \mathbb{1}_{\{v|B_i'[v]=1\}}(v) - nq}{p - q} \quad (2)$$

where $\mathbb{1}_Y(v)$ is an indicator function that

$$\mathbb{1}_X(v) = \begin{cases} 1, & \text{if } v \in Y \\ 0, & \text{if } v \notin Y \end{cases}$$

To transfer a length-d binary vector, the communication cost is $O(d)$ for each user, which is expensive when d is large.

Random sampling. The method of RAPPOR above deals with the assumption that each user holds exactly one item.

When each user possesses a set of values (assume that the size of the set is fixed to be l) and the collector wants to estimate the frequency of each item, one naive method which changes RAPPOR slightly is as follows. Each user generates a length-d bit vector B', similar to that in RAPPOR, with exactly l ones in B'. Then the two probability p and q in Equation (1) changes into $p = \frac{e^{\epsilon/(2l)}}{1+e^{\epsilon/(2l)}}$ and $q = \frac{1}{1+e^{\epsilon/(2l)}}$. The aggregation step is the same as that in RAPPOR.

Qin *et al.* [34] claim that the naive method introduce high noise in the results, and they propose the method of random sampling. Instead of reporting all the items, each user samples one item from the set randomly and applies RAPPOR on that item. To solve the bias caused by the random sampling, the collector multiplies the frequency by l during the stage of aggregation. The method of random sampling narrows down the error in the results significantly.

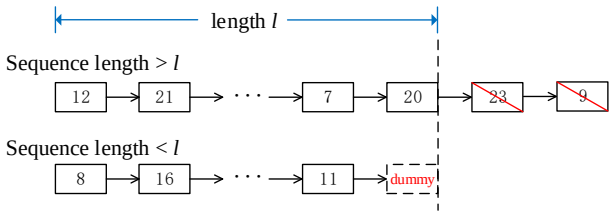


Figure 1: Sequence padding and truncation

4 Problem Definition

This paper focuses on mining frequent sequential patterns among a set of sequences. The system contains n users and one collector. The collector collects sequential data from the users with local differential privacy and estimates the frequency of sequential patterns.

Formally, the sequential data of user u_i is a list of items $s_i = \langle v_1, v_2, ..., v_l \rangle$. $v_1, ..., v_l$ are the items possessed by the user and we assume that each item only appears once in a sequence. For simplicity, we assume that the number of items in the sequential data of each user is fixed as l. In real applications, if the user has less than l items, he pads his sequence with dummy items (l dummy items are needed), which will be ignored by the collector. If the user has more than l items, he truncates his sequence to the length of l. An example of the padding and truncation of sequences is shown in Figure 1. The loss of information may introduce bias to the results, so l should be carefully chosen such that most of the users have less than l items. We denote the domain of the items (including the dummy items) as X. We assume that the number of real items in X is d, then the size of X is $d + l$.

The collector focuses on sequential patterns of the sequences, which is defined as follows.

Definition 2. (Sequential Pattern). *A sequence* $sp = \langle v_{n_1}, v_{n_2}, ..., v_{n_k} \rangle$ *is a sequential pattern of sequence* $s = \langle v_1, v_2, ..., v_l \rangle$, *iff sp can be derived from s by deleting some or no elements without changing the order of the remaining elements.*

The length of a sequential pattern is defined as the number of elements in it. We assume that the collector focuses on sequential patterns of fixed length and the length is denoted as m,*i.e.*, there are m items in the sequential pattern, and we call it length-m sequential pattern.

Let Φ_i be the set of all length-m sequential patterns in the sequence s_i. Let f_{sp} be the frequency of a sequential pattern sp. Formally,

$$f_{sp} = \frac{|\{u_i | sp \in \Phi_i, 1 \le i \le n\}|}{n}$$

The goal of the collector is to estimate the frequencies of all the length-m sequential patterns and find the top-k patterns with the highest frequencies, while satisfying LDP. We assume that the results only contain the top few patterns. The accuracy of the results should consider both the ranking of the top-k patterns and their frequencies.

5 Proposed Method

In this section, we describe the solution for mining top-k length-m sequential patterns while satisfying LDP, including the randomization step for each user and the frequency estimation for the collector.

5.1 Data Randomization

Most of the existing ϵ-LDP mechanisms encode the data into bit vectors. However, bit vector can not encode the order of items in sequences. Wang *et al.* [36] analyzed set-valued data with the method of exponential mechanism, which is a common solution for traditional differential privacy. We follow the idea and solve our task of LDP sequential pattern mining with the exponential mechanism.

Let Ω be the set of all the length-m sequences combined by the items in X. For a user u_i with a sequence s_i, he randomly responds with a subset $t \subseteq \Omega$ and the size of t is fixed as w. Let Φ_i be the set of all the length-m sequential patterns in s_i. The probability that the subset t is chosen is as follows,

$$\Pr[K(s_i) = t] \propto \exp(\epsilon \cdot \frac{u(t, \Phi_i)}{\Delta u}) \quad (3)$$

where $u(t, \Phi_i)$ is a utility function which represents the similarity between t and Φ_i. We define $u(t, \Phi_i)$ as the number of common elements between t and Φ_i, *i.e.*, $u(t, \Phi_i) = |t \cap \Phi_i|$. Δu is the sensitivity of the utility function which is defined as

$$\Delta u = \max_{i,j \in [1,n]} |u(t, \Phi_i) - u(t, \Phi_j)|$$

Thus we have $\Delta u = w$ based on our definition of $u(t, \Phi_i)$. Let λ and φ be the size of Ω and Φ, *i.e.*, $\lambda = |\Omega|$ and $\varphi = |\Phi|$. We give the detailed process of the data randomization in Algorithm 1.

In the algorithm, given a user's sequence s and the domain of items X, Ω denotes the domain of all possible length-m sequences (line 2) and Φ denotes the set of all length-m sequential patterns in s (line 3). B is computed as the normalizer (line 6). Specifically, we have

$$B = \sum_{t \in \Omega} \exp(\epsilon \cdot \frac{|t \cap \Phi|}{w}) = \sum_{j=0}^{w} \binom{\varphi}{j} \binom{\lambda - \varphi}{w - j} \exp(\frac{\epsilon \cdot j}{w}) \quad (4)$$

Note that the value of B is constant as we assume that the input sequences have the same length. The probability that the user samples a set t of sequences from Ω is

$$\Pr[K(s_i) = t] = \exp(\epsilon \cdot \frac{u(t, \Phi)}{w})/B \quad (5)$$

The random variable r (line 7) is used to determine the value of $int = |t \cup \Phi|$, *i.e.*, the size of the intersection of t and Φ, which is computed in line 11-12. At last, the set t is sampled with int sequences coming from Φ and $w - int$ sequences coming from Ω (line 14-15). The function sample(Y, n) samples n items from Y without

Algorithm 1 Randomization Step of LDPSPM

Input: $\boldsymbol{s} = \langle v_1, v_2, ..., v_l \rangle$: sequential data, X: domain of items, w: size of the output set.

Output: t: a set of w length-m sequences.

$t = \emptyset$
$\Omega = \{< v_1, v_2, ..., v_m > | v_1, v_2, ..., v_m \in X\}$
$\Phi = \{\text{all length-}m \text{ sequential patterns in } s\}$
$\lambda = |\Omega|$
$\varphi = |\Phi|$
$B = \sum_{j=0}^{w} \binom{\varphi}{j} \binom{\lambda-\varphi}{w-j} \exp(\frac{\epsilon \cdot j}{w})$
$r = \text{uniform}(0, 1)$
$int = 0$
$prob = \binom{\lambda-\varphi}{w} / B$
while $prob < r$ **do**
  $int = int + 1$
  $prob = prob + \binom{\varphi}{int} \binom{\lambda-\varphi}{w-int} \exp(\frac{\epsilon \cdot int}{w})/B$
end while
$t = t \cup \text{sample}(\Phi, int)$
$t = t \cup \text{sample}(\Omega, w - int)$
return t

replacement. The sample function can be done in $O(w^2)$. As we find in the experiments, the optimal w is 1, and thus the randomization step for each user can be finished in $O(1)$ time. Each user responds with a set of w length-m sequences. An item costs d bits during the communication, so the communication cost for each user to transfer w length-m sequences is $O(wmd)$.

Theorem 2. *The randomization step of LDPSPM in Algorithm 1 satisfies ϵ-LDP.*

Proof. For any two possible sequences s_1 and s_2, and any output t, we have

$$\begin{aligned}\frac{\Pr[K(s_1) = t]}{\Pr[K(s_2) = t]} &= \frac{\exp(\epsilon \cdot u(t, \Phi_1)/w)/B}{\exp(\epsilon \cdot u(t, \Phi_2)/w)/B} \\ &= \exp(\epsilon \cdot \frac{u(t, \Phi_1) - u(t, \Phi_2)}{w}) \\ &\leq \exp(\epsilon \cdot \Delta u/w) \\ &= \exp(\epsilon)\end{aligned}$$

□

5.2 Frequency Estimation

During the process of data randomization, a user u_i with sequence s_i sends a randomized set of sequences t_i to the collector. The collector aims to find the top-k length-m sequential patterns of the original sequences based on the responses $\{t_1, t_2, ..., t_n\}$ collected from the users.

Consider a length-m sequence $sp_m \in \Omega$, a user's sequence s and the random response t, we define two probability values P and Q as follows,

$$\begin{aligned}\Pr[sp_m \in t \,|\, sp_m \in \Phi_s] &= P, \\ \Pr[sp_m \in t \,|\, sp_m \notin \Phi_s] &= Q.\end{aligned}$$

where Φ_s denotes the set of all the length-m sequential patterns in s. Specifically, we have

$$P = \sum_{j=1}^{w} \binom{\varphi-1}{j-1} \binom{\lambda-\varphi}{w-j} \exp(\frac{\epsilon \cdot j}{w})/B \tag{6}$$

$$Q = \sum_{j=0}^{w-1} \binom{\varphi}{j} \binom{\lambda-\varphi-1}{w-j-1} \exp(\frac{\epsilon \cdot j}{w})/B \tag{7}$$

The values of P and Q remain the same for any s and sp_m, as we assume that the length of the users' sequences are the same and any element only appears once in a sequence.

Let c_{sp} be the number of times the sequential pattern sp occurs in the responses $\{t_1, t_2, ..., t_n\}$. Specifically, we have

$$c_{sp} = \sum_{i=1}^{n} \mathbb{1}_{t_i}(sp) \tag{8}$$

The frequency of sp existing as a sequential pattern in the original sequences $\{s_1, s_2, ..., s_n\}$ can be estimated as:

$$\tilde{f}_{sp} = \frac{c_{sp} - nQ}{n(P - Q)} \tag{9}$$

Theorem 3. *$\tilde{f}_{sp}$ is an unbiased estimation of f_{sp} for any sequential pattern sp,* i.e., *$\forall_{sp} \mathrm{E}[\tilde{f}_{sp}] = f_{sp}$, where f_{sp} is the true frequency of sp in the original sequences $\{s_1, s_2, ..., s_n\}$.*

Proof.

$$\begin{aligned}\mathrm{E}[\tilde{f}_{sp}] &= \mathrm{E}\left[\frac{c_{sp} - nQ}{n(P - Q)}\right] \\ &= \frac{\mathrm{E}[c_{sp}] - nQ}{n(P - Q)} \\ &= \frac{nf_{sp}P + n(1 - f_{sp})Q - nQ}{n(P - Q)} \\ &= \frac{n(f_{sp}P + Q - f_{sp}Q - Q)}{n(P - Q)} \\ &= f_{sp}\end{aligned}$$

□

After getting the estimated frequencies of all the sequential patterns, the collector sorts all the patterns and finds the top-k sequential patterns. The frequency estimation component of LDPSPM for the collector is shown as Algorithm 2.

5.3 Utility Analysis

5.3.1 Error Bound

We use mean squared error to measure the error of the estimated frequencies of sequential patterns. For a sequential pattern sp with frequency f_{sp}, the estimated frequency is denoted as $\tilde{f}_{sp}$. The mean squared error of the estimated frequency is measured as $\mathrm{E}[|f_{sp} - \tilde{f}_{sp}|^2]$.

Algorithm 2 Frequency Estimation Step of LDPSPM

Input: $T = \{t_1, t_2, ..., t_n\}$: responses from n users.
Output: F_{top_k}: the top-k sequential patterns along with their frequencies.

$C = \{c_{sp_1} = 0, c_{sp_2} = 0, ..., c_{sp_\lambda} = 0\}$
$F = \{f_{sp_1} = 0.0, f_{sp_2} = 0.0, ..., f_{sp_\lambda} = 0.0\}$
for $t \in T$ **do**
  for $sp \in t$ **do**
    $c_{sp} = c_{sp} + 1$
  end for
end for
for $sp \in \Omega$ **do**
  $\tilde{f}_{sp} = \frac{c_{sp} - nQ}{n(P-Q)}$
end for
Sort F and get the top-k frequencies F_{top_k}
return F_{top_k}

Theorem 4. *For a sequential pattern sp, the mean squared error of the estimated frequency $\tilde{f}_{sp}$ is:*

$$\mathrm{E}[|f_{sp} - \tilde{f}_{sp}|^2] = \frac{f_{sp}P(1-P) + (1-f_{sp})Q(1-Q)}{n(P-Q)^2} \quad (10)$$

Proof. The variable $\tilde{f}_{sp}$ in Equation (9) is a linear transformation of c_{sp}. According to Equation (8), the variable c_{sp} is the summation of n independent Bernoulli random variables. Specifically, nf_{sp} (resp. $(1-f_{sp})n$) of them are drawn from the Bernoulli distribution with parameter P (resp. Q). $\tilde{f}_{sp}$ is an unbiased estimation of f_{sp} as shown in Theorem 2, thus we have

$$\begin{aligned}
\mathrm{E}[|f_{sp} - \tilde{f}_{sp}|^2] &= \mathrm{Var}[\tilde{f}_{sp}] \\
&= \mathrm{Var}\left[\frac{\sum_{i=1}^{n} \mathbb{1}_{t_i}(sp) - nQ}{n(P-Q)}\right] \\
&= \frac{\sum_{i=1}^{n} \mathrm{Var}[\mathbb{1}_{t_i}(sp)]}{n^2(P-Q)^2} \\
&= \frac{f_{sp}P(1-P) + (1-f_{sp})Q(1-Q)}{n(P-Q)^2}
\end{aligned}$$

□

5.3.2 Choosing w

We assume that each user responds with a set of w length-m sequential patterns in Algorithm 1. The value of w should be determined to minimize the error of the estimation. We aim to minimize the sum of the mean squared errors of the estimated frequency of all the sequential patterns, *i.e.*,

$$\sum_{sp} \mathrm{E}[|f_{sp} - \tilde{f}_{sp}|^2] = \frac{\varphi P(1-P) + (\lambda - \varphi)Q(1-Q)}{n(P-Q)^2} \quad (11)$$

We numerically compute the sum of errors in Equation (11) for every $w \in [1, \lambda]$ and choose the value of w which minimizes the sum.

6 Prune the Domain

In this section, we propose an improved mechanism, called P-LDPSPM, which further improves the performance of LDPSPM.

As we assumed in Section 4, the collector is interested in only a few top-k sequential patterns. When the item domain is large, massive noises will be introduced in responses, due to those items which are not contained in the top-k sequential patterns, and thus affect the accuracy of the estimated frequencies. As the result shown in Equation (10), the mean squared error of the estimated frequency is proportional to the size of Ω, and thus has an exponential relationship with the size of the item domain, which is confirmed by numerical computation. If we can narrow down the size of the domain, the accuracy of the results will be improved.

As we observe in real-life sequential datasets, if we treat the sequences as set-valued data (*i.e.*, ignore the order of the items in a sequence and treat them as a set), the items that contained in the top sequential patterns show higher frequencies than the others. Based on this observation, we design a step of pruning before the random response to extract several candidate items. Due to the privacy concern, this step should also meet the demand of LDP, which can be done using existing LDP solutions on set-valued data, *e.g.*, naive RAPPOR and random sampling we described in Section 3. We choose random sampling based on RAPPOR because of its accuracy, and we show the process in Algorithm 3.

Algorithm 3 Sampling RAPPOR

Input: $\boldsymbol{s} = \langle v_1, v_2, ..., v_l \rangle$: sequential data, ϵ_1: the privacy budget for pruning.
Output: B': the output binary vector.

$B = \mathbf{0}$, $B' = \mathbf{0}$
Uniformly choose an item from the sequence s
$B[v] = 1$
Randomize B' as follows:

$$\Pr[B'[j] = 1] = \begin{cases} \frac{e^{\epsilon_1/2}}{1+e^{\epsilon_1/2}}, & \text{if } B[j] = 1 \\ \frac{1}{1+e^{\epsilon_1/2}}, & \text{if } B[j] = 0 \end{cases}$$

return B'

Theorem 5. *The sampling RAPPOR algorithm satisfies ϵ_1-LDP* [34].

One solution for the whole system to achieve LDP is splitting the privacy budget for the pruning step based on the property of sequential composition in Theorem 1. Another solution is dividing the users into two groups, one group for the pruning step, and the other group for the LDPSPM algorithm. Both two groups use the full privacy budget. We choose the second solution, as it is proven [37] that the method of dividing users has a better utility than dividing the privacy budget. We divide the

users into two groups. The first group contains 20% of the population and finishes the pruning task. The second group contains 80% of the population and finishes the LDPSPM task.

After collecting the bit vectors from the first group, the collector computes the frequencies as in Equation (2), and multiplies the results by l to get an unbiased estimation of the frequency of each item. Then the collector picks a candidate set of k_{max} items with the highest frequencies (excluding the dummies). Based on the observation on real-life datasets, we set $k_{max} = km$ in our work, which can cover most of the items in the top-k sequential patterns. It's worth mentioning that when k exceeds some value, the candidates will cover all items and the pruning step loses the advantages. Thus we remove the pruning step if $km > d$ and all users are engaged in LDPSPM. The collector then sends the k_{max} candidate items to the second group. For each user in the second group, he adjusts his sequence by replacing those items which are not in the candidate set by dummy items. At last, each user in the second group adopts the random response in Algorithm 1 for mining frequent sequential patterns.

We summarize the mechanism of our pruning based LDP sequential pattern mining (P-LDPSPM) as follows:

Step 1: (User) Random sampling. Each user in the first group randomly samples one item from his sequence and adopts random sampling on it. Then he sends the noisy result to the collector.

Step 2: (Collector) Prune the domain. The collector picks a set of candidates to narrow down the domain based on the results collected from Step 1.

Step 3: (User) Randomization of LDPSPM. Each user in the second group adjusts his sequence based on the candidates and adopts random response in Algorithm 1.

Step 4: (Collector) Frequency estimation. The collector estimates the frequency of each length-m sequential pattern and gets the top-k sequential patterns.

7 Evaluation

In this section, we conduct experiments to evaluate LDPSPM and P-LDPSPM on both synthetic and real datasets. Specifically, we seek to answer the following questions. First, how key parameters affect the results of the estimation of top-k sequential patterns. Second, how the pruning step improves the results of LDPSPM.

7.1 Datasets

Synthetic Datasets: The synthetic datasets are sequences generated by the IBM Quest data generator. We vary the parameters and generate several datasets to see the impact of different parameters.

Real Dataset: The dataset is constructed by sequences of page views on msnbc.com of different users for the entire day of September, 28, 1999 [5].

Each sequence recorded the categories of pages that each user requested. We modify the original dataset to make sure that each category appears at most once in each sequence. The modified dataset contains 388,434 sequences. Each sequence is truncated or padded to a fixed length of 9.

7.2 Parameters

There are several key parameters that may affect the effectiveness of the algorithms.

Number of users (n). The ϵ-LDP mechanisms proposed in this paper are essentially based on random response and large noises are contained in the results. A large population can effectively remove the bias introduced by the noise.

Privacy budget (ϵ). The privacy budget determines the amount of noise to be added, and thus affect the accuracy of the estimation of the collector.

Number of top sequential patterns (k). We expect that the results will be better when k is small because the top sequential patterns have higher frequencies and thus can resist the noise to some extent.

In the experiments, we evaluate the impact of different parameters on the effectiveness of our algorithms. The optimized value of w is always 1 when we change the parameters in our experiments, thus we set $w = 1$ by default in all the next experiments.

7.3 Metrics

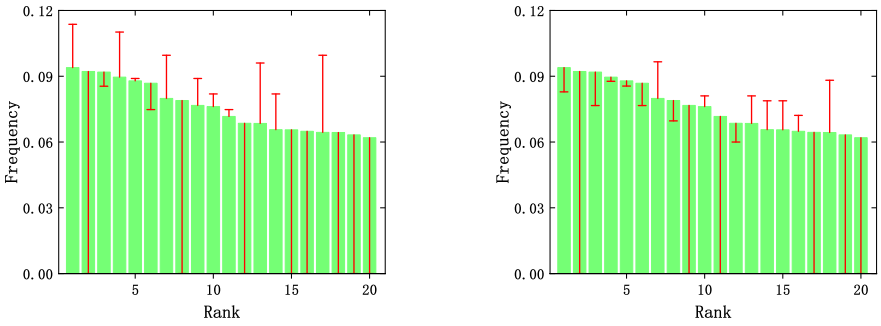
The results of the top-k sequential pattern mining contain two aspects, *i.e.*, the rank of the top-k sequential patterns and the corresponding frequencies. Thus the metrics should cover both the two aspects.

Define sp_i as the i-th most frequent length-m sequential pattern in the original dataset. We denote the ground truth of the top-k length-m sequential patterns as $\mathbf{x}_t = \{sp_1, sp_2, ..., sp_k\}$. There are two metrics we use:

1) Discounted cumulative gain (DCG). The DCG measures the quality of the estimated rank of the sequential patterns. Let rel_{sp_i} be the relevance of a sequential pattern sp_i, which is defined as:

$$rel_{sp_i} = \begin{cases} \log_2(k - \tilde{r}_{sp_i}), & \text{if } k - \tilde{r}_{sp_i} > 0 \\ 0, & \text{if } k - \tilde{r}_{sp_i} \leq 0 \end{cases}$$

where $\tilde{r}_{sp_i} = |rank_{actual}(sp_i) - rank_{estimated}(sp_i)|$ is the relative error of the rank of sp_i.



(a) LDPSPM, n=100,000 (b) P-LDPSPM, n=100,000

(c) LDPSPM, n=500,000 (d) P-LDPSPM, n=500,000

Figure 2: Impact of the number of users (synthetic datasets)

The DCG of the estimated ranked list of sequential patterns is computed as follows,

$$DCG_k = rel_{sp_1} + \sum_{i=2}^{k} \frac{rel_{sp_i}}{\log_2(i)} \tag{12}$$

The factor $\log_2(i)$ in the denominator gives more weight of the sequential patterns with higher ranks. Finally, we normalize the DCG of a ranked list by dividing with the ideal DCG (IDCG), which is the DCG value when the estimated ranks are the same as the actual ranks. Then we get the Normalized DCG (NDCG):

$$NDCG_k = \frac{DCG_k}{IDCG_k} \tag{13}$$

The value of NDCG is always between 0 and 1, and we can compare the results of the top-k sequential pattern mining across different k.

2) Mean relative error (MRE). We measure the accuracy of the estimated frequencies with the mean relative error between the actual frequency and the estimated frequency. MRE is computed as follows:

$$MRE_k = \frac{1}{k} \sum_{sp \in \mathbf{x}_t} \frac{|f_{sp} - \tilde{f}_{sp}|}{f_{sp}} \tag{14}$$

7.4 Results

7.4.1 Impact of n

First, we evaluate the impact of the number of users on the effectiveness of our mechanisms. We conduct experiments on two synthetic datasets. The numbers of users in the two datasets are 100,000 and 500,000, respectively. There are 60 real items for both datasets. Each sequence is truncated or padded to a fixed length of 13, thus there are 13 dummy items. We set $m = 2$ and $k = 20$ in the experiments, and the privacy budget is set with $\epsilon = 3$. Figure 2 shows several representative results. The results contain the true and the estimated frequencies of the true top-20 sequential patterns. The green bars show their actual frequencies, and the red candlestick bars show the estimated frequencies. Specially, if a sequential pattern with actual rank in top-20 is missed by the algorithm (i.e., not included in the estimated top-20 sequential patterns), we set the candlestick bar to 0, even though the estimated frequency is not 0. In Figure 2(a), when $n = 100,000$, LDPSPM fails to capture many top-k sequential patterns, and the estimated frequencies are far away from the actual ones. When we increase the number of users to $n = 500,000$, as shown in Figure 2(c), the results are much better. The accuracy of the estimated frequencies gets improved. The number of missed sequential patterns decreases, especially for those patterns with high ranks (*e.g.*, ranks higher than 10). Meanwhile, Figure 2(b) and Figure 2(d) show the results of the improved algorithm P-LDPSPM with $n = 100,000$ and $n = 500,000$ respectively. Compared with the results in Figure 2(a) and Figure 2(c), with the pruning step, the number of missed sequential patterns decreases, and the estimated frequencies are closer to the actual ones.

The results in Figure 2 reveal an intrinsic challenge of LDP sequential pattern mining that the accuracy of the

Table 1: Experimental results while varying m, #real items and #users

#real items, #dummies	#users	m=2		m=3		m=4	
		NDCG	MRE	NDCG	MRE	NDCG	MRE
$d = 10$, $l = 8$	$n = 1 \times 10^5$	0.98	0.03	0.76	0.13	0.04	1.11
	$n = 5 \times 10^5$	0.99	0.02	0.92	0.07	0.15	0.59
	$n = 1 \times 10^6$	0.99	0.01	0.96	0.05	0.19	0.37
	$n = 5 \times 10^6$	0.99	0.01	0.97	0.03	0.52	0.19
	$n = 1 \times 10^7$	1	0.01	0.98	0.02	0.71	0.15
	$n = 5 \times 10^7$	1	0.01	0.99	0.02	0.88	0.06
$d = 20$, $l = 8$	$n = 1 \times 10^5$	0.97	0.05	0.24	0.43	0	15.68
	$n = 5 \times 10^5$	0.98	0.02	0.57	0.21	0	2.59
	$n = 1 \times 10^6$	0.99	0.02	0.74	0.13	0.01	1.93
	$n = 5 \times 10^6$	0.99	0.01	0.94	0.05	0.01	0.94
	$n = 1 \times 10^7$	1	0.01	0.96	0.04	0.02	0.61
	$n = 5 \times 10^7$	1	0.01	0.99	0.02	0.13	0.27

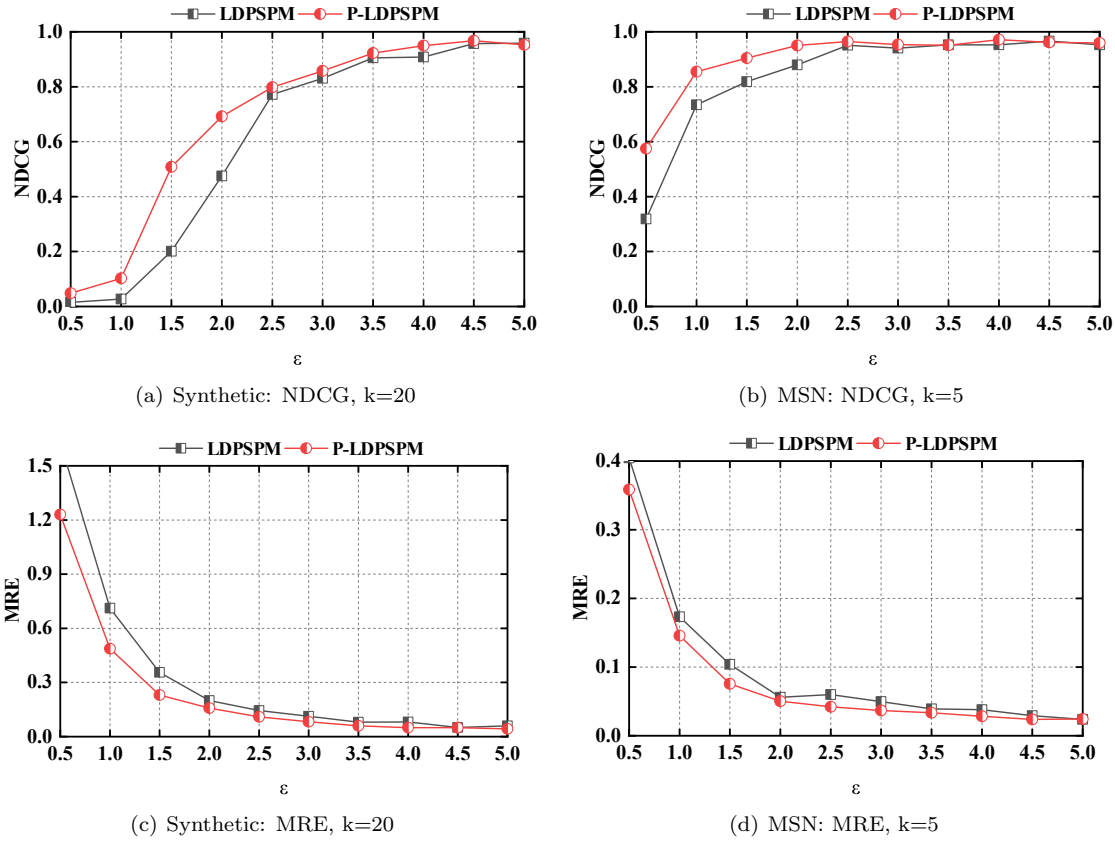
estimation will be poor if the population of users engaged in the algorithm is insufficient. The LDP algorithms are essentially based on random response that each user responds with a set of sequential patterns. The precision of the results will be affected by the size of the domain of the sequential patterns. The responses are more scattered when the size of the domain is larger, which leads to poor performance of the estimation of the frequencies. This is serious when the number of items and the length of targeted sequential patterns increase. The number of all real and dummy items is $d + l$ and the length of the target sequential patterns is m, then the number of all length-m patterns is A^m_{d+l}, which increases exponentially with the increase of d, l and m.

We conduct experiments on synthetic datasets to explore the correlation between the quality of the results and the length of the target sequential patterns, the number of items and the number of users. We generate datasets with different numbers of items and users. Each sequence is truncated or padded to a fixed length $l = 8$. The number of dummy items is l, as well. We fix $\epsilon = 3$ and $k = 20$ and explore the performance of the algorithm when m is set with 2, 3 and 4. The results are shown in Table 1 where each experiment is conducted 10 times to get an average value. Under the setting in Table 1, the P-LDPSPM algorithm degrades into LDPSPM since the pruning step is omitted when $km > d$ as described in Section 6, so the results in Table 1 are conducted with LDPSPM.

In Table 1, we can see that the results are much better when m is smaller. The performance of the algorithm drops dramatically when m increases. Besides, the results are better when the number of items is smaller. Increasing the population of users improves the quality of the results. For example, when the number of the items is 10 and $m = 4$, increasing the number of users from 1×10^5 to 5×10^7 leads to the increase of NDCG from 0.04 to 0.88 and the decrease of MRE from 1.11 to 0.06. Moreover, when the number of users increases to a certain level, the metrics no longer changes much and tends to a stable value. For example, considering the number of items as 10 and $m = 2$, the value of MRE remains 0.01 when the number of users increases from 1×10^6 to 5×10^7. This is caused by the padding and truncation of the original sequences as described in Section 4, which leads to deviations from the original sequences. Furthermore, we find that the quality of the results is closely related with the number of all length-m sequential patterns, *i.e.*, A^m_{d+l}. The number of users needed to be engaged in the algorithm to reach the stable metric is in direct proportion to the number of all length-m patterns. For example, the result under the parameters $d = 10$, $l = 8$, $m = 2$, $n = 1 \times 10^6$ has a similar quality as the result under the parameters $d = 20$, $l = 8$, $m = 3$, $n = 5 \times 10^7$. The limitation of the algorithm comes out that many more users are needed to maintain the quality of the results when the number of items and the length of target patterns increase.

7.4.2 Impact of ϵ

In this experiment, we evaluate the impact of the privacy budget with the synthetic and the MSN dataset. The synthetic dataset contains 500,000 users and the length of each sequence is 13. There are 60 real items and 13 dummy items. Due to the difference in the number of items of the two datasets, we set $k = 20$ for the synthetic dataset and $k = 5$ for the MSN dataset. As we describe in Section 7.4.1, many more users are needed to maintain the quality of the results when the number of items and the length of patterns increase. Due to the limited number of the users in the real-world MSN dataset, we set m with a relatively small value (i.e., $m = 2$) in the following experiments. Figure 3 shows the results for the two datasets along with the change of ϵ. It is clear that the results get better (with higher NDCG and lower MRE) when ϵ increases. P-LDPSPM gets higher NDCG and lower MRE than LDPSPM on both two datasets. More specifically, P-LDPSPM gets more improvement than LDPSPM when ϵ is relatively small. The reason is that when ϵ increases, fewer noises are contained in the responses, and thus the



(a) Synthetic: NDCG, k=20

(b) MSN: NDCG, k=5

(c) Synthetic: MRE, k=20

(d) MSN: MRE, k=5

Figure 3: Experimental results while varying ϵ

estimations of the collector get closer to the real values, even without the pruning step.

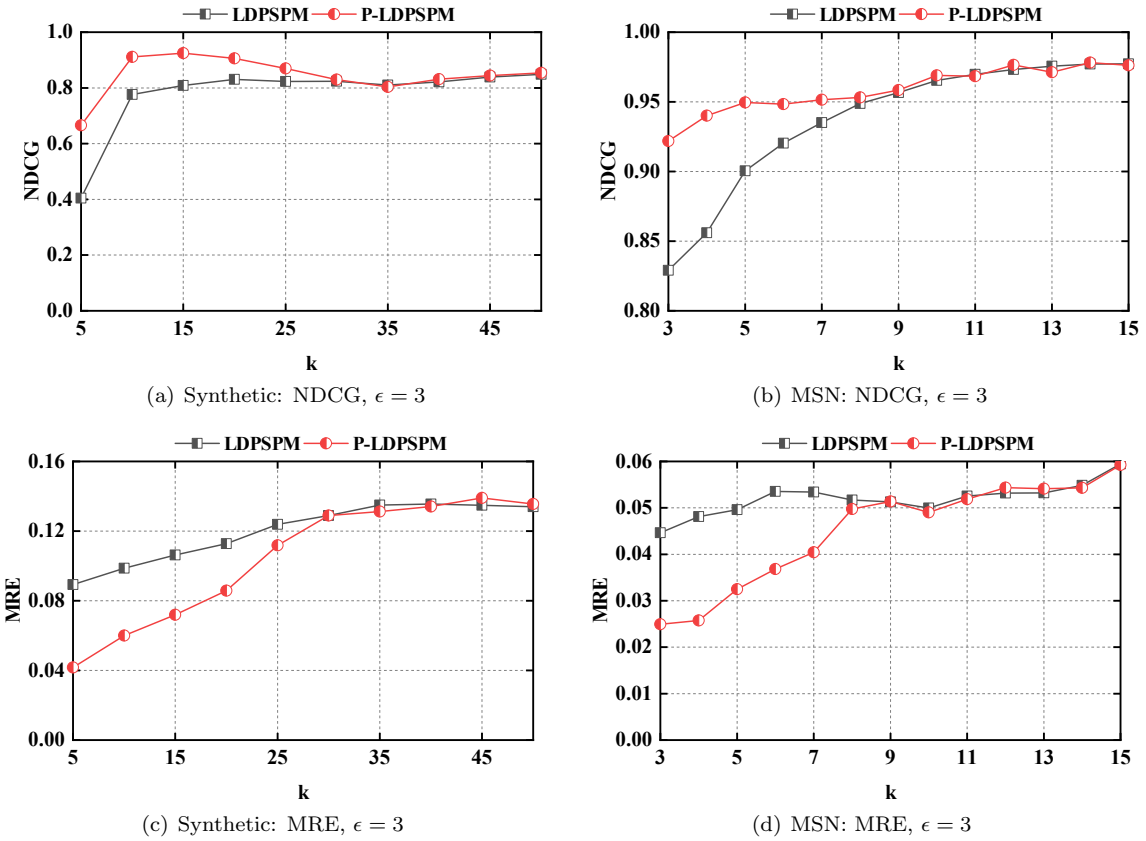
7.4.3 Impact of k

Next, we evaluate the impact of the number of reported sequential patterns with the synthetic and the MSN dataset. The setting of the synthetic dataset is the same as that in Section 7.4.2. We fix $\epsilon = 3$ and $m = 2$ for both two datasets. The results are shown in Figure 4. The NDCG increases on the whole when k increases. This is because when k increases, more true top-k sequential patterns are captured in the estimated top-k patterns, and the influence of the wrong ranks of the top sequential patterns on the NDCG decreases. The MRE increases when k increases, *i.e.*, the average accuracy of the frequencies decreases when more sequential patterns are identified. It is obvious that the sequential patterns with higher frequencies appear more in the responses of the users, and thus the estimated frequencies of them are closer to the true frequencies. Also, the advantage of P-LDPSPM compared to LDPSPM is much more significant when k is small, and the results of two algorithms get closer when k increases. When k increases, the ratio of the candidates identified in the pruning step to the whole item domain gets higher, and the utility gain due to the pruned domain is offset by the utility loss caused by the split of user groups. At some values of k (*e.g.*, when $k > 30$ in the synthetic dataset and $k > 8$ in the MSN dataset), the candidates cover the whole item domain, as we explained in Section 6, the pruning step is removed and the results become the same for LDPSPM and P-LDPSPM.

8 Conclusions

In this paper, we study the problem of mining frequent sequential patterns under local differential privacy. We propose an efficient and effective mechanism called LDPSPM. Each user randomly responds with a set of sequential patterns with an optimal size. We further propose a mechanism P-LDPSPM, which adds a pruning step for LDPSPM and further improves the performance by reducing the impact of nonsignificant items. Both theoretical analysis and extensive experiments show the effectiveness and efficiency of our methods.

The limitation of this work lies in that the algorithms are only suitable for sequential patterns of fixed lengths. Besides, large population of users are needed to maintain the quality of the results when the number of items and the length of the target patterns are big. Future works can focus on mechanisms that support frequency estimation of sequential patterns of various lengths and reduce the dependence on the number of users.



(a) Synthetic: NDCG, $\epsilon = 3$

(b) MSN: NDCG, $\epsilon = 3$

(c) Synthetic: MRE, $\epsilon = 3$

(d) MSN: MRE, $\epsilon = 3$

Figure 4: Experimental results while varying k

Acknowledgments

The research was supported by the National Key R&D Program of China 2018YFB0803400, 2018YFB2100300, National Natural Science Foundation of China under Grant No.61972369, No.61572453, No.61520106007, No.61572454, and the Fundamental Research Funds for the Central Universities, No.WK2150110009.

References

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, pp. 3–14, 1995.

[2] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 429–435, 2002.

[3] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pp. 127–135, 2015.

[4] L. Bonomi and L. Xiong, "A two-phase algorithm for mining sequential patterns with differential privacy," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 269–278, 2013.

[5] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Visualization of navigation patterns on a web site using model-based clustering," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 280–284, 2000.

[6] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 638–649, 2012.

[7] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Information Sciences*, vol. 231, pp. 83–97, 2013.

[8] R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin, "Private spatial data aggregation in the local setting," in *IEEE 32nd International Conference on Data Engineering (ICDE'16)*, pp. 289–300, 2016.

[9] Y. C. Chen, W. L. Wang, M. S. Hwang, "RFID authentication protocol for anti-counterfeiting and privacy protection", in *The 9th International Conference on Advanced Communication Technology*, pp. 255–259, 2007.

[10] X. Cheng, S. Su, S. Xu, and *et al.*, "Differentially private maximal frequent sequence mining," *Computers & Security*, vol. 55, pp. 175–192, 2015.

[11] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *The AAAI Press*, 2011. (`http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#ChengCLS11`)

[12] C. H. Chuang, W. F. Lu, Y. C. Lin, and J. C. Chen, "Visual exploration using improved moving average methods for time series datasets," *International Journal of Electronics and Information Engineering*, vol. 9, no. 1, pp. 46–60, 2018.

[13] P. S. Chung, C. W. Liu, and M. S. Hwang, "A study of attribute-based proxy re-encryption scheme in cloud environments", *International Journal of Network Security*, vol. 16, no. 1, pp. 1-13, 2014.

[14] G. Cormode, T. Kulkarni, and D. Srivastava, "Answering range queries under local differential privacy," in *Proceedings of International Conference on Management of Data*, vol. 12, no. 10, pp. 1126–1138, 2019.

[15] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy, data processing inequalities, and statistical minimax rates," *Statistics Theory*, 2013. arXiv:1302.3203.

[16] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, pp. 1–19, 2008.

[17] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Cryptography and Security*, pp. 1054–1067, 2014.

[18] G. Fanti, V. Pihur, and Ú. Erlingsson, "Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 3, pp. 41–61, 2016.

[19] P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, "Fast vertical mining of sequential patterns using co-occurrence information," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 40–52, 2014.

[20] A. S. M. T. Hasan and Q. Jiang, "A general framework for privacy preserving sequential data publishing," in *The 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA'17)*, pp. 519–524, 2017.

[21] W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of public auditing for secure data storage in cloud computing," *International Journal Network Security*, vol. 18, no. 1, pp. 133–142, 2016.

[22] M. S. Hwang, C. C. Lee, and P. S. Chung, "An efficient IC-lock self-reader data security in cloud computing," *Applied Mathematics & Information Sciences*, vol. 9, no. 4, p. 2099, 2015.

[23] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Journal of Machine Learning Research*, pp. 2879–2887, 2014.

[24] J. W. Kim, D. H. Kim, and B. Jang, "Application of local differential privacy to collection of indoor positioning data," *IEEE Access*, vol. 6, pp. 4276–4286, 2018.

[25] H. H. Le, M. Kushima, K. Araki, and H. Yokota, "Differentially private sequential pattern mining considering time interval for electronic medical record systems," in *Proceedings of the 23rd International Database Applications & Engineering Symposium*, pp. 1–9, 2019.

[26] C. T. Li, M. S. Hwang, Y. P. Chu, "Further improvement on a novel privacy preserving authentication and access control scheme for pervasive computing environments", *Computer Communications*, vol. 31, no. 18, pp. 4255–4258, Dec. 2008.

[27] Y. Li, G. Wang, Y. Yuan, X. Cao, L. Yuan, and X. Lin, "PrivTS: Differentially private frequent time-constrained sequential pattern mining," in *International Conference on Database Systems for Advanced Applications*, pp. 92–111, 2018.

[28] C. W. Liu, W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of attribute-based access control with user revocation in cloud data storage," *Internatioanl Journal Network Security*, vol. 18, no. 5, pp. 900–916, 2016.

[29] C. W. Liu, W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of public auditing for shared data storage with user revocation in cloud computing", *International Journal of Network Security*, vol. 18, no. 4, pp. 650–666, 2016.

[30] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 19–30, 2009.

[31] N. Mohammed, B. C. M. Fung, and M. Debbabi, "Walking in the crowd: Anonymizing trajectory data for pattern analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1441–1444, 2009.

[32] T. T. Nguyên, X. Xiao, Y. Yang, and *et al.*, "Collecting and analyzing data from smart device users with local differential privacy," *Databases*, 2016. arXiv:1606.05053.

[33] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu, "Mining sequential patterns by pattern-growth: The prefixspan approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1424–1440, 2004.

[34] Z. Qin, Y. Yang, T. Yu, and *et al.*, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 192–203, 2016.

[35] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in *International Conference on Extending Database Technology*, pp. 1–17, 1996.

[36] S. Wang, L. Huang, Y. Nie, P. Wang, H. Xu, and W. Yang, "Privset: Set-valued data analyses with locale differential privacy," in *IEEE Conference on Computer Communications*, pp. 1088–1096, 2018.

[37] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *USENIX Security 17*, pp. 729–745, 2017.

[38] T. Wang, N. Li, and S. Jha, "Locally differentially private frequent itemset mining," in *IEEE Symposium on Security and Privacy (SP'18)*, pp. 127–143, 2018.

[39] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[40] J. T. Wilson and M. H. Goldstein, *History-based Tracking of User Preference Settings*, 2014. US-8793614.

[41] S. Xu, X. Cheng, S. Su, and *et al.*, "Differentially private frequent sequence mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2910–2926, 2016.

[42] Z. Yang and M. Kitsuregawa, "LAPIN-SPAM: An improved algorithm for mining sequential pattern," in *The 21st International Conference on Data Engineering Workshops*, pp. 1222–1222, 2005.

[43] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine Learning*, vol. 42, no. 1-2, pp. 31–60, 2001.

Biography

Huihua Xia received the B.S. degree in computer science from University of Science and Technology of China in 2014. He is currently working towards the Ph.D. degree at the Department of Computer Science and Technology, University of Science and Technology of China. His current research interests include data privacy and information security.

Wenchao Huang received the B.S. and Ph.D degrees in computer science from University of Science and Technology of China in 2006 and 2011, respectively. He is an associate professor in School of Computer Science and Technology, University of Science and Technology of China. His current research interests include information security, formal methods and mobile computing.

Yan Xiong received the B.S., M.S., and Ph.D degrees from University of Science and Technology of China in 1983, 1986 and 1990 respectively. He is a professor in School of Computer Science and Technology, University of Science and Technology of China. His main research interests include distributed processing, mobile computing, computer network and information security.

Fuyou Miao received his Ph.D of computer science from University of Science and Technology of China in 2003. He is an associate professor in the School of Computer Science and Technology, University of Science and Technology of China. His research interests include applied cryptography, trusted computing and mobile computing.

Research on Detection and Defense of Malicious Code Under Network Security

Xiaoli Xiong and Yongguang Hou
(Corresponding author: Yongguang Hou)

Weinan Normal University
Weinan Normal University, Middle Chaoyang Road, Weinan, Shaanxi 714099, China
Email: yongjiao736401335@163.com
(Received May 24, 2019; Revised and Accepted Feb. 12, 2021; First Online Aug. 14, 2021)

Abstract

This paper mainly analyzed the detection and defense methods of malicious code, proposed to extract features by variable-length N-gram, and used weighted IG for feature selection. Finally, the performance of naive Bayesian (NB), random forest (RF), and support vector machine (SVM) classification models was compared. The results showed that the variable-length N-gram had a good performance in feature extraction, and weighted IG was better than IG in feature selection. Furthermore, the SVM model had the best performance in classifying malicious code and normal code, with an F1 score of 0.9367 and a log loss of 0.0321, which were better than the other two models. The results verify the reliability of the proposed method in the detection and defense of malicious code, which can be further applied in practice.

Keywords: Feature Extraction; Malicious Code; Network Security; Variable-Length N-Gram; Weighted Information Gain

1 Introduction

With the development of the network, the types and number of attacks and threats are also growing. More and more complex attacks and threats have brought serious harm to the network, among which malicious code is one of the important factors [4]. With the development of malicious code, its threat to the network has become more serious [2]. How to detect and defend malicious code has become a key and difficult problem in network security [7, 12]. Li *et al.* [9] used the AutoEncoder method to reduce data dimension, extract features, and used the deep trust network (DBN) to detect malicious code. The experiment showed that the method had a high detection accuracy and a low time complexity.

Acarturk *et al.* [1] designed a method framework and detected malicious code by analyzing run trace outputs by long short term memory (LSTM). Two models, instruction as a sequence model (ISM) and basic block as a sequence model (BSM), were obtained by establishing data sets through run traces of files. The accuracy of ISM and BSM was 87.51% and 99.26%, respectively. Cui *et al.* [5] proposed a method based on deep learning. Firstly, malicious code was transformed into a gray image, and then it was identified by a convolutional neural network (CNN). The experiment on the Vision Research Lab data set showed that the proposed method had good accuracy and speed. Nikolopoulos *et al.* [11] proposed a graph-based method, which used a system to call a correlation graph (ScD graph) to detect whether unknown samples were malicious or benign. Based on the concept of similarity, the performance of the model was improved. After evaluation, it was found that this method had a great potential in detecting malware. In this study, for the detection and defense of malicious code, features were extracted through the variable-length N-gram and selected through weighted IG. Finally, the performance of three models in detecting malicious code was compared. This study makes some contributions to achieve network security better.

2 Detection and Defense of Malicious Code

Malicious code refers to a kind of code deliberately written to achieve some malicious functions, which is usually embedded into the program without authorization to steal users' data and trade secrets. Driven by interests, the current malicious code is more covert and purposeful, has various communication modes, and has been widely spread on mobile platforms [10].

Generally speaking, the attack of malicious code includes four steps:

1) Searching for targets, such as local files, storage devices, *etc.*;

2) Saving to the target: viruses, worms can actively save themselves in the target, and Trojans need to cheat

users to download;

3) Triggering, the same as saving, includes active triggering and passive triggering;

4) Surviving for a long time: it can exist statically, such as exe, dll files, *etc.*, and can also exist dynamically, such as services, ports, *etc.*

In the detection and defense of malicious code, there are mainly two methods.

Signature-based method [6].] It detects the binary string of the program. This method is based on the known signature database and can not detect the unknown program.

Heuristic-based method [17].] Analysts extract the heuristic rules of known code and use them to find new malicious code. This method can also not find the new unknown code in time.

With the development of data mining technology, the detection and defense technology of malicious code based on it has been widely recognized [8]. It can effectively make up for the shortcomings of the above two methods and make a great contribution to the realization of network security. This paper mainly analyzed the application of some data mining methods.

3 Detection and Defense of Malicious Code

3.1 Feature Extraction

In the malicious code, there is an instruction code different from the normal code, which is the difference between the malicious code and the normal code. These instruction codes that can be used to distinguish are called features. In the malicious code, there are many kinds of features, such as instruction sequence, string, PE file header, application programming interface (API) function, *etc.* In this study, the code file is disassembled by IDA Pro to get the byte sequence of machine code, which is used as the expression form of features. Then, features are extracted by the N-gram-based method.

N-gram model [15], also known as the first-order Markov chain, can extract many potential and difficult features in feature extraction. This method also has two disadvantages. First, it may produce no edge matching for byte sequences of different lengths, resulting in failure of feature extraction; second, extraction by N-gram will obtain a larger feature set, which has a high requirement for the storage space. To make up for the defects of N-gram, this paper uses variable-length N-gram for feature extraction.

Compared with N-gram, the length of variable-length N-gram is not fixed, preventing meaningful sequences from being disassembled. In variable-length N-gram, the first step is to find breakpoints, and the continuous sequence between two breakpoints is the possible feature sequence. This paper uses the expert voting algorithm. This method contains two attributes, frequency and entropy. For frequency, the higher the frequency of subsequence in a paragraph, the more likely it is to contain a breakpoint; for entropy, the greater the entropy is, the more likely there is a breakpoint after the sequence. After calculating the frequency and entropy of each position, the scores of the two positions with the largest frequency and entropy are added by one. Finally, the results are synthesized. According to the accumulated score, the possible breakpoints are judged. The continuous sequence between the two breakpoints is a feature.

In the specific calculation, this paper uses the Trie tree of $d = 4$ to achieve expert voting, D means the depth of Trie.

Suppose there is a sequence: D F G D F H, the Trie tree is shown in Figure 1.

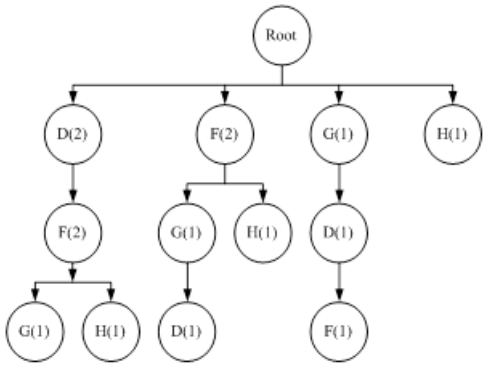


Figure 1: Trie tree

In Figure 2, the number in brackets refers to the number of times it appears. When searching for breakpoints, the window slides through the sequence in order. The window passes through "D, F, G" first; the first possible breakpoint may is between D and F. The frequency and entropy are calculated. The frequency of every position is the sum of the frequencies of the front and back sequences in its window, for example, $f(D) = f(D) + f(FG)$. The frequency of the node is:

$$p(x) = \frac{f(x)}{f(parent(x))}$$

The entropy of every position is the entropy of its left node, i.e., $e(x) = -\sum_{x \in X} p(x) \log p(x)$. Before calculating the cumulative fraction of the position, the two values should be standardized to eliminate the dimensional relationship. After the window traverses the whole sequence, the maximum value is found according to the fraction; then, the breakpoint is obtained.

3.2 Feature Selection

After extracting features with the variable-length N-gram, as the feature dimension is large, dimension reduction is needed. In this study, a feature selection method based on weighted information gain (weighted IG) [14] is used. For samples, the larger the IG value of an attribute is, the greater the amount of information is, i.e., the larger the discrimination degree is. However, IG ignores the frequency of features; thus, it needs to be improved. In a feature Ng (N-gram), its weighted IG is:

$$WIG(Ng) = \lambda \sum_{V_{Ng}\in\{0,1\}} \sum_{C\in\{C_i\}} \rho(V_{Ng}, C) \log \frac{\rho(V_{Ng}, C)}{\rho(V_{Ng})\rho(C)}$$

$$\lambda = \begin{cases} \log(1 + g(\frac{f_M K_N}{f_N K_M})), & f_M \neq 0, f_N \neq 0 \\ \log(1 + \frac{f_M}{K_M}), & f_M \neq 0, f_N = 0 \\ \log(1 + \frac{f_N}{K_N}), & f_M = 0, f_N \neq 0 \end{cases}$$

$$g(x) = \begin{cases} x, & x \geq 1 \\ \frac{1}{x}, & 0 < x < 1, \end{cases}$$

where V_{Ng} stands for the value of feature V_{Ng} (if the feature appears in the sample, its value is 1; otherwise, it is 0); C stands for sample class, $\rho(V_{Ng}, C)$ stands for the proportion of class C in Ng, $\rho(V_{Ng})$ stands for the proportion of Ng in the sample, $\rho(C)$ stands for the proportion of class C in the sample, λ stands for the feature weight, f_M and f_N are the total number of times the feature appears in malicious code and normal code, and K_M and K_N are the number of malicious code and normal code in the sample.

The larger the calculated WIG value is, the more effective the feature in distinguishing normal code from malicious code is.

3.3 Classification Model

Malicious code detection is to distinguish normal code from malicious code, which requires a classification model. The model can classify new unknown samples after learning the training samples with class labels. This paper mainly compares the performance of three models.

Naive Bayes (NB) [16]. This method is based on the Bayes theorem, with low computational complexity. It is assumed that the number of samples is N and the dimension of a feature is d. The result of the classification is represented by $y \in \{0, 1\}$, $y = 0$ for the normal code and $y = 1$ for the malicious code. Let the conditional probability of classification be $p(x|y)$; then,

$$p(y = 1|X) = \frac{p(X|y = 1)}{p(X|y = 1) + p(X|y = 0)}$$

According to the set threshold, when $p(y = 1|X)$ exceeds the threshold, malicious code can be classified.

Random forest (RF) [3].] The algorithm samples data through bootstrap. It is assumed that the number of samples is N. N samples are randomly sampled every time to train the decision tree. Then, m ($m < M$) variables are randomly selected to find out the attribute that can achieve the best segmentation effect. Without pruning, a single decision tree is generated. Finally, every decision tree generates a prediction result, and the modal number is taken as the final classification result, i.e., the class that is selected most by the tree is the class of samples.

Support vector machine (SVM) [13]. It maps the output data into a high-dimensional space and establishes a set of hyperplanes to maximize the interval between classes. The classification function is:

$$f(x) = sgn\left(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b\right),$$

where a is a Lagrange multiplier, K is a kernel function, and b is is a threshold. If $f(x) > 0$, it is a normal code; otherwise, it is a malicious code.

4 Experimental Analysis

4.1 Experimental Data

From https://virusshare.com and www.malware-traffic-analysis.net, Windows malicious PE file set was downloaded, and there was a total of 5000 samples. The normal sample set was also downloaded from the Baidu app store and scanned through security software to confirm that 3000 samples were benign. The total experimental data were 8000 samples. 70% of the samples were taken as training samples, and 30% as test samples, as shown in Table 1.

Table 1: Experimental data set

	Malicious code	Normal code
Training sample	3500	2100
Test sample	1500	900

4.2 Evaluation Index

In this study, the performance of the algorithm was evaluated by the F1 score and logloss. F1 score was the harmonic average of the accuracy rate and recall rate, 1 for the best and 0 for the worst. For the confusion matrix (Table 2), the F1 score is calculated as follows.

The accuracy is: $A = \frac{TP+TN}{TP+TN+FP+FN}$;

The precision is: $P = \frac{TP}{TP+FP}$;

The recall rate is: $R = \frac{TP}{TP+FN}$;

The F1 score is: $F1 = \frac{2PR}{P+R}$

Logloss refers to the logarithmic loss, which measures the uncertainty of classification by the degree of difference from the actual situation. Its calculation method is:

$$logloss = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{C}[y_{ij}\log(\rho_{ij}) + (1-y_{ij})\log(1-\rho_{ij})],$$

where N is the number of samples, C is the number of classes, y_{ij} refers to whether the i-th sample belongs to class j or not (1 if it does and 0 if it does not), and ρ_{ij} refers to the probability of the i-th sample being classified as class j.

Table 2: Confusion matrix

		Actual situation	
		Malicious code	Normal code
Classification results	Malicious code	TP	FP
	Normal code	FN	TN

4.3 Experimental Results

Firstly, the performance of variable-length N-gram was analyzed and compared with the traditional N-gram. The naive Bayesian (NB) model was used as the classification model. The results of different feature extraction methods are shown in Figure 2.

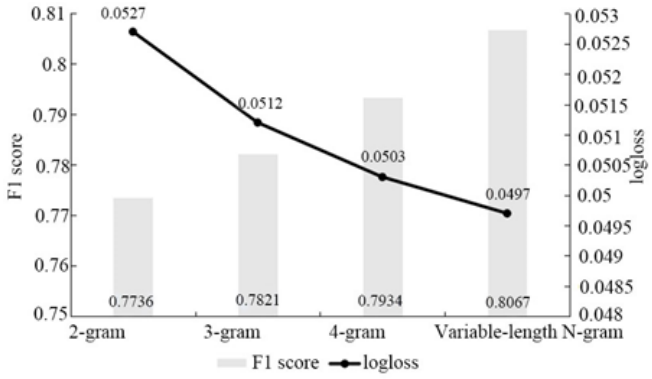


Figure 2: Comparison of different feature extraction methods

It was seen from Figure 2 that when using the traditional N-gram for feature extraction in the detection and defense of malicious code, the F1 score was 2-gram < 3-gram < 4-gram, the logloss was 2-gram > 3-gram > 4-gram, the F1 score of N-gram was lower than 80%, and the logloss was greater than 0.05; when using variable-length N-gram for feature extraction, the F1 score was 0.8067, which was 0.0133 larger than 4-gram, and the logloss was 0.0497, which was 0.0006 smaller than 4-gram. The comparison results showed that variable-length N-gram had better performance and was more conducive to detecting and defending malicious code.

Then, the performance of weighted IG was analyzed. The feature was extracted with variable-length N-gram, and the NB model was used for classification. IG and weighted IG were compared, and the results are shown in Figure 3.

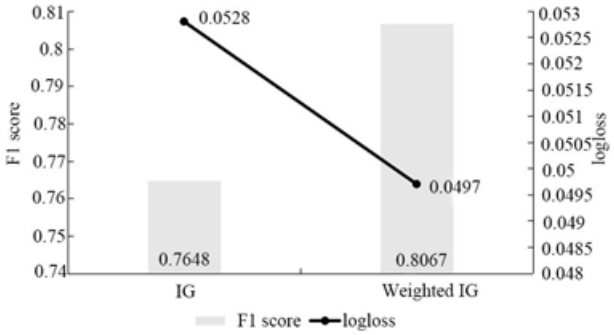


Figure 3: Comparison of different feature selection methods

It was seen from Figure 3 that when IG was used for feature selection, the F1 score of the algorithm was 0.7648, and the logloss was 0.0528; when weighted IG was used for feature selection, the F1 score of the algorithm was 0.8067, which was 0.0419 larger than IG, and the logloss was 0.0497, which was 0.0031 smaller than IG. The results showed that weighted IG had better performance and obtained better detection results for malicious code.

Finally, the performance of the three models was compared by using the variable-length N-gram + weighted IG method, and the results are shown in Figure 3.

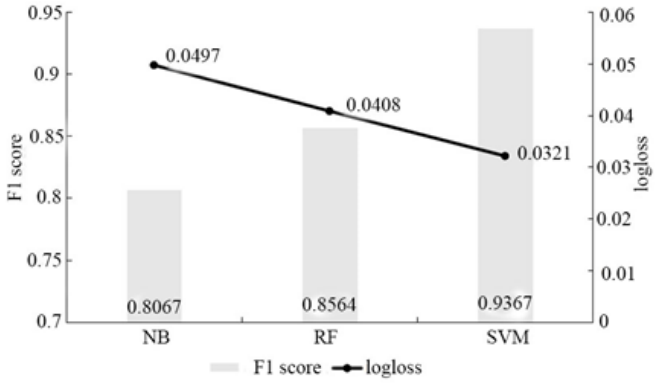


Figure 4: Comparison of different classification models

It was seen from Figure 4 that NB < RF < SVM in the comparison of the F1 score, i.e., SVM had the highest F1 score, 0.9367, which was 0.13 larger than NB and 0.0803 larger than RF; in the comparison of logloss, NB > RF > SVM, i.e., SVM had the lowest logloss, 0.0321, which

was 0.0176 smaller than NB and 0.0087 smaller than RF. The results showed that SVM had the best performance and best classification effect in the detection and defense of malicious code.

5 Conclusion

This paper mainly analyzed the detection and defense methods of malicious code, extracts features with variable-length N-gram, and then selects features with weighted IG. Finally, the performance of three different models was compared. The results showed that:

1) Variable-length N-gram has better performance than traditional N-gram;

2) Weighted IG was better than IG in feature selection;

3) Among the three models, SVM had the best performance, with the highest F1 score (0.9367) and the lowest logloss (0.0321).

The results showed that the method of malicious code detection and defense designed in this paper is effective and can be further promoted and applied in practice, which is conducive to realize network security.

References

[1] C. Acarturk, M. Sirlanci, P. G. Balikcioglu, D. Demirci, N. Sahin, O. A. Kucuk, "Malicious code detection: Run trace output analysis by LSTM," *IEEE Access*, vol. PP, no. 99, pp. 1-1, 2021.

[2] I. Ahn, H. C. Oh, J. Park, "Investigation of the C-SEIRA model for controlling malicious code infection in computer networks," *Applied Mathematical Modelling*, vol. 39, no. 14, pp. 4121-4133, 2015.

[3] M. Belgiu, L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, no. -, pp. 24-31, 2016.

[4] Z. Cui, L. Du, P. Wang, X. Cai, W. Zhang, "Malicious code detection based on CNNs and multi-objective algorithm," *Journal of Parallel and Distributed Computing*, vol. 129, pp. 50-58, 2019.

[5] Z. Cui, F. Xue, X. Cai, Y. Cao, G. Wang, J. Chen, "Detection of malicious code variants based on deep learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3187-3196, 2018.

[6] J. W. Ha, H. K. Kim, J. I. Lim, "Research on malicious code hidden website detection method through whitelist-based malicious code behavior analysis," *Journal of the Korea Institute of Information Security & Cryptology*, vol. 21, no. 4, pp. 61-75, 2011.

[7] S. Islam, H. Ali, A. Habib, N. Nobi, M. Alam, and D. Hossain, "Threat minimization by design and deployment of secured networking model," *International Journal of Electronics and Information Engineering*, vol. 8, no. 2, pp. 135–144, 2018.

[8] T. H. Lee, H. K. Kwang, "A study on detection of small size malicious code using data mining method," *Jouranl of Information and Security*, vol. 19, no. 1, pp. 11-17, 2019.

[9] Y. Li, R. Ma, R. Jiao, "A hybrid malicious code detection method based on deep learning," *International Journal of Software Engineering and Its Applications*, vol. 9, no. 5, pp. 205-216, 2015.

[10] A. Narayanan, M. Chandramohan, L. Chen, Y. Liu, "A multi-view context-aware approach to android malware detection and malicious code localization," *Empirical Software Engineering*, vol. 23, pp. 1222–1274, 2018.

[11] S. D. Nikolopoulos, I. Polenakis, "A graph-based model for malicious code detection exploiting dependencies of system-call groups," in *Proceedings of the 16th International Conference on Computer Systems and Technologies (CompSysTech'15)*, pp. 228-235, 2015.

[12] E. U. Opara and O. J. Dieli, "Enterprise cyber security challenges to medium and large firms: An analysis," *International Journal of Electronics and Information Engineering*, vol. 13, no. 2, pp. 77–85, 2021.

[13] K. Schittkowski, "Optimal parameter selection in support vector machines," *Journal of Industrial & Management Optimization*, vol. 1, no. 4, pp. 465-476, 2017.

[14] G. Singer, R. Anuar, I. Ben-Gal, "A weighted information-gain measure for ordinal classification trees," *Expert Systems with Applications*, vol. 152, pp. 113375, 2020.

[15] A. Tripathy, A. Agrawal, S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems With Applications*, pp. 117-126, 2016.

[16] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, C. Zhang, "Self-adaptive attribute weighting for naive Bayes classification," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1487-1502, 2015.

[17] B. Zhang, Q. Li, Y. Ma, "Research on dynamic heuristic scanning technique and the application of the malicious code detection model," *Information Processing Letters*, vol. 117, pp. 19-24, 2017.

Biography

Xiaoli Xiong, born in 1984, graduated from central China Normal University in 2010. She is a a doctoral candidate. She is working in Weinan Normal University as an associate professor. She is interested in design and development of educational software.

Yongguang Hou, born in 1985, received the master's degree from Guangzhou University in 2010. He is a senior engineer. He is interested in the design and development of educational software.

Malicious Atack Pevention Model of Internet of Vehicles Based on IOV-SIRS

Peng-shou Xie, Cheng Fu, Xin Wang, Tao Feng, and Yan Yan
(Corresponding author: Cheng Fu)

School of Computer and Communications, Lanzhou University of Technology
287 Lan-Gong-Ping Road, Lanzhou, Gansu 730050, China
Email: 452708186@qq.com

(Received Apr. 15, 2020; Revised and Accepted Dec. 10, 2020; First Online Aug. 14, 2021)

Abstract

With the wide application of multiple wireless communication technologies, vehicle nodes realize the connection of various networks such as WiFi, Bluetooth, 802.11p, LTE-V2X, and 5G. Therefore, the attacker accesses the car's internal network through wireless communication and uses malicious viruses for malicious attacks. These malicious viruses interfere with normal vehicle communication, spoofing, or tamper information, which will seriously threaten the security of the Internet of Vehicles. Therefore, this paper studies propagation dynamics on the Internet of Vehicles establishes an IOV-SIRS prevention model, and calculates the relevant data such as the threshold of virus spread and the degree of an outbreak. Finally, we proved in simulation experiments that the IOV-SIRS model has a good inhibitory effect on malicious virus spread.

Keywords: Difference of Individual Awareness; Epidemic Model; IOV-SIRS Model; Malicious Virus; Security of Internet of Vehicles

1 Introduction

In the United States, the Internet of vehicles(IoV) research is primarily based on Wireless Access in Vehicular Environment (WAVE) [10, 11], the wireless communications standard for the Internet of vehicles. WAVE protocol stack developed from Dedicated Short Range Communications (DSRC) standard, including IEEE 802.11p and IEEE P1609. At present, it is the most promising wireless communication standard for the IoV, which takes into account the characteristics of actual Internet of vehicle communication, and can achieve efficient transmission of information under high-speed mobile conditions. However, if we want to use WAVE successfully, we need to build a dedicated service base station for the IoV. As a result, this dramatically limits the popularity of the IoV. But in China, Huawei company has successfully built an LTE-V network and developed communication chips that can provide real-time communications to the vehicle by loading a SIM card into it. The relevant information depends on the application layer program and is finally presented to the IoV users through cellular wireless communication technology.

IoV is a part of wireless communication. Wireless communication is generally integrated in vehicle systems, the IOV-SIRS (IoV-Susceptible-Infected-Recovered-Susceptible) model proposed in this paper can also be integrated to protect the safe of IoV. But attackers installing malware can cause malicious attack to IoV. Malicious viruses used by malware can cause huge harm to the IoV.For example, malicious viruses can interfere or block communication, causing vehicle nodes to fail to establish communication within the receiving range; Malicious viruses faked the relevant information and sent it, causing the vehicle to receive the wrong information, causing the driver to make abnormal behaviors, posing a certain threat to driving; Malicious viruses also can tamper information, each vehicle in IoV can be used as a terminal or relay node, information sent or received by them may be tampered, this will bring more scams and cause huge losses to the user [3]. Overall, those malicious viruses which from malware will affect the normal function of the system, seriously affect driving safety, and even cause traffic accidents [9].

In the research of WAVE technology of IoV. Reference [7] analysed the microscopic effects of wireless propagation model which can be insisted on estimating the more accurate system performance and ensuring the traffic safety in WAVE technology. Reference [14] based on 802.11p/WAVE, analyzed the channel competition situation of Internet of Vehicles MAC layer vehicles accessing the Internet through RSU equipment, and proposed a RSU unit network throughput model suitable for highway traffic scenarios.Reference [6] designed a message format for the Internet of Vehicles to effectively control the flooding of broadcast messages, which provides an important theoretical basis for the design of the upper protocol of the Internet of Vehicles WAVE protocol stack. Researchers

also have done a lot of research on defending against malicious attacks. Reference [19] proposed a Markov game model of mimic Defense. The balance of offensive and defensive games is calculated by a non-linear planning program to determine the best prevention strategy considering defensive costs.

Reference [8] in order to analyze the virus spread under the road environment mixed with Cooperative Adaptive Cruise Control (CACC) vehicles andcommon vehicles, considering the interaction among traffic flow?information flow and virus propagation, CACC vehicle virus infection probability is calculated and the dynamic model of virus spread is built. Reference [18] proposed the stochastic spread model of worms in Internet of vehicles based on stochastic process theory. Reference [1] propsed a user selection and belief propagation based dual defense scheme for large-scale intrusion. Reference [2] based on the reality of network attack-defense confrontation, the influence of bounded rationality on attack-defense stochastic game is analyzed. Under the constraints of bounded rationality, a stochastic game model is constructed. Aiming at the problem of network state explosion, a method of extracting network state and attack-defense action based on attack-defense graph is proposed, which the game state space is effectively reduced. Reference [16] in view of the problem that the existing honeypots often fail to resist the penetration attack due to the lack of confidentiality, an active deception defense method based on dynamic camouflage network was presented.

Although researchers have done a lot of research on defense techniques for malicious attacks [4, 17, 20, 21, 27], but these methods are passive defenses.The disadvantage of this method is obvious: the update speed of security software always lags behind virus updates. From the 2017 WannaCry incident, we can know that if this kind of defense technology is used in the IoV, it will not be able to better prevent new malicious attacks. Therefore, the goal of this paper is to design a method that can achieve "active defense" for the IoV with the help of complex networks and propagation dynamics [12, 13, 22, 26].

The main technical contributions of this paper are as follows. First, the IOV-SIRS prevention model is proposed. This model provides analysis and research ideas for the spread and control of malicious viruses, and is an important means to suppress the spread of viruses in the IoV, so as to achieve the goal of preventing the spread of viruses. At the same time, also provides support for further research on the prevention technology of malicious attacks on the IoV. Second, the data obtained through the calculation of the IOV-SIRS model provides a theoretical basis for the security department, which is of great significance for formulating scientific and effective preventive measures. Based on the research data, people can master the law of virus spread, assess the speed of infection and the scale of outbreaks, so as to realize the "active defense" of malicious attacks by IoV.

2 Building Malicious Attack Detection Modle of IoV Based on IOV-SIRS

2.1 Process of Malicious Attack Prevention of IoV-SIRS

The main work of this section is to improve the SIRS model to obtain a new model. The number of Susceptible, Infected and Recovered can be obtained through anti-virus software reporting, and intrusion detection system and other methods. According to the data analysis of the security center, the infection rate and cure rate of the current connected node of IoV are calculated, so as to obtain the effective spread rate λ_{rt}. According to the basic regeneration number theory, if λ_{rt} is greater than the spread threshold of of the IOV-SIRS model , the network is infected. Sustained malicious attacks will cause the entire network to be infected. When the λ_{rt} is less than the spread threshold of of the IOV-SIRS model, the network is healthy, which means small malicious attack will disappear automatically finally. With the development of malicious attacks on the IoV,when approaches the propagation threshold, the security center can proactively adjust the security scheme and deploy precautionary measures one step in advance to prevent the malicious attacks that are about to expand. The process is shown in Figure 1.

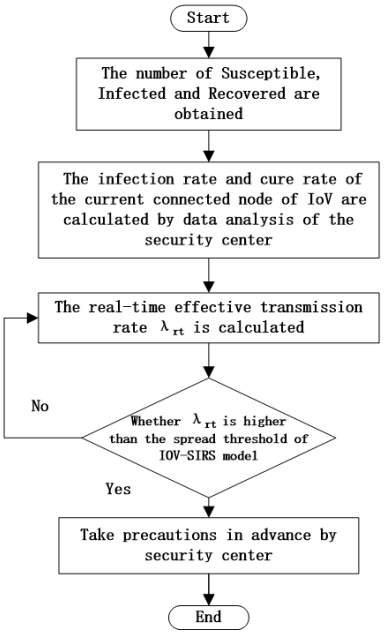


Figure 1: Process of malicious attack prevention of IoV-SIRS

2.2 Classic SIRS Model

First, need to briefly introduce the SIRS model and some proper nouns.Statistics show that the Internet is a scale-free network and exhibits a strong power-law distribution,

so it is a inhomogeneous network [15]. Like the Internet, the IoV is a typical complex network with structures ranging from simple to complex, constantly evolving, and has complex propagation dynamics behavior.

According to the SIRS model [5] define three kinds of nodes of IoV:

Susceptible (S). The normal node in IoV, and contact with infected nodes may be infected.

Infected (I). A node infected by a malicious attack virus in IoV, and it will infect a susceptible node with a certain probability.

Recovered (R). A node in IoV that has been cured or directly immune, and that node will not be infected.

The Susceptible state can be changed to the Infected state by the Infected with the infection rate α; Infected state is converted to Recovered state with cure rate β; The Recovered state is reconverted to the Susceptible state with the immune loss rate γ. As shown in Figure 2.

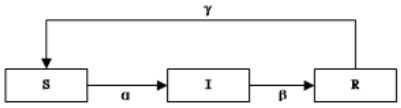


Figure 2: Classic SIRS model

Since the degree of node with power-law distribution does not have a feature scale in a scale-free network [24], the inhomogeneous nodes is considered. $S_k(t)$, $I_k(t)$ and $R_k(t)$ represent the density of the three types of nodes of degree k at time t, respectively, as shown in Equation (1).

$$S_k(t) + I_k(t) + R_k(t) = 1 \tag{1}$$

At time t, the ratio of Infected [24] is shown in Equation (2).

$$I(t) = \sum_k I_k P(k) \tag{2}$$

Where $\theta(t)$ is the probability that one edge is randomly connected to Infected at time t. In a scale-free network with uncorrelated node degrees, $\theta(t)$ is independent of the degree of the node, so it can be expressed as Equation (3) [23].

$$\theta(\mathrm{t}) = \frac{1}{\langle k \rangle} \sum_k k P(k) I_k(t) \tag{3}$$

Under the scale-free network, the nonlinear field theory of propagation dynamics can obtain the nonlinear differential equations of the relative density of Susceptible, Infected and Recovered with time t, as shown in Equation (4).

$$\begin{cases} \frac{dS_k(t)}{dt} = -\alpha k S_k(t)\theta(t) + \gamma R_k(t) - \delta S_k(t) \\ \frac{dI_k(t)}{dt} = -\beta I_k(t) + \alpha k S_k(t)\theta(t) \\ \frac{dR_k(t)}{dt} = -\gamma R_k(t) + \beta I_k(t) + \delta S_k(t) \end{cases} \tag{4}$$

For the SIRS model, there is a variable R_0, which indicates the maximum number of people who can be infected during the average infectious period when all individuals in a group are susceptible. This variable is called the basic reproductive number [15]. When R_0 <1, the number of infections of a single source of infection during the average infection period is less than 1, so the disease has only a disease-free balance point, and the disease-free balance is stable, and the infectious disease can disappear without control. When R_0 >1, the infectious disease not only has a disease-free balance, but also an endemic balance, which means that the infectious disease will always exist in this group and further evolve into endemic disease.

Let $\lambda = \frac{\alpha}{\beta}$ be the effective spread rate [12]. According to the basic reproductive number theory, it can be known that if only the solution of $\frac{\alpha\langle k^2 \rangle}{\beta\langle k \rangle} > 1$ exists, the $\lambda > \frac{\langle k \rangle}{\langle k^2 \rangle}$ is required. Let λ_c be the spread threshold of the SIRS model on a scale-free network, as shown in Equation (5).

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \tag{5}$$

When $\lambda > \lambda_c$, if the disease is not controlled, it will explode on a large scale and become an endemic disease, and gradually converge to the equilibrium point of the endemic disease state. When $\lambda < \lambda_c$, it will automatically die without controlling the disease.

2.3 Improved SIRS Model

In this section, we improved the SIRS model to built an IOV-SIRS model. In the complex network,when researchers study spread of viruses , they do not distinguish between research objects. Usually, things that can be spread in the network are called infectious diseases.There are many similarities between viruses that carry out malicious attacks on the IoV and infectious diseases, such as infectivity, destructiveness, variability, latency, *etc.* Therefore, the transmission mode of the virus in the Internet of Vehicles is similar to the infectious disease model. The classic SIRS model considers the transition mechanism between state nodes, which is consistent with the propagation characteristics of malicious viruses in the IoV. However, the differences of individual awareness and direct immunization have not been considered.

The IOV-SIRS model we built is shown in Figure 2. In the IOV-SIRS model, the total number of IoV nodes is N. Assuming that N is constant, the proportion of no vigilance Susceptible is P($0 < P < 1$) in all Susceptible, and is infected as Infected with infection rate α. The proportion of vigilance Susceptible accounts for 1-P in all Susceptible?and is infected as Infected with infection rate α_1. Infected changes to Recovered with the cure rate β. Recovered changes to Susceptible with the immune loss rate γ. Susceptible changes to Recovered with the direct immunity rate δ. The IOV-SIRS model is shown in Figure 3.

The outbreak of malicious attacks of IoV will inevitably lead to the behavioral responses of individual nodes, and

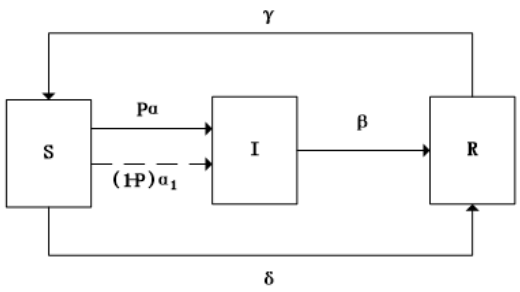


Figure 3: IOV-SIRS model

the behavior of these individual nodes will in turn affect the spread of the overall aggressive virus [23]. Therefore, two factors indicating the difference of individual awareness are made in this paper.

1) The factor of Vigilance Awareness(VA). The VA is the size of index of the vigilance awareness degree of Susceptible. Because the nodes in the scale-free network are unevenly distributed, the number of neighbor infections of each Susceptible is different,so the vigilance degree of Susceptible is different.

   Let n_{inf} be the number of Infected neighbors of Susceptible with degree k.

   The intensity of the VA of the Susceptible with degree k in the scale-free network is shown in Equation (6).

$$D(VA, n_{\inf}) = 1 - (1 - VA)^{n_{\inf}} \tag{6}$$

   The larger the VA, the easier it is for Susceptible to take preventive measures to reduce the risk of infection, and the infection rate of Susceptible with VA is shown in Equation (7).

$$\alpha_1 = \alpha(1 - D) = \alpha(1 - VA)^{n_{\inf}} \tag{7}$$

   Where VA$\in$ $(0, 1)$, when VA=0, represents that all Susceptible have no VA and the infection rate $\alpha_1 = \alpha$ is transformed into Infected; When VA=1, $\alpha_1 = \alpha = 1$ Representing Susceptible do not translate into Infected under ideal conditions.

2) The factor of Prevention Awareness (PA). The PA is the size of index of the prevention awareness degree of the Recovered. From the perspective of reality, Recovered have undergone healing,or are transformed from Susceptible by directly immune, and they all have a certain prevention awareness.

Let n_{sus} be the number of Susceptible neighbors of Recovered with degree k, The intensity of the PA of Recovered with degree k in the scale-free network is shown in Equation (8).

$$Q(PA, n_{\text{sus}}) = 1 - (1 - PA)^{n_{\text{sus}}} \tag{8}$$

The larger the PA, the easier it is for Recovered to update the protective measures to reduce the immune loss, and the immune loss rate of Recovered with the PA is as shown in Equation (9).

$$\gamma(1 - Q) = \gamma(1 - PA)^{n_{\text{sus}}} \tag{9}$$

Where PA$\in$ $(0, 1)$, when PA=0 means that Recovered have no PA, and will convert the normal immune loss rate γ into Susceptible.when PA=1, γ=0 means Recovered is ideal, the Recovered will not be reversed into Susceptible.

3 Data and Stability Analysis

3.1 Analysis of Spread Threshold

This section analyzes the data and stability of the IOV-SIRS model. IoV is a typical inhomogeneous network with node degrees uncorrelated [24, 25], so this paper only calculates the spread threshold under the node degrees uncorrelated inhomogeneous network. The spread threshold under the scale-free network decreases with the increase of the scale of the network, and the threshold approaches zero when the network scale is infinite. However, the scale of the IoV is limited in reality, so it is of practical significance to analyze the spread threshold in IoV.

According to the average field theory and the transformation process of each node in Figure 3, the model of IOV-SIRS on the scale-free network can be obtained as shown in Equation (10).

Under steady state conditions, the initial density of relative density changes with time to 0, satisfying Equation (11).

Let E_0 be the disease-free balance point, and finally stabilize at E_0, indicating that the malicious virus disappears; E_1 is the endemic balance, and finally stable at E1 means that the malicious virus does not disappear. The VA and the PA are introduced, and the equilibrium solutions E0 and E1 of the above equations are obtained by calculation, as shown in Equations (12) and (13).

Equations (12) and (13) are brought into Equation (3), resulting in Equation (14).

Obviously $\theta(\infty)$=0 is a trivial solution to the equation. If $f(\theta(\infty))$ is continuously differentiable, it can be proved that it is strictly monotonically increasing with respect to $\theta(\infty)$, so that the equation has an extraordinary solution of $\theta < \theta(\infty) < 1$, then the right side satisfies Equation (15).

Equation (16) is obtained.

Thus the effective spread rate λ is Equation (17).

Let λ_c be the spread threshold and obtain the spread threshold of the IOV-SIRS model, as shown in Equation (18).

The result show that when PA=VA=δ=0, the the spread threshold becomes $\frac{\langle k \rangle}{\langle k^2 \rangle}$, which is consistent with the SIRS model spread threshold. When $\lambda < \lambda_c$, they do not need to control the malicious virus, they will die automatically. When $\lambda > \lambda_c$, if the malicious virus is not controlled, it will cause a large-scale outbreak. The specific proof process is proved in experiments.

$$\begin{cases} \frac{dS_k(t)}{dt} &= -(1-P)\alpha_1 k S_k(t)\theta(t) - P\alpha k S_k(t)\theta(t) + \gamma R_k(t) - \delta S_k(t) \\ \frac{dI_k(t)}{dt} &= (1-P)\alpha_1 k S_k(t)\theta(t) + P\alpha k S_k(t)\theta(t) - \beta I_k(t) \\ \frac{dR_k(t)}{dt} &= \beta I_k(t) + \delta S_k(t) - \gamma R_k(t) \end{cases} \tag{10}$$

$$\frac{dS_k(t)}{dt} = 0, \frac{dI_k(t)}{dt} = 0, \frac{dR_k(t)}{dt} = 0 \tag{11}$$

$$E_0 = (S_k(t), 0) = \left(\frac{\gamma(1-PA)^{n_{sus}}}{\gamma(1-PA)^{n_{sus}} + \delta + \left[(1-P)(1-VA)^{n_{inf}} + P\right] \alpha k \theta(t)}, 0 \right) \tag{12}$$

$$E_1 = (S_k(t), I_k(t)) \tag{13}$$

$$= \frac{\left[\gamma(1-PA)^{n_{sus}}\right]\beta}{\left[\gamma(1-PA)^{n_{sus}} + \delta\right]\beta + \left[\gamma(1-PA)^{n_{sus}} + \beta\right]\left[(1-P)(1-VA)^{n_{inf}} + P\right]\alpha k \theta(t)}$$

$$= \frac{\left[\gamma(1-PA)^{n_{sus}}\right]\left[P(1-VA)^{n_{inf}} + 1 - P\right]\alpha_2 k \theta(t)}{\left[\gamma(1-PA)^{n_{sus}} + \delta\right]\beta + \left[\gamma(1-PA)^{n_{sus}} + \beta\right]\left[(1-P)(1-VA)^{n_{inf}} + P\right]\alpha k \theta(t)}$$

$$\theta(\infty) = \frac{1}{\langle k \rangle} \sum_k k P(k) I_k(\infty) = f(\theta(\infty)) = \frac{1}{\langle k \rangle} \sum_k k^2 P(k) \tag{14}$$

$$* \frac{\left[\gamma(1-PA)^{n_{sus}}\right]\left[P(1-VA)^{n_{inf}} + 1 - P\right]\alpha_2 \theta(\infty)}{\left[\gamma(1-PA)^{n_{sus}} + \delta\right]\beta + \left[\gamma(1-PA)^{n_{sus}} + \beta\right]\left[(1-P)(1-VA)^{n_{inf}} + P\right]\alpha k \theta(\infty)}$$

$$\left.\frac{df(\theta(\infty))}{d\theta(\infty)}\right|_{\theta(\infty)=0} > 1 \tag{15}$$

$$\left.\frac{df(\theta(\infty))}{d\theta(\infty)}\right|_{\theta(\infty)=0} = \frac{\langle k^2 \rangle}{\langle k \rangle} \frac{\left[\gamma(1-PA)^{n_{sus}}\right]\left[(1-P)(1-VA)^{n_{inf}} + P\right]\alpha}{\left[\gamma(1-PA)^{n_{sus}} + \delta\right]\beta} > 1 \tag{16}$$

$$\lambda = \frac{\alpha}{\beta} > \frac{\langle k \rangle}{\langle k^2 \rangle} \frac{\left[\gamma(1-PA)^{n_{sus}} + \delta\right]}{\left[\gamma(1-PA)^{n_{sus}}\right]\left[(1-P)(1-VA)^{n_{inf}} + P\right]} \tag{17}$$

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \frac{\left[\gamma(1-PA)^{n_{sus}} + \delta\right]}{\left[\gamma(1-PA)^{n_{sus}}\right]\left[(1-P)(1-VA)^{n_{inf}} + P\right]} \tag{18}$$

3.2 Analysis of Disease-Free Equilibrium

This section analyzes the disease-free equilibrium of the IOV-SIRS model and calculates the density of the final immune nodes. For the IOV-SIRS model, when the malicious virus of IoV is cleared and reaches a disease-free steady state, it finally stabilizes at the disease-free equilibrium point E_0,and the normalized equation $R(\infty) = 1 - S_k(\infty) - I_k(t)$, and $I_k(\infty) = 0$ and Equation (19) is obtained.

The final infection range of the standard SIRS model with direct immunization is $\sum_k P(k)\left[1 - \frac{\gamma}{\gamma+\delta+\alpha k\theta(t)}\right]$. The analysis shows that the infectious range value of the IOV-SIRS model is smaller than the classical SIRS model.The specific results are demonstrated in simulation experiments.

3.3 Analysis of Endemic Equilibrium

This section analyzes the endemic equilibrium of the IOV-SIRS model and calculates the final infection density. For the IOV-SIRS model, when malicious virus of IoV breaks out and reaches the endemic steady state, it finally stabilizes at the endemic equilibrium point E_1. In the scale-free network, the average degree and degree distribution [15] are shown in Equation (20).

Where m is the minimum number of connected edges in the network, and Equation (20) is substituted into Equation (3) to obtain Equation (21).

Equation (22) is obtained by integrating k on both sides.

The final infection density Equation (23) is obtained from $I(t) = \sum_k I_k P(k)$.

The specific results are demonstrated in simulation experiments.

4 Simulation Experiment

In this section, all the experiments are implemented in python3 platform on a computer with Intel(R) Core (TM) i5-7500HQ CPU @2.50GHz, 8G RAM, Win10 (64 bit) operating system. Experiments set different parameter values to observe the effect of these values on the density of three nodes. The initial parameters are:the number of node of IoV is set to 10000, the average degree is 3, the basic infection rate is α=0.1, the cure rate is β=1, the immune loss rate is γ=0.2, and the direct immunity rate is δ=0.2; The initial susceptible node At 8000, the infected node is 1000 and the recovered node is 1000.

4.1 Comparison with SIRS and IOV-SIRS

The experiments in this section are used to discuss the influence of PA and VA on the infection level of the

$$R(\infty) = \sum_k P(k)\left(1 - S_k(\infty) - I_k(\infty)\right) \tag{19}$$

$$= \sum_k P(k) * \left(1 - \frac{\gamma(1-PA)^{n_{sus}}}{\gamma(1-PA)^{n_{sus}} + \delta + [(1-P)(1-VA)^{n_{inf}+P}]\alpha k\theta(t)}\right)$$

$$\langle k \rangle = \int_m^{\infty} kP(k) = 2m, P(k) = \frac{2m^2}{k^3} \tag{20}$$

$$\theta(\infty) = f(\theta(\infty)) = \frac{1}{\langle k \rangle}\sum_k kP(k)I_k(\infty) = \frac{1}{2m}\sum_k k^2\frac{2m^2}{k^3} * I_k(\infty) \tag{21}$$

$$I_k(\infty) = \left\{\frac{\left[\gamma(1-PA)^{n_{sus}}\right][P(1-VA)^{n_{inf}}+1-P]\alpha_2\theta(\infty)}{\left[\gamma(1-PA)^{n_{sus}}+\delta\right]\beta + \left[\gamma(1-PA)^{n_{sus}}+\beta\right]\left[(1-P)(1-VA)^{n_{inf}}+P\right]\alpha k\theta(\infty)}\right\}$$

$$\theta(\infty) = \left(\frac{\left[\delta+\gamma(1-PA)^{n_{sus}}\right]\beta}{m\left[\gamma(1-PA)^{n_{sus}}+\beta\right]\left[(1-P)(1-VA)^{n_{inf}}+P\right]\alpha}\right) \tag{22}$$

$$* \left(\exp\left(\frac{\left[\delta+\gamma(1-PA)^{n_{sus}}\right]\beta}{m\left[\gamma(1-PA)^{n_{sus}}+\beta\right]\left[(1-P)(1-VA)^{n_{inf}}+P\right]\alpha}\right) - 1\right)^{-1}$$

$$I(\infty) = \frac{2\gamma(1-PA)^{n_{\text{sus}}}}{\gamma(1-PA)^{n_{\text{sus}}+\beta}} \tag{23}$$

$$* \left(\exp\left(\frac{\left[\delta+\gamma(1-PA)^{n_{sus}}\right]\beta}{m\left[\gamma(1-PA)^{n_{sus}}+\beta\right]\left[(1-P)(1-VA)^{n_{inf}}+P\right]\alpha}\right) - 1\right)^{-1} * (1-\theta(\infty))$$

Infected and the immunity level of the Recovered. The SIRS model did not have vigilance awareness and prevention awareness, so its VA and PA are 0. Setting P = 0.8 means that only a small part of the susceptible nodes have vigilance awareness. The specific parameters are set as follows:

- The SIRS model are set to P=1, VA=0, PA=0, δ=0;
- The IOV-SIRS model are set to P = 0.8, VA = 0.2, PA= 0.2, δ = 0.2;

After multiple simulations were taken to obtain the mean value, the change of the density of the Infected with time t is I(t), and the change of the density of the recovered with time t is R(t). they were all obtained and shown in Figures 4 and 5.

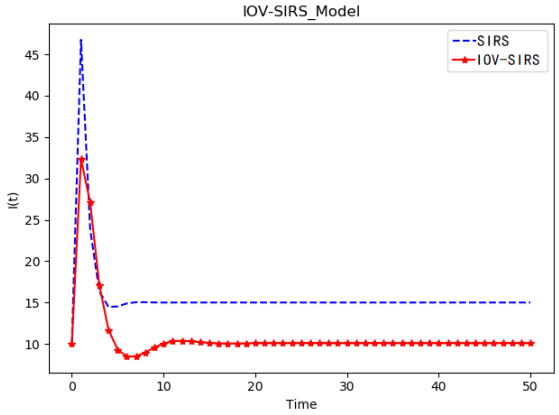


Figure 4: I(t) changes with time t

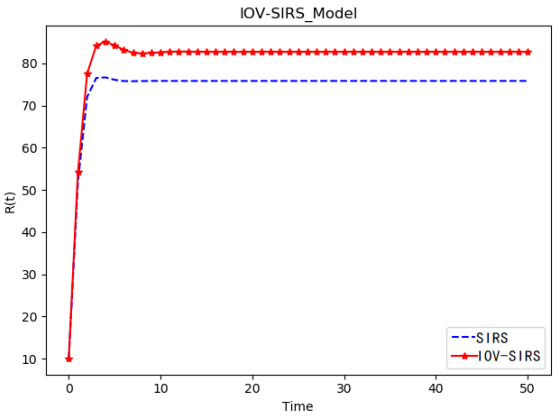


Figure 5: R(t) changes with time t

The curve in Figure 4 shows that after the malicious viruses outbreak, the density of Infected increased rapidly. After reaching the maximum infection scale, it gradually decreased after healing, and finally stabilized.The simulation results show that the IOV-SIRS model is superior to the SIRS model in controlling infection.

The curve in Figure 5 shows that after the virus outbreak, the Infected becomes an Recovered after being cured. After reaching the maximum density, some Recovered lose their immunity and transform into Susceptible, resulting in a slight decrease in the density of Recovered and eventually reaching a steady state.The Simulation results show that the IOV-SIRS model is superior to the SIRS model in maintaining immunity.

4.2 Influence of VA on IOV-SIRS Model

The experiments in this section are used to discuss the effect of VA on Susceptible and Infected. The specific parameters are set as follows:

The IOV-SIRS model without VA are set to P=1, VA=0, PA=0, δ=0.2.

The IOV-SIRS model with VA are set to P=0.8, VA=0.2, PA=0, δ=0.2.

After multiple simulations were taken to obtain the mean value, the change of the density of the Susceptible with time t is S(t), and the change of the density of the Infected with time t is I(t). They were all obtained and shown in Figures 6 and 7.

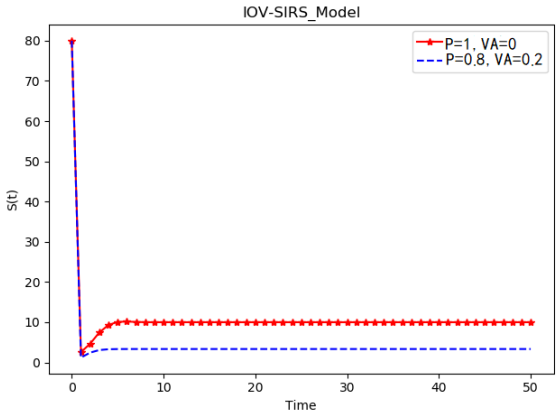


Figure 6: S(t) changes with time t

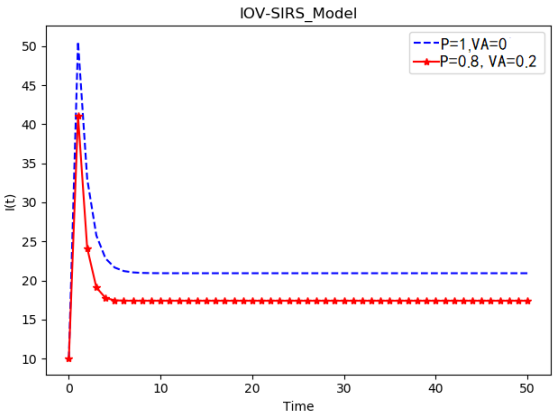


Figure 7: I(t) changes with time t

The curve in Figure 6 shows that after the malicious viruses outbreak, the Susceptible are infected by the Infected, resulting in a rapid decline in the density of Susceptible. When the density of Susceptible reaches the lowest, because some Recovered lose their immunity and become Susceptible again, the density of Susceptible increases slightly and eventually stabilizes. As shown in Figure 6.

The results in Figures 6 and 7 show that in the IOV-SIRS model, the smaller the P value is, the larger the VA is, which makes the density of Susceptible increase in the steady state.Therefore, the lower the probability of Susceptible being transformed into Infected, and the density of Infected at steady state is also reduced as the the VA increases.The simulation results prove that VA inhibits the spread of malicious attacks.

4.3 Influence of PA on IOV-SIRS Model

The experiments in this section are used to discuss the effect of PA on Recovered. The specific parameters are set as follows:

The IOV-SIRS model without PA are set to P=0.8, VA=0.2, PA=0, δ=0.2;

The IOV-SIRS model with PA are set to P=0.8, VA=0.2, PA=0.2, δ=0.2;

After multiple simulations were taken to obtain the mean value, the change of the density of the Recovered with time t is R(t),it was obtained and shown in Figure 8.

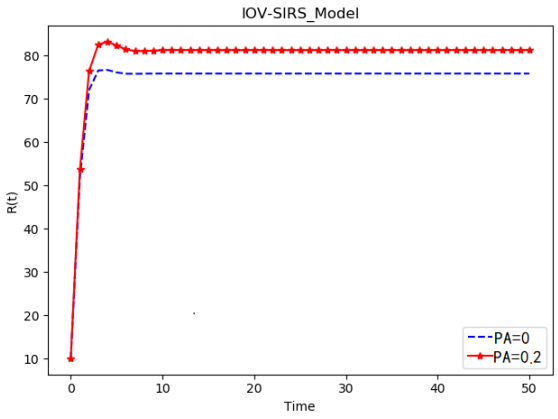


Figure 8: R(t) changes with time t

The results in Figure 8 show that in the IOV-SIRS model, the larger the PA, the smaller the probability of immune loss of Recovered, therefor, the density of Recovered is higher in the steady state.The simulation results show that increasing PA can effectively reduce immune loss.

4.4 Influence of delta on IOV-SIRS Model

The experiments in this section are used to discuss the effect of the direct immune rate ? on Infected and Recovered. The specific parameters are set as follows:

The IOV-SIRS model with high direct immunity rate are set to P=0.8, VA=0.2, PA=0.2, δ=0.1;

The IOV-SIRS model with low direct immunity rate are set to Set P=0.8, VA=0.2, PA=0.2, δ=0.3;

After multiple simulations take the mean value, the change of the density of the recovered node R(t) with time t and the change of the infected node I(t) with time t are obtained, as shown in Figures 9 and 10.

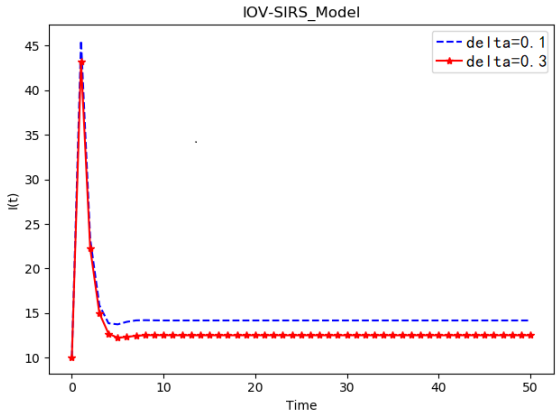


Figure 9: I(t) changes with time t

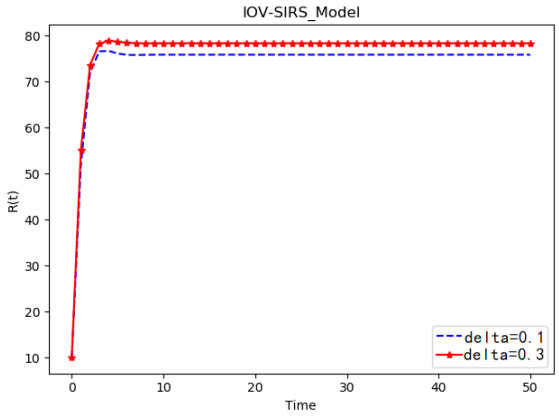


Figure 10: R(t) changes with time t

The results of Figures 9 and 10 show that in the IOV-SIRS model, increasing the direct immunization rate can slightly reduce the density of Infected in steady state and slightly increase the density of Recovered.The simulation results show that although increasing the direct immunity rate can slightly suppress the spread of malicious attacks, the overall effect is not very obvious.

5 Preventive Measures

Before the malicious viruses spread or outbreak , different levels of measures can be implemented to suppress the malicious viruses spread and outbreak. This section lists some security measures as a reference.

5.1 Increase the VA

Increase the dissemination of malicious attacks of IoV and the promotion of infection routes, with the aim of increasing the VA for IOV users. When the Susceptible in high VA, it will reduce high-risk individual behavior.

Specific measures include: Increasing publicity, refusing to receive unfamiliar files; Secure access system authentication; Increasing personal random factor signature authentication; Password hardening; Trusted access and data transmission,etc.

5.2 Increase the PA

The security department can increase the protection propaganda to increase the PA.In high PA,the Recovered can open the protection measures and methods one step earlier to effectively reduce the immune loss.

Specific measures include: Increasing secondary publicity; Isolating susceptible populations; Updating immune patches in a timely manner; Strengthening system protection; When the virus breaks out, reduce the contact between the Recovered and other nodes, *etc.*

5.3 Increase the Direct Immunization Rate

The security department releases and updates the immunization patch in a timely manner, and promptly pushing the information to the user can increase the direct immunization rate.

Specific measures include: Push installation of immune patches and security software updates; Close sensitive ports; Block unfamiliar IP; Filter low-security TCP/UDP protocol, *etc.*

6 Conclusion

At present, few researchers use the idea of complex networks to design prevention models for IoV, and passive defense methods are not enough to cope with the ever-changing attacks of new malicious viruses. Therefore, this paper proposed the IOV-SIRS model designed for IoV. First, establish the IOV-SIRS model based on IoV; Then, based on the differential equation, the average field theory method is used to obtain the spread threshold; Finally, the disease-free balance point and the endemic disease balance point are calculated, and the relevant data are obtained to further obtain the final spread of the virus Scope and degree of immunity. The experimental results verify the validity of the theory proposed by the IOV-SIRS model. At the same time, the data calculated by the IOV-SIRS model allows the security department can grasp the overall situation of the virus spread trend, and deploy the different levels of security measures when the virus is about to break out. This paper points out that before the virus breaks out,mastering the method of virus spread in advance is the most effective way to protect the IoV from large-scale malicious attacks.

The disadvantage of this paper is the lack of consideration of the impact of real environmental factors, which will be the next problem to be solved. The real IoV environment is more complicated, and it also contains realistic factors such as the migration of out-of-group individuals, non-linear infection rate, *etc.* The IOV-SIRS model mentioned in this paper only considers the the difference of

individual awareness and direct immunity, which is not enough Cope with complex and changing real environment. The next step is to add some factors that consider the real environment to the IOV-SIRS model, such as the non-linear infection rate, the migration of out-of-group individuals, and the horizontal and vertical transmission of viruses.This can further improve the accuracy of model predictions and data.

Acknowledgments

This research is supported by the National Natural Science Foundations of China under Grants No.61862040, No.61762059 and No.61762060. The authors gratefully acknow-ledge the anonymous reviewers for their helpful comments and suggestions.

References

[1] Z. Baoyu, J. Wei, C. Qingqing, "Dual defense scheme for large-scale intrusion in cognitive radio networks," *System Engineering and Electronics*, vol. 41, no. 10, pp. 2352–2358, 2019.

[2] Z. Chuanfu, Y. Junnan, Z. Hongqi, "Network defense decision-making method based on stochastic game and improved wolf-phc," *Journal of Computer Research and Development*, vol. 56, no. 5, pp. 942–954, 2019.

[3] A. Dewanje and K. A. Kumar, "A new malware detection model using emerging machine learning algorithms," *International Journal of Electronics and Information Engineering*, vol. 13, no. 1, pp. 24–32, 2021.

[4] Y. Farhaoui, "Design and implementation of an intrusion prevention system," *International Journal of Network Security*, vol. 19, no. 5, pp. 675–683, 2017.

[5] C. Gaoqiang, Y. Zhuzhi, X. Chengyi, L. Zhongxin, "Sirs epidemic model with direct immunization on complex networks," *Control and Decision*, vol. 23 no. 4, pp. 468–472, 2008.

[6] W. Guoxin, L. Ye, "Research on connectivity model and application of internet of vehicles based on 802.11p/wave," *Journal on Communications*, vol. 34, no. 6, pp. 85–91, 2013.

[7] D. H. Ha, "Radio channel model suitable for the next generation wireless access vehicular environment technology," *The Journal of KIIT*, vol. 12, no. 4, pp. 43–50, 2014.

[8] Q. Hongkun, Y. Guizhen, W. Lei, W. Yunpeng, "A research on security of information propagation of cacc vehicles under internet of vehicles environment," *Automotive Engineering*, vol. 41, no. 3, pp. 252–258, 2019.

[9] M. Inam, Z. Li, A. Ali, and A. Zahoor, "A novel protocol for vehicle cluster formation and vehicle head selection in vehicular ad-hoc networks," *International Journal of Electronics and Information Engineering*, vol. 10, no. 2, pp. 103–119, 2019.

[10] L. Jingfeng, *Research on 5G-based Internet of Vehicles Road Safety Information Transmission Mechanism*, Shanghai Jiaotong University, 2018 (in Chinese).

[11] B. Junling, *The Prototype and Simulation of WAVE Standard for VANET*, Dalian University of Technology, 2014 (in Chinese).

[12] Y. Li, Y. Deng, Y. Xiao, and J. Wu, "Attack and defense strategies in complex networks based on game theory," *Journal of Systems Science and Complexity*, vol. 32, no. 6, pp. 1630–1640, 2019.

[13] L. Li, Z. Wei, W. Weichuan, Z. Lin, L. Linjun, "Malware propagation model based on time delay in wireless sensor networks," *Advanced Engineering Sciences*, vol. 51, no. 3, pp. 167–174, 2019.

[14] Z. Long, L. Ye, L. Linfeng, "Study on the downslink performance of roadside unit in vehicular ad-hoc networks," *Journal of Software*, vol. 26, no. 7, pp. 1700–1710, 2015.

[15] W. Mengzhang, *Analysis of Stability for Some Classes of Epidemic Model*, China: Hunan University, 2018 (in Chinese).

[16] P. Qingqi, T. Guangming, W. Yang, L. Xiaohu, W. Shuo, W. Jianhua, "Active deception defense method based on dynamic camouflage network," *Journal of Communications*, vol. 41, no. 2, pp. 97–111, 2020.

[17] Y. Ren, S. Wang, X. Zhang, and M. S. Hwang, "An efficient batch verifying scheme for detecting illegal signatures," *International Journal Network Security*, vol. 17, no. 4, pp. 463–470, 2015.

[18] F. Runze, X. Junkun, W. Ming, G. Wei, Z. Hanxun, Y. Yang, "Stochastic propagation model of worms in internet of vehicles," *Automotive Engineering*, vol. 13, no. 9, pp. 1524–1533, 2019.

[19] W. Shuai, S. Jianliang, Z. Xingming, G. Zeyu, "Markov game modeling of mimic defense and defense strategy determination," *Journal on Communications*, vol. 39, no. 10, pp. 143–154, 2018.

[20] J. Xia, Z. Cai, G. Hu, and M. Xu, "An active defense solution for ARP spoofing in openflow network," *Chinese Journal of Electronics*, vol. 28, no. 1, pp. 172–178, 2019.

[21] P. Xie, T. Xiao, and H. J. Fan, "An algorithm of the privacy security protection based on location service in the internet of vehicles," *International Journal Network Security*, vol. 21, no. 4, pp. 556–565, 2019.

[22] Z. Xu and D. Chen, "An sis epidemic model with diffusion," *Applied Mathematics-A Journal of Chinese Universities*, vol. 32, no. 2, pp. 127–146, 2017.

[23] L. Yanling, *Epidemic Spreading in Complex Networks based on Human Behavior*, China:Nanjing University of Posts and Telecommunications, 2015 (in Chinese).

[24] W. Yibo, F. Ruguo, "Seir-based covid-19 transmission model and inflection point prediction analysis," *Journal of University of Electronic Science and Technology of China*, vol. 49, no. 3, pp. 369–374, 2020.

[25] G. Yinjing, Q. Zhihong, T. Chunsheng, "The key technology and development of intelligent and connected transportation system," *Journal of Electronics and Information*, vol. 42, no. 1, pp. 2–19, 2020.

[26] H. Yuli, L. Jianhua, C. Qiying, S. Shigen, F. En, "Reliability evaluation for wsns with malware spread," *Acta Electronica Sinica*, vol. 46, no. 1, pp. 75–81, 2018.

[27] M. Zhang, C. Shen, N. He, S. Han, Q. Li, Q. Wang, and X. H. Guan, "False data injection attacks against smart gird state estimation: Construction, detection and defense," *Science China Technological Sciences*, vol. 62, pp. 2077–2087, 2019.

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 Security on Internet of Things. E-mail: xiepsh_lut@163.com.

Cheng Fu was born in June 1991. He is a master student at Lanzhou University of Technology. His major research field is network and information security. E-mail: 452708186@qq.com.

Xin Wang was born in Jan.1995. She is a master student at Lanzhou University of Technology. She major research field is network and information security. E-mail: 415711979@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.

Yan Yan was born in Oct. 1980. She is a associate professor and a supervisor of master student at Lanzhou University of Technology. Her major research field is privacy protection, multimedia information security.E-mail: yanyan@lut.cn.

Role-Engineering Optimization with User-Oriented Cardinality Constraints in Role-Based Access Control

Wei Sun, Xiaoya Yuan, and Hui Su
(Corresponding author: Wei Sun)

Center of Network Information and Computing, Xinyang Normal University
237 Nanhu Road, Shihe District, Xinyang, Henan 464000, China
Email: sunny810715@xynu.edu.cn

(Received Dec. 4, 2020; Revised and Accepted May 6, 2021; First Online Aug. 14, 2021)

Abstract

Role-based access control (RBAC) is a popular security mechanism used by many organizations because it provides various constraint policies, such as cardinality constraints and separation-of-duty constraints. Role-engineering technology is an effective method for constructing RBAC systems. However, the mining scales are large, and the management burdens of systems are weighty. Furthermore, conventional role-engineering methods do not consider cardinality constraints. To address these issues, this paper proposes a novel method, called role-engineering optimization, with user-oriented cardinality constraints on roles (REO_UCCR). First, to reduce the mining scales and alleviate the management burdens of systems, we convert basic role mining into a clustering problem using the Hamming distance technique and four tuples of clusters. Then, we implement role mining while constructing an unconstrained role engineering system. Second, to verify whether the given cardinality constraints can be satisfied in the constructed system, we present a role optimization algorithm to reconstruct a constrained RBAC system. The experiments using synthetic and real datasets demonstrate the effectiveness of the proposed method and show encouraging results.

Keywords: Role Engineering; Role Mining; Role Optimization; User-Oriented Cardinality Constraints

1 Introduction

With the rapid development and comprehensive application of network information technology, a large amount of information storage and exchanges are required in large-scale and complex information-management systems [11, 22]. An increasing number of enterprises and organizations have adopted role-based access control (RBAC) as their main access-control mechanism over the last three decades, as it makes security administration more flexible and manageable [2, 7, 16, 17]. With the successful implementation of RBAC systems, devising an accurate and effective set of roles and constructing a good RBAC system, which can satisfy actual application requirements, have become critical tasks. The bottom-up role-engineering technology [1, 6, 14] aims to migrate from non-RBAC systems to RBAC systems. It starts from the original user-permission assignments and aggregates them into roles by applying data mining techniques, which is also known as role mining and has gained considerable attention in recent years.

In fact, role mining is the task of clustering users with identical or similar permissions and constructing different roles with these permissions [19]. Roles containing several identical permissions are frequently assigned to users. Usable roles can frequently facilitate the management and maintenance of the system and decompose the set of users into clusters of users with different attribute properties. To enhance the interpretability of role mining, it is indeed necessary to cluster users with the same attribute properties. However, due to the diversity of the attribute properties of entities and the variability of accesses, the mining scales are large, and the management burdens of systems are very heavy using conventional role-mining methods.

A key characteristic of RBAC is that it allows for the specification and enforcement of various types of security policies [15,18], such as separation-of-duty constraints and cardinality constraints, which reflect the security requirements of organizations and can ensure the security of RBAC systems. There are four different types of cardinality constraints among users, roles, and permissions, and they limit the maximum number of roles related to users or to permissions, the maximum number of permissions a role can have, or the maximum number of users to which a role can be assigned [12]. For example, the general-manager role in a company must be assigned to only one person, and ordinary users should not have too many roles; otherwise, there is the possibility that users

will abuse their privileges. In terms of the approaches to the construction of role engineering, however, most of the existing methods cannot determine whether the given cardinality constraints are satisfied in a constructed RBAC system.

To address the abovementioned issues, this paper proposes a novel method, called role-engineering optimization, with user-oriented cardinality constraints on roles (REO_UCCR). In summary, the main contributions of this work are as follows:

1) To reduce the mining scales and alleviate the management burdens of systems, we adopt the Hamming distance technique to rearrange an original access matrix, generate user clusters, and then implement role mining, while constructing an unconstrained role engineering system.

2) To verify whether the given cardinality constraints can be satisfied in the constructed system, we first present the definition of the role-engineering optimization problem and then propose a role optimization algorithm to reconstruct a constrained RBAC system.

The rest of the paper is organized as follows. In Section 2, we discuss the related work and present some necessary preliminaries. In Section 3, we propose a novel research method and present several algorithms and running examples. We show the experimental evaluations in Section 4. Section 5 concludes the paper and discusses future work.

2 Related Work and Preliminaries

2.1 Methods of Role Engineering

To discover interesting roles in existing permission assignment relationships, two algorithms, called the Complete Miner and Fast Miner, were proposed [19]. Both the two algorithms use subset enumeration and allow for overlapping roles. While the first algorithm enumerates all potential roles, its computational complexity is exponential. The second algorithm improves the mining process, and its computational complexity is remarkably reduced. Vaidya *et al.* [20] converted role mining into a matrix decomposition problem and presented a definition of a basic role mining problem (basic RMP). The basic RMP has been proven to be NP-complete, and several existing studies have already been conducted to find efficient solutions. To avoid an abuse of privileges, Blundo *et al.* [3] proposed a heuristic capable of returning a complete set of roles, which limited the number of permissions assigned to a role. John *et al.* proposed two alternative approaches for restricting the number of roles assigned to a user: The role priority-based approach (RPA) and the coverage of permissions-based approach (CPA). The RPA prioritizes roles based on the number of permissions and assigns optimal roles to users, according to the priority order. The CPA chooses roles by iteratively picking the role with the largest number of permissions that are yet to be uncovered and then ensures that no user is assigned more than a given number of roles [9]. To simultaneously limit the maximum number of roles assigned to a user and a related permission, Harika *et al.* proposed two role-optimization methods: Post processing and concurrent processing. In the first method, roles are initially mined without taking the constraints into account. The user-role and role-permission assignments are then checked for constraint violation in the optimization process and appropriately re-assigned, if necessary [8]. The concurrent processing method implements optimization with double constraints during the process of role mining. Wang *et al.* [21] proposed two kinds of role mining algorithms in order to satisfy the permission cardinality constraints. The first algorithm discovered roles by decomposing a sorted access control matrix, and the second intersected the permissions of adjacent users in the access control matrix to generate candidate roles. Blundo *et al.* [4] focused on cardinality constraints, defined the constrained role mining problem for each constraint type, and presented efficient heuristics for these problems. In addition, to satisfy separation-of-duty constraints and ensure authorization security, we proposed a method, called role-mining optimization, with separation-of-duty constraints and security detection for authorizations [13].

Two main limitations are apparent in the existing studies. The first limitation is that the role-mining scales are very large, and the management burdens of systems are very heavy. The second limitation is that most conventional role-mining methods do not consider whether or not the number of roles related to a user is restricted. If the number of roles assigned to a user exceeds a particular value, then there is the possibility of an abuse of privileges, and the system is not secure. Hence, we propose a novel role-engineering optimization method in order to alleviate the management burdens, while ensuring the system security. We also evaluate the performance of the proposed method on the synthetic and real datasets.

2.2 Preliminaries

2.2.1 Basic Components of Role Engineering

According to the NIST standard of RBAC, conventional role engineering for RBAC consists of the following basic components:

1) U, P, and R are the basic elements of RBAC, which represent the sets of users, permissions, and roles, respectively;

2) $UPA \subseteqq U \times P$ represents a many-to-many mapping relationship of user-permission assignments;

3) $URA \subseteqq U \times R$ represents a many-to-many mapping relationship of user-role assignments;

4) $RPA \subseteqq R \times P$ represents a many-to-many mapping relationship of role-permission assignments;

5) $user\_roles(u) = \{r|\exists r \in R : ((u,r) \in URA)\}$, which represents a set of roles assigned to user u;

6) $user\_perms(u) = \{p|\exists p \in P, \exists r \in R : ((u,r) \in URA) \wedge ((r,p) \in RPA)\}$, which represents a set of permissions assigned to user u.

2.2.2 The Basic RMP Problem and Fast Miner Method

The basic RMP [20] can be formally represented as follows:

$$\begin{cases} \min |R| \\ URA \otimes RPA = UPA. \end{cases} \quad (1)$$

For convenience, the UPA, URA, and RPA are used to represent their respective assignment relationships, as well as the corresponding matrices. The Fast Miner method [19] mainly consists of the following steps:

1) According to the hash mapping rule, a given access control matrix is converted into the user-permission assignment relationship;

2) To reduce the size of the original data set, different users who have the same permissions in the permission assignments are grouped together, and an initial set of roles is constructed;

3) All the potentially interesting roles are identified by implementing intersections between any pair of the initial roles. New roles are generated, and the number of users associated with any new role is counted.

2.2.3 Hamming Distance

Since the access control matrix, UPA, is a Boolean matrix, each row (or each column) can be regarded as a binary vector of the same length. The well-known technique of Hamming distance [5] is widely used to measure the distance between two different equal-length vectors. It states that given two equal-length Boolean vectors, x and y, the distance between x and y, denoted as $Dis(x,y)$, is the number of positions, where the vectors take different values for the same column position.

For instance, given two row vectors, $x =$ "100110" and $y =$ "110011", $Dis(x,y) = 3$. Clearly, the distance between any two rows in UPA increases as the number of column positions taking different values increases.

2.2.4 User-Oriented Cardinality Constraints on Roles (UCCR)

The UCCR [12] states that, given set U of users, set R of roles, and threshold MRC_{user}, the number of roles assigned to any user should not exceed MRC_{user}. This can be formalized as follows:

$$\forall u \in U : |user\_roles(u) \cap R| \leq MRC_{user} \quad (2)$$

In addition, there are another three cardinality constraints in RBAC, and they are not discussed in this work.

3 Proposed Method

In this section, we propose a novel research method, named REO_UCCR, which includes three aspects: 1) The generation of user clusters, 2) construction of unconstrained role engineering, and 3) role-engineering optimization with UCCR. Specifically, we adopt the Hamming distance technique to rearrange an original access matrix and generate user clusters in the preprocessing stage. Subsequently, we construct an unconstrained role engineering system in the role mining stage. Last, to verify whether the given cardinality constraints can be satisfied in the constructed system, we present a role optimization algorithm to reconstruct a constrained RBAC system. An overall view of the proposed framework is shown in Figure 1.

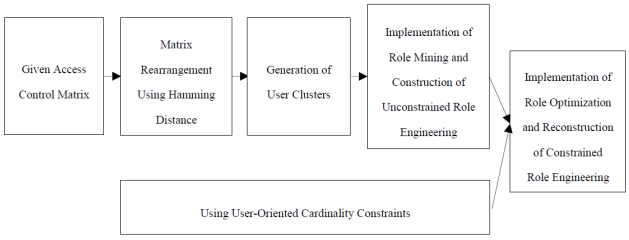


Figure 1: Overview of the proposed role-engineering optimization framework

3.1 Generation of User Clusters

To intuitively represent the matrix, UPA, we use the Hamming distance to rearrange it, as defined below.

Definition 1. *(Matrix rearrangement problem with Hamming distance) Given an original matrix UPA, and a Hamming distance list D between any two rows of UPA, find a rearranged matrix UPA', such that the sum of distances between the adjacent rows of UPA' is minimal, which can be formalized as follows:*

$$\min(\sum_i Dis(UPA'[i], UPA'[i+1])), \quad \forall Dis(UPA'[i], UPA'[i+1]) \in D. \quad (3)$$

According to Definition 1, we present the process of matrix rearrangement in Algorithm 1.

Algorithm 1 Matrix rearrangement.

Input: original matrix *UPA*
Output: rearranged matrix *UPA'*
1: Initialize *UPA'*= *UPA*;
2: Represent *UPA'* as a list of row vectors: *UPA'*[1], *UPA'*[2], ...;
3: Identify matrix D_r of the Hamming distances between any two row vectors, such that $\forall i,j$: $D_r[i][j]$= *Dis*(*UPA'*[*i*], *UPA'*[*j*]);
4: **for** each *UPA'*[*i*] in *UPA'* **do**
5: **if** ($\exists$*UPA'*[*j*]: $D_r[i][j] < D_r[i][i+1]$) **then**
6: *swap*(*UPA'*[*i*+1], *UPA'*[*j*]);
7: **end if**
8: **end for**

It can be seen, from Algorithm 1, that different users with the same permissions are grouped together, which can be regarded as a user group. To reduce the mining scale, we represent the groups as different user clusters and adopt four tuples to store them, as well as other properties. This is easy to implement, and we present its definition as follows.

Definition 2. *(Four tuples of user clusters) The users with the same permissions, as well as their properties, are group, which is denoted as a four-tuple form <c, user_perms(c), count_users(c), count_unperms(c)>, where c is a user cluster, C is a set of different clusters, user_perms(c) is the permission set associated with c, count_users(c) is the number of users included in c, and count_unperms(c) is the number of permissions uncovered recently in c.*

It is apparent that count_unperms(c) is equal to the value of $|user\_perms(c)|$, before role mining.

3.2 Construction of Unconstrained Role Engineering

To alleviate the management burdens of RBAC systems, it is necessary to make the cluster–permission assignments relationship as sparse as possible. Based on Definition 2, we choose the user cluster that involves the maximum number of users and regard it and its whole permission set as the candidate cluster and role, respectively. The construction process is presented in Algorithm 2.

In Algorithm 2, we first create and initialize several variables in Lines 1-4, including C', Cand_Roles, CRA, maxcount_users, and cand_cluster. For each cluster in C', we calculate the number of users derived from the cluster hierarchies and then identify the candidate cluster and its maximum number of users in Lines 5-15. Lines 16-18 update the Cand_Roles, CRA, and remove the candidate cluster. Next, for each cluster, we remove the permissions assigned to cand_cluster, which are covered by other clusters, and remove the clusters, when all of the permissions of those clusters have been covered by C' (Lines 19-28).

Algorithm 2. Construction of unconstrained role engineering.

Input: set C of the preprocessed results
Output: immediate cluster set C', candidate role set *Cand_Roles*, and cluster-role assignments *CRA*
1: Create and initialize $C'= C$;
2: Create and initialize *Cand_Roles*= Φ, and *CRA*= Φ;
3: Create and initialize *maxcount_users*=0, which represents the maximum number of users included in a cluster;
4: Create and initialize *cand_cluster*=Φ, which represents the cluster involving the maximum number of users;
5: **for** each c_i in C' **do**
6: **for** each c_j in $C' \setminus \{c_i\}$ **do**
7: **if** $user\_perms(c_i) \subseteq user\_perms(c_j)$ **then**
8: $count\_users(c_i)$+= $count\_users(c_j)$;
9: **end if**
10: **if** *maxcount_users*< $count\_users(c_i)$ **then**
11: *maxcount_users*= $count\_users(c_i)$;
12: *cand_cluster*= c_i;
13: **end if**
14: **end for**
15: **end for**
16: *Cand_Roles*= *Cand_Roles* ∪ {*user_perms*(*cand_cluster*)};
17: *CRA*= *CRA* ∪ {(*cand_cluster*, *user_perms*(*cand_cluster*))};
18: C'=$C' \setminus$ {*cand_cluster*};
19: **for** (each c_k in C') ∧ ($user\_perms(cand\_cluster) \subseteq user\_perms(c_k)$) **do**
20: **for** each p in *user_perms*(*cand_cluster*) **do**
21: $user\_perms(c_k)$=$user\_perms(c_k) \setminus \{p\}$;
22: $count\_unperms(c)$- -;
23: **end for**
24: *CRA*= *CRA* ∪ {c_k, *user_perms*(*cand_cluster*)};
25: **if** $count\_unperms(c_k)$= =0 **then**
26: C'=$C' \setminus \{c_k\}$;
27: **end if**
28: **end for**

3.3 Role-Engineering Optimization with UCCR

To further satisfy the constraint requirements for user clusters in RBAC systems, while enhancing the interpretation of the mining results, the UCCR should be taken into consideration in the role engineering. Specifically, the unconstrained mining results are checked to identify whether they violate the given cardinality constraints on roles. If there are no constraint violations, they are regarded as efficient solutions. First, we define the role-engineering optimization problem as follows.

Definition 3. *(Role-engineering optimization problem) Given a cluster-permission assignment matrix CPA, and a particular constraint threshold MRC_{user}, find a set Optim_Roles of roles and the corresponding decomposed matrices CRA and RPA, such that the CRA and RPA are consistent with the CPA, the number of roles assigned to any user is less than or equal to MRC_{user}, and the number of the optimal roles is minimized. This process can be formalized as follows:*

$$\begin{cases} \min |Optim\_Roles| \\ CRA \otimes RPA = CPA \\ |user\_roles(c) \cap Optim\_Roles| \leq MRC_{user}. \forall c \in C. \end{cases} \tag{4}$$

According to the mining results from Algorithm 2, we present the optimization process in Algorithm 3.

In Algorithm 3, the unconstrained mining results, Cand_Roles, CRA and C', are considered as inputs, and we output the optimized results, including Optim_Roles and the updated CRA. We make some initializations in the first lines. Next, Lines 5 and 6 indicate that, if the number of roles in c_i equals MRC_{user}–1, and there exist other permissions that are uncovered in c_i, then a new role is generated. If another cluster, c_j, includes role temp, while satisfying the cardinality constraint, then Optim_Roles and CRA are updated in Lines 7-9; otherwise, only the role, $\{user\_roles(c_i) \cup temp\}$, is assigned to c_i in Lines 10-12. In addition, we call Algorithm 2 again in order to revise C' and the relationship of its assignments in Line 15.

Algorithm 3. Role-engineering optimization.

Input: unconstrained mining results *Cand_Roles*, *CRA*, and C' and the threshold MRC_{user}
Output: constrained role set *Optim_Roles* and updated *CRA*
1: Create and initialize *Optim_Roles*= *Cand_Roles*;
2: Create temporary role *temp*;
3: **while** $C' \neq \Phi$ **do**
4: $\forall c \in C'$: *user_roles*(c)={$r \in$ *Optim_Roles* | $(c,r) \in CRA$};
5: **if** $\exists c_i \in C'$: (| *user_roles*(c_i) |== MRC_{user} - 1) $\wedge$ (*count_unperms*(c_i) != 0) **then**
6: *temp*= *user_perms*(c_i);
7: **if** $\exists c_j \in \{C' \setminus c_i\}$: (| *user_roles*($c_j$) | < MRC_{user}) $\wedge$ (*temp* $\subseteq$ *user_perms*(c_j)) **then**
8: *Optim_Roles*= *Optim_Roles* $\cup$ {*temp*};
9: *CRA*= *CRA* $\cup$ {(c_i,*temp*),(c_j,*temp*)};
10: **else**
11: *Optim_Roles*= *Optim_Roles* $\cup$ {*user_roles*(c_i) $\cup$ *temp*};
12: $CRA = (CRA \setminus \cup_{r \in user\_roles(c_i)} \{(c_i, r)\}) \cup \{c_i, user\_roles(c_i) \cup temp\}$;
13: **end if**
14: **end if**
15: call Algorithm 2;
16: **end while**

3.4 Running Examples

In this subsection, we present an illustrative example to demonstrate the effectiveness of the REO_UCCR in the following.

Example 1. *Consider the matrix UPA of an original assignment, which is comprised of 15 users and 4 permissions, as shown in Table 1, and $MRC_{user} = 2$.*

In the preprocessing stage, we first identify the distance matrix, D_r, for the original matrix, as shown in Table 2, where both the rows and columns correspond to the row vectors, and the values of the cells are the Hamming distances between any two rows. It is seen that $D_r[2][4] == D_r[2][5] == D_r[2][13] == D_r[2][14] == 0$, $D_r[3][8] == D_r[3][9] == 0$, $D_r[6][7] == D_r[6][15] == 0$, and $D_r[10][11] == 0$. According to Algorithm 1, the same (or similar) row vectors are clustered by choosing the minimal distances and swapping different rows. Similarly, we can also cluster the same (or similar) column vectors in order to further rearrange the matrix, and the result is shown in Table 3. Subsequently, in the generation of user clusters, we can identify the cluster-permission assignments, CPA, and cluster tuples, as shown in Tables 4 and 5, respectively. It is apparent that the compressed CPA is easier to use than the original assignments UPA, which can reduce the mining scale. Indeed, it is convenient and feasible to analyze and handle the compressed cluster set.

Then, we repeatedly call Algorithm 2 and Algorithm 3 in the role mining and optimization stages, and Table 6 presents the optimization process, which does not stop until C' is empty. It is seen from the table that, after the second step in the role mining, only user cluster c_1 remains in C', while the permissions, p_1 and p_2, which are uncovered, are included. Obviously, the number of roles in c_1 is less than 2. However, if we regard $\{p_1, p_2\}$ as a candidate role, then it can be only assigned to c_1 and is not associated with any other cluster, which increases the engineering cost. Thus, we remove the role $\{p_4\}$ that has been assigned to c_1 and assign a new role $\{p_1, p_2, p_4\}$ to c_1 in order to reduce the management burden. In addition, we load and implement our method in the regular mining tool, RMiner [10], as shown in Figure 2, and compare its performance with that of the existing mining methods, as shown in Table 7. It is seen from the table that, however, the user clusters, c_1 and c_2, using the enumeration method [19], violate the given constraint.

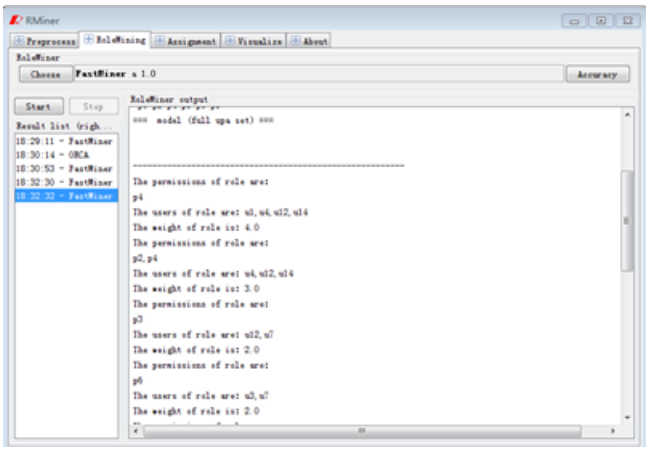


Figure 2: The mining tool, RMiner

Table 1: Original matrix UPA

	p_1	p_2	p_3	p_4
u_1	0	0	0	0
u_2	1	1	0	1
u_3	0	1	1	0
u_4	1	1	0	1
u_5	1	1	0	1
u_6	0	1	1	1
u_7	0	1	1	1
u_8	0	1	1	0
u_9	0	1	1	0
u_{10}	0	0	0	1
u_{11}	0	0	0	1
u_{12}	0	0	0	0
u_{13}	1	1	0	1
u_{14}	1	1	0	1
u_{15}	0	1	1	1

4 Experiments and Analyses

In this section, we perform two groups of experiments with REO_UCCR. The first group of experiments is used to evaluate its performance with respect to different values of constraints. The second group is to compare its performance with the existing methods. We consider four

Table 2: Distance matrix Dr

	UPA' [2]	UPA' [3]	UPA' [4]	UPA' [5]	UPA' [6]	UPA' [7]	UPA' [8]	UPA' [9]	UPA' [10]	UPA' [11]	UPA' [13]	UPA' [14]	UPA' [15]
UPA'[2]	0	3	0	0	2	2	3	3	2	2	0	0	2
UPA'[3]	3	0	3	3	1	1	0	0	3	3	3	3	1
UPA'[4]	0	3	0	0	2	2	3	3	2	2	0	0	2
UPA'[5]	0	3	0	0	2	2	3	3	2	2	0	0	2
UPA'[6]	2	1	2	2	0	0	1	1	2	2	2	2	0
UPA'[7]	2	1	2	2	0	0	1	1	2	2	2	2	0
UPA'[8]	3	0	3	3	1	1	0	0	3	3	3	3	1
UPA'[9]	3	0	3	3	1	1	0	0	3	3	3	3	1
UPA'[10]	2	3	2	2	1	2	3	3	0	0	2	2	2
UPA'[11]	2	3	2	2	2	2	3	3	0	0	2	2	2
UPA'[13]	0	3	0	0	2	2	3	3	2	2	0	0	2
UPA'[14]	0	3	0	0	2	2	3	3	2	2	0	0	2
UPA'[15]	2	1	2	2	0	0	1	1	2	2	2	2	0

Table 3: Rearranged matrix UPA'

	p_1	p_2	p_4	p_3
u_2	1	1	1	0
u_4	1	1	1	0
u_5	1	1	1	0
u_{13}	1	1	1	0
u_{14}	1	1	1	0
u_6	0	1	1	1
u_7	0	1	1	1
u_{15}	0	1	1	1
u_3	0	1	0	1
u_8	0	1	0	1
u_9	0	1	0	1
u_{10}	0	0	1	0
u_{11}	0	0	1	0

real datasets from the work in [12] and adopt the mining tool, RMiner, to evaluate the performance of the unconstrained role mining. The original datasets also include the density of each dataset, the number of the candidate role sets, Cand_Roles, and the execution time, as shown in Table 8. All experiments are implemented on a standard desktop PC, with an Intel i5–7400 CPU, 4 GB RAM, and 160 GB hard disks, running a 64-bit Windows 7 operating system. All simulations are compiled and executed under the Java environment.

Table 4: CPA

	p_1	p_2	p_4	p_3
c_1	1	1	1	0
c_2	0	1	1	1
c_3	0	1	0	1
c_4	0	0	1	0

4.1 Performance Evaluations of the REO_UCCR

To evaluate the effectiveness of our method in the role optimization process, we consider the number of the optimized roles, Optim_Roles, and the size of the assignments, URA, as measures.

Taking the dataset, Firewall 1, as an example for implementing the experiments, the value of the constraint, MRC_{user}, varies from 2 to 8, with a step of 2. We implement the experiments 5 times and take their average values. The results are shown in Figures 3 and 4. In Figure 3, the lateral axis represents MRC_{user}, and the vertical axis represents the number of Optim_Roles. In Figure 4, the lateral axis represents MRC_{user}, and the vertical axis represents the size of the URA.

Figure 3 shows that the number of roles tends to decrease slightly as the value of MRC_{user} increases. When the number of roles is considered as a unique measure, the value of MRC_{user} is greater, and the redundancies of the mining results are fewer. Figure 4 shows that, however, the size of the URA tends to increase remarkably as MRC_{user} increases, which is contrary to the variation tendency in Figure 3. This is because the greater the value of MRC_{user}, the weaker the constraint, and the greater the number of roles assigned to users. The value of $|URA|$ is up to 1516, particularly when MRC_{user} equals 8, which increases the burdens of the system management. On the contrary, the smaller the MRC_{user}, the stronger the constraint will be. Furthermore, the results of the same experiments using the datasets, Firewall 2, Domino, and Healthcare are shown in Figures 5 to 10. It is observed from the tables that, for Firewall 2, the number of roles decreases from 11 to 10 with the increasing value of MRC_{user}, while the value of $|URA|$ is up to 877 when MRC_{user} equals 8. For Domino, the number

Table 5: Four tuples of clusters

User Cluster (c)	Original Permissions (user_permissions (c))	Original User Number (count_users (c))	Uncovered Permission Number (count_unperms (c))
$c_1 = \{u_2, u_4, u_5, u_{13}, u_{14}\}$	$\{p_1, p_2, p_4\}$	5	3
$c_2 = \{u_6, u_7, u_{15}\}$	$\{p_2, p_3, p_4\}$	3	3
$c_3 = \{u_3, u_8, u_9\}$	$\{p_2, p_3\}$	3	2
$c_4 = \{u_{10}, u_{11}\}$	$\{p_4\}$	2	1

Table 6: The optimization process

Step	Optim_Roles	CRA	Updated C'	count_unperms(c)
1	$\{\{p_4\}\}$	$\{(c_4, \{p_4\}), (c_1, \{p_4\}), (c_2, \{p_4\})\}$	$\{c_1, c_2, c_3\}$	$\{p_1, p_2, p_3\}$
2	$\{\{p_4\}, \{p_2, p_3\}\}$	$\{(c_4, \{p_4\}), (c_1, \{p_4\}), (c_2, \{p_4\}), (c_2, \{p_2, p_3\}), (c_3, \{p_2, p_3\})\}$	$\{c_1\}$	$\{p_1, p_2\}$
3 (finish)	$\{\{p_4\}, \{p_2, p_3\}, \{p_1, p_2, p_4\}\}$	$\{(c_4, \{p_4\}), (c_2, \{p_4\}), (c_2, \{p_2, p_3\}), (c_3, \{p_2, p_3\}), (c_1, \{p_1, p_2, p_4\})\}$	ϕ	ϕ

Table 7: Comparison of mining results

User Cluster	Enumeration Method [19]	Blundo [3]	REO_UCCR
c_1	$\{p_1, p_2, p_4\}, \{p_2, p_4\}, \{p_2\}, \{p_4\}$	$\{p_4\}, \{p_1, p_2\}$	$\{p_1, p_2, p_4\}$
c_2	$\{p_2, p_3, p_4\}, \{p_2, p_3\}, \{p_2, p_4\}, \{p_2\}, \{p_4\}$	$\{p_4\}, \{p_2, p_3\}$	$\{p_4\}, \{p_2, p_3\}$
c_3	$\{p_2, p_3\}, \{p_2\}$	$\{p_2, p_3\}$	$\{p_2, p_3\}$
c_4	$\{p_4\}$	$\{p_4\}$	$\{p_4\}$

of roles decreases from 23 to 22 as the value of MRC_{user} increases, while the value of $|URA|$ is up to 169. For Healthcare, the number of roles decreases from 18 to 17 as the value of MRC_{user} increases, while the value of $|URA|$ is up to 143 when MRC_{user} equals 8.

According to the above analyses, we present the optimized results for different datasets when MRC_{user} equals 2, as shown in Table 9. It can be seen that the value of $|URA|$ is less than that of the enumeration method. Therefore, the optimized results using our method not only satisfy the security requirements, but also alleviate the burdens of the system management.

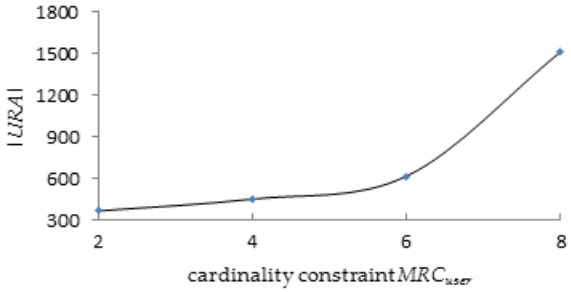


Figure 4: Results of the user-role assignments using Firewall 1

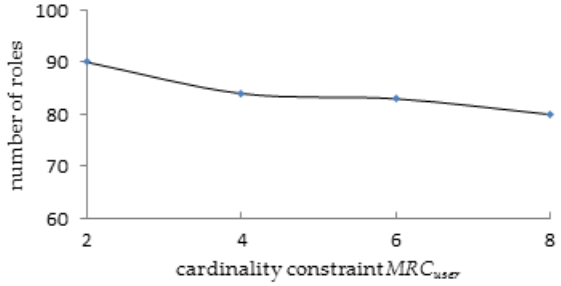


Figure 3: Results of the optimized roles using Firewall 1

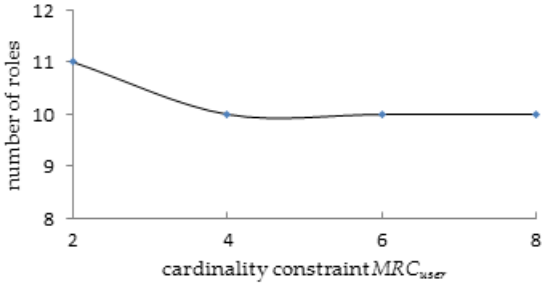


Figure 5: Results of the optimized roles using Firewall 2

Table 8: Original datasets

Dataset	|U|	|P|	|UPA|	Density	|Cand_Roles|	Execution Time(s)
Domino	79	231	730	4%	20	0.01
Healthcare	46	46	1,486	70%	15	0.01
Firewall 1	365	709	31,951	12.3%	69	0.11
Firewall 2	325	590	36,428	19%	10	0.15

Table 9: Comparison of the mining results

	Enumeration Method [19]		REO_UCCR	
Dataset	$\|R\|$	$\|URA\|$	|Optim_Roles|	$\|URA\|$
Domino	20	110	23	79
Healthcare	15	106	18	46
Firewall 1	69	874	90	36
5 Firewall 2	10	434	11	325

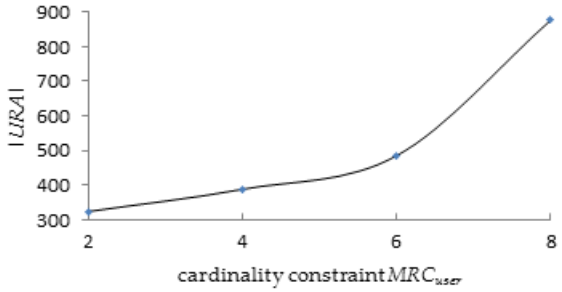


Figure 6: Results of the user-role assignments using Firewall 2

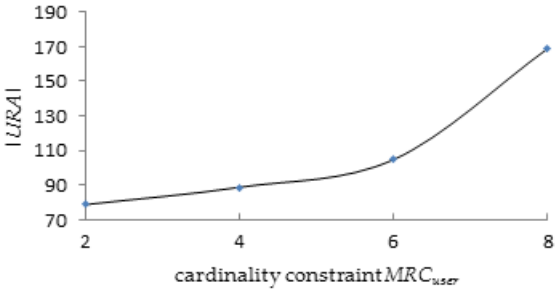


Figure 8: Results of the user-role assignments using Domino

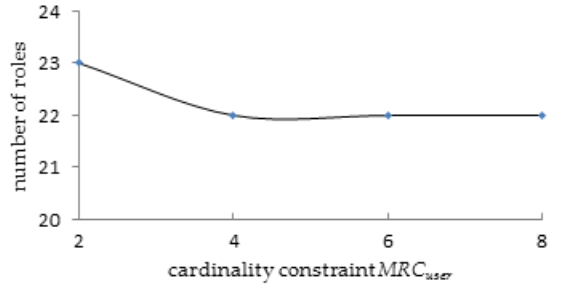


Figure 7: Results of the optimized roles using Domino

4.2 Performance Comparisons with the Existing Methods

To evaluate the efficiency of the REO_UCCR, we implement experiments with the datasets, Domino and Healthcare, as shown in Table 8, and compare its performance with the results of the representative methods, RPA and CPA [9], which are shown in Figures 11 and 12, respectively, where the lateral axis represents MRC_{user}, and the vertical axis represents the number of the optimized roles.

It can be observed, from Figure 11, that the number of roles decreases as MRC_{user} increases for the REO_UCCR, which tends to be stable as MRC_{user} increases to a certain value. Specifically, the number of roles does not obviously vary and remains close to 20 when the value of MRC_{user} exceeds 8. A further observation is that the number of roles first varies slightly and then increases significantly as MRC_{user} decreases. This is because the greater the value of MRC_{user}, the weaker the constraint, and the more roles assigned to any user. In other words, with a greater value of MRC_{user}, regular roles are more applicable and can be utilized more fre-

quently. Thus, fewer irregular roles need to be created, and the number of roles does not vary considerably. For the RPA and CPA, however, the number of roles tends to increase as MRC_{user} increases from 1 to 4. This is because the Domino dataset contains exclusive permissions and produces exclusive roles in the presence of constraints. As shown in the figure, the maximum number of roles is close to 30 when MRC_{user} equals 4, while the minimum number of roles is 23 when MRC_{user} equals 1. Therefore, our method outperforms the RPA and CPA for the dataset, Domino. Similarly, it can be observed, from Figure 12, that the number of roles also decreases as MRC_{user} increases for the REO_UCCR, which tends to be stable and remains close to 15 when MRC_{user} increases to a certain value. However, the variations of the results of both the RPA and CPA are simple. The RPA generates 15 roles that remain unchanged when MRC_{user} exceeds 1, while the number of roles is 18 when MRC_{user} equals 1; and the CPA generates 18 roles that remain unchanged as MRC_{user} varies. Therefore, our method only outperforms the CPA for the dataset, Healthcare.

4.3 Advantages and Shortcomings of the REO_UCCR

From the above analyses, we find the REO_UCCR has the following main advantages:

1) In the initial role-engineering construction, it can reduce the mining scales and alleviate the burdens of system management by using the Hamming distance technique and cluster tuples;

2) In the role-engineering optimization, it can restrict the maximum number of roles assigned to any user and ensure system security by reconstructing a constrained RBAC system, based on the previous mining results.

Compared with the existing studies, the security characteristics of the proposed method are shown in Table 10, where a tick V indicates that the characteristic is available. It can be seen from table that our proposed method still has shortcomings: It does not satisfy the other three cardinality constraints or the separation of duty constraint.

5 Conclusions

A novel role-engineering method, called REO_UCCR, was proposed in this paper. We first converted the basic role mining problem into a clustering problem based on the Hamming distance technique and four tuples of clusters and implemented role mining, while constructing an unconstrained role engineering system. Then, we implemented the role optimization algorithm to reconstruct a constrained RBAC system in order to verify whether the given cardinality constraints can be satisfied in the constructed system. The experiments demonstrated that the

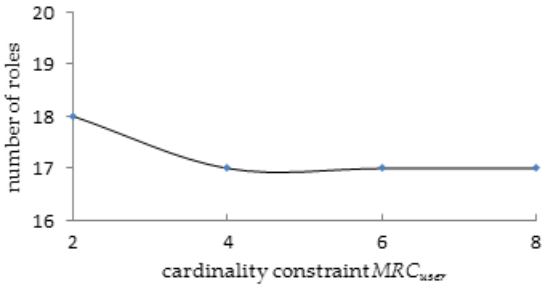


Figure 9: Results of the optimized roles using Healthcare

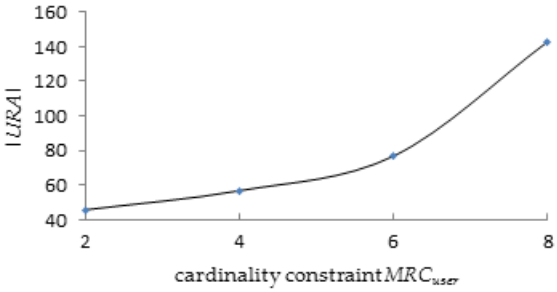


Figure 10: Results of the user-role assignments using Healthcare

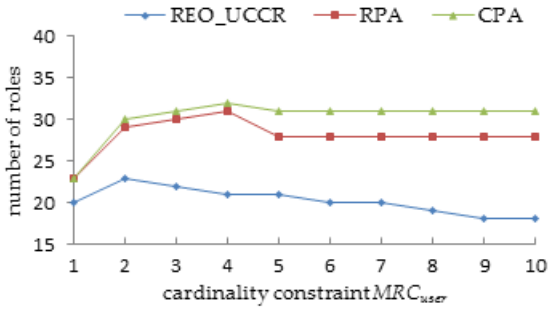


Figure 11: Performance comparison using Domino

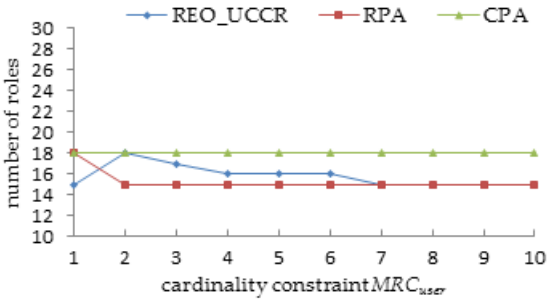


Figure 12: Performance comparison using Healthcare

Table 10: Comparison of security characteristics

Characteristic	Blundo *et al.* [3]	John *et al.* [9]	Harika *et al.* [8]	Wang *et al.* [21]	Blundo *et al.* [4]	Sun *et al.* [13]	Proposed Method
UCCR		V	V		V		V
Other cardinality constraints	V		V	V	V		
Separation of duty constraint						V	
Reducing the mining scales						V	V

proposed method not only alleviates management burdens, but also ensures system security. However, a few interesting issues remain to be solved. To further enhance the interpretability of mining results, one issue for future work is how to implement the other cardinality constraints for role-engineering optimization.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China (61501393), the Natural Science Foundation of Henan Province of China (182300410145, 182102210132), and the Foundation of Henan Educational Committee, under Contract No. 20B520031.

References

[1] W. Bai, Z. Pan, S. Guo, and Z. Chen, "RMMDI: A novel framework for role mining based on the multi-domain information," *Security and Communication Network*, 2019. (`https://doi.org/10.1155/2019/8085303`)

[2] 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.

[3] C. Blundo, and S. Cimato, "Constrained role mining," in *Proceedings of the Security and Trust Management-8th International Workshop*, pp. 289–304, Sep. 2012.

[4] C. Blundo, S. Cimato, and L. Siniscalchi, "Managing constraints in role based access control," *IEEE Access*, vol. 8, pp. 140497–140511, 2020.

[5] J. Ernvall, J. Katajainen, and M. Penttonen, "NP-completeness of the Hamming salesman problem," *BIT Numerical Mathematics*, vol. 25, no. 1, pp. 289–292, 1985.

[6] N. Gal-Oz, Y. Gonen, and E. Gudes, "Mining meaningful and rare roles from web application usage patterns," *Computers & Security*, vol. 82, pp. 296–313, 2019.

[7] M. Ghafoorian, D. Abbasinezhad-Mood, and H. Shakeri, "A thorough trust and reputation based RBAC model for secure data storage in the cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 4, pp. 778–788, 2018.

[8] P. Harika, M. Nagajyothi, J. C. John, S. Sural, J. Vaidya, and V. Atluri, "Meeting cardinality constraints in role mining," *IEEE Transactions on Dependable and Secure Computing*, vol. 12, no. 1,pp. 71–84, 2015.

[9] J. C. John, S. Sural, V. Atluri, and J. Vaidya, "Role mining under role-usage cardinality constraint," in *Proceedings of the 27th IFIP TC 11 Information Security and Privacy Conference on Information Security and Privacy Research*, pp. 150–161, June 2012.

[10] R. Li, H. Li, W. Wei, X. Ma, and X. Gu, "RMiner: A tool set for role mining," in *Proceedings of the 18th ACM Symposium on Access Control Models and Technologies*, pp. 193–196, June 2013.

[11] L. Liu, Z. Cao, and C. Mao, "A note on one outsourcing scheme for big data access control in cloud," *International Journal of Electronics and Information Engineering*, vol. 9, no. 1, pp. 29–35, 2018.

[12] B. Mitra, S. Sural, J. Vaidya, and V. Atluri, "A survey of role mining," *ACM Computing Surveys*, vol. 48, no. 4, pp. 1–37, 2016.

[13] W. Sun, S. Wei, H. Guo, and H. Liu, "Role-mining optimization with separation-of-duty constraints and security detections for authorizations," *Future Internet*, vol. 11, no. 9, pp. 201, 2019.

[14] S. D. Stoller, and T. Bui, "Mining hierarchical temporal roles with multiple metrics," *Journal of Computer Security*, vol. 26, no. 1, pp. 121–142, 2018.

[15] 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, pp. 164, 2020.

[16] A. Thakare, E. Lee, A. Kumar, V. B. Nikam, and Y. G. Kim, "PARBAC: Priority-attribute-based RBAC model for azure IoT cloud," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2890–2900, 2020.

[17] Y. Tian, Y. Peng, G. Gao, and X. Peng, "Role-based access control for body area networks using attribute-based encryption in cloud storage," *International Journal of Network Security*, vol. 19, no. 5, pp. 720–726, 2017.

[18] J. D. Ultra, and S. Pancho-Festin, "A simple model of separation of duty for access control models," *Computers & Security*, vol. 68, pp. 69–80, 2017.

[19] J. Vaidya, V. Atluri, and Q. Guo, "The role mining problem: A formal perspective," *ACM Transactions on Information and System Security*, vol. 13, no. 3, pp. 1–31, 2010.

[20] J. Vaidya, V. Atluri, and Q. Guo, "The role mining problem: Finding a minimal descriptive set of roles,"

in *Proceedings of the 12th ACM Symposium on Access Control Models and Technologies*, pp. 175–184, June 2007.

[21] J. Wang, J. Dong, and Y. Tan, "Role mining algorithms satisfied the permission cardinality constraint," *International Journal of Network Security*, vol. 22, no. 3, pp. 371–380, 2020.

[22] J. Wang, J. Liu, and H. Zhang, "Access control based resource allocation in cloud computing environment," *International Journal of Network Security*, vol. 19, no. 2, pp. 236–243, 2017.

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 working in the Center of Network Information and Computing, Xinyang Normal University. His current research interests include access control and system security.

Xiaoya Yuan received her B.S. and M.S. degrees from the School of Computer and Information Technology, Xinyang Normal University, China, in 2003 and 2011, respectively. She is currently working in the Center of Network Information and Computing, Xinyang Normal University. Her current research interests include image processing and pattern recognition.

Hui Su received his B.S. and M.S. degrees from the School of Computer and Information Technology, Xinyang Normal University, China, in 1992 and 2008, respectively. He is currently working in the Center of Network Information and Computing, Xinyang Normal University. His current research interests include image processing and pattern recognition.

An Improved FH-CP-ABE Scheme with Flexible Attribute Management and Efficient User Decryption

Junliang Zhang and Weiyou Zhang
(Corresponding author: Junliang Zhang)
School of Mechatronic Engineering, Dezhou University
Dezhou, Shandong, China
(Email: zhangdzu@126.com)
(Received June 15, 2020; Revised and Accepted Dec. 10, 2020; First Online Aug. 14, 2021)

Abstract

File hierarchy ciphertext-policy attribute-based encryption (FH-CP-ABE) is a promising method to support data sharing with a multilevel hierarchy structure. This paper proposes an improved FH-CP-ABE scheme to support dynamic privilege management and efficient user decryption. The construction introduces a match-control mechanism between decryption keys and ciphertexts. It achieves flexible and arbitrary attribute alteration and privilege management. Additionally, this improvement increases user performance by delegating most of the decryption operations to the cloud. Finally, the instance and analysis show that this improvement is secure and practical for mobile devices in the hybrid cloud.

Keywords: Access Control; CP-ABE; File Hierarchy; Hybrid Cloud Computing; Privilege Management

1 Introduction

Ciphertext-policy attribute-based encryption (CP-ABE) [2, 4, 9] is a promising mechanism to implement fine-grained data access control on encrypted data. Recently, Wang *et al.* [21] proposed a traditional file hierarchy ciphertext-policy attribute-based encryption (FH-CP-ABE) scheme in cloud computing. They constructed an efficient solution to share data files with the characteristic of multilevel hierarchy, particularly in the area of finance, healthcare, government and military.

However, attribute authorization or revocation (*i.e.*, attribute alteration) is a practical and crucial function in organization management systems. In addition, user performance (*i.e.*, decryption calculation) is also a practical and critical challenge in mobile wireless networks. Focusing on improving the practicability of CP-ABE solutions, this paper introduces a variant of the traditional FH-CP-ABE scheme with flexible system management and efficient user performance.

1.1 Contribution

This paper proposes an improved FH-CP-ABE scheme to introduce the methods of flexible system management and efficient user performance. The main contributions of this study are described below:

- This proposal presents a version-based attribute management. This contribution helps to achieve flexible system privilege management. It allows arbitrary attribute alteration (authorization or revocation) without updating user keys or existing ciphertexts. It additionally provides both forward and backward security for access strategies. It is achieved by embedded a version tag like a timestamp in decryption keys and ciphertexts, respectively.

- It presents a proxy-based decryption. This contribution helps to achieve efficient performance of user decryption. It reduces most of the decryption calculation burden for users. Particularly for mobile users, there are a few calculations for mobile devices. It is profited from version-based attribute management. Most of the calculations are executed by the proxy (the attribute manager) who is responsible for matching the version tag between user keys and ciphertexts.

Based upon the above contributions, this study further designs a specific implementation environment for the proposed scheme. Finally, the analysis shows the security guarantee, system flexibility and user performance for mobile data sharing in hybrid cloud.

1.2 Organization

The remaining of this paper is organized as follows. Related works are introduced in Section 2. Section 3 presents the construction of the improved FH-CP-ABE. A specific application system for the proposed scheme is described

in Section 4. In addition, Section 5 analyzes the security and practicability of the proposal. Finally, Section 6 concludes this paper.

2 Related Work

Attribute-based encryption [6–8, 12, 13, 23, 24, 26, 27, 29, 30] has been studied and researched for decades since Sahai and Waters' fuzzy identity-based encryption [1]. Ciphertext-policy [3, 9, 26] and key-policy [17, 19, 25] are two directions that attribute-based encryption develops. Particularly in CP-ABE schemes, decryption key is generated by a set of authorized attributes and ciphertext is produced by integrating an access policy.

In order to efficiently share the hierarchy files in cloud computing, Wang *et al.* proposed a variant of CP-ABE scheme [20, 21]. In their proposal, a layered model of access structure is constructed to solve the problem of multiple hierarchy file sharing. Files are encrypted with one integrated access structure. The security of their proposal is formally proved to resist chosen plaintext attacks (CPA) under the Decisional Bilinear Diffie-Hellman (DBDH) assumption, and the storage and computation cost are saved.

However, the attribute management is a critical problem for a practical application. There are many solutions to support attribute revocation in ABE schemes [5,11,16, 18,22,28]. Most solutions are based upon the methods of key updating or ciphertext updating. Therefore, key updating makes interaction difficulties for most of common users, and ciphertext updating brings huge burden for system resource cost. Moreover, the computation cost is another pivotal issue for users. Although there exist some methods to outsource the decryption to clouds [10,14,15], the computation burden is still heavy for mobile devices and cannot be directly applied in other variant of ABE schemes. Specially due to the limitation of computation and energy resource, mobile users cannot load too many computation tasks. Thus proxy-based decryption should be integrated in CP-ABE schemes to offload the burden of client computation.

In order to overcome the above difficulties, this paper improves the traditional FH-CP-ABE scheme [21] to enhance the practicability and efficiency.

3 Proposal

This section firstly introduces some preliminaries used in this paper, including innovative idea, model and definitions. Then the improved file hierarchy ciphertext-policy attribute-based encryption scheme is described in details.

3.1 Preliminary

This part shows some preliminaries used in this paper. More instructions can consult the references of original FH-CP-ABE [21] and CP-ABE [9] scheme.

3.1.1 Innovative Idea

The innovative idea of this proposal is devoted to improve the practicability of the tradtional FH-CP-ABE scheme [21], and the detailed measures are illustrated below:

Arbitrary attribute alteration. In order to achieve flexible privilege management, this proposal introduces the concept of version tag like timestamp for access control. The version tag is embedded in decryption keys and ciphertexts, respectively. The version interception between ciphertexts and decryption keys decides the access control. Specifically, encrypted files can be correctly decrypted if there is a version match between decryption keys and ciphertexts. As a result, version control needs a split of decryption key tuple: a control key, version keys and a secret key. The control key plays a role of pulling the decryption key version to the location of ciphertext version on the version axis. If corresponding attributes are valid, the version keys can be privileged to ciphertext version. In addition, the version control process is executed by the attribute authority automatically, *i.e.*, by a proxy (maybe a server, a private cloud and so on) according to the access control list. Therefore, users' attributes can be arbitrary changed without updating users' decryption keys or existing ciphertexts.

Efficient decryption performance. Benefitting from the authority's version control on the privilege management, this construction outsources partial decryption to the attribute authority. Most of the decryption calculation are executed by the attribute authority. The version control and partial decryption are synchronously processed at the same time. The user decryption only needs the last secret key and takes very few calculations without sacrificing data security. It largely increases the decryption performance for users. It is more important for mobile users with limited computation and energy resources.

3.1.2 Definition

Definition 1. Construction model. *The FH-CP-ABE scheme consists of four phases: Setup, KeyGen, Encrypt and Decrypt. It is described below:*

Setup $(PK, MSK) \leftarrow Setup(1^{\kappa})$. *The probabilistic operation takes a security parameter* κ *as input and outputs public keys* PK *and a master secret key* MSK.

KeyGen $DK \leftarrow KeyGen(PK, MSK, S)$. *It inputs* PK, MSK *and a set of attributes* S, *and creates a decryption key tuple* DK.

Encrypt $CT \leftarrow Encrypt(PK, ck, \mathcal{T})$. *The operation inputs* PK, $ck = \{ck_1, ..., ck_k\}$ *and a hierarchical access tree* $\mathcal{T}$. *It creates an integrated ciphertext of content keys* CT.

Decrypt $ck_i(i \in [1,k]) \leftarrow Decrypt(PK, CT, DK)$. *The algorithm inputs PK, CT with an integrated access structure $\mathcal{T}$. DK described by a set of attributes S. If S matches part of $\mathcal{T}$, some content keys $ck_i(i \in [1,k])$ can be decrypted. If it matches the whole $\mathcal{T}$, all the content keys ck can be decrypted.*

In addition, the content key ck is used for symmetric encryption. The message $M = \{m_1, m_2, ..., m_k\}$ is encrypted by $ck = \{ck_1, ck_2, ..., ck_k\}$, i.e., *$m_i = D_{ck_i}(E_{ck_i}(m_i)), i \in [1,k]$, where E and D are designated symmetric encryption and decryption algorithm, respectively.*

Definition 2. Bilinear maps. *Let $\mathbb{G}_0$ and $\mathbb{G}_T$ be two groups of prime order p, g is the generator of $\mathbb{G}_0$. $e : \mathbb{G}_0 \times \mathbb{G}_0 \mapsto \mathbb{G}_T$ is a bilinear mapping if it satisfies:*

- *Bilinearity: $\forall u, v \in \mathbb{G}_0, \forall a, b \in \mathbb{Z}_p \Longrightarrow e(u^a, v^b) = e(u, v)^{ab}$.*
- *Non-degeneracy: $\exists u, v \in \mathbb{G}_0 \Longrightarrow e(u, v) \neq 1$.*
- *Computability: $\forall u, v \in \mathbb{G}_0$, the computation $e(u, v)$ is efficient.*

Definition 3. DBDH assumption. *A challenger chooses a group $\mathbb{G}_0$ of prime order p based on the security parameter of system. Let g be a generator of $\mathbb{G}_0$, and $a, b, c \in \mathbb{Z}_p$ be randomly chosen. With (g, g^a, g^b, g^c), the adversary must distinguish a valid tuple $e(g,g)^{abc} \in \mathbb{G}_T$ from a random element $R \in \mathbb{G}_T$.*

An algorithm $\mathcal{B}$ that outputs a guess $\mu \in \{0,1\}$ has advantage ϵ in solving DBDH in $\mathbb{G}_0$ if the following condition satisfies.

$$|Pro[\mathcal{B}(g, g^a, g^b, g^c, T = e(g,g)^{abc}) = 0] - Pro[\mathcal{B}(g, g^a, g^b, g^c, T = R) = 0]| \geq \epsilon$$

DBDH assumption holds if no polynomial algorithm has a non-negligible advantage in solving the DBDH problem.

Definition 4. Version tags. *A system initializes a public version management mechanism. It publishes a benchmark version $v_0 = v$ as the initialized version. Then the version should be updated to a new version $v_{t\pm1}$ from v_t when an attribute is altered (authorized or revoked). As shown in Figure 1, this paper defines $t \in \mathbb{Z}_p$ as the interference between an appointed version v_t and the base version v_0 (benchmark). More intuitively, this research defines $v_t - v_0 = t\Delta v$, where $v_t \in \mathbb{Z}_p$ is an ascending order array, v_0 is the zero element and Δv is the identity element. In the version axis, there is a control key playing a role of vernier. Like a vernier caliper, the vernier can move from one position (tag) to another position (tag) if the privilege on the target tag is authorized. This mechanism provides perfect and flexible access control with forward and backward security (controlled by an honest-but-curious manger,* e.g., *system manager, proxy manger and so on).*

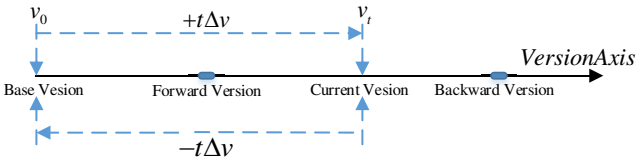


Figure 1: Version tag

Definition 5. Access structure. *Access Structure: Let $\{P_1, P_2, ..., P_n\}$ be a set of $n \in \mathbb{N}^+$ parties. A collection $\mathbb{A} \subseteq 2^{\{P_1, P_2, ..., P_n\}}$ is monotone for $\forall B$ and C, if $B \in \mathbb{A}, B \subseteq C$, then $C \in \mathbb{A}$. An access structure (respectively, monotone access structure) is a collection (respectively, monotone collection) $\mathbb{A}$ of nonempty subsets of $\{P_1, P_2, ..., P_n\}$,* i.e., *$\mathbb{A} \in 2^{\{P_1, P_2, ..., P_n\}} \backslash \{\emptyset\}$. The sets in $\mathbb{A}$ are called authorized sets, and the sets not in $\mathbb{A}$ are called unauthorized sets.*

Definition 6. Hierarchy access tree. *Let $\mathcal{T}$ be a hierarchical tree representing an access structure which is divided into k access levels. Nodes of the tree are denoted as (x, y). The symbol x represents the node's row in $\mathcal{T}$ (from top to bottom), and y represents the node's column in $\mathcal{T}$ (from left to right). In Figure 2, the nodes can be denoted as: $R = (1,1), A = (2,1), B = (2,2), C = (3,1), D = (3,2), E = (4,1), F = (4,2), G = (4,3)$. To facilitate description of the access tree, several functions and terms are defined below:*

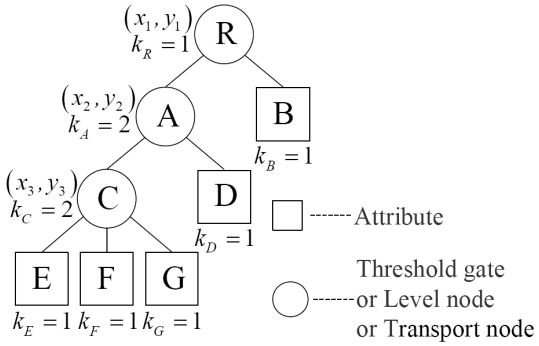


Figure 2: Hierarchy access tree

- *(x, y) denotes a node of tree $\mathcal{T}$. If (x, y) is a leaf node, it denotes an attribute. If (x, y) is a non-leaf node, it denotes a threshold gate, such as* **AND**, **OR**, *n-of $m (n < m)$. For example, the nodes A and E denote a threshold gate and an attribute in Figure 2.*
- *$num_{(x,y)}$ denotes the number of (x, y)'s children in $\mathcal{T}$. For example, $num_R = 2$ in Figure 2.*
- *$k_{(x,y)}$ denotes the threshold value of node (x, y), where $0 < k_{(x,y)} \leq num_{(x,y)}$. When $k_{(x,y)} = 1$ and (x, y) is a non-leaf node, (x, y) is an OR gate. When $k_{(x,y)} = num_{(x,y)}$ and (x, y) is a non-leaf node, it is an AND gate. In particular, if (x, y) is a leaf node, $k_{(x,y)} = 1$. For example, $k_A = 2$ denotes an AND gate in Figure 2.*

- $(x_i, y_i)(i \in [1,k])$ *denotes a level node of* $\mathcal{T}$. *In this work, access tree* $\mathcal{T}$ *is divided into* k *access levels. And the hierarchy of the nodes is sorted in descending order. That is,* (x_1, y_1) *is the highest hierarchy, and* (x_k, y_k) *is the lowest hierarchy. For example,* $(x_2, y_2) = A$ *is the second hierarchy in Figure 2.*

- $parent_{(x,y)}$ *represents the parent of the node* (x, y) *in* $\mathcal{T}$. *For example,* $parent_{(2,1)} = parent(A) = R$ *in Figure 2.*

- *Node* (x, y) *is a transport node if one of the children of* (x, y) *contains at least one threshold gate. For example,* R *and* A *are transport nodes in Figure 2.*

- $TN - CT(x, y)$ *represents a threshold gate set of transport node* (x, y)*'s children in* $\mathcal{T}$. *It is marked as*

$$TN - CT(x, y) = \{child_1, child_2, ...\}.$$

  For example, $TN - CT(A) = \{C\}$ *in Figure 2.*

- $att(x, y)$ *denotes an attribute associated with the leaf node* (x, y) *in Figure 2.*

- $index(x, y)$ *returns an unique value associated with the node* (x, y), *where the value is assigned to* (x, y) *for a given key in an arbitrary manner.*

- $\mathcal{T}_R$ *denotes a tree diagram, where root node of the tree is* R.

- $\mathcal{T}_{(x,y)}$ *denotes the sub-tree of* $\mathcal{T}$ *rooted at the node* (x, y). *If an attribute set* S *satisfies* $\mathcal{T}_{(x,y)}$, *this study denotes it as* $T_{(x,y)}(S) = 1$. $T_{(x,y)}(S)$ *is recursively computed as follows. If* (x, y) *is a non-leaf node,* $T_{(x,y)}(S)$ *returns* 1 *if and only if at least* $k_{(x,y)}$ *children return* 1. *If* (x, y) *is a leaf node, then* $T_{(x,y)}(S)$ *returns* 1 *if and only if* $att(x, y) \in S$.

3.2 Construction

The construction is composed of setup phase, key generation phase, encryption phase and decryption phase. The detailed operations are illustrated below:

3.2.1 $Setup\{1^\kappa\} \mapsto \{PK, MSK\}$

The authority runs the operation which inputs a security parameter κ and chooses random numbers $\alpha, \beta, v, \Delta v \in \mathbb{Z}_p$. It outputs PK and MSK below:

$$\begin{aligned} PK &= \{\mathbb{G}_0, g, e(g,g)^\alpha, h = g^\beta, g^v, g^{\Delta v}\}, \\ MSK &= \{g^\alpha, \beta\}, \end{aligned}$$

where $v_0 = v$.

3.2.2 $KeyGen\{PK, MSK, S\} \mapsto \{DK\}$

The authority executes this key generation algorithm which inputs a set of attributes $S(S \subseteq \tilde{A})$ and creates a decryption key tuple $DK = \{SK, CK, D, \forall j \in S : D_j, D'_j\}$ as follows:

$$\begin{aligned} SK &= sk, \\ CK &= g^{\Delta v r_j}, \\ D &= g^{\alpha/sk} h^r, \\ \forall j \in S : D_j &= g^r H_1(j)^{r_j} g^{v r_j}, \\ D'_j &= h^{r_j}, \end{aligned}$$

where $\tilde{A}$ is the system attribute set, $H_1 : \{0,1\}^* \to \mathbb{G}_0$ is a hash function, $r', r'_j, sk \in \mathbb{Z}_p$ are randomly selected and designated $r = r'/sk, r_j = r'_j/sk$.

Note that the decryption key is composed of three parts SK, CK and $D, \forall j \in S : D_j, D'_j$ mastered by user and manager respectively. In details, SK is belonged to a user who executes the final decryption of FH-CP-ABE. CK and $D, \forall j \in S : D_j, D'_j$ are owned by system/attribute manager who is responsible for privilege and policy management. The manager securely stores users' keys marked with respective attribute version in the access control list as shown in Table 1.

3.2.3 $Encrypt\{PK, M, \mathcal{T}\} \mapsto \{CT\}$

Assume that a data owner shares $k-$hierarchy files

$$M = \{m_1, m_2, ..., m_k\}$$

with k access levels. Every file is encrypted by a specific symmetric encryption algorithm which securely generates corresponding content keys

$$ck = \{ck_1, ck_2, ..., ck_k\}.$$

Then, the above content keys are encrypted as follows:

- Data owner sets level nodes $(x_i, y_i)(i = 1, 2, ..., k)$ in $\mathcal{T}$, and selects k random numbers $s_1, ..., s_k$ in $\mathbb{Z}_p$. Then it computes $\tilde{C}_i$ and C'_i for all $i = 1, 2, ..., k$ as follows:

$$\begin{aligned} \tilde{C}_i &= ck_i e(g,g)^{\alpha s_i}, \\ C'_i &= g^{s_i}. \end{aligned}$$

- Polynomial structure rule: It is same with the traditional description, more details can be referred in [21].

- Beginning from the root node R, data owner sets

$$q_R(0) = q_{(x_1,y_1)}(0) = s_1$$

  and chooses d_R other points of the polynomial q_R to define it completely, where the points are made of two types of nodes. The one are level nodes which

are children of R. The other are remaining nodes randomly selected. For each non-root node (x,y), it sets $q_{(x,y)}(0) = q_{(x_i,y_i)}(0) = s_i$ if (x,y) is a level node. Otherwise, $q_{(x,y)}(0) = q_{parent(x,y)}(index(x,y))$. The other $d_{(x,y)}$ points of $q_{(x,y)}$ are made of the level nodes of the children of (x,y) and the remaining nodes randomly selected. Let Y be the set of leaf nodes in $\mathcal{T}$. Then, data owner computes $C_{(x,y)}$ and $C'_{(x,y)}$for all nodes (x,y) in the set of Y as follows:

$$\begin{aligned} C_{(x,y)} &= h^{q_{(x,y)}(0)}, \\ C'_{(x,y)} &= H_1(att(x,y))^{q_{(x,y)}(0)} g^{v_t q_{(x,y)}(0)}, \end{aligned}$$

where $g^{v_t} = g^{v_0} g^{t\Delta v} = g^{v_0 + t\Delta v}$ and t is current version timestamp.

- In $\mathcal{T}$, let X be the set of transport nodes, and $TN - CT(x,y)$ be the threshold gate set of transport node (x,y)'s children, where

$$TN - CT(x,y) = \{child_1, ..., child_j, ...\}.$$

Then, data owner computes $\hat{C}_{(x,y),j}$ for each node (x,y) in the set of X and all $j = 1, 2, ...$ as follows:

$$\hat{C}_{(x,y),j} = e(g,g)^{\alpha(q_{(x,y)}(0) + q_{child_j}(0))} \cdot H_2(e(g,g)^{\alpha q_{(x,y)}(0)}),$$

where $H_2 : \{0,1\}^* \to \mathbb{G}_T$ is a hash function.

- Data owner outputs the integrated ciphertext CT as below:

$$CT = \{version, \mathcal{T}, \tilde{C}_i, C'_i, C_{(x,y)}, C'_{(x,y)}, \hat{C}_{(x,y),j}\}.$$

3.2.4 $Decrypt\{PK, CT, DK\} \mapsto \{M\}$

Assume that a data user (consumer) needs the public key PK and DK described by S to decrypt CT. Similarly to CP-ABE [9], a recursive operation

$$DecryptNode(CT, DK, (x,y))$$

should be defined firstly. The detailed descriptions of decryption algorithm are illustrated below:

1) If (x,y) is a leaf node, let $j = att(x,y)$ and define $DecryptNode(CT, DK, (x,y))$ as below. If $i \notin S$, then $DecryptNode(CT, DK, (x,y)) = null$. Otherwise, the operation $DecryptNode(CT, DK, (x,y))$ is obtained by following calculations:

$$\begin{aligned} & DecryptNode(CT, DK, (x,y)) \\ &= \frac{e(D_i, C_{(x,y)})}{e(D'_i, C'_{(x,y)})} \\ &= \frac{e(g^r H_1(i)^{r_i} g^{v_t r_i}, h^{q_{(x,y)}(0)})}{e(h^{r_i}, H_1(att(x,y))^{q_{(x,y)}(0)} g^{v_t q_{(x,y)}(0)})} \\ &= \frac{e(g,h)^{r q_{(x,y)}(0)} \cdot e(H_1(i), h)^{r_i q_{(x,y)}(0)} \cdot e(g,h)^{v_t r_i q_{(x,y)}(0)}}{e(h, H_1(att(x,y)))^{r_i q_{(x,y)}(0)} \cdot e(h,g)^{v_t r_i q_{(x,y)}(0)}} \\ &= e(g,h)^{r q_{(x,y)}(0)} \\ &= e(g,g)^{r\beta q_{(x,y)}(0)}. \end{aligned}$$

Note that system manager executes the version-based access control and computes D_i as below:

$$\begin{aligned} g^r H_1(i)^{r_i} g^{v_t r_i} &= g^r H_1(i)^{r_i} g^{v r_i} \cdot (g^{\Delta v r_i})^t \\ &= g^r H_1(i)^{r_i} g^{(v + t\Delta v) r_i} \\ &= g^r H_1(i)^{r_i} g^{v_t r_i}, \end{aligned}$$

where $v_t = v + t\Delta v$, v is the key version in the access control list and v_t is the ciphertext version integrated in the access policy. In the above computation, D_i can be updated to the ciphertext version if and only if the attribute (x,y) is a valid under the condition of version control.

2) If (x,y) is a non-leaf node, $DecryptNode(CT, DK, (x,y))$ is defined as below. For all nodes z that are children of (x,y), it runs $DecryptNode(CT, DK, z)$ and stores the output as F_z. Let $S_{(x,y)}$ be an arbitrary $k_{(x,y)}$-sized child nodes set z, and then $F_z \neq null$. If the set does not exist, $F_z = null$. Otherwise, $F_{(x,y)}$ is computed as below:

$$\begin{aligned} F_{(x,y)} &= \prod_{z \in S_{(x,y)}} F_z^{\Delta_{i,S'_{(x,y)}(0)}} \\ &= \prod_{z \in S_{(x,y)}} (e(g,g)^{r\beta q_z(0)})^{\Delta_{i,S'_{(x,y)}(0)}} \\ &= \prod_{z \in S_{(x,y)}} (e(g,g)^{r\beta q_{(x,y)}(i)})^{\Delta_{i,S'_{(x,y)}(0)}} \\ &= e(g,g)^{r\beta q_{(x,y)}(0)}, \end{aligned}$$

where $S'_{(x,y)} = \{index(z) : z \in S_{(x,y)}\}, i = index(z)$. Then the procedures of decryption algorithm are described below:

- If the attribute set S satisfies part or the whole $\mathcal{T}$, that is S satisfies part or the whole level nodes, $e(g,g)^{r\beta s_i} (i \in [1,k])$ can be obtained by the following recursive operation:

$$\begin{aligned} A_i &= DecryptNode(CT, DK, (x_i, y_i)) \\ &= e(g,g)^{r\beta q_{(x_i,y_i)}(0)} \\ &= e(g,g)^{r\beta s_i} (i \in [1,k]). \end{aligned}$$

- Next, $e(g,g)^{\alpha s_i}$ can be computed by the following calculation:

$$\begin{aligned} F'_i &= \frac{e(C'_i, D)}{A_i} \\ &= \frac{e(g^{s_i}, g^{\alpha/sk} h^r)}{e(g,g)^{r\beta s_i}} \\ &= e(g,g)^{\alpha/sk \cdot s_i} (i \in [1,k]). \end{aligned}$$

- Then F_i can be recovered by calculating:

$$F_i = (F'_i)^{sk} = e(g,g)^{\alpha s_i} (i \in [1,k]). \quad (1)$$

- Based on the hierarchical nodes, if S includes the lower authorization nodes, it can recursively calculate all of the authorization's level nodes with the values of transport nodes $\hat{C}_{(x,y),j}(j = 1, 2, ...)$ by using the following calculation. Therefore,

$$F_{(i+1),j}, ..., F_{k,j}$$

are obtained sequentially. That is, the values

$$e(g,g)^{\alpha s_i}, e(g,g)^{\alpha s_{i+1}}, ..., e(g,g)^{\alpha s_k},$$

are got.

$$\begin{aligned} F_{(i+1),j} &= \frac{\hat{C}_{(x_i,y_i),j}}{F_i \cdot H_2(F_i)} \\ &= \frac{e(g,g)^{\alpha(s_i + q_{child_j}(0))} \cdot H_2(e(g,g)^{\alpha s_i})}{e(g,g)^{\alpha s_i} \cdot H_2(e(g,g)^{\alpha s_i})} \\ &= e(g,g)^{\alpha q_{child_j}(0)} (j = 1, 2, ...). \end{aligned}$$

- Then the corresponding content keys $ck_i, ..., ck_k$ are decrypted by the following calculation:

$$\frac{\tilde{C}_i}{F_i} = \frac{ck_i e(g,g)^{\alpha s_i}}{e(g,g)^{\alpha s_i}} = ck_i (i \in [1, k]).$$

- At last, the authorized files $\{m_i, ..., m_k\}$ are decrypted with $\{ck_i, ..., ck_k\}$ using the corresponding symmetric decryption algorithm.

4 Application

This section introduces a PHR case study to apply this improvement in real applications.

4.1 Scenario

This part introduces the background in medical cloud systems.

4.1.1 System Introduction

Taking medical system as an example, to securely share the personal health record (PHR) information in cloud systems, PHR information M contains patient's personal information m_1 (*e.g.*, name, address, social number and telephone number) and medical information m_2 (*e.g.*, medical history, medical test, diagnosis and treatment record). In the traditional scheme, the access structure is shown in Figure 3. In details, the access structure is set m_1 as: $\mathcal{T}_1$ = {("Cardiology" AND "Researcher") AND "Attending Physician"}. Similarly, m_2 is termed as T_2 = {("Cardiology" AND "Researcher").} The two access structures have hierarchical relationships where the access structure $\mathcal{T}_1$ is the extension of $\mathcal{T}_2$ [20]. The two structures are integrated into one structure $\mathcal{T}$ as shown in Figure 3.

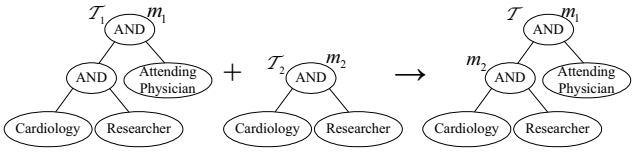


Figure 3: Hierarchy access structure

4.1.2 Participant Interaction

There are three participants involved in medical system application, including user, private cloud and public cloud.

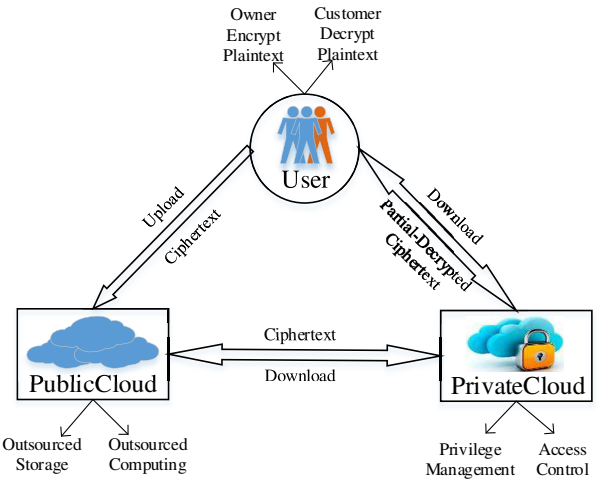


Figure 4: System model

In details, user is consisted of data owner and data customer. Data owner encrypts the plaintext according to the encryption algorithm and uploads the generated ciphertext to the public cloud. Data customer downloads the partially decrypted ciphertext from the private cloud and executes the final decryption calculation.

The public cloud provides services of outsourced storage and computing for its employers, and responds the legal request according to the service protocol. The public cloud can honestly execute the service protocol, but cannot guarantee the data security and organization privacy.

The private cloud is managed by system manager who is also in charge of managing attribute authorization and making access policies. More specifically, there is a system manager, also named attribute authority, who initializes the medical system and employs the public cloud to execute secure hierarchy file sharing for all participants. It executes partial decryption algorithm to support privilege management and access control. Additionally, this manager is responsible for attribute revocation and authorization, and system policy making. It is honest to execute system privilege management and privacy protection, but is still curious about data content.

There are two interactions between users and clouds. Firstly, data owner simply interacts with public cloud to

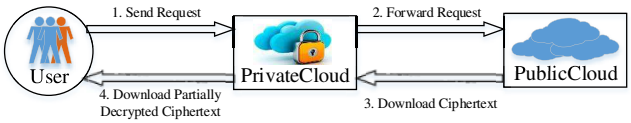


Figure 5: Decryption interaction

upload the ciphertext. Secondly, data customer interacts with public cloud through private cloud to achieve data access control, privilege management and efficiency enhancement.

4.2 Implementation

This section describes an implementation utilizing this proposed improvement. The implementation is applied in hybrid cloud computing for medical PHR application. The main purpose of this implementation is to construct a practical and secure hierarchy PHR sharing system. It aims to achieve flexible privilege management and efficient user decryption. This implementation provides flexible privilege management that supports arbitrary attribute alteration and strategy making for system. It also provides secure proxy-based decryption for user efficiency. The construction is depends upon the private cloud that plays an important role as a proxy for organization management. The proxy helps to achieve privilege management and pre-decryption. Figure 4 shows the participants and their roles in this system. In details, there are two types of user: data owner and data consumer. Data owner owns the plaintext, makes the access policy and uploads the ciphertext to public cloud for sharing. Data consumer requests the ciphertext to obtain the target information if and only if his/her attributes satisfies the access policy. Usually, this paper uses 'user' to replace data owner and data consumer. Additionally, the private cloud is managed by the system authority who is responsible for privilege and access control management. It not only ensures the privacy of personal attribute and organization structure, but also helps to execute partial decryption calculations for users. The public cloud provides services of storage and computing. For security and privacy concern, this research assumes that the public cloud is honest but curious about data content and organization privacy, the private cloud is honest but just curious about data content, and users are only responsible for securing data security. The detailed implementation is consisted of five phases, and described as below.

4.2.1 System Setup

System setup is the first step to initialize a hierarchy file sharing system for medical PHR. This phase is executed by the authority who manages the medical system. The authority runs the algorithm $Setup\{1^\kappa\}$ described in 3.2.1. In details, the authority inputs a security parameter κ. The algorithm randomly generates $\alpha, \beta, v, \Delta v \in \mathbb{Z}_p$, and outputs the public key PK and the master secret key MSK as follows:

$$PK = \{\mathbb{G}_0, g, e(g,g)^\alpha, h = g^\beta, g^v, g^{\Delta v}\},$$

$$MSK = \{g^\alpha, \beta\}.$$

The authority makes PK public as system parameters, and stores MSK privately.

In addition, the authority initializes the system attribute set $\tilde{A}$ and maintains an access control list that records legal users with authorized attributes and version tags. This access control list is stored in the private cloud and shown as in Table 1. Suppose there exist m users and n attributes managed by the authority. The access control list consists of user's identity with corresponding values CK, DK and attribute key tuples $\{D_i, D_i', version\_tag\}, i = 1, 2, ..., n$. Specifically, the $version\_tag$ is marked by different symbols as below.

- $\perp$ denotes a fully authorized attribute and the version could be any value.
- $\top$ denotes an unauthorized attribute and the version must be $NULL$.
- $\vdash$ denotes a partially (forward) authorized attribute and the version could be any value after v_j.
- $\dashv$ also denotes a partially (backward) authorized attribute and the version could be any value before v_k.

As shown in Table 1, it takes 'user_2' and 'user_3' as examples.

- 'user_2' has the privileges of attribute Att_1 with version $v_t, t = 0, 1, 2, ...$, and attribute Att_n with version $v_t, t = j, j+1, j+2, ...$
- 'user_3' does not have the privilege of attribute Att_1, but she/he has the privilege of attribute Att_n with version $v_t, t = 0, 1, ..., k$.

4.2.2 Key Generation

After finishing system setup phase, attribute authority generates the tuple of decryption keys based upon the access control list, and then distributes them to every user. Suppose that a set of attributes S is authorized to a user and S is a subset of system attribute set $\tilde{A}$. The authority utilizes MSK to generates DK for S according to this proposal in Section 3.2.2 as follows:

$$\begin{aligned} SK &= sk, \\ CK &= g^{\Delta v r_j}, \\ D &= g^{\frac{\alpha}{sk}} h^r, \\ \forall j \in S: D_j &= g^r H_1(j)^{r_j} g^{v r_j}, \\ D_j' &= h^{r_j}, \end{aligned}$$

where the authority randomly chooses $r', r_j', sk \in \mathbb{Z}_p$ and calculates $r = \frac{r'}{sk}, r_j = \frac{r_j'}{sk}$. Then the authority securely

Table 1: Access list

ID	CK	DK	Att_1	$\cdots$	Att_n
user_1	CK	DK	$\{D_1, D'_1, \perp\}$	$\cdots$	$\{D_n, D'_n, \perp\}$
user_2	CK	DK	$\{D_1, D'_1, \perp\}$	$\cdots$	$\{D_n, D'_n, (\vdash, v_j)\}$
user_3	CK	DK	$\{\top, \top, NULL\}$	$\cdots$	$\{D_n, D'_n, (\dashv, v_k)\}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
user_m	CK	DK	$\{D_1, D'_1, \perp\}$	$\cdots$	$\{D_n, D'_n, \perp\}$

erases all the random parameters r', r'_j, r, r_j and stores $CK, D, \forall j \in S : D_j, D_{,j}$ in the access control list. According to the regulation, the unauthorized attributes are set:

$$ATT_k = (\top, \top, NULL), \forall k \in (\tilde{A} - S).$$

Next, the authority issues the secret decryption key SK to the user through a secure channel, *e.g.*, face-to-face, encrypted email and so on. Finally, the user stores his/her secret key SK privately.

Note that, this study introduces a passive key generation mechanism that SK is generated and known by the authority. However, there exists an active key generation mechanism that r', r'_j and sk are generated by the user and $r = \frac{r'}{sk}, r_j = \frac{r'_j}{sk}$ are calculated by the user. Then the user securely sends r and r_j to the authority for key generation. This mechanism provides an ideal solution for single key pass that needs standard guidelines.

4.2.3 Message Encryption

When a data owner wants to share a hierarchy file with k access levels, he/she needs to execute the program described in Section 3.2.3. Sketchy description, data owner firstly confirms the hierarchy of shared files and enacts the access structure. Then data owner picks up the content key, and encrypts the hierarchy file with corresponding content key for every file in different levels. The next step encrypts the content keys using this proposed FH-CP-ABE method. Finally, as shown in Figure 4, data owner uploads the content ciphertext, key ciphertext, access structure and version tag to public cloud.

4.2.4 Message Decryption

When a data customer wants to obtain an encrypted file sharing in the public cloud, he/she should send a request to the private cloud. Additionally, the private cloud is managed by the organization manager and responsible for authenticating the identity of access users. Once the private cloud identifies the request is from a legal user, it forwards this request to the public cloud with its signature. The public cloud authenticates and replies the forwarded request. If the signature is legal, the requested files are sent back to the private cloud. The private cloud executes partial decryption operations described in Section 3.2.4. The partial decryption is executed till the calculation F'_i. The rest calculation is executed by data customer. According to the calculation, the private cloud replies the data customer with the partial decrypted files $F_i, \hat{C}_{(x_i,y_i),j}, \tilde{C}_i, i \in S$ and the content ciphertext. Once the data customer receives the reply, he/she starts executing the final decryption from the calculation of F_i using the secret decryption key SK. In other words, the execution of decryption is divided into two parts. The private cloud executes the calculation before the Equation 1, and the data customer starts the calculation from the Equation 1. Finally, the data consumer obtains the plaintext under the help of private cloud.

4.2.5 System Management

Besides the common management of identity and attribute authorization, there are several more practical issues for system management in the application of FH-CP-ABE scheme. For example, different organizations develop various access strategies. Additionally, organization structures and user attributes are privacy for system management.

Firstly, organizations can take full advantage of private cloud to achieve flexible access control management. It supports specific and fine-grained access strategy with both forward and backward control using version tag. This mechanism is maintained by system manager who sets the access strategy. It regulars privileges of the revoked and new authorized attributes. For example, the new authorized attributes can be authorized or unauthorized to access the ciphertexts before, and the revoked attributes is in a similar way. The access strategy is set independently and managed by version tags depending upon different applications in mobile data sharing system.

Secondly, the system management of version tag is an important segment. Usually, the authority must update the version tag when an attribute is revoked or authorized. In the real application, it could be practical if there are more situations that need to update the version tag. For example, there exists a regularly mechanism that requires updating the version tag periodically for secure concerns. Moreover, the system update may also require updating the version tag for consistence. The management of version tag helps to achieve more flexibilities for system practicability.

Thirdly, the privacy of organization structure and user attributes is critical in medical, military and financial domains. The management of participants with attributes is flexible for the authority who initializes the system and manages the private cloud. The private cloud can achieve

anonymous identity authentication for all the access requests. In addition, the private cloud can also protect the organization structure and user attributes using the mechanisms of alias and secure communication. All the sensitive information can be processed in the system of private cloud so as to maintain privacy for organizations and users.

Finally, the private cloud can achieve resource management by controlling the priority level of requests. It provides the mechanism of load balancing through the identity authentication with attribute hierarchy in the access control list. When the system receives too many requests, the private cloud can also set the partition of hardware resources that responds different level requests that want to access the sharing files in public cloud.

5 Analysis

This section analyzes the proposal in the aspects of security and practicability.

5.1 Security

Similar to the traditional security proof [21], this FH-CP-ABE scheme is secure against CPA under DBDH assumption. Note that the security of symmetric encryption is out of consideration. It assumes that the symmetric encryption is secure and the content keys cannot be obtained except decrypting the ciphertext in this construction.

Theorem 1. *Suppose DBDH assumption holds, then no polynomial adversary can selectively break the proposed system.*

This section provides a direct demonstration to prove the scheme is secure similar to the traditional scheme [21]. The different construction between this scheme and the traditional proposal is that there are several more public parameters $g^v, g^{\Delta v}$ and the decryption keys are split into three parts. These differences contribute to the flexibility and practicability and bring the differences of encryption and decryption calculations. The different constructions are consisted of random numbers. Firstly, the version tag introduces several more parameters related to $g^v, g^{\Delta v}$ which are randomly confused by r_j. Secondly, decryption keys are randomly confused by SK, and can be recovered by random parameters SK, CK and g^{vr_j}. Finally, the security proof of the proposal depends upon the traditional scheme that adds random challenges in the traditional security proof.

5.2 Practicability

Considering the real application scenario, a practical scheme should provide flexible system management and efficient user performance. This section analyzes the practicability of the improvement from the aspects of flexibility of system management and efficiency of user performance.

5.2.1 Flexibility

From the angle of system management, flexibility is necessary to implement various management strategies. For example, different organization needs different access policy to deal with human resource affairs. Some organizations require that revoked attributes cannot access files before, and the others require that revoked attributes can access files before. In order to provide flexible privilege management for system manager, this improvement introduces version tag to regular fine-grained access strategies.

Moreover, the privacy management is also self-defined by system managers. It can be set anonymous or real to adapt hardware resources and application requirements. As an instance, the attributes of inner employees are sensitive information for the system manager and attributes should be protect for privacy. In another alias system, the attributes of outer users can be public for simplify system management and hardware cost.

Finally, there are several more flexibilities that need to custom-tailor in an real scenario for specific applications.

5.2.2 Efficiency

From the angle of efficiency, this improvement increases user decryption efficiency by the method of proxy-based decryption. The exponent and paring calculations are the expensive calculations in the decryption phase. Most of the expensive calculations are executed by the private cloud (as a proxy) who is in charge of privilege management. The rest of calculations are related to the secret decryption key SK that controlled by the user in a secure manner. User calculations are simple (multiplication, division and hash operations) and suitable for resource limited mobile devices.

6 Conclusion

This paper presents an improved FH-CP-ABE scheme with more practicality. The improvement introduces the mechanism of version match to synchronize decryption keys and ciphertexts. It helps to achieve both forward and backward access control for dynamic attribute management, and reduce most of the computation burden for final user(client) decryption. Moreover, this paper describes an implementation utilizing the improved scheme in hybrid cloud environment for mobile users. The implementation achieves flexible arbitrary attribute alteration without updating user keys or ciphertexts. Additionally, it greatly increases efficiency for user decryption. Finally, the security and performance analysis demonstrate that this proposal is suitable for mobile on-line hierarchy file access control in cloud computing.

References

[1] S. Amit and B. Waters, "Fuzzy identity-based encryption," in *Annual International Conference on*

the Theory and Applications of Cryptographic Techniques, vol. 3494, pp. 457–473,2005.

[2] Z. Cao, L. Liu, Z. Guo, "Ruminations on attribute-based encryption," *International Journal of Electronics and Information Engineering*, vol. 8, no. 1, pp. 9–19, 2018.

[3] L. Cheung, and C. Newport, "Provably secure ciphertext policy ABE," in *Proceedings of the 14th ACM Conference on Computer and Communications Security*, pp. 456–465, 2007.

[4] P. S. Chung, C. W. Liu, and M. S. Hwang, "A study of attribute-based proxy re-encryption scheme in cloud environments", *International Journal of Network Security*, vol. 16, no. 1, pp. 1-13, 2014.

[5] W. Guojun, *et al.*, "Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers," *Computers & Security*, vol. 30, no. 5, pp. 320–331, 2011.

[6] W. Guojun, L. Qin, and W. Jie, "Hierarchical attribute-based encryption for fine-grained access control in cloud storage services," in *Proceedings of the 17th ACM Conference on Computer and Communications Security*, vol. 30, pp. 735–737, 2010.

[7] M. Hui, *et al.*, "Verifiable and exculpable outsourced attribute-based encryption for access control in cloud computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 6, pp. 679–692, 2015.

[8] H. Jinguang, *et al.*, "Improving privacy and security in decentralized ciphertext-policy attribute-based encryption," *IEEE transactions on Information Forensics and Security*, vol. 10, no. 3, pp.665–678, 2014.

[9] B. John, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," *IEEE Symposium on Security and Privacy (SP'07)*, 2007. DOI: 10.1109/SP.2007.11.

[10] L. Junzuo, *et al.*, "Attribute-based encryption with verifiable outsourced decryption," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1343–1354, 2013.

[11] H. Junbeom, and D. K. Noh, "Attribute-based access control with efficient revocation in data outsourcing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 7, pp. 1214–1221, 2010.

[12] Y. Kang, W. Guohua, *et al.*, "Attribute based encryption with efficient revocation from lattices," *International Journal of Network Security*, vol. 22, no. 1, pp. 161-170, 2020.

[13] Z. leyou, Y. Hongjian, "Recipient anonymous ciphertext-policy attribute-based broadcast encryption," *International Journal of Network Security*, vol. 20, no.1, pp. 168-176, 2018.

[14] Q. M. Malluhi, A. Shikfa, and V. C. Trinh, "A ciphertext-policy attribute-based encryption scheme with optimized ciphertext size and fast decryption," in *Proceedings of ACM on Asia Conference on Computer and Communications Security*, pp. 230–240, 2017.

[15] G. Matthew, S. Hohenberger, and B. Waters, "Outsourcing the decryption of abe ciphertexts," *USENIX Security Symposium.*, vol. 2011, no. 3, 2011.

[16] A. Nuttapong, and H. Imai, "Attribute-based encryption supporting direct/indirect revocation modes," *IMA International Conference on Cryptography and Coding*, vol. 5921, pp. 278–300, 2009.

[17] A. Nuttapong, B. Libert, and E. D. Panafieu, "Expressive key-policy attribute-based encryption with constant-size ciphertexts," *International Workshop on Public Key Cryptography*, vol. 6571, pp. 90–108, 2011.

[18] W. Pengpian, F. Dengguo, and Z. Liwu, "Towards attribute revocation in key-policy attribute based encryption," in *International Conference on Cryptology and Network Security*, vol. 7092, pp. 272–291, 2011.

[19] Y. Shucheng, *et al.*, "Achieving secure, scalable, and fine-grained data access control in cloud computing," *Proceedings IEEE INFOCOM*, 2010. DOI: 10.1109/INFCOM.2010.5462174.

[20] W. Shulan, *et al.*, "A novel file hierarchy access control scheme using attribute-based encryption," *Applied Mechanics and Materials*, vol. 701–702, pp. 911–918, 2015.

[21] W. Shulan, *et al.*, "An efficient file hierarchy attribute-based encryption scheme in cloud computing," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1265-1277, 2016.

[22] Y. Shucheng, *et al.*, "Attribute based data sharing with attribute revocation," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, pp. 261–270, 2010.

[23] L. Suqing, *et al.*, "Revisiting attribute-based encryption with verifiable outsourced decryption," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 10, pp. 2119–2130, 2015.

[24] J. Taeho, *et al.*, "Control cloud data access privilege and anonymity with fully anonymous attribute-based encryption," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 190–199, 2014.

[25] G. Vipul, *et al.*, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM conference on Computer and Communications Security*, pp. 89–98, 2006.

[26] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," *International Workshop on Public Key Cryptography*, vol. 6571, pp. 53–70, 2011.

[27] T. Weiliang, C. Yafeng, *et al.*, "Security analyses of a data collaboration scheme with hierarchical attribute-based encryption in cloud computing," *International Journal of Network Security*, vol. 22, no. 2, pp. 212–217, 2020.

[28] L. Wen-Min, *et al.*, "Flexible CP-ABE based access control on encrypted data for mobile users in hybrid cloud system," *Journal of Computer Science and Technology*, vol. 32, no. 5, pp. 974–990, 2017.

[29] F. Xingbing, *et al.*, "Large universe attribute based access control with efficient decryption in cloud storage system," *Journal of Systems and Software*, vol. 135, pp. 157-164, 2018.

[30] H. Xiong, and S. Jianfei, "Comments on Verifiable and exculpable outsourced attribute-based encryption for access control in cloud computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 4, pp. 461–462, 2017.

Biography

Junliang Zhang Junliang Zhang is an assistant professor at the department of mechanical and electrical engineering of Dezhou University, Shandon, China. He received his B.S. and M.S. degree in 1999 and 2004 at Shandong University of Technology, and PH.D. degree in mechanical and electrical engineering at China University of Petroleum in 2007, respectively. He finishes his post-doctoral researches in the topics of the completion oil recovery tools, intelligent switch for layered water plugging and Intelligent Robot for Oil Field Workover Work. Now he is doing research in the areas of control and automation of electromechanical system, access control of electronic and information security.

Weiyou Zhang Weiyou Zhang is a senior engineer at the department of mechanical and electrical engineering of Dezhou University, Shandon, China. He is employed as a special professor for the research and development of mechanical and electrical engineering. He received his B.S. degree in machine building technics and equipment subject in 1994 at Shandong University, Jinan, China. Now he is doing research in the areas of control and automation of electromechanical system.

A Blind Signature-based Location Privacy Protection Scheme for Mobile Social Networks

Xin Xu, Mi Wen, and Liangliang Wang
(Corresponding author: Xin Xu)
College of Computer Science and Technology, Shanghai University of Electric Power
Shanghai, 200090, China
Email: 18108035@mail.shiep.edu.cn
(Received June 6, 2020; Revised and Accepted Feb. 10, 2021; First Online Aug. 14, 2021)

Abstract

Location-based services (LBS) have gradually become an integral part of people's lives. In mobile social networks, we hope to protect our location privacy information in different environments and get services in time. However, scholars have proposed various effective location privacy protection strategies, such as k-anonymity, fuzzy location, etc. However, these privacy protection technologies are based on the location information obtained, and it is rarely achieved from the stage of location acquisition. In this article, we propose a privacy protection scheme based on blind signatures [9] for mobile locations. The solution uses a blind signature and pseudonym to verify the user's identity anonymously. It adds false information to form k-anonymity, which can flexibly protect the user's relevant information in different environments and achieve the two-stage privacy protection of LBS. Simulation results show that this method has better performance and higher security compared with other existing approaches, and it can be applied to different types of mobile environments.

Keywords: Blind Signature; Location Privacy Protection; Mobile Social Networks; User Collaboration

1 Introduction

With the popularity of mobile devices, LBS has become an essential part of human life. It is widely used in all aspects of people's lives and bring great convenience to people. For example, when we are chatting with our friends, we could send our own location information to them for sharing our life. We can also send our location information to the relevant application to get the local weather, and plan the perfect route information for travel. Unfortunately, when we enjoy the convenience of LBS, we also face some challenges. If the service provider is untrusted, it may leak the user's location information. Some attackers can further steal users' privacy data (such as salary and bank card number) based on users' location information and social engineering, so as to obtain more benefits from targeted attacks [1, 8, 15, 24, 27].

The LBS mainly contains two stages:

1) Location acquisition stage: The stage where the mobile device acquires its current location through GPS or a third-party network.

2) Service acquisition stage: The stage in which the mobile device sends the location information and the inquired points of interest to the LBS provider, and the LBS provider performs the inquiry and returns the service information to the mobile device [5].

As shown in Figure 1. Unfortunately, the existing privacy protection methods of LBS cannot be applied to both stages of LBS at the same time. In the service acquisition phase, the user directly submits their location information, and privacy protection technology enables the LBS provider to provide users with corresponding services without knowing the user's exact location by anonymization and generalization of the location information [14, 23, 25]. While in the location acquisition phase, the user submits the current location fingerprint, and the Location Provider(LP) estimates the specific location based on the fingerprint [11,17,20], but the location fingerprint represents a definite location, which cannot be processed by obfuscation technology.

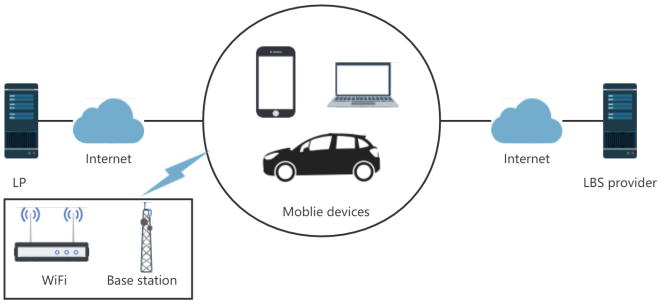


Figure 1: Location-based service architecture

In mobile social networks, we hope to be able to obtain services in time while protecting location privacy. However, since users may be in different environments and scenarios, the demand for anonymity levels may also be different [21]. If we use the same level of privacy protection methods, it may affect the quality of the service, such as the real-time nature of the message; it may also affect the strength of privacy protection, such as the confidentiality of data and the anonymity of user identity and location [22]. To solve the above problems, this paper proposes a mobile location privacy protection scheme based on a blind signature [7]. At the same time, the combination of k-anonymity and virtual information technology enables the solution to be implemented flexibly according to the number of users in the collaboration group. Therefore, the security of the solution will not vary greatly in different environments.

Our main contributions are listed as follows:

1) In order to fully protect the privacy of the two stages of LBS, we study the similarity of information transmission between the two stages, and use blind signature technology to separate the user identity and related request information, so as to achieve location privacy protection of both stages of LBS.

2) Secondly, we use k-anonymity and false information to protect the location information and keep the diversity of requested information according to the needs of users and the differences in the surrounding environment.

3) Thirdly, we analyze the security strength and privacy protection capabilities of BSLPP (Blind Signature-based Location Privacy Protection). In particular, we use provable security technology to formally prove that it is safe to protect users' private information under man-in-the-middle attacks. Through performance analysis, we prove that BSLPP is indeed more effective than the scheme mentioned in [16,18]. Compared with the scheme mentioned in [11], the safety and applicability of our scheme have also been improved.

The rest of the paper is organized as follows: Related work is reviewed in Section 2. We introduce our system model, security requirements, and our design goal in Section 3. In Section 4, we introduce our scheme design. Section 5 shows the security analysis of the scheme. Section 6 conducts performance evaluation. We conclude this paper in Section 7.

2 Related Work

For the protection of the third LP's privacy, Damiani and Cuijpers [2] first pointed out that when users use the third-party network location, the user will submit the information of nearby access points(APs) to the third-party LP, and the LP will calculate the location information. In this case, the LP will obtain their location information before the user, which causes the user's location information to be leaked. Sun *et al.* [11] used homomorphic encryption to perform location processing in the ciphertext space to prevent the access point information in the service from being threatened by privacy. However, the balance between system overhead and quality of service has become a disadvantage of this solution. Wang *et al.* [20] proposed a method to add virtual information to the location request so that the location server cannot distinguish the user's real location information from the virtual information. However, the virtual location generated by this strategy may be recognized by the location server, which greatly reduces the security of the solution. Song *et al.* [17] applied the location privacy protection strategy to complete fingerprint matching on the client. In this approach, the client matches its location fingerprint with the fingerprint data received within this range to obtain its location information. But the fingerprint data received in this range is provided by LP, it still knows the location range of the client, so the protection of user identity and specific location is not high.

In the service acquisition phase, in order to deal with untrusted third-party anonymous servers, Peng *et al.* [14] added a function generation server on the basis of the anonymous server structure to aggregate users with the same value to achieve k -anonymity. In addition, scholars have also proposed techniques that combine k-anonymity with other technologies, such as autonomous learning [13] and clustering [26]. Ye *et al.* [23] introduced pseudo-queries in LBS query requests to effectively resist query probability statistical attacks and continuous attacks, and prevented attackers from mapping the specific content of query requests based on user identity. Zhao *et al.* [25] combined user privacy with geographic location information, and generated corresponding fake locations to protect user privacy based on the user's different access probabilities to different points of interest. In order to obtain higher query accuracy and privacy protection level, there are still some works are based on the cryptographic techniques [10, 28]. Liao *et al.* [12] pointed out Qi's registration agreement [4] may not delete the linkability of the real ID and authorized anonymous ID, so they proposed an improved registration and re-obfuscation protocol that prevented administrators from obtaining unauthorized anonymity and true identity. Maede *et al.* [18] proposed a new privacy protection protocol by using blind signature technology. Instead of excessively protecting the user's ID, they encrypted the query information to achieve user identity and security. The separation of messages protects the user's location privacy, but once the database colludes with others to leak the shared key, it will the user's identity be misused to make illegal queries. Researchers also proposed differential privacy technique [3,6,19] to protect the user's location information. However, all of these solutions are bringing a large computation burden on the user side, which makes it is not suitable for mobile devices. Junggab *et al.* [16] developed a location privacy protection strategy based on pseudonyms. He functional-

ized the pseudonyms and used secret sharing technology to share locations with designated friends, giving a lot of calculations to the service is more suitable for mobile social networks, but server failure is still the bottleneck of the solution.

3 System Model, Security Requirements and Design Goal

In this section, we formalize the system model, system design goal, and system security requirements. This paper adopts a distributed peer-to-peer model, and proposes an untrusted environment-oriented architecture consisting of mobile users, LP, and LBS providers. The architecture and message flow are shown in Figure 2.

3.1 System Model

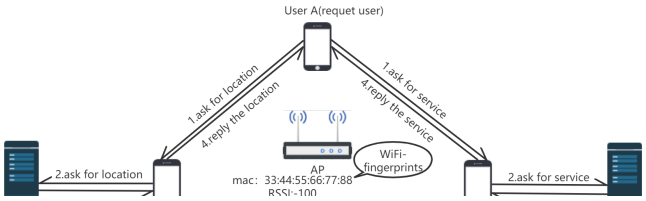


Figure 2: System model

Mobile User: It can communicate with other users or servers through the AP. The mobile user may be a user requesting a service or a member of a collaboration group. As a user requesting a service, he interacts with members of the collaboration group and lets them initiate requests to the server instead of himself. As members of the collaboration group, they interact with the server on behalf of users who request services, complete location queries and LBS requests, and return the results of the requests to users who request services.

AP: APs provide communication channels for collaborative users, WiFi fingerprints for users who need them, and basic information for the location.

LP: The LP calculates location information based on the Wi-Fi fingerprint sent by the user, and returns the location result to the corresponding user.

LBS Provider: The LBS provider returns the corresponding service information to the user based on the location provided by the user and the requested service.

3.2 System Security Requirements

Security is critical to the success of location privacy protection. In our security model, we consider that the LP and the LBS provider are untrusted, and those collaborative users are also at risk of leaking information. First, we send location-related information and request types to the LPs and the LBS providers, who may leak our related privacy information to criminals. For collaborative users, they may also be mixed with attackers, and cooperate with other collaborative users to modify our information or analyze it through space-time correlation analysis to leak our private information. In addition to the above two types of insecure factors, there are also man-in-the-middle attacks and analyze attacks launched against us by external attackers. Therefore, in order to prevent the above-mentioned insecure factors, the following security requirements should be satisfied in the process of location privacy protection.

The Data Confidentiality. Protect personal location privacy-related information from attackers, that is, even if communication is eavesdropping during collaboration, the content of the message cannot be identified. In this way, the user's privacy data protection can be satisfied.

The Anonymity of the User's Identity and Location. Even if the LPs and LBS providers get the real location information and the requested content, they cannot distinguish which user it comes from.

Authentication and Data Integrity. Authenticate the encrypted information sent by legitimate cooperative users that have not been tampered with during transmission, that is, if an attacker forges and/or modifies information, malicious operations should be detected. The collaborator only completes the corresponding service for receiving correct and credible messages.

3.3 Design Goal

Under the above system model and security requirements, our design goal is to provide a location-based service with strong applicability, high security and responsiveness. Specifically, the following two goals should be achieved.

Suitable for Various Environments. Due to the mobility of users, we may be in a sparsely populated area, so the confidentiality of some privacy protection algorithms may be greatly reduced. We want to protect our location privacy wherever we are.

Ensure the Safety and Timeliness of Services. We want users' privacy not to be known to anyone, even collaborative users. On the basis of ensuring security, we also hope that it will not affect the user's service experience.

4 The Proposed BSLPP Scheme

This paper protects the privacy of the user's location by user collaboration, and gets rid of the bottleneck of

using anonymous servers. The solution is divided into three parts including establish an anonymous collaborative group, protect the privacy of location services, and protect the privacy of LBS. Table 1 lists the notations used throughout the description of the scheme for ease of reference.

Table 1: Notation

U_A	Mobile user A
$ID_{A_{num}}$	Pseudonym calculated using the MAC of user A
k	The number k of anonymity
k_{min}	the minimum value of k that meets the need for anonymity
R	The number of hops in the anonymous zone
$H(m)$	Hash message m
MAC_{other}	Client's hardware address
$Figure_{num}$	A WiFi fingerprint
$r_{A_{num}}$	A random number generated by user A
t_{num}	Timestamp to prevent replay attacks
$Location_{num}$	A series of location information
$Type_{num}$	Multiple types of request services
S_{L-P}	Provider of location
S_{LBS-P}	Provider of LBS
$PubK_A$	Public key pair of user A
$PriK_A$	Private key pair of user A
$\{m\}_{PubK_A}$	Encrypt message m with the public key of user A
$\{m\}_{PribK_A}$	Sign message m with the private key of user A
K_{AB}	Shared secret key between users A and$B$
$\{m\}_{K_{AB}}$	Encrypt message m with the secret shared key K_{AB}
$C(x)$	Blind message x
$C^{-1}(x)$	Unblind message x

4.1 Establish Anonymous Collaborative Group

Our solution uses a point-to-point communication method and establishes a k-anonymous collaboration group to prepare for the next service acquisition. We describe the proposed scheme as follows.

1) User A asks whether users within the R-hops range are willing to participate in collaborative work through broadcasting.

2) If the user agrees to cooperate, the user who receives the request message sends the reply to user A with his public key, user identifier ID and hops R' from user A. If they does not agree, user A ignores this message.

3) After receiving $2k$ user's replies, user A selects k users arbitrarily, lists them (user identifier, user public key, recent usage times, and hop count). If the number of users receiving the reply is less than k and greater than k_{min}, randomly select n ($n < k$) users, and add k-n messages as dummy messages from the recent historical query information. If the number of users who received the reply is less than k_{min}, expanded hop count R and continue to repeat 1) - 2).

4) In the next query, if the difference between the number of hops with the collaborating user is less than R, the interaction continues. If the difference is greater than R, then user A deletes the corresponding user in the list to save time and costs. After deletion, if the number of collaborative users is less than k, please repeat 1) -3).

The pseudo code of establishing anonymous collaborative group is elaborated in Algorithm 1.

Algorithm 1 Establish anonymous collaborative group

1: $U_A \rightarrow U_{other}$: request for the collaborative work within the range of hops R
2: **if** U_{other} agree **then**
3: Compute ID_{other}=$H(MAC_{other})$*r_{other_1}+r_{other_2};
4: Send $\{ID_{other}$,$PubK_{other}$,$H(ID_{other}$,$PubK_{other})\}$ to U_A;
5: **end if**
6: **if** the number of users exceeds $2k$ **then**
7: Select k users randomly in $2k$ users and add $\{ ID_{other}$,$PubK_{other}$,$R \}$ into list L;
8: **end if**
9: **if** the number of users less than k and exceeds k_{min} **then**
10: Select n users randomly, and k-n messages in the historical query information.
11: **end if**
12: **if** the number of users less than k_{min} **then**
13: Expanded hop count R or reduce k and return *step 1*;
14: **end if**
15: **while** their collaboration distance differs by more than R **do**
16: Delete the corresponding user from list L;.
17: **if** the number of cooperative users is less than k **then**
18: Return *step 1*
19: **end if**
20: **end while**
21: End

Figure 3 shows the step of establishing an anonymous collaborative group, where user A executes the algorithm. The transmission range of user A is represented by a dotted circle. User A sends a broadcast request for collaboration, and then receives *16* near peers represented by white circles (Figure 3(a)). User A then randomly selects *8* collaborative users represented by black dots and

add them in list L(Figure 3(b)). But user A receives the number of near peers represented by white circles less than *8* and exceeds *4* (Figure 3(c)). User A then randomly selects *5* collaborative users represented by black dots and add them in list L and sends *3*different messages in the next service request algorithm (Figure 3(d)). When the number of received replies is less than *4*, expand the number of hops of the route or reduce the number of anonymities as required (Figure 3(e)- 3(f)).

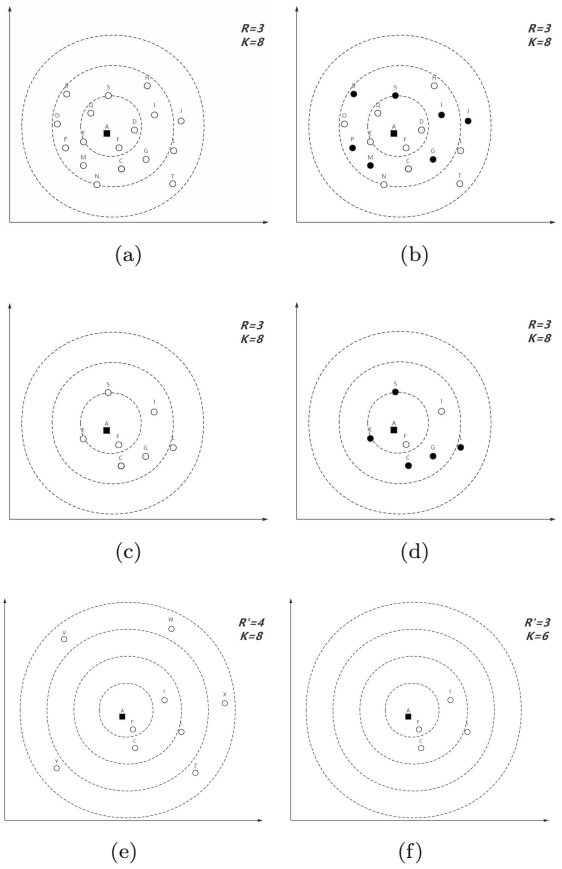


Figure 3: Components of an anonymous collaboration group

4.2 Protect the Privacy of Location Service

To ensure that users obtain the privacy of their location through the network. On the one hand, we use blind signature technology to blind the WiFi fingerprint information, so that the user's identity is separated from the request. On the other hand, we mix some dummy location information to confuse the sight of the collaborative users and LPs, but also make the solution available in sparsely populated places.

1) User A sends a WiFi fingerprint request to the surrounding collaborative users.

2) The collaborative user randomly selects zero to three WiFi fingerprint information around him, signs and sends it using his private key to user A.

3) User A receives the message and verifies it with the corresponding public key, and records it as $Figure_1$, $Figure_2$...

4) Then user A randomly selects a user in a collaboration group and records the most recent trials in his list, such as user B. User A uses his pseudonym ID_{A_1} to send his public key and blinded WiFi fingerprint information to user B. To ensure security, the fingerprint information here may be mixed with WiFi fingerprints requested from other users. Even if someone queries the location through the location server, they don't know if it is the user's real location.

5) User B authenticates the sent message. If the verification is successful, the blinded message is signed, otherwise the request from user A is rejected.

6) User A verifies the message that sent back. If the verification is successful, the message is unblinded. Otherwise, ignore the message and reselect a new user, repeat 4).

7) User A encrypt the message that including the query request, the signed and unblinded message and session key with User B's public key and sends to User user B using pseudonym ID_{A_2}.

8) After received the message, user B decrypts and verifies the validity of the signature. If the verification is passed, the location information query is performed on behalf of the user A, otherwise the request is rejected.

9) User B submits the fingerprint information to the LP, and the LP performs calculation based on the fingerprint information and returns the location information to user B. User B encrypts the result and the signed and unblinded message with the session key, then returns it to user A.

The pseudo code of protecting the privacy of location service is elaborated in Algorithm 2.

4.3 Protect the Privacy of LBS

In terms of obtaining location-based services, we must not only consider protecting user location information, but also protect the type of information requested. Because some users are more sensitive to where they are and some users pay more attention to the type of message requested. Therefore, we have taken these two points into consideration while integrating the privacy protection of location services to provide users with more comprehensive protection.

1) User A randomly selects users where except user B in a collaboration group, such as user C. And records the most recent trials of user C in his list. User A uses his pseudonym ID_{A_3} to send his public key and

Algorithm 2 Protect the privacy of location service

1: $U_A \rightarrow U_{other}$: request for the WiFi fingerprint
2: **if** U_{other} agree **then**
3: Send $M_1 = \{ Figure, t_1, m_1 = \{ Figure\}_{PriK_{other}}\}$ to U_A;
4: **end if**
5: **if** $\{m_1\}_{PubK_{other}} == Figure$ is (TRUE) **then**
6: U_A add it in list FP;
7: **end if**
8: U_A computes x=$H(Q$=$(Figure_1$,$Figure_2$,...,$Figure_n)$ to ID_B;
9: U_A sends M_2=$\{ ID_{A_1}, PubK_A, t_2, C(x), m_2$=$\{ H(ID_{A_1}$,$t_2$,$C(x), PubK_A)\}_{PriK_A}\}$ to ID_B;
10: U_B decrypts $\{m_2\}_{PubK_A}$
11: **if** H$(ID'_{A_1}, PubK'_A, t'_2$,$C(x)) == m_2$ is (TRUE) **then**
12: U_B sends M_3=$\{ ID_B, t_3, m_3 = \{C(x)\}_{PriK_B} \}$ to ID_{A_1};
13: **end if**
14: **if** $\{m_3\}_{PubK_B} == C(x)$ is (TRUE) **then**
15: U_A unblinds $s = C^{-1}(m_3)$ and sends M_4=$\{ ID_{A_2}, t_4, m_4$ =$\{ ID_{A_2}, t_4, s, Q, K_{AB}\}_{PubK_B}\}$ to ID_B;
16: **end if**
17: U_B decrypts $\{m_4\}_{PriK_B}$
18: **if** $H(Q') == \{s'\}_{PubK_B}$ is (TRUE) **then**
19: U_B sends Q to S_{L-P};
20: **end if**
21: After received $\{Result\}$ from S_{L-P}, U_B sends $\{Result, s\}_{K_{AB}}$ to ID_{A_2}.
22: End

blinded request information based on location services to user C. In order to meet the demands of different users, we propose two formats of request information here. When the user is in a sensitive location, he sends a collection of locations containing other nearby locations. When the user requests a more sensitive message type, he sends a collection of information of multiple request types.

2) User C authenticates the sent message. If the verification is successful, the blinded message is signed. Otherwise the request from user A is rejected.

3) User A verifies the message that sent back. If the verification is successful, the message is unblinded. Otherwise, ignore the message and reselect a new user, repeat 1).

4) User A encrypt the message that including the query request, the signed and unblinded message and session key with user C's public key and sends to user C using pseudonym ID_{A_4}.

5) After received the message, user C decrypts and verifies the validity of the signature. If the verification is passed, the location information query is performed on behalf of the user A, otherwise the request is rejected.

6) User C submits the request information to the LBS provider, and the LBS provider returns a result set based on the requested information to user C. User C encrypts the result and the signed and unblinded message with the session key, then returns it to user A.

The pseudo code of protecting the privacy of LBS is elaborated in Algorithm 3.

Algorithm 3 Protect the privacy of LBS

1: U_A computes y=$H(Q$=$(Location_1, Location_2, ..., Location_n, Type)$ or Q=$(Type_1$,$Type_2, ...,Type_n, Location))$ to ID_C;
2: U_A sends M_5=$\{ ID_{A_3}, PubK_A, t_5, C(y), m_5$=$\{ H(ID_{A_3}, t_5$,$C(y),$ to ID_C
3: U_C decrypts $\{m_5\}_{PubK_A}$
4: **if** $m_5 ==$ H$(ID'_{A_3}, PubK'_A, t'_5$,$C(y))$ is (TRUE) **then**
5: Send M_6=$\{ ID_C, t_6, m_6$=$\{C(y)\}_{PriK_C} \}$ to ID_{A_3};
6: **end if**
7: **if** $\{m_6\}_{PubK_C} == C(y)$ is (TRUE) **then**
8: U_A unblinds $s = C^{-1}(m_6)$ and send M_7=$\{ ID_{A_4}, t_7, m_7$=$\{ ID_{A_4}, t_7, s, Q, K_{AC}\}_{PubK_C}\}$ to ID_C;
9: **end if**
10: U_C decrypts $\{m_7\}_{PriK_C}$
11: **if** $H(Q') == \{s'\}_{PubK_C}$ is (TRUE) **then**
12: Sends Q to S_{LBS-P};
13: **end if**
14: After received $\{Result\}$ from S_{LBS-P}, U_C sends $\{Result, s\}_{K_{AC}}$ to ID_{A_4}.
15: End

For security reasons, users must use different collaborative users to serve them during two different queries. After the user's identity as a request service and a collaboration service is converted, he must be using different public-private key pairs. In order to prevent correlation attacks and analysis attacks based on historical information, users often need to maintain and change their public and private keys.

5 Security Analysis

In this section, we will analyze the security of the proposed BSLPP scheme. In particular, according to the security requirements discussed earlier, our analysis will focus on data confidentiality and anonymity of user identities, and authentication and data integrity.

1) The confidentiality of data and anonymity of user identities are achieved in the proposed BSLPP scheme. In the proposed BSLPP scheme, when the user sends request information to the collaborating user, the information Q is first hashed to form x = $H(Q)$, and then the summary information x is

blinded as follows:

$$C_1 = r_1 * x^a$$
$$C_2 = r_2 * x^b \quad (1)$$

After receiving the encrypted blind message, the collaborative user first decrypts the private key to obtain C_1, C_2, and then signs C_1, C_2 as follows::

$$C_1' = {C_1}^d \bmod n$$
$$C_2' = {C_2}^d \bmod n \quad (2)$$

After the user obtains the signed data C_1', C_2', he unblinds it as

$$S_1 = C_1' * {r_1}^{-1} \bmod n$$
$$S_2 = C_2' * {r_2}^{-1} \bmod n \quad (3)$$
$$S = {S_1}^k * {S_2}^l \bmod n$$

accroding to Equations (1) (2) (3), we can get Equation (4)

$$\begin{aligned} S^e &= {S_1}^k * {S_2}^{le} \bmod n \\ &= ((C_1' * {r_1}^{-1})^k * (C_2' * {r_2}^{-1})^l)^e \bmod n \\ &= (({C_1}^d * {r_1}^{-1})^k *, ({C_2}^d * {r_2}^{-1})^l)^e \bmod n \\ &= (x^{d*(ak+bl)})^e \bmod n \\ &= (x^d)^e \bmod n \\ &= x \bmod n. \end{aligned} \quad (4)$$

On the one hand, the user obtains the signature data of the collaborative user without disclosing the content of the request message; on the other hand, even if the user publishes the signature, the collaborative user cannot track the signature data. Because the collaborative user retains a set of data (C_1, C_2, C_1', C_2'), but he has no way to know (r_1, r_2, a, k, b, l) from S. In the two interaction phases of request signature and request service, users use different pseudonyms and public and private keys to interact with the collaborative user, so the collaborative user cannot associate the original request data Q with the user's identity. Thus, the confidentiality of data and anonymity of user identities is achieved between users and collaborative users.

2) The authentication and data integrity of between users and collaborative users are also achieved in the proposed BSLPP scheme. In the proposed BSLPP scheme, The user sends the signed data together with the original data to the collaborating user. The collaborating user first hashed to form $x_i' = H(Q_i')$, then sign the formed abstract to get S_i'. Compare S_i' with S_i. If they are the same, they are the users who have completed the signature before, and thus complete the identity verification as users in the collaboration group. In each information exchange process, we hash the message. If the data is changed, we will easily find out. Since then, the integrity of the data has been verified.

6 Performance Evaluation

In this section, we evaluate the performance of our framework in terms of computation cost on the involved parties as well as the communication cost. The main computation operations in our scheme include exponentiation and multiplication in G and GT, and pairing, besides, it also contains the symmetric and the asymmetric encryption and decryption cost. Table 2 presents the meaning of notations used in this section.

Table 2: Notations used in the performance evaluation

Notations	Description
M	The number of records which satisfies the original query
N	The number of information in the server
N'	The number of information submitted
T_{LE}	The time of lagrangian interpolation construction
T_{LD}	The time of lagrangian interpolation decryption time
T_{SE}	The time of symmetric encryption
T_{SD}	The time of symmetric decryption
T_{BM}	The time of blind message
T_{UM}	The time of unblind message
T_{PE}	The time of RSA encryption
T_{PD}	The time of RSA decryption
T_p	The time of pairing
${T_e}^G$	The time of exponentiation in group G
${T_m}^G$	The time of multiplication in G

6.1 Efficiency Analysis

In this section, we introduce the calculation and communication comparison between [11, 18] and our scheme. Since [11,18] are two different stages in LBS, we will compare the stages corresponding to our scheme with them separately. For simplicity, we have omitted some fixed costs in all three frameworks. See Table 3 for details.

In [11], the system uses a client-server model, so there is no agent consumption. The cost of the client is the number of APs collected in the current building and the exponentiation and modular multiplication of homomorphic encryption and decryption costs. The cost of the server lies in the number of APs submitted in the database and the exponentiation and modular multiplication of the state encryption and decryption costs.

The calculation cost on the user side depends only on the number of POI records that satisfy the original query. Compared with the third part of our framework, the computational cost of the user side in [18] saves one public key encryption time; But in terms of the computational cost on the agent. in [18], it needs to match the corresponding token pair to the blinded message is signed, and our solution is to use the unified key of the agent to sign. So

Table 3: Computation comparisons with other schemes

Schemes	User	Proxy	Service Provider
priWEL [11]	$2N*T_e^G+4N*T_m^G$	0	$N*(N'+1)*T_e^G+3N*(2N'+3)*T_m^G$
Ours-stage2	$T_{BM}+2T_{PE}+2T_{PD}+T_{UM}+T_{SD}$	$2T_{PD}+3T_{PE}+T_{SE}$	0
BlindLocation [18]	$T_{BM}+2T_{PE}+T_{PD}+T_{UM}+T_{SD}*M$	$T_{PD}+T_{PE}+T_p+T_{SE}*M$	0
functional pseudonym [16]	$6T_e^G+(M+3)*T_m^G+2T_{LE}$	$N*T_e^G+4N*T_m^G+N*T_{LD}$	0
Ours-stage3	$T_{BM}+2T_{PE}+T_{PD}+T_{UM}+T_{SD}*M$	$2T_{PD}+T_{PE}+T_{SE}*M$	0

as the number of agents receiving tasks increases, our solution will be slightly lower than the communication cost in [18].

6.2 Experiment

In this section, we will divide the evaluation into three stages according to the proposed BSLPP scheme. The encryption public key and private key we use here are respectively 128 bytes, the key length used by the blinding function is 128 bytes, and the session key length is 16 bytes. We assume that each identifier ID (including the pseudonym) and HMAC string are 20 bytes; the timestamp size is 3 bytes, and the WiFi fingerprint size is 22 bytes. Broadcast message length is 16 bytes; based on the request information sent by the location service, the size of the response location information and service result information is 20 bytes. According to the knowledge of cryptography, when the length of the plaintext is greater than the key length (bytes) -11, it is necessary to implement fragment encryption in RSA algorithm encryption. When segmentation is not required, the ciphertext length is equal to the key length; otherwise, the ciphertext length is equal to the key length multiplied by the number of slices. We use python language to implement blind signature technology and public and private key encryption process, and use AES-128 scheme to achieve session process protection.

6.2.1 Cost of the Establish Anonymous Collaborative Group

We use the pseudonym calculation time and information transmission time to evaluate the cost of the first stage. It can be seen from Figure 4 that the calculation of pseudonyms does not come from the same user, but different collaborative users calculate their own pseudonyms, so the pseudonym calculation time here will not increase with the increase of users.

Time to measure the pseudonym calculation time that User A needs to wait. It can be seen from Figure 5 that the information interaction between user A and the collaborative user only includes the broadcast information sent by user A and the pseudonym and public key transmitted by the collaborative user. We assume here that user A can only collect 3 to 4 people for each broadcast. If you want to increase collaboration, users must resend the broadcast again. Although the cost of information

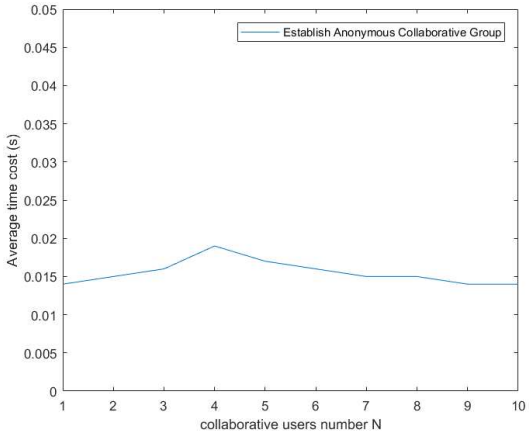


Figure 4: Time spent on encrypting information

transmission will increase with the increase in the number of users, it will also ensure the security of the following because of the increase in the number of users.

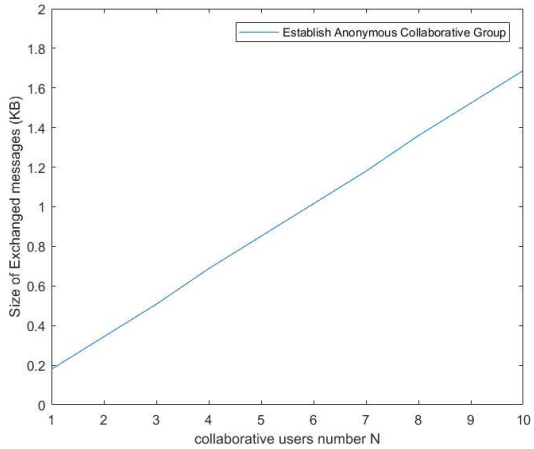


Figure 5: Time spent on exchanging information

6.2.2 Cost of the Protect the Privacy of Location Service

We implemented the priWEL designed in [11] to compare the BSLLP at the stage of protecting the privacy of location services. Based on the comparison of the time spent encrypting the information and the size of the information exchanged, the results are shown in Figure 6 and Figure 7.

In terms of information encryption time, the less infor-

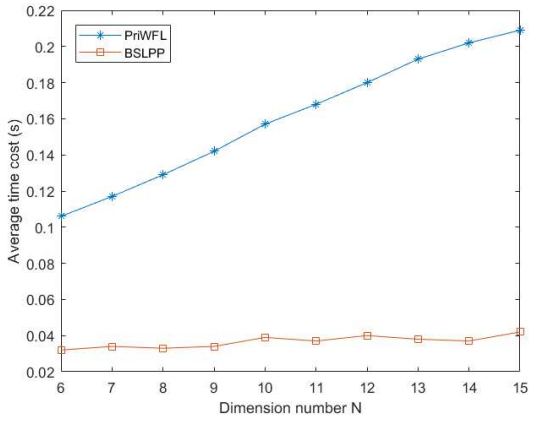


Figure 6: Time spent on encrypting information

mation encryption time, the less time the user spends on location privacy protection. It can be seen from Figure 6 that compared with the priWEL and the BLSSP, it takes less time under the same conditions, and the algorithm is more stable. The increase in the number changes greatly, so it takes less time to encrypt the information, and the responsiveness of the location request is better.

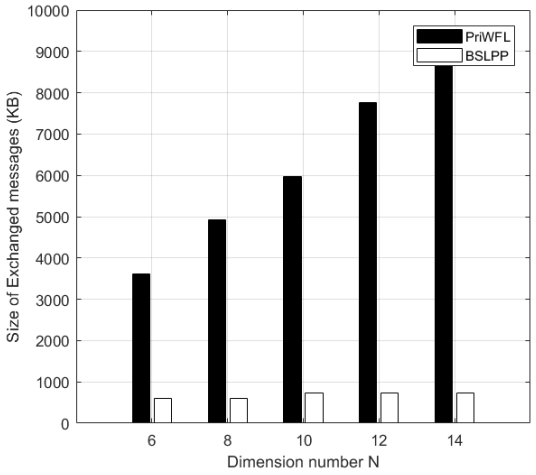


Figure 7: Time spent on exchanging information

In terms of the size of the information exchanged, the smaller the information exchanged, the less time it takes for the location privacy protection process, and the user can get a better experience. As can be seen from Figure 7, under the same conditions of BSLPP and priWEL, as the number of APs increases, the size of the information exchanged by the BSLPP has always been smaller than the priWEL, and will not increase with the number of APs And volatility. In the BSLPP, the hash algorithm we use not only protects the authenticity of the information, but also greatly reduces the bandwidth consumed by the information exchange. At the same time, two methods of the dummy and k-anonymity are used in the solution to ensure the anonymity of the information. The more users communicate with each other at the same time, the higher the anonymity of the information.

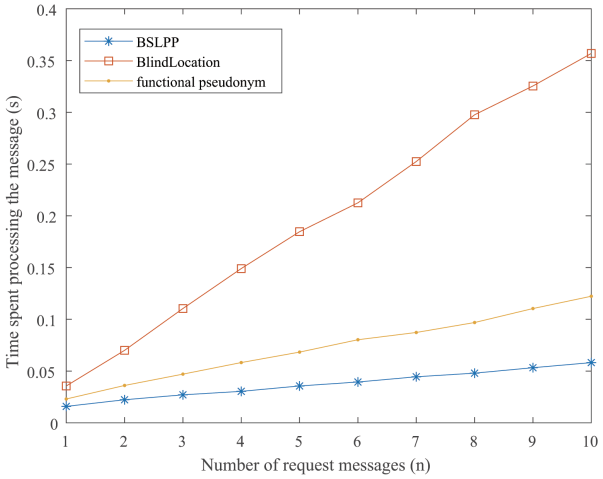


Figure 8: Time spent on location-based service

6.2.3 Cost of the Protect the Privacy of LBS

In the privacy protection stage based on location services, In terms of calculation cost and communication cost, we compare the solution of this paper with the two solutions of BlindLocation [18] and functional pseudonym [16]. In terms of calculation cost, we use CPU running time to represent it, as shown in Figure 8. As far as communication cost is concerned, we measure the size of the exchange message to be displayed, and the specific results are shown in Table 4.

Table 4: Efficency analysis

	Size of Exchanged messages
BSLPP	355 bytes
BlindLocation	227 bytes
functional pseudonym	756 bytes

As can be seen from Table 4, in the privacy protection stage based on location services, we have good performance in terms of encryption and decryption overhead and network overhead, and smartphones can easily implement blind privacy-based location privacy protection solutions. In addition, the scheme provides an acceptable trade-off between privacy and efficiency. We compare this scheme with the existing protocols in Table 5.

As shown in Table 5, in the privacy protection phase of location-based services, our solution has relatively small information exchange, and smartphones can easily implement a location privacy protection solution based on blind signatures. Our scheme does not need any special condition to support privacy, while some of the previous works will do. For example, the remaining solutions cannot achieve the privacy protection of location services and LBSs at the same time. BlindLocation, Obfuscation, and PIR don't protect location privacy in sparsely populated places.

Table 5: Comparison

	BSLPP	BlindLocation	functional pseudonym
Full process protection	Yes	NO	NO
Supporting Anonymity	Yes	Yes	Yes
Missing Quality	NO	NO	NO
Computing Cost	Acceptable	Acceptable	High
Supporting mobile users	Yes	NO	Yes
Untrusted third party server	YES	NO	NO
Supporting sparsely populated areas	Yes	NO	NO

7 Conclusion

This paper considers the issue of mobile user location privacy during the use of location services and location-based queries. We propose a mobile location privacy solution that is compatible with sparsely populated areas. The security analysis of the protocol and the performance analysis and comparison with the existing protocols prove that the scheme has a good performance in terms of user location privacy. The protocol does not need to rely on a trusted or semi-trusted third-party anonymous server to achieve user anonymity, but completes anonymous queries through blind signature technology and guarantees service quality.

In future work, we hope to optimize the algorithm, perform finer-grained verification management for users, and assist query work more efficiently.

Acknowledgments

This article was supported by the National Natural Science Foundation of China Grant (No. 61872230, U1936213, No. 61802248 and No. 61802249), Shanghai Science and Technology Innovation Action Plan (19511103700), Shanghai university young teacher training support program Grant(No. ZZsdl18006). We are very grateful to the reviewers for their valuable comments, and it helpsus to improve the quality of this paper.

References

[1] X. J. Chen and Y. Mu, "Preserving user location privacy for location-based service," in *Green, Pervasive, and Cloud Computing*, pp. 290–300, 2016.

[2] M. L. Damiani and C. Cuijpers, "Privacy challenges in third-party location services," in *IEEE 14th International Conference on Mobile Data Management*, pp. 63–66, June 2013.

[3] Y. Gong, C. Zhang, Y. Fang, and J. Sun, "Protecting location privacy for task allocation in ad hoc mobile cloud computing," *IEEE Transactions on Emerging Topics in Computing*, vol. 6, no. 1, pp. 110–121, 2018.

[4] Q. He, D. P. Wu, and P. Khosla, "The quest for personal control over mobile location privacy," *IEEE Communications Magazine*, vol. 42, no. 5, pp. 130–136, 2004.

[5] M. Hou, H. Zhang, and Y. Wang, "OFC: An approach for protecting location privacy from location provider in location-based services," in *IEEE Third International Conference on Data Science in Cyberspace (DSC'18)*, pp. 917–922, June 2018.

[6] M. S. Hwang, E. F. Cahyadi, H. W. Yang, and C. Y. Yang, "An improvement of the remote authentication scheme for anonymous users using elliptic curves cryptosystem," *IEEE 4th International Conference on Computer and Communications (ICCC'18)*, 2018. DOI: 10.1109/CompComm.2018.8780891.

[7] M. S. Hwang, C. C. Lee, Y. C. Lai, "An untraceable blind signature scheme", *IEICE Transactions on Foundations*, vol. E86-A, no. 7, pp. 1902–1906, July 2003.

[8] C. C. Lee, M. S. Hwang, and I. E. Liao, "Security enhancement on a new authentication scheme with anonymity for wireless environments," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 1683–1687, 2006.

[9] C. C. Lee, M. S. Hwang, and W. P. Yang, "A new blind signature based on the discrete logarithmproblem for untraceability," *Applied Mathematics and Computation*, vol. 164, no. 3, pp. 837–841, 2005.

[10] L. Li, R. Lu, and C. Huang, "EPLQ: Efficient privacy-preserving location-based query over outsourced encrypted data," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 206–218, 2016.

[11] H. Li, L. Sun, H. Zhu, X. Lu, and X. Cheng, "Achieving privacy preservation in wifi fingerprint-based localization," in *IEEE Conference on Computer Communications*, pp. 2337–2345, Apr. 2014.

[12] J. Liao, Y. H. Qi, P. W. Huang, M. T. Rong, and S. H. Li, "Protection of mobile location privacy by using blind signature," *Journal of Zhejiang University Science*, vol. 7A, no. 6, pp. 984–989, 2006.

[13] G. Natesan and J. G. Liu, "An adaptive learning model for k-anonymity location privacy protection," in *IEEE 39th Annual Computer Software and Applications Conference*, 2015. DOI: 10.1109/COMPSAC.2015.281.

[14] T. Peng, Q. Liu, and G. Wang, "Enhanced location privacy preserving scheme in location-based services," *IEEE Systems Journal*, vol. 11, no. 1, pp. 219–230, 2017.

[15] J. H. Qin and H. L. Luo, "User privacy disclosure and protection in location-based services (in chinese)," *Computer Programming Skills and Maintenance*, vol. 08, no. 113–114, 2015.

[16] J. Son, D. Kim, M. Z. A. Bhuiyan, R. Tashakkori, J. Seo, and D. H. Lee, "Privacy enhanced location sharing for mobile online social networks," *IEEE Transactions on Sustainable Computing*, vol. 5, no. 2, pp. 279–290, 2018.

[17] Q. G. Song and J. Wang, "Wifi location fingerprinting positioning method of privacy protection technology research and implementation," Technical Report, June 2017.

[18] M. A. Talouki, A. B. Dastjerdi, and N. Movahedinia, "Blindlocation: Supporting user location privacy using blind signature," in *The 7th International Conference on Computer and Knowledge Engineering (ICCKE'17)*, pp. 53–59, Oct. 2017.

[19] L. Y. Wang, D. Q. Yang, and X. Han, "Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 627–636, 2017.

[20] Y. H. Wang, H. L. Zhang, and X. Z. Yu, "KAP: Location privacy-preserving approach in location services (in chinese)," *Journal of Communications*, vol. 35, no. 11, pp. 182–190, 2014.

[21] S. S. Wu, J. B. Xiong, G. H. Ye, and Z. Q. Yao, "Research on location privacy protection based on fake location in mobile internet environment (in chinese)," *Information Network Security*, vol. 10, no. 54–59, 2016.

[22] R. Y. Yu, Z. H. Bai, L. Y. Yang, P. F. Wang, and Y. H. Liu, "A location cloaking algorithm based on combinatorial optimization for location-based services in 5G networks," *IEEE Access*, vol. 4, pp. 6515–6527, 2017.

[23] A. Y. Ye, Y. C. Li, and L. Xu, "A novel location privacy-preserving scheme based on l-queries for continuous LBS," *Computer Communications*, vol. 98, no. 1–10, 2016.

[24] X. J. Zhang, X. L. Gui, and Z. D. Wu, "A survey of location service privacy protection (in chinese)," *Journal of Software*, vol. 26, no. 9, pp. 2373–2395, 2015.

[25] D. P. Zhao, J. S. Ma, X. L. Wang, and X. X. Tian, "Personalized location anonymity - A kernel density estimation approach," in *Web-Age Information Management*, pp. 52–64, June 2016.

[26] L. J. Zheng, H. H. Yue, L. H. Zhang, and X. Pan, "A new location privacy protection algorithm," *IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing*, 2017. DOI: 10.1109/CSE-EUC.2017.253.

[27] Y. J. Zhu, "Analysis and assessment of privacy disclosure risks in location-based services (in chinese)," *Guizhou University*, vol. 03, no. 1–73, 2016.

[28] H. Zhu, F. Liu, and H. Li, "Efficient and privacy-preserving polygons spatial query framework for location-based services," *IEEE Internet Things*, vol. 4, no. 2, pp. 536–545, 2017.

Biography

Xin Xu Graduate. College of Computer Science and Technology in Shanghai University of Electric Power. He research interests mainly focus on the security and privacy protection for the location.

Mi Wen(M'10) Received the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2005, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008, both in computer science. She is currently an Associated Professor of the College of Computer Science and Technology with Shanghai University of Electric Power, Shanghai, China. From May 2012 to May 2013, she was a Visiting Scholar at the University of Waterloo, Waterloo, ON, Canada. Her research interests include privacy preserving in wireless sensor network, smart grid, etc. Dr. Wen serves as an Associate Editor of Peer-to-Peer Networking and Applications (Springer). She acts as the TPC Member of some flagship conferences such as the IEEE INFOCOM, the IEEE ICC, and the IEEE GLOBECOM.

Liangliang Wang Received his Ph.D. degree from Shanghai Jiao Tong University, in 2016. Currently, he is a assistant professor in College of Computer Science and Technology, Shanghai University of Electric Power. His research interests include information security and smart grid, etc.

Comments on A Remote User Authentication Scheme for Multi-server 5G Networks

Jiaqing Mo and Zhongwang Hu
(Corresponding author: Jiaqing Mo)
School of Computer Science and Software, Zhaoqing University, Zhaoqing, Guangdong Province, 526061, China
Email: mojiaqing@126.com
(Received June 2, 2020; Revised and Accepted Feb. 11, 2021; First Online June 5, 2021)

Abstract

Recently, Ying-Nayak proposed a lightweight and untraceable authentication scheme for multi-server making use of self-certified public key cryptography (LASPC) based on elliptic curve cryptography in 5G networks environment. With provision of a formal security proof and an informal security analysis, they stated that their protocol is secure and robust enough to resist known attacks. Nevertheless, in this work, we show that their proposal is unable to defend against replay attack, traceability attack, and smart card loss attack, which results in user impersonation attack. Furthermore, some necessary countermeasures are provided to address these issues. Besides, we point out that an inherent design flaw makes their scheme fail to provide mutual authentication, which prevents their scheme from putting into practice.

Keywords: Design Flaw; Multi-server; Mutual Authentication; Replay Attack; Smart Card Loss Attack

1 Introduction

5G communication technology has become the core infrastructure of industrial Internet system in the future because of its ability to achieve wider coverage, greater bandwidth, higher rates, more access capacity, and achieve the effect of low delay and high reliability [13,17,20]. Accordingly, due to the diversity of services and the rapid growth of data that need to be stored, the single server architecture with limited storage and communication is no longer suitable for 5G wireless networks environment. To overcome these defects, a multi-server structure suitable for mobile users to access anytime and anywhere has emerged [8,17,21]. Because the wireless network channel is open, the attackerss can intercept, eavesdrop, modify and replay messages on wireless channels, and launch various attacks [2,18]. Very recently, to ensure communication security and user privacy in multi-server architecture, Ying-Nayak [19] proposed a lightweight user authentication scheme employing self-certified public key cryptography (LASPC) based on elliptic curve cryptography [4,11,12], which emphasizes mutual authentication between the user and server can be achieved without the participation of the third party in 5G wireless networks environment. Furthermore, they also provided a security proof via the random oracle model and some security analysis, for which they were confident that their proposal is able to thwart various attacks with some admirable security attributes.

However, when scrutinizing Ying-Nayak's LASPC scheme, it is regretful to find that the LASPC scheme is not as robust as they stated. We demonstrate that the LASPC scheme is vulnerable to replay attack, traceability attack, smart card loss attack, which may lead to user impersonation attack. Moreover, we find that their LASPC scheme has a fatal design flaw for which LASPC fails to provide mutual authentication between the user and the server during the mutual authentication phase.

The rest of this paper is prepared as follows: We present the adversary model in Section 2, and review the LASPC scheme in Section 3; In Section 4, we make a cryptanalysis of LASPC scheme and show its vulnerabilities. Finally, we draw a conclusion in Section 5.

2 Adversary Model

When cryptanalyzing a user authentication protocol, it is necessary to define the adversary model to suppose the capacities of the attacker. Based on the previous works, the adversary model is illustrated in the following items.

1) The attacker can fully control the open channel, *i.e.* the attacker can eavesdrop, intercept, modify, delete, replay the messages exchanged among the communication participants over the public channel [1, 3, 5, 7, 16].
2) The attacker can somehow obtain the lost/stolen smart card and extract its secret data using analysis methods in [6, 10].
3) The attacker knows how to perform the user authentication protocol, which means that the protocol is public [1, 9].

3 Review of the LASPC Scheme

In this section, we brief Ying-Nayak's LASPC scheme, which is made up of four phases: Setup phase, registration phase, mutual authentication phase, and password change phase. We skip the password change phase because it is not related to our work. For ease of description, we list symbols involved in LASPC in Table 1.

Table 1: Notations

Symbol	Description
RC	Registration center
s_{rc}	RC's private key
Pub_{rc}	RC's public key
U_i	i^{th} user
ID_i	U_i's identity
S_j	j^{th} server
ID_j	S_j's identity
PW_i	U_i's password
$\oplus$	Operation of bitwise XOR
$\|\|$	Operation of concatenation
$\Rightarrow$	The secure channel
$\rightarrow$	The insecure channel

3.1 Setup Phase

RC picks up two large prime numbers p and q, a non-regular elliptical curve E, a generator P of E with order q of the group G, the system's private key $s_{rc} \in Z_q^*$, and computes the corresponding public key $Pub_{rc} = s_{rc}P$. Afterwards, RC chooses two one-way hash function $H_1(), H_2()$, and publishes parameters $\{P, p, q, Pub_{rc}, H_1(), H_2()\}$ and keeps s_{rc}.

3.2 User Registration Phase

The user U_i registers in RC as follows.

1) $U_i \Rightarrow RC : \{ID_i, \beta_i\}$, where $\beta_i = H_l(ID_i \| PW_i \| \mathrm{c}_0)$, c_0 is a random number.
2) $RC \Rightarrow U_i : \{\varphi_i, N_i, M_i\}$, where $\{\varphi_i = H_2(\beta_i \| A_i)$, $A_i = c_1 P, N_i = A_i \oplus H_l(ID_i)$, $M_i = H_2(A_i) s_{rc} \bmod q + c_1$, and $c_1 \in Z_q^*$ is a random number.
3) U_i stores $\{\varphi_i, N_i, M_i, c_0\}$ into the smart card.

3.3 Server Registration Phase

To obtain a private key, S_j registers in RC as follows.

1) $\mathrm{S_j} \Rightarrow RC : \{ID_j\}$.
2) $RC \Rightarrow S_j : \{B_j, D_j\}$, where $B_j = c_2 P, D_j = H_2(ID_j \| B_j) s_{rc} \bmod q + c_2$, and c_2 is a random nonce.
3) $S_j : S_j$ computes $D_j P$, and checks if $D_j P = H_2(ID_j \| B_j) Pub_{rc} + B_j$. If it is true, S_j keeps $\{D_j, B_j\}$.

3.4 Mutual Authentication Phase

1) $U_i \rightarrow S_j : \{\delta_i, DID_i, A_i^*, F_i\}$. U_i inputs ID_i^* and PW_i^*, the smart card calculates $\beta_i^* = H_I(ID_i^* \| PW_i^* \| c_0), A_i^* = N_i \oplus H_1(ID_i^*)$. If $\varphi_i = H_2(\beta_i \| A_i^*)$, selects two random nonces $c_3, c_4 \in Z_q^*$, computes $DID_i = H_1(ID_i^* \| c_3), F_i = c_4 P$, $R_i = H_2(DID_i \| F_i), \delta_i = M_i + R_i c_4 \bmod q$, and sends $\{\delta_i, DID_i, A_i^*, F_i\}$ to S_j. Otherwise, the card aborts the session.
2) $S_j \rightarrow U_i : \{\delta_j, ID_j, B_j, F_j\}$. S_j verifies the condition $\delta_i P = H_2(A_i^*) Pub_{rc} + A_i^* + H_2(DID_i \| F_i) F_i$. If it holds, S_j selects a random number $c_5 \in Z_q^*$, computes $F_j = c_5 P, R_j = H_2(DID_i \| ID_j \| B_j), \delta_j = D_j s_{rc} \bmod q + R_j c_j$, and sends $\{\delta_j, ID_j, B_j, F_j\}$ to U_i. Otherwise, S_j rejects the session.
3) U_i checks if $\delta_j P \stackrel{?}{=} H_2(ID_j \| B_j) Pub_{rc} + B_j + H_2(DID_i \| ID_j \| B_j) F_j$. If it holds, U_i and S_j build a session key that is shared between them.

4 Cryptanalysis of LASPC Scheme

This section shows that Ying-Nayak's LASPC scheme for multi-server in 5G wireless networks suffers from replay attack, traceability attack, smart card attack. In addition, this section demonstrates that there is a serious design flaw in LASPC scheme.

4.1 Replay Attack

It is very important for a user authentication scheme to withstand replay attack. Unfortunately, LASPC scheme cannot resist replay attack. If an attacker captures the login message $\{\delta_i, DID_i, A_i^*, F_i\}$ and retransmits it to server S_j the latter will pass the verification by checking $\delta_i P? = H_2(A_i^*) Pub_{rc} + A_i^* + H_2(DID_i \| F_i) F_i$, then selects a new random number $c_5' \in Z_q^*$ and performs the subsequent process of LASPC. This is because server S_j does not have a mechanism to check if the message is fresh, which eventually results in the server's computing resources being consumed. Though Ying-Nayak argues that the replay messages can be found since the server will check the validation of δ_i due to the freshness of c_4. To achieve this purpose, the server needs to check whether the messages are fresh by maintaining a table that stores all the received messages from various users, which makes their scheme impractical because the number of messages increases very quickly in 5G environment, and the query process in the database will take a lot of time.

Remark 1. *To resist replay attack, it is necessary to adopt a mechanism which is used to verify whether the messages are fresh, specifically with timestamp technique. Meanwhile, to prevent an attacker from changing the timestamp in the retransmitted message at will, the hash function also needs to be used to protect the timestamp. If*

the attacker tries to retransmit the previous messages to launch replay attack, the receiver can hinder this attack via checking the timestamp and the hashed authenticator.

4.2 Traceability Attack

User untraceability is significant for an authentication protocol in protecting user privacy. It means that the attacker cannot identify which user sent the message, nor can he confirm messages are from the same user. In 5G wireless networks, the user transmits their messages by means of broadcast. Thus, the user untraceability is a security requirement that cannot be ignored. However, in LASPC scheme, we observe the user's login request message contains a user-specific parameter A_i^*, which is transmitted over the public channel. Because A_i^* is a fixed value and it is different from the Ai generated by RC for other users when they register in RC, the attacker can identify the request message containing A_i^* from a large number of messages in the public channel, and links them to a specific user. The malicious attacker even will analyze the user's access time and behaviors to gain a better understanding of the user's habits for other terrible purposes, and so on. Thus, in this regard, LASPC is prone to traceability attack.

Remark 2. *To thwart the traceability attack, the authentication scheme needs to be able to prevent the attacker from distinguishing whether these messages are sent by the same user from different authentication sessions. Though Ying-Nayak use a dynamic identity $DID_i = H_1(ID_i^*||c_3)$ in the login message to confuse the attacker, parameter A_i is never changed in every login request, and therefore the attacker will get to know that the messages are transmitted by the same user, helping him easily trace the user. To cope with this issue, their scheme needs using encryption technology to prevent attackers from identifying particular parameters. It can also prevent the traceability attack by using $\oplus$ on A_i and a secret data shared between U_i and S_j to conceal A_i.*

4.3 Smart Card Loss Attack

In Ying-Nayak's LASPC scheme, they argue that their protocol can resist known threats even if the lost/stolen card is acquired by the attacker. However, we demonstrate that the LASPC scheme suffers from smart card loss attack which leads to impersonation attack.

Sopposed that an attacker has captured the user's login request $\{\delta_i, DID_i, A_i^*, F_i\}$ in the process of mutual authentication phase, and the stolen/lost smart card is somehow acquired by the attacker, he can launch an offline identity guessing attack as follows.

Step 1. The attacker extracts the secret information $\{\varphi_i, N_i, M_i, c_0\}$ from the card.

Step 2. The attacker selects an item $ID_i{}^*$ from the dictionary space D_{ID}.

Step 3. The attacker verifies the correctness of ID_i by checking the condition $H_1(ID_i^*)? = N_i \oplus A_i^*$. If it is true, the attacker has found the correct identity of the user U_i. Otherwise, the attacker repeats Steps $2-3$ until $H_1(ID_i^*) = N_i \oplus A_i^*$.

The time complexity of the above attack is $\mathrm{O}(T_h * |D_{ID}|)$, where T_h is the running time of hash function and $|D_{ID}|$ is the size of D_{ID}. Because T_h is negligible, the time required for the attacker to perform the above attack process is linear to $\mathrm{O}(|D_{ID}|)$. According to [14,15], $|D_{ID}|$ is rather limited in practice with $|D_{ID}| \leq 10^6$, it is very effective for an attacker to perform an offline identity guessing attack.

Once the attacker acquires the user's identity ID_i, he may conduct user impersonation attack with the fixed parameter as follows:

Step 1. The attacker selects two random nonces c_3' and c_4', and computes a dynamic identity $DID_i' = H_1(ID_i \| c_3')$, $F_i' = c_4' P, R_i' = H_2(DID_i' \| F_i')$.

Step 2. The attacker computes a certificate $\delta_i' = M_i + R_i' c_4'$ mod q, and sends $\{\delta_i', DID_i', A_i{}^*, F_i'\}$ to S_j

On receipt of the login request message, S_j checks whether $\delta_i' P = H_2(A_i^*) Pub_{rc} + A_i^* + H_2(DID_i' \| F_i') F_i'$. It is clear that the condition is valid, so the server S_j treats the attacker as a legitimate user and performs the subsequent procedure to try to build a session key with him.

We can see that the LASPC scheme will suffer from smart card loss attack if the card is obtained by the attacker, who can conduct user impersonation attack successfully even without knowing the user's password.

Remark 3. *Actually, the root cause of smart card loss attack on the LASPC scheme is also related to the use of the fixed value of A_i in U_i 's login message. Thus, the solution to the traceability attack in Remark 2 is also suitable for this smart card loss attack.*

4.4 Design Flaw

In LASPC scheme, during the mutual authentication phase, when U_i receives the response message $\{\delta_j, ID_j, B_j F_j\}$, it will verify the legitimacy of server S_j by checking if $\delta_j P = H_2(ID_j \| B_j) Pub_{rc} + B_f + H_2(DID_i \| ID_j \| B_j) F_j$. If the condition is true, U_i will pass the legitimacy authentication of the server and performs the subsequent procedure. However, we find that U_i will always reject server S_j because the equation never holds, which is shown as follows:

From the left side of the equation, we conclude that

$$\begin{aligned}\delta_j P &= (D_j s_{rc} \bmod q + R_j c_5) P \\ &= D_j s_{rc} P \bmod q + R_j c_5 P \\ &= (H_2(ID_j \| B_j) s_{rc} \bmod q + c_2) s_{rc} P \bmod q \\ &\quad + H_2(DID_i \| ID_j \| B_j) c_5 P \\ &= (H_2(ID_i \| B_j) s_{rc} P \bmod q + c_2 P) s_{rc} \bmod q \\ &\quad + H_2(DID_i \| ID_i \| B_j) F_i \\ &= (H_2(ID_j \| B_j) Pub_{rc} + B_j) s_{rc} \\ &\quad + H_2(DID_i \| ID_j \| B_j) F_j\end{aligned}$$

And the right side of the equation is $H_2(ID_j \| B_j) Pub_{rc} + B_j + H_2(DID_i \| ID_j \| B_j) F_j$.

Obviously, $(H_2(ID_j \| B_j) Pub_{rc} + B_j) s_{rc}$ is not equal to $H_2(ID_j \| B_j) Pub_{rc} + B_j$, thus the equation does not hold.

From these analysis, it is evident that user U_i always fails to verify the validity of server S_j, so LASPC scheme cannot provide mutual authentication between U_i and S_j during the mutual authentication phase. As a result, a shared session key that guarantees communication cannot be negotiated between the user and the server.

Therefore, we believe that there is a fatal inherent design flaw in LASPC scheme.

5 Conclusion

In this paper, we have shown that the LASPC scheme proposed by Ying-Nayak cannot withstand replay attack, traceability attack and smart card loss attack as they claim. An attacker can easily conduct replay attack and traceability attack on their scheme. Furthermore, if the attacker captures the login message and obtains the stolen/lost smart card, he can disclose the user's identity and launch user impersonation attack. We also present some countermeasures to cope with these weaknesses. Besides, we point out that their LASPC scheme has a serious design flaw.

Acknowledgments

The study was supported by the Special Innovation Project of Colleges and Universities in Education Department of Guangdong Province 2019 of China (No. 2019KTSCX198).

References

[1] R. Amin, S. H. Islam, G. Biswas, M. K. Khan, L. Leng and N. Kumar, "Design of an anonymity-preserving three-factor authenticated key exchange protocol for wireless sensor networks," *Computer Networks*, vol. 101, pp. 42–62, 2016.

[2] Z. Guo, "Cryptanalysis of a certificateless conditional privacy-preserving authentication scheme for wireless body area networks," *International Journal of Electronics and Information Engineering*, vol. 11, no. 1, pp. 1–8, 2019.

[3] M. S. Hwang, E. F. Cahyadi, Y. C. Chou and C. Y. Yang, "Cryptanalysis of kumar's remote user authentication scheme with smart card," in *The 14th International Conference on Computational Intelligence and Security (CIS'18)*, pp. 416–420, 2018.

[4] M. S. Hwang, E. F. Cahyadi, C. Y. Yang and S. F. Chiou, "An improvement of the remote authentication scheme for anonymous users using an elliptic curve cryptosystem," in *IEEE 4th International Conference on Computer and Communications (ICCC'18)*, pp. 1872–1877, 2018.

[5] M. S. Hwang, Li-Hua Li, "A New Remote User Authentication Scheme Using Smart Cards", *IEEE Transactions on Consumer Electronics*, vol. 46, no. 1, pp. 28–30, Feb. 2000.

[6] T. H. Kim, C. Kim and I. Park, "Side channel analysis attacks using am demodulation on commercial smart cards with seed," *Journal of Systems & Software*, vol. 85, no. 12, pp. 2899–2908, 2012.

[7] C. C. Lee, C. H. Liu, M. S. Hwang, "Guessing attacks on strong-password authentication protocol", *International Journal of Network Security*, vol. 15, no. 1, pp. 64–67, 2013.

[8] T. W. Lin and C. L. Hsu, "Anonymous group key agreement protocol for multi-server and mobile environments based on chebyshev chaotic maps," *Journal of Supercomputing*, vol. 74, no. 9, pp. 4521–4541, 2018.

[9] C. H. Ling, C. C. Lee, C. C. Yang, and M. S. Hwang, "A secure and efficient one-time password authentication scheme for WSN", *International Journal of Network Security*, vol. 19, no. 2, pp. 177-181, Mar. 2017.

[10] S. Mangard, E. Oswald and T. Popp, *Power Analysis Attacks: Revealing the Secrets of Smart Cards*, 2007. ISBN:978-0-387-30857-9.

[11] J. Mo and H. Chen, "A lightweight secure user authentication and key agreement protocol for wireless sensor networks," *Security and Communication Networks*, vol. 2019, 2019. (`https://doi.org/10.1155/2019/2136506`)

[12] J. Mo, Z. Hu, H. Chen and W. Shen, "An efficient and provably secure anonymous user authentication and key agreement for mobile cloud computing," *Wireless Communications and Mobile Computing*, vol. 2019, pp. 1–12, 2019.

[13] S. Shin and T. Kwon, "A privacy-preserving authentication, authorization, and key agreement scheme for wireless sensor networks in 5g-integrated internet of things," *IEEE Access*, vol. 8, no. 99, pp. 67555–67571, 2020.

[14] D. Wang and P. Wang, "Two birds with one stone: Two-factor authentication with security beyond conventional bound," *IEEE Transactions on Dependable & Secure Computing*, vol. 15, no. 4, pp. 708–722, 2016.

[15] D. Wang and P. Wang, "Understanding security failures of two-factor authentication schemes for real-time applications in hierarchical wireless sensor networks," *Ad Hoc Networks*, vol. 20, no. 2, pp. 1–15, 2014.

[16] C. Wang, G. Xu and J. Sun, "An enhanced three-factor user authentication scheme using elliptic curve cryptosystem for wireless sensor networks," *Sensors*, vol. 17, no. 12, pp. 2946, 2017.

[17] A. M. K. Wong, C. L. Hsu, T. V. Le, M. C. Hsieh and T. W. Lin, "Three-factor fast authentication scheme with time bound and user anonymity for multi-server e-health systems in 5g-based wireless sensor networks," *Sensors*, vol. 20, no. 9, pp. 2511, 2020.

[18] H. W. Yang, J. Z. Lee and M. S. Hwang, "A taxonomy of bluetooth security," *International Journal of Electronics and Information Engineering*, vol. 12, no. 2, pp. 43–65, 2020.

[19] B. Ying and A. Nayak, "Lightweight remote user authentication protocol for multi-server 5g networks using self-certified public key cryptography," *Journal of Network and Computer Applications*, vol. 131, pp. 66–74, 2019.

[20] Y. Zhang, R. Deng, E. Bertino and D. Zheng, "Robust and universal seamless handover authentication in 5g hetnets," *IEEE Transactions on Dependable and Secure Computing*, 2019. (`https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8758145`)

[21] S. Zhou, Q. Gan and X. Wang, "Authentication scheme based on smart card in multi-server environment," *Wireless Networks*, vol. 26, no. 2, pp. 855–863, 2020.

Biography

Jiaqing Mo is an associate professor in school of computer science and software, Zhaoqing University, China. He has been published more than twenty papers in journals and conferences. His main research interests include information security, wireless mobile communication, trusted computing. Contact him at mojiaqing@126.com.

Zhongwang Hu joined the school of computer at Zhaoqing University in July 2009 as a professor. He holds a M.S. degree in computer science from National University of Defense Technology of China in 2003. His current research includes network security, mobile computing. Contact him at zhongwanghu@126.com.

An Independent Variable Swinging Coupled Chaotic System for a Pseudorandom Bit Generator

Qi Wu

(Corresponding author: Qi Wu)

Department of Electronic Commerce, Jiangxi University of Finance & Economics
Shuanggang Street, 330013 Nanchang, China
Email: wuqiocjzd@126.com

(Received May 1, 2020; Revised and Accepted Feb. 10, 2021; First Online Aug. 15, 2021)

Abstract

Though coupling is a simple method of constructing high-dimensional chaotic systems, it's forsaken by most chaos researchers. We then proposed a novel way of coupling, i.e., Independent Variable Coupling, in our former papers. In this paper, we introduce Independent Variable Attracting Coupling and Independent Variable Pushing Coupling based on the results obtained from applying the coupling function. Here, we remedy it with a missing piece, i.e., Independent Variable Swinging Coupling. After applying it to skew-tent mapping to form a new 2D chaotic system, we devise a corresponding Pseudorandom bit Generator, whose performance is the 3rd best compared with the prior research outcomes and can thus be applied to some other scenarios.

Keywords: Chaotic System; Independent Variable Swinging Coupling; Pseudorandom Bit Generator

1 Introduction

In 1949, Shannon proposed that cryptosystems should possess two properties, confusion and diffusion [1]. Chaotic systems move so irregularly that patterns could hardly be recognized, which is similar to confusion. Chaotic systems are extremely keen to parameters and initial values that minute variation in them will be amplified enormously, which is analogous to diffusion. Therefore, many efforts have been made to apply chaos to cryptology, one of which is designing pseudorandom bit generators (usually abbreviated as PRBG) based on chaotic systems [2–18].

Coupling is an easy approach for constructing new chaotic systems [2–17]. In [15], methods of coupling are divided into 4 categories, i.e. perturbation coupling [2–5], independent variable-dependent variable coupling [6], dependent variable coupling [7–12], and independent variable coupling [13–17]. As mentioned previously, the former two are deemed suitable for chaos control, while the latter two proper for chaos anti-control, which is made available for cryptographic applications. Although most researchers nowadays propose new high dimensional chaotic systems directly instead of starting from low dimensional ones step by step, I still insist that coupling shouldn't be forsaken. On the contrary, due attention should be paid to coupling which is a stable method of constructing high dimensional chaotic systems.

In this paper, the author propose a new kind of independent variable coupling, i.e. independent variable swinging coupling, then apply it to skew-tent mapping to obtain a 2D discrete chaotic system. Afterwards, a PRBG is devised based on it. Results of statistical tests testify that the generated sequences hold good pseudorandomness. At last, linear complexity and cipher space are calculated as well. All the experiments illustrate that the proposed PRBG is a wonderful candidate.

The upcoming parts of this paper are organized as follows. In Section 2, a chaotic system is constructed and analyzed. In Section 3, we devise a PRBG and analyze the pseudorandomness of generated sequences via 5 statistic tests and the linear complexity is computed as well. In Section 4, the cipher space is calculated. Section 5 concludes.

2 Independent Variable Swinging Coupled Chaotic System

First, let's review the concept of independent variable coupling.

Given two skew tent mappings defined as:

$$x_(i+1) = f_\alpha(x_i) = \begin{cases} \frac{x_i}{\alpha} & x_i \in [0,\alpha] \\ \frac{1-x_i}{1-\alpha} & x_i \in (\alpha,1] \end{cases} \quad (1)$$

$$y_(i+1) = f_\alpha(y_i) = \begin{cases} \frac{y_i}{\alpha} & y_i \in [0,\alpha] \\ \frac{1-y_i}{1-\alpha} & y_i \in (\alpha,1] \end{cases} \quad (2)$$

Importing a coupling function g, independent variable coupling applies g first and f next, i.e. the new chaotic system becomes

$$\begin{cases} x_{i+1} = f_\alpha(g_\mu(x_i, y_i)) \\ y_{i+1} = f_\alpha(g_\mu(y_i, x_i). \end{cases} \tag{3}$$

Here, μ determines the extent of cohesion (or coupling, as could be seen in the following equations).

In [13], g (the subscript μ is sometimes omitted, if without ambiguity) is defined as:

$$g_1(x_i, y_i) = \mu x_i + (1-\mu) y_i. \tag{4}$$

In [14], g is defined as:

$$g_2(x_i, y_i) = x_i^{\mu} y_i^{(1-\mu)}. \tag{5}$$

In [15], g is defined as:

$$g_3(x_i, y_i) = \frac{1}{\frac{\mu}{x_i} + \frac{1-\mu}{y_i}} \tag{6}$$

Obviously, the above 3 functions are weighted arithmetic, geometric, and harmonic average respectively. No matter what x_i and y_i are, the output of g comes between x_i and y_i. Informally, we could say that x_i and y_i attract each other after applying g. We may call this category of coupling Independent Variable Attracting Coupling.

In [16], g is defined as:

$$g_4(x_i, y_i) = \begin{cases} x_i + \frac{1-x_i}{x_i} y_i \mu & x_i \geq y_i \\ x_i - \frac{x_i}{1-x_i}(1-y_i)\mu & x_i < y_i \end{cases} \tag{7}$$

In [17], g is defined as:

$$g_5(x_i, y_i) = \begin{cases} x_i + \frac{1-x_i}{x_i}(x_i - y_i)\mu & x_i \geq y_i \\ x_i - \frac{x_i}{1-x_i}(y_i - x_i)\mu & x_i < y_i \end{cases} \tag{8}$$

Apparently, no matter what x_i and y_i are, the output of $g4/g5$ goes outside the interval (x_i, y_i) or $({}_y i, x_i)$: when $y_i \in [0, x_i]$, $g4/g5$ lies in $[x_i, 1]$; when $y_i \in (x_i, 1]$, $g4/g5$ lies in $[0, x_i]$. Informally, we could say that x_i and y_i exclude each other after applying $g4/g5$. We may call this category of coupling Independent Variable Pushing Coupling.

In this paper, we propose a new coupling function $g6$, defined as:

$$g_6(x_i, y_i) = \begin{cases} y_i - \frac{1-x_i}{1-y_i} y_i \mu & x_i \geq y_i \\ y_i + \frac{x_i}{y_i}(1-y_i)\mu & x_i < y_i \end{cases} \tag{9}$$

As same as $g4/g5$, the output of $g6$ will go outside the interval (x_i, y_i) or (y_i, x_i). The difference is when $y_i \in [0, x_i]$, $g6$ lies in $[0, y_i]$; when $y_i \in (x_i, 1]$, $g6$ lies in $[y_i, 1]$. Informally, we could say that y_i swings x_i to the other side. (Imagine that, x_i and y_i are two persons standing side by side. Now, y_i pulls x_i so hard that x_i goes to the other side of y_i. As far as I know, swing is the best verb for expressing this.) We may call this category of coupling Independent Variable Swinging Coupling. Therefore, we name the new chaotic system as Independent Variable Swinging Coupled Chaotic System (abbreviated as IVSCCS hereafter). Next, we analyze its chaotic properties via bifurcation diagram & Lyapunov exponent spectrum, respectively.

Let $x_0 = 0.6$, $y_0 = 0.3$, $\alpha = 0.5$, μ goes from 0 to 1 with step 0.001. (The values are set as same as my previous works [10–12, 16, 17]). For the 1001 parameters, iterate IVSCCS 500 times, filtering the first 200 times, and depict the values of x for the last 300 times, as shown in Figure 1.

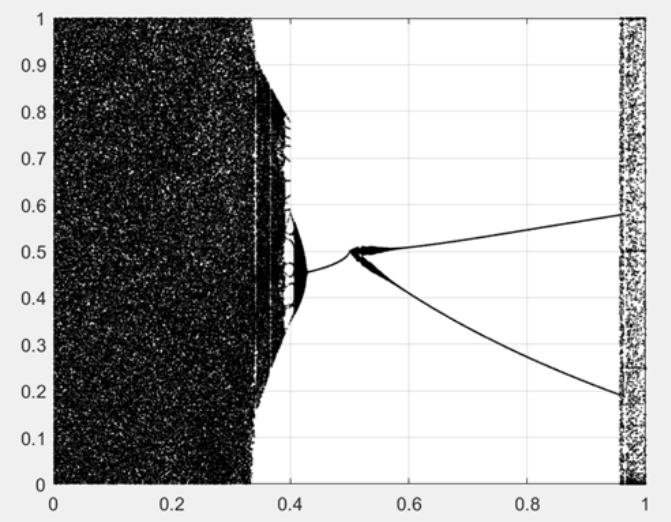


Figure 1: Bifurcation diagram of IVSCCS (μ as horizontal axis)

From Figure 1, it could be seen that there is no apparent periodic area for μ in intervals [0, 0.33] & [0.96, 1], where possibility for cryptographic applications may exist. Let $x_0 = 0.6$, $y_0 = 0.3$, $\mu = 0.2$, α goes from 0 to 1 with step 0.001. For the 1001 parameters, iterate IVSCCS 500 times, filtering the first 200 times, and depict the values of x for the last 300 times, as shown in Figure 2.

From Figure 2, it could be seen that there is no apparent periodic for the entire scope of α, which is superb for cryptographic applications.

Next, recall the definition for Lyapunov exponent of 2D discrete dynamic system. Given a 2D discrete dynamic system defined as:

$$\begin{cases} x_{i+1} = f_1(x_i, y_i) \\ y_{i+1} = f_2(x_i, y_i). \end{cases} \tag{10}$$

Its corresponding Jacobian matrix is

$$J_i = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} & \frac{\partial f_1}{\partial y_i} \\ \frac{\partial f_2}{\partial x_i} & \frac{\partial f_2}{\partial y_i} \end{bmatrix} \tag{11}$$

then the Lyapunov exponents for System (10) are

$$\lambda = lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \ln |\mu_i|, \tag{12}$$

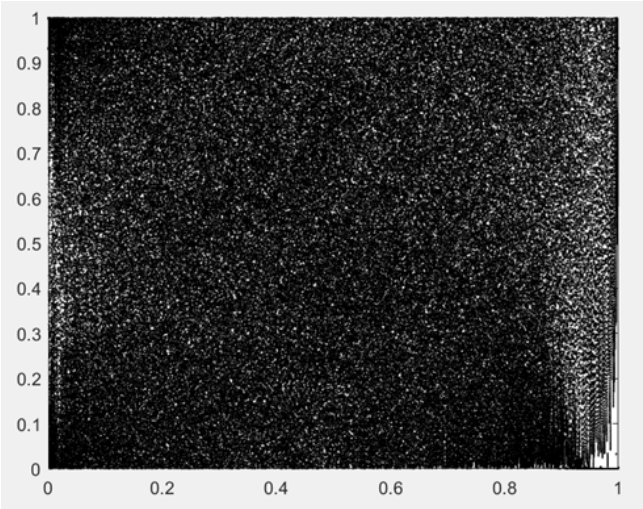


Figure 2: Bifurcation diagram of IVSCCS (α as horizontal axis)

where μ_i is the eigenvalue of J_i. J_i has 2 eigenvalues, so System (10) owns 2 Lyapunov exponents, obtained by calculating Equation (12) respectively for each eigenvalue of J_i in every iteration.

Let $x_0 = 0.6$, $y_0 = 0.3$, $\alpha = 0.5$, μ goes from 0 to 1 with step 0.001. For the 1001 parameters, iterate IVSCCS 6000 times, filter the first 5000 times, and depict the Lyapunov exponents for the last 1000 times, as shown in Figure 3.

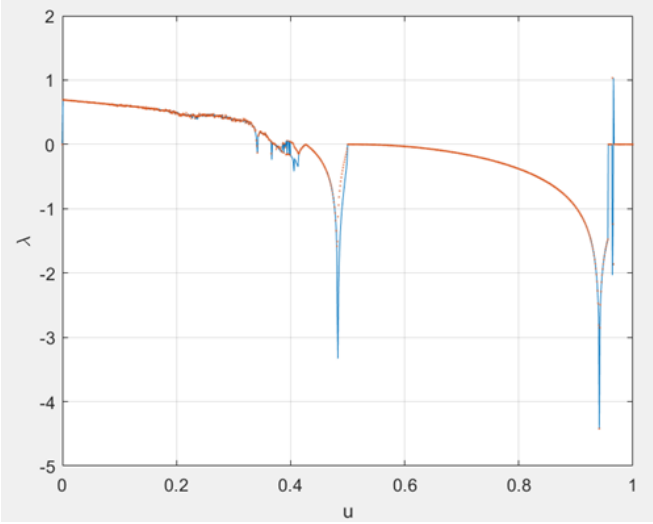


Figure 3: Lyapunov exponent spectrum of IVSCCS (μ as horizontal axis)

From Figure 3, it could be seen that both Lyapunov exponents are positive for μ in interval [0, 0.34], where candidates for cryptographic applications might reside. Let $x_0 = 0.6$, $y_0 = 0.3$, $\mu = 0.2$, α goes from 0 to 1 with step 0.001. For the 1001 parameters, iterate IVSCCS 6000 times, filter the first 5000 times, and depict the Lyapunov exponents for the last 1000 times, as shown in Figure 4.

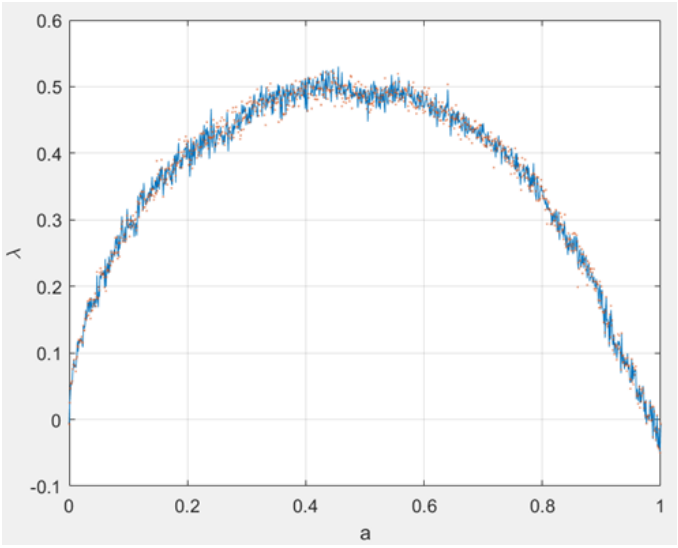


Figure 4: Lyapunov exponent spectrum of IVSCCS (α as horizontal axis)

From Figure 4, it could be seen that for most scope of α, both Lyapunov exponents are positive, which is excellent for cryptographic applications.

Next, we devise a PRBG based on IVSCCS.

3 Design & Analysis of PRBG Based on IVSCCS

Here, we follow the framework of [18] to design a simple PRBG. Given x_0, y_0, α, μ, after each iteration of IVSCCS, emit a bit si via comparing x_i and y_i:

$$s_i = \begin{cases} 0 & x_i < y_i \\ 1 & x_i \geq y_i \end{cases} \tag{13}$$

When both α and μ go from 0 to 1 with step 0.0001, for the 100020001 pairs of parameters, 39929 $\alpha - \mu$ ones pass all 5 tests for pseudo randomness.

Let's take a look at where this result dwells in all my works (Table 1).

The result is the 3rd best achieved thus far (worse than 73548 in [17] and 73097 in [11]).

Next, we illustrate the results of tests under level of significance 0.05 for 3 sequences of length 50000 generated when $x_0 = 0.49$, $y_0 = 0.55$, α and μ set to (0.483, 0.112), (0.491,0.053), (0.511,0.002), respectively (see Table 2 $\sim$ Table 7). Readers unfamiliear with those tests could refer to [10–17] for some basic knowledge.

As BM algorithm is too time-consuming [19], when calculating the linear complexity, we reduce the length of the 3 sequences to 1000 bits, with other conditions unchanged.

Table 1: Number of qualified $\alpha - \mu$ pairs in my papers

Papers	Qualified $\alpha - \mu$ pairs
[10]	7735
[11]	73097
[12]	38424
[13]	2700
[14]	2800
[15]	7638
[16]	38709
[17]	73548
The proposed	39929

Table 2: Results of monobit test

α	μ	χ^2	Critical Value
0.484	0.108	0.4500	3.84
0.491	0.057	0.5917	
0.511	0.002	1.0765	

Table 3: Results of serial test

α	μ	χ^2	Critical Value
0.484	0.108	4.6271	5.99
0.491	0.057	0.9700	
0.511	0.002	4.7077	

Table 4: Results of poker test

α	μ	χ^2 $(m = 4)$	Critical Value
0.484	0.108	24.4211	25
0.491	0.057	18.3232	
0.511	0.002	12.7194	

Table 5: Results of runs test

α	μ	χ^2	Critical Value
0.484	0.108	26.5634	31.4
0.491	0.057	24.7061	
0.511	0.002	25.4193	

Table 6: Results of auto-correlation test

α	μ	$\lvert\chi^2\rvert$ $(d = 10000)$	Critical Value
0.484	0.108	0.45	1.96
0.491	0.057	0.37	
0.511	0.002	1.22	

Table 7: Results of linear complexity

α	μ	LC	$N/2$
0.484	0.108	500	500
0.491	0.057	501	
0.511	0.002	500	

From Table 2 $\sim$ Table 7, it could be seen that, all these 3 sequences have passed the 5 categories of pseudo-random tests and their linear complexity are close to half of their length (similar to the output ofBinary Symmetric Source). They're quite suitable for cryptographic scenarios that demand pseudorandom numbers, such as image encryption, hash function, etc.

4 Analysis of Cipher Space

When an adversary offenses with precision 10^{-8}, then for my 4 system parameters, the cipher space K is:

$$K = (10^8)^4 = 10^3 2 \approx 2^{106.3}.$$

Namely, attacking the proposed generator brutally is slightly easier than attacking 128-bit AES. When the precision rises, the difficulty for the adversary increases as well.

5 Conclusion

Different from our former efforts [10–17], this paper proposes a novel way for coupling, i.e. Independent Variable Swinging Coupling, which is applied to skew tent mapping to construct a 2-dimensional chaotic system. Bifurcation diagram and Lyapunov exponent spectrum demonstrate that, the system possesses a wide chaotic area and is quite suitable for cryptographic applications. Next, when applied to design of PRBG, it overwhelms most of our results achieved so far. Moreover, it owns large cipher space and could withstand brute force attacks.

Acknowledgments

This research is financially supported by the Science and Technology Project of Provincial Education Department of Jiangxi for Youth (GJJ180288). Thanks go to Professor Zhihong Guan, Meng Jia and Jian Shu for their helpful advice.

References

[1] B. L. Hao, *Starting with Parabolas*, Shanghai Science and Technology Education Press, Shanghai, 2003.

[2] Z. Liu, Y. Tian, Y. Song, "The synchronization condition of linearly coupled chaotic system," *Acta Physica Sinica*, vol. 55, no. 8, pp. 3945-3949, 2006.
[3] G. Bao, N. Mandula, T. Xin, Eredencang, "Dynamic behavior of complete synchronization of coupled chaotic oscillators," *Acta Physica Sinica*, vol. 56, no. 4, pp. 1971-1974, 2007.
[4] H. Yu, Y. Liu, "Synchronization of symmetrically nonlinear-coupled chaotic systems," *Acta Physica Sinica*, vol. 54, no. 7, 2005.
[5] J. Ma, X. Wu, H. Qin, "Realization of synchronization between hyperchaotic systems by using a scheme of intermittent linear," *Acta Physica Sinica*, vol. 62, no. 5, pp. 1-8, 2013.
[6] W. Wu, M. Xia, Q. Pang, Y. Fan, "Simulated study of coupled chaotic systems' spatiotemporal dynamics," *Chinese Journal of Electron Devices*, vol. 30, no. 4, pp. 1384-1386, 2007.
[7] J. Liu, X. Fu, "Spatiotemporal chaotic one-way hash function construction based on coupled tent maps," *Journal on Communications*, vol. 28, no. 6, pp. 30-38, 2007.
[8] J. Liu, "One-way hash function based on integer coupled tent maps and its performance analysis," *Journal of Computer Research & Development*, vol. 45, no. 3, pp. 563-569, 2008.
[9] S. Luo, S. Qiu, X. Chen, "Spatiotemporal chaotic one-way hash function construction based on coupled tent maps," *Journal of Shenzhen University Science and Engineering*, vol. 29, no. 4, pp. 335-340, 2012.
[10] Q. Wu, "A dependent variable harmonically coupled chaotic system for a pseudorandom bit generator," in *Proceedings International Conference on Smart Materials, Intelligent Manufacturing and Automation*, 2018.
[11] Q. Wu, "A dependent variable exclusively coupled chaotic system for a pseudorandom bit generator," in *Proceedings International Conference on Network and Information Systems for Computers*, 2018.
[12] Q. Wu, "A pseudorandom bit generator based on a dependent variable exclusively coupled chaotic system," in *Proceedings International Conference on Intelligent Information Processing*, pp. 11-16, 2018.
[13] Z. W. Tan, Q. Wu, "Study of linearly cross-coupled chaotic systems for a random bit generator," in *Proceedings International Conference on Computational Intelligence and Security*, 2008.
[14] Z. W. Tan, Q. Wu, "Study of exponentially cross-coupled chaotic systems for a random bit generator," in *Proceedings International Symposium on Intelligent Information Technology Application*, pp. 224-227, 2008.
[15] Q. Wu, Z. W. Tan, C. X. Wan, "A harmonically coupled chaotic system for a pseudo-random bit generator," *Journal of Chinese Computer System*, vol. 32, no. 4, pp. 639-643, 2011.
[16] Q. Wu, "An independent variable exclusively coupled chaotic system for a pseudorandom bit generator," in *Proceedings International Conference on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration*, pp. 341-344, 2016.
[17] Q. Wu, "Independent variable exclusively coupled chaotic pseudorandom bit generator," *Computer Engineering & Science*, vol. 38, no. 11, pp. 2197-2201, 2016.
[18] S. Li, X. Mou, Y. Cai, "Pseudorandom bit generator based on couple chaotic systems and its applications in stream-cipher cryptography," in *Proceedings Progress in Cryptology*, pp. 316-329, 2001.
[19] A. J. Menezes, P. C. Oorschot, S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1996.

Biography

Qi Wu, born in 1984, is a Phd & a lecturer of Jiangxi University of Finance & Economics. His research field is information security & chaos theory. He has published more than twenty papers and has finished several projects.

A New Fast Matching Method for Dummy K-anonymous Location Privacy Protection in Location Based Services

Xiaohui Zhu and Renlong Qi
(Corresponding author: Renlong Qi)

School of Electrical Engineering, Zhengzhou University of Science and Technology
Zhengzhou 450000, China
Email: qirenlong@163.com

(Received May 29, 2020; Revised and Accepted Feb. 10, 2021; First Online Aug. 15, 2021)

Abstract

This paper proposes a dummy k-anonymous location privacy protection based on a fast-matching method that adopts the space coordinate transformation algorithm. First, the 2-D coordinates are converted to binary Morton code. With the fast matching method, non-adjacent position points distributed in different grids are selected as candidate sets of dummy positions. Then, the semantic similarity of place name information of position points in the candidate sets is calculated using the edit distance, and k-1 position points with the smallest semantic similarity are selected as dummy positions. While satisfying the semantic l-diversity and physical dispersion, this method can improve the generation efficiency of dummy locations and further improve the quality of location service.

Keywords: Dummy K-Anonymous Location Privacy Protection; Fast Matching; Physical Dispersion; Semantic Similarity

1 Introduction

With the popularization of smart mobile devices and the development of GPS, location-based service (LBS) has become one of the most promising services for mobile users [9, 19, 29]. LBS has brought great convenience to people's daily life and social activities. At the same time, people also pay more attention to the issue of sensitive information leakage when using LBS. Because service acquisition requires interaction between different location services providers (LSP), users' location data must be shared among them. Untrusted third parties can easily obtain some important personal information of users by analyzing and comparing the above location information [7]. For example, obtaining the user's behavior near the hospital can reveal the user's health status. If the user's recent places of departure and termination are analyzed, the user's home address, work place and work nature can be known. Once the private information falls into the illegal system, it will seriously threaten the security of users. Therefore, it is necessary to protect the privacy of personal location and try not to fall into the hands of businessmen and agents [28].

In order to prevent the disclosure of private information, scholars have proposed many methods of location privacy protection, including fuzzy method, encryption method and strategy-based method. Fuzzy method is the primary choice of location privacy protection, which is mainly realized by spatial anonymity or dummy location. Spatial anonymity method usually requires the fully-trusted third party (TTP) to complete the privacy protection work [22]. When a user needs to obtain location services, a k-anonymous area containing the user's location is generated by TTP and sent to the LBS server for query. In this method, when the area of the anonymous area is very large, it not only consumes more time, but also reduces the accuracy of the query. TTP is easy to become the bottleneck of the system. However, the method of dummy position does not need to build an anonymous area, and does not need the assistance of TTP. In the mobile client, the method of dummy position is used to achieve position anonymity by generating dummy position, which can make up for the deficiency of the spatial anonymity method.

In the location privacy protection based on dummy location, the generation efficiency of dummy location affects the quality of location service. The indistinguishability between dummy location and real location affects the effect of privacy protection. In this paper, a dummy k-anonymous location privacy protection method based on approximate fast matching is proposed by fully considering the location geographic semantic information features. This method uses the spatial coordinate conversion algorithm to convert the two-dimensional position coordinates into binary Morton codes. Through approximate matching, the position points distributed in different grids and

not adjacent to each other are selected as candidate sets of dummy positions. Then the semantic similarity of the place name information in the candidate set is calculated by using the edit distance. $k-1$ position points with the smallest semantic similarity degree are selected as dummy positions. While satisfying the semantic l-diversity and physical dispersion, this proposed method can improve the efficiency of dummy location generation and further improve the quality of location service.

2 Related Works

Location privacy protection methods are divided into two categories according to the architecture: Peer-to-peer (P2P)-based distributed structure and TTP-based central server structure. In the distributed structure, the user collaborates with the neighboring user on the client side to achieve anonymity, or fake the dummy location to achieve ambiguity, so as to achieve location privacy protection. Naghizadeh [12] proposed a P2P-based spatial anonymity method, which adopts the location information of neighbor nodes to achieve k-anonymity privacy protection, but it ignored the security of neighbor nodes. P2P-based scheme has the advantages of simplicity and flexibility, but it increases the cost of all software and hardware resources, communication overhead of smart phones. In the central server structure, reference [3] used TTP to generalize or blur the exact location sent from the mobile terminal to achieve the purpose of location privacy protection. This structural pattern can achieve good privacy protection effect. However, safety precautions should be taken for TTP.

Benisch [1] introduced an information caching mechanism, which reduced the probability of information disclosure by reducing the number of users accessing TTP, but it increased the burden on mobile clients. In addition, Panaousis [15] proposed an independent structure model, which enabled users to protect location privacy according to their own abilities and knowledge. This method was simple in structure and easy to be combined with other structures, but it had high requirements on clients. Grissa [5] proposed the architecture of multi-location server, which divided users into different subsets according to security requirements, improved the concealment of location information, and was suitable for application in social networks. Yang [24] proposed a location privacy protection method based on privacy information retrieval, and implemented location privacy protection by hard disk data retrieval and encryption way. This method simplified the system structure and had a good effect of privacy protection, but it increased the communication and hardware overhead and reduced the service quality. With the maturity and popularity of cloud service technology, Teng [20] proposed a searchable and encrypted location privacy protection method, and implemented a data access mode without displaying data access to ensure the confidentiality of encrypted data and user query records, but the query efficiency and the accuracy of query results needed to be improved.

Currently, k-anonymous method is still adopted by the location privacy. K-anonymous method uses the generalization and fuzzy technology to deal with key attribute values in the database, so that any one record can not be distinguished from k records. The location privacy protection method based on k-anonymous area is mainly divided into space and dummy position anonymous [11]. Niu [14] realized location privacy protection by constructing a core anonymous region, which should satisfy two conditions:

1) The region area reached a certain value;

2) The region should contain k users.

Um [21] proposed the grid division method and provided top-down grid cloaking and bottom-up grid algorithm for different privacy requirements. Xu [23] proved that the size of k-anonymous region had a great impact on the accuracy of query results, which provided guidance for the research on the method of anonymous region division. On this basis, references [16, 25, 27] proposed various anonymous region construction methods based on different geometric shapes, but these methods had two serious defects:

1) They must rely on TTP, which is not absolutely safe and easy to become a system bottleneck.

2) The size of the anonymous area and the accuracy of the query result are a pair of contradictions.

If the anonymous area is larger, the privacy protection effect is better, but the accuracy of the query result will be decreased.

Due to the serious shortcomings of the spatial anonymous region construction method, the dummy location method has been widely studied because it does not have the system bottleneck problem and has the advantages of high query accuracy. Zhang [26] first introduced the method of dummy location into the location privacy protection. Instead of requiring an anonymous server, several dummy locations were generated by the mobile client and sent to the LBS server together with the real location for query. Prince [17] proposed a randomization method to add dummy locations, so that users could add dummy locations in circular or rectangular areas to implement location privacy protection according to their requirements. Considering that the adversary with background information may steal the location privacy, Niu *et al.* [13] proposed a dummy location selection (DLS) algorithm and improved DLS algorithm by selecting a dummy location based on entropy measurement.

The above methods choose the dummy position from the aspects of query probability without considering the geographic semantic information of the position. But the attacker may determine the user's real location by analyzing the geo-semantic information. The solid triangle A in Figure 1 represents the real position. Hollow dots represent a set of candidates for dummy positions. Solid

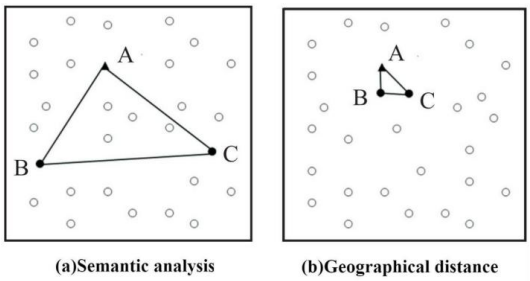


Figure 1: Position similarity attack

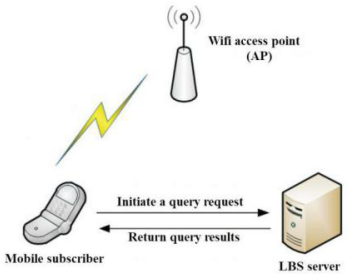


Figure 2: System structure model

dots B and C represent the selected dummy positions. In the dummy location selection shown in Figure 1(a), the three location points A, B and C represent hospitals, so the attacker can easily identify the possible health problems of query users through semantic analysis. In the dummy location selection shown in Figure 1(b), the selected dummy location is too close to the real location, so the attacker can easily find the specific location of the query user through the geographical distance. Therefore, the selection of dummy locations should consider the geographic semantic information of location points as far as possible, so as to ensure the physical dispersion and semantic diversity in all location points including the real location to effectively improve the protection effect of location privacy.

To solve the problem of possible semantic attack, Bergstra [2] proposed a dummy position selection method based on semantic perception, which ensured the physical dispersion and semantic diversity between dummy position points.

However, the use of Euclidean distance to calculate the physical distance between two points is inefficient when the data volume is large. Moreover, the construction of a semantic tree in Wi-Fi APs to calculate the semantic distance between two position points increases the burden of Wi-Fi APs, prolongs the preprocessing time, and thus reduces the service quality.

Aiming at the above problems, this paper proposes an approximate fast matching method for dummy position selection, which divides the selected area into $m \times m$ grids, and calculates the Morton code of the grid where each position point is located.

And then through the approximate matching, the position points in different grids are selected as candidate sets of dummy positions. By the geographical name information, the semantic similarity between two places in the candidate set is calculated. $k-1$ positions with the smallest semantic similarity are selected as the dummy position points. Then we send the real location and $k-1$ dummy location points to the LBS server for query. Experiments show that this method can reduce the burden of Wi-Fi APs and simplify the pretreatment process. While ensuring the physical dispersion and semantic diversity of dummy position points, the time efficiency of dummy position generation is improved.

Figure 3: Selected location area

3 System Model

In the TTP-based central server model, when more users initiate queries, TTP is easy to become a system bottleneck, and it is not absolutely safe and reliable. If the TTP is attacked, all location privacy will be compromised. Therefore, the TTP-free system model is adopted in this paper. The generation of dummy position points and the sending of query requests are completed by mobile clients. The system structure model is shown in Figure 2.

According to the system model in Figure 2, mobile users can obtain the location geographic information of an area including the current location as shown in Figure 3 through Wi-Fi APs.

Figure 4 shows that the mobile client divides the area into a square grid with $m \times m$. By using the approximate matching algorithm of grid position and geographical semantics respectively, $k-1$ position points that are not adjacent to each other in different grids with the smallest semantic similarity are selected as dummy positions. Finally, $k-1$ dummy locations and real locations are sent to LBS server for query.

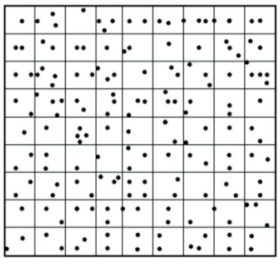


Figure 4: Grid partition

4 Proposed Location Privacy Protection Method

Definition 1. *Let Rs represent the selected rectangular area and Rs can be defined as* $Rs = m \times m, S$. *Where m represents the rows and columns number of the grid in Rs region. S represents the position points set contained within the Rs region.*

Definition 2. *Rc represents the regions included in each grid.* l_{phi} *is the physical distance between any two position points.* l_{sem} *represents the semantic similarity between any two position points. The* l_{grid} *represents the distance between any two grids.*

Definition 3. S_1 *represents the candidate set* $l_1, l_2, \cdots, l_n$ *of position points that satisfy the physical dispersion.* S_2 *represents the dummy position points set* $l_1, l_2, \cdots, l_{n-1}$ *that satisfy the condition. Dummy position result set* RS_{t_i} *includes dummy position set* $l_1, l_2, \cdots, l_{n-1}$ *and real location* l_{real}.

Definition 4. *If the semantic similarity between* l_i *and* l_j *satisfies the condition: If* $1 - |SEM_{t_i}|/C_k^2 \geq \theta$, *where* $SEM_{t_i} = l_{sem}|l_{sem}(l_i, l_j) \leq l$, $k = RS_{t_i}$, C_k^2 *is the combination formula, l is the preset semantic diversity threshold, then the result set* RS_{t_i} *is a θ-security set. The purpose of privacy protection is to get the maximum value* $\theta = 1$. *Meanwhile, the semantic similarity between?*l_i *and* l_j *is less than or equal to 1.*

The proposed dummy position generation method in this paper is realized by the following two algorithms. Algorithm 1 partitions the selected region as many grids. All position points in the grid are converted to Morton code. Through approximate matching, the position points which are not adjacent to each other and in different grids are selected to generate the dummy positions candidate set S_1. Algorithm 2 uses the edit distance to calculate the semantic similarity of the dummy position points in the candidate set, and selects $k-1$ positions with the smallest semantic similarity to generate the dummy position result set S_2.

We get the location geographic information of the square area and divide the area into $m \times m$ grid. According to the grid line, all the points are converted into Morton code. The Morton code of points in the same grid must be the same. So the Hamming distance between the positions is 0. Only one location point is selected in the same grid, and the selected grid is not adjacent to each other to ensure the dispersion of physical locations, as shown in Figure 5.

In the semantic similarity calculation of geographical name, according to the characteristics of Chinese geographical name, the method of "prefix words" is firstly used to eliminate the same "prefix words" in the geographical name information according to the principle of forward matching. Then for the rest of the geographical name, it calculates the edit distance to improve the matching efficiency and accuracy of semantic similarity. For example, "Shenyang No.2 middle school" and "Shenyang Zhicheng middle school", the string "Shenyang" for semantic similarity calculation has little sense, it also can affect the accuracy of the calculation results.

Algorithm 1 Generate a dummy position candidate set S_1

1: $Rs, S, m, (x_i, y_i) = p_i$.
2: Partition the region Rs as $m \times m$ gird.
3: According to the grid line, along the abscissa from left to right, the up line is 1, the down line is 0; Along the vertical from top to down, the left line is 1, the right line is 0; The x-value is encoded in the odd bit, the y-value is encoded in the even bit.
4: Repeat step 3, convert all position points in the region including the real position, and save the result in the double-end queue D.
5: Perform XOR operation on the binary code in queue D, and the Hamming distance is the physical distance l_{phi} between each position point.
6: By physical distance, randomly select position points that are not adjacent to each other and distributed in different grids, and store them in S_1.
7: Generate dummy position candidate set S_1.

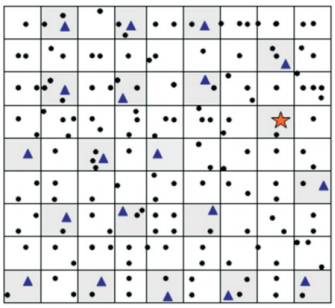


Figure 5: Generating dummy position candidate set

$D[i, j]$ is the edit distance of dynamic programming matrix. In the calculation of edit distance, the cost of edit operation is between [0,1], and different values are set according to the requirements. The values set in this paper are 0 and 1, when $a_i = b_i$, the replacement cost is 0; Otherwise, the cost of all editing operations is 1. The following formula represents the dynamic programming matrix D of the editing distance between the calculation string "A=No.2 middle school" and "B = Zhicheng middle school".

$$D = \begin{Bmatrix} 0\ 1\ 2\ 3\ 4 \\ 1\ 1\ 2\ 3\ 4 \\ 2\ 2\ 2\ 3\ 4 \\ 3\ 3\ 3\ 2\ 3 \\ 4\ 4\ 4\ 3\ 2 \end{Bmatrix} \quad (1)$$

Through the calculation of dynamic programming matrix (1), the edit distance between the strings is $D[i, j] =$

Algorithm 2 Generate dummy position result set S_2

Obtain the geographical information of each location point in candidate set S_1.

2: The geographical information is matched from frontward to backward. If the matching value is the same, then the same consecutive characters are ignored. So we can obtain two new geographical strings A and B.

Suppose that the A string contains i characters, denoted as $A = a_1a_2a_3La_i$. B string contains j characters, expressed as $B = b_1b_2b_3Lb_j$.

4: Constructing a dynamic programming matrix with $i+1$ column and $j+1$ row, the last element value obtained from $D[i,j]$ is $ed(A,B)$.
If $j = 0$, then return i; If $i = 0$, return j.

6: The first row is initialized as $0, 1, \cdots, i$. The first column is initialized as $0, 1, \cdots, j$.
Assign values to the elements in the matrix. If $a_i = b_i$, then $D[i,j] = D[i-1, j-1]$; If $a_i \neq b_i$, $D[i,j] = 1 + min(D[i-1, j-1], D[i-1, j], D[i, j-1])$.

8: Repeat step 7 until all values in the matrix are obtained. The final edit distance is $D[i,j]$.
The similarity matching index $S(A,B)$ is calculated by $D[i,j]$, namely, semantic similarity.

10: $k-1$ position points with minimal semantic similarity are selected to generate dummy position result set S_2.

$D[4,4] = 2$.

Calculating the semantic similarity between the name information strings as follows:

$$S(A,B) = 1 - \frac{D[i,j]}{max|A|,|B|} = 0.5. \quad (2)$$

Where $|A|$ and $|B|$ represent the length of two strings respectively, and the maximum value of the string length is taken to participate in the calculation of similarity matching index.

According to formula (2), the semantic similarity between the candidate position points can be calculated, and then the $k-1$ position points with the smallest semantic similarity can be calculated according to the following formula:

$$Argmin(S(l_i, l_j)). \quad (3)$$

5 Experiments and Analysis

Experiments use true location map data of Shenyang, getting 55 position geographic information points in $8 \times 8km^2$ rectangular area, which is divided into 16×16 rectangular grids. The main parameter k is between [2, 30].

The hardware environment of the experiment is: 3.2HHz, Intel Core i7 processor, 16GB memory. The operating system is Windows10, which is developed on MyEclipse platform and implemented in Java programming language. Table 1 is the default parameter configuration of the experiment.

Table 1: Experiment parameters

Parameter	Value
k	$\geq$ 2
l	0.2
Grid number	16×16
Position point set	10000
Space range/km^2	8×8
WiFi APs coverage area/m	800

5.1 Dummy Generation Efficiency

The results show the average time of generating the dummy position with MaxMinDistDS method, SimpMaxMinDistDS method [4] and the proposed method as shown in Table 2.

As can be seen from Table 2, with the increase of k, the MaxMinDistDS method takes more time than the proposed method. The proposed method is more efficient in generating dummy locations. When $k \leq 4$, the SimpMaxMinDistDS method has higher time efficiency than the proposed method. When $k \geq 5$, the efficiency of the proposed method is higher. With the increase of k, the efficiency advantage of this method becomes more and more obvious.

Figure 6 shows the average time efficiency comparison for generating dummy positions with FADBM [10], UMS [18], ST [8], KTL [6]. As can be seen from Figure 6, with the increase of k, the time of the five methods is increased too. However, the proposed method and the FADBM method cost less time than the other three methods, while the FADBM method takes the least time and the KTL method takes the most time. As can be seen from Figure 6, when $k \leq 5$, the proposed method takes relatively more time. When k=6,7, they have the similar time. When $k \geq 8$, the proposed method takes more time than the FADBM method, but less than the other three methods.

According to the efficiency comparison experiment, when the privacy measurement value selected by the user is small, the advantage of this method are not obvious. However, with the increase of k value, except the random method, the proposed method is more efficient than other methods. It can be seen that in the case of massive data and high degree of anonymity, the method in this paper has higher efficiency and can further improve the level of location service.

5.2 Semantic Diversity Comparison

We make semantic diversity comparison with different methods. As can be seen from Figure 7, the values of the MaxMinDistDS method and the SimpMaxMinDistDS method basically do not change with the increase of k value, and are always infinitely close to 1. The θ of the proposed method is 1, which can meet the requirements of semantic diversity. The θ of DLS method is

Table 2: Average time of generating the dummy position/s

Method	k=2	k=3	k=4	k=5	k=6	k=7	k=8
MaxMinDistDS	0.08	1.71	13.01	106.32	295.42	592.93	899.46
SimpMaxMinDistDS	0.03	0.031	0.045	0.058	0.0145	0.187	0.23
Proposed method	0.045	0.046	0.051	0.051	0.058	0.063	0.067

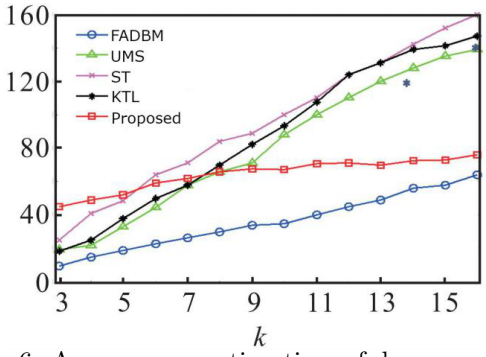


Figure 6: Average generation time of dummy position

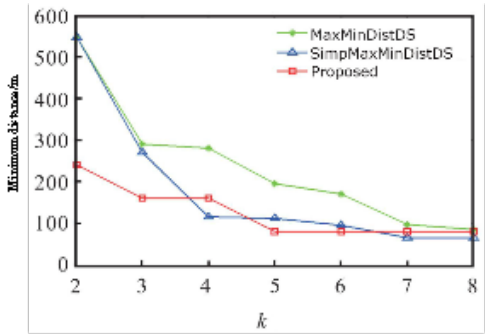


Figure 7: The minimum distance between the k points

smaller. Because MaxMinDistDS method, SimpMaxMinDistDS method and the proposed method consider the location of the semantic diversity, and DLS method only takes the query probability of dummy position candidate set into account, without the location point that may have the same characteristics of semantic information. Also, a query point on a larger probability often in hot spots, the semantic information between these locations is very similar, and thus has more semantic similarity. Therefore, the θ of the DLS method is smaller.

6 Conclusions

Aiming at the widely used anonymous location privacy protection method based on dummy location, this paper proposes a k-anonymous location privacy protection method based on approximate fast matching in order to solve the problems such as low generation efficiency of false location, complex preprocessing process and insufficient consideration of geographic semantic information characteristics. This method converts spatial position coordinates into binary strings. Through fast matching, the approximate geographical distance between position points is calculated, the efficiency of false position points is improved. The method of editing distance is used to calculate the semantic similarity between position points. The experimental results show that this method can reduce the generation time of false location and effectively improve the protection effect of location privacy under the premise of satisfying the physical dispersion and semantic diversity of dummy location.

Acknowledgments

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

References

[1] M. Benisch, P. G. Kelley, N. M. Sadeh, *et al.*, "The impact of expressiveness on the effectiveness of privacy mechanisms for location sharing," in *Symposium on Usable Privacy & Security*, 2009.

[2] A. D. Bergstra, B. Bert, B. Alex, *et al.*, "The mediating role of risk perception in the association between industry-related air pollution and health," *Plos One*, vol. 13, no. 5, pp. e0196783, 2018.

[3] Z. Changli, M. Chunguang, Y. Songtao, "Location privacy-preserving method for LBS continuous KNN query in road networks," *Journal of Computer Research and Development*, vol. 52, no. 11, pp. 2628-2644, 2015.

[4] S. Chen, H. Shen, "Semantic-aware dummy selection for location privacy preservation," *IEEE Trustcom/BigDataSE/ISPA*, 2016.

[5] M. Grissa, A. A. Yavuz and B. Hamdaoui, "When the hammer meets the nail: Multi-server PIR for database-driven CRN with location privacy assurance," in *IEEE Conference on Communications and Network Security (CNS'17)*, pp. 1-9, 2017.

[6] L. Guo, Y. Zhu, H. Yang, *et al.*, "A k-nearest neighbor query method based on trust and location privacy protection," *Concurrency and Computation Practice and Experience*, vol. 22, pp. e5766, 2020.

[7] C. M. Huang, D. T. Dao, C. M. Mai. "Location-Based Service (LBS) data sharing using the k-member-limited clustering mechanism over the 4G and Wi-Fi hybrid wireless mobile networks," in *International Conference on Information Networking*, 2017.

[8] Y. Li, S. Li, "A real-time location privacy protection method based on space transformation," in *The 14th International Conference on Computational Intelligence and Security (CIS'18)*, 2018.

[9] T. Lin, L. Hang, L. Jie, Y. Shoulin, "An efficient and secure Cipher-Text retrieval scheme based on mixed homomorphic encryption and multi-attribute sorting method under cloud environment," *International Journal of Network Security*, vol. 20, no. 5, pp. 872-878, 2018.

[10] J. Liu, X. Jiang, S. Zhang, *et al.*, "FADBM: Frequency-aware dummy-based method in long-term location privacy protection," in *IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS'19)*, 2019.

[11] J. Liu, S. L. Yin, H. Li and L. Teng, "A density-based clustering method for k-anonymity privacy protection," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 1, pp. 12-18, Jan. 2017.

[12] A. Naghizadeh, S. Berenjian, E. Meamari, *et al.*, "Structural-based tunneling: Preserving mutual anonymity for circular P2P networks," *International Journal of Communication Systems*, vol. 29, no. 3, pp. 602-619, 2016.

[13] B. Niu, Q. Li, X. Zhu, G. Cao and H. Li, "Achieving k-anonymity in privacy-aware location-based services," in *IEEE Conference on Computer Communications*, pp. 754-762, 2014.

[14] J. Niu, X. Zhu, L. Shi and J. Ma, "Time-aware dummy-based privacy protection for continuous LBSs," in *International Conference on Networking and Network Applications*, pp. 15-20, 2019.

[15] E. Panaousis, A. Laszka, J. Pohl, A. Noack and T. Alpcan, "Game-theoretic model of incentivizing privacy-aware users to consent to location tracking," in *IEEE Trustcom/BigDataSE/ISPA*, pp. 1006-1013, 2015.

[16] L. Peng, Z. Chen, L. T. Yang, *et al.*, "Deep convolutional computation model for feature learning on big data in internet of things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 790-798, 2018.

[17] P. K. Sahu, S. K. Chandra, "Location privacy using user anonymity and dummy locations," *International Journal of Innovative Technology & Creative Engineering*, vol. 1, no. 10, 2011.

[18] H. Shuhei, A. Daichi, H. Takahiro, *et al.*, "Dummy generation based on user-movement estimation for location privacy protection," *IEEE Access*, vol. 6, pp. 22958-22969, 2018.

[19] L. Teng, H. Li, "A high-efficiency discrete logarithm-based multi-proxy blind signature scheme," *International Journal of Network Security*, vol. 20, no. 6, pp. 1200-1205, 2018.

[20] L. Teng, H. Li and S. Yin, "IM-Mobishare: An improved privacy preserving scheme based on asymmetric encryption and bloom filter for users location sharing in social network," *Journal of Computers (Taiwan)*, vol. 30, no. 3, pp. 59-71, 2019.

[21] J. Um, H. Kim, Y. Choi and J. Chang, "A new grid-based cloaking algorithm for privacy protection in location-based services," in *The 11th IEEE International Conference on High Performance Computing and Communications*, pp. 362-368, 2009.

[22] H. Wang, H. Huang, *et al.*, "Efficient location privacy-preserving k-anonymity method based on the credible chain," *Isprs International Journal of Geo-Information*, vol. 6, no. 6, pp. 163, 2017.

[23] J. Xu, X. Tang, H. Hu and J. Du, "Privacy-conscious location-based queries in mobile environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 3, pp. 313-326, 2010

[24] Y. Yang, "Perceived k-value location privacy protection method based on LBS in augmented reality," *International Journal of Security & Its Applications*, vol. 9, no. 4, pp. 25-32, 2015.

[25] S. Yin, H. Li, and 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.

[26] W. Zhang, X. Cui, D. Li, D. Yuan and M. Wang, "The location privacy protection research in location-based service," in *The 18th International Conference on Geoinformatics*, pp. 1-4, 2010.

[27] L. Zou, X. Wang, S. Yin, "A data sorting and searching scheme based on distributed asymmetric searchable encryption," *International Journal of Network Security*, vol. 20, no. 3, pp. 502-508, 2018.

[28] Y. Zhang, Q. Zhang, Z. Li, *et al.*, "A k-anonymous location privacy protection method of dummy based on approximate matching," *Control and Decision*, vol. 35, no. 1, pp. 65-73, 2020.

[29] C. Zhao, Y. Wang, J. Wang, *et al.*, "Study on key technologies of location-based service (LBS) for forest resource management," *Njas Wageningen Journal of Life Sciences*, vol. 10, no. 1-2, pp. 292-300, 2012.

Biography

Zhu Xiaohui, female, master degree, graduated from Sichuan University of Science and Technology in 2008, majoring in pattern recognition and intelligent system. She is now a full-time teacher in the teaching and Research Section of electronic Science and Technology in the School of Electrical engineering of Zhengzhou University of Science and Technology. She is mainly responsible for the courses of SCM principle and applied microcomputer principle and interface technology.

Renlong Qi was born in Zhengzhou City, Henan Province. He is now with the School of Electronic and Electrical Engineering in Zhengzhou University of Science and Technology. His interests include image processing and electrical engineering.

Security Analysis and Enhancements of a User Authentication Scheme

Wan-Rong Liu[1], Xin He[2], and Zhi-Yong Ji[1]
(Corresponding author: Zhi-Yong Ji)
Shanghai Sixth People's Hospital East affiliated to Shanghai University of Medicine, Health Sciences, Shanghai 201306, China[1]
Department of Engineering Science and Technology, Shanghai Ocean University, Shanghai 201306,China[2]
This work was supported in part by the 2018 "Research and Development and Application of Limb Local Drug Delivery Dialysis Device" of the seed fund program of Shanghai university of medicine, health Sciences (SFP-18-21-14-001)
Email: joyer99@126.com
(Received June 15, 2020; Revised and Accepted May 6, 2021; First Online Aug. 15, 2021)

Abstract

With the rapid development of the Internet, telemedicine information system is more and more around us. Still, the security of people's information is one of the biggest limiting factors for the widespread use of telemedicine information systems. Aslam *et al.* suggest that Amin *et al.*'s authentication protocol is their analysis of the three-factor authentication protocol is one of the best. Still, through our analysis, we find that Amin *et al.*'s protocol is susceptible to the agreement of privilege internal attack, replay attack. So on, we base on the agreement Amin *et al.*'s protocol propose an improved three-factor authentication protocol verified by BAN logic, the performance, and efficiency compared with the agreement of our agreement in the increase in a small amount of calculation has higher security.

Keywords: Anonymity; Authentication; Telecare Medicine Information System

1 Introduction

Telecare Medicine Information Systems (TMIS) is an information system that adopts network technology and can carry out consultation, monitoring and other special medical activities for patients in any location [14, 25]. This system is of great significance both from the perspective of patients with physical disabilities and from the perspective of the prevention and treatment of severe infectious diseases [7, 9]. By implanting or wearing sensors on the patient to collect physiological data of patients, continuously monitor their health status, and send the data to the hospital in real time, so that hospital professionals can diagnose patients and figure out the next treatment plan. It not only saves the commuting cost for patients, improves the time utilization rate, but also effectively reduces the direct contact between medical staff and patients during the prevention and control of severe infectious diseases, greatly reduces the risk of infection, and maximizes the therapeutic effect of patients. However, the application of telemedicine information system has produced a large amount of physiological information of patients, which is easy to be intercepted or modified by attackers if it is transmitted in an insecure channel. If the doctor gets the wrong information about the patient, he may make a wrong diagnosis. If the information is intercepted, the patient cannot get timely treatment, which may endanger the patient's life in serious cases. In this case, identity authentication is particularly important [13, 30].

Identity authentication refers to through certain means, complete the identification of the user's identity, the purpose is to confirm that the current claimed as a certain identity of the user, is indeed the claimed user [15, 16, 21, 22, 26]. Considering the number of parties in the authentication protocol, the authentication schemes can be divided into three types: one-way, mutual, and group authentication [27]. Considering the number of factors in the authentication agreement, the authentication scheme can also be divided into three categories: one, two, and three factor. Many authentication protocols have been proposed for telemedicine information system [11, 18, 32]. The first remote computer authentication scheme was proposed by Lamport [20]. In the beginning, the authentication protocol is mainly based on single-factor authentication, such as static password, where the user sets a string of static data, and the static password will remain unchanged until the user changes it. However, the security of static password has many shortcomings, although users can often change the password to improve the security, password will remain unchanged for a period of time, the single-factor authentication method

has not been able to meet the needs of the Internet for identity authentication security. The first two-factor authentication scheme was proposed by Hwang in 1990 [17]. So far, scholars have done a lot of research on two-factor authentication protocol [1, 12, 28, 33].

In the early 21st century, the three-factor authentication schemes were proposed. In three-factor authentication, the user needs to provide his/her biometric information in addition to smart card, ID and password. The biometric information of each person is unique to himself. The digital information converted from the biometric information has a high entropy value and does not require the user to remember, which makes it difficult for the attacker to guess the user's biometric information and keep the information secret. Although biometrics has good characteristics, users cannot guarantee that the biometrics information input is exactly the same every time, such as fingerprints, and a slight deviation will lead to failure and rejection.

In order to solve the problem of failed rejection, Jin *et al.* [19] proposed a authentication protocol with fingerprint data and marked random numbers. To achieve this, biological hash functions were created, a technique that combines tagging random numbers with biometric recognition. However, not all experts adopt the biological hash function to reduce the failure rejection rate. They believe that users cannot guarantee that the input of biometric information is exactly the same every time, and accept the input biometric information as long as it is within a certain error range, such as Arshad and Nikooghadam [5]'s authentication protocol. But Lu *et al.* [35] have shown that Arshad and Nikooghadam's authentication protocol has shortcomings such as offline password guessing attacks. In 2013, Chang *et al.* [31] proposed one of the first three-factor authentication scheme for TMIS, and their scheme depends on the biometric information of the user as the third layer of the security. In the same year, Das *et al.* [8] exposed some weaknesses in Chang *et al.*'s scheme. Liu and Chung [23] proposed a user authentication scheme for wireless healthcare sensor networks in 2017.

Challa *et al.* [10] proposed an improved protocol of Liu and Chung's scheme. But Liu and Chung" scheme and Challa *et al.*" scheme power consumption are greatly increased, which is not suitable for telemedicine information systems. In 2015, Xu *et al.* [34] proposed a user authentication scheme preserving uniqueness and anonymity for connected health care. Amin *et al.* [3] proved that Xu *et al.*'s scheme has a design flaw and proposed a secure three-factor user authentication and key agreement protocol for TMIS with user anonymity. Meanwhile, Aslam *et al.* [6] thought Amin *et al.*'s scheme was the best among all the three-factor authentication methods in their survey. In 2016, Niloofar *et al.* [29] pointed out some weaknesses in Amin *et al.*'s scheme, such as the inability to defend against replay attack. We think the agreement structure of Amin et al is good, so further analyze Amin *et al.*'s scheme in detail, propose an improved three-factor authentication protocol for TMIS. The protocol is based on Computational Diffie-Hellman (CDH) problem and timestamp mechanism [24]. The CDH problem based on DH problem is a discrete logarithm problem based on finite field, which obtains the calculation results indirectly instead of solving the discrete logarithm problem directly.

1) Discrete logarithm problem: given $P, aP \in E/Fq$, for unknown $a \in Zn*$, the probability of success of finding the value of a is negligible.

2) Computational Diffie-Hellman problem: given $P, aP, bP, P \in E/Fq$, for unknown $a, b \in Zn*$, the probability of success of finding the value of abP is negligible.

A timestamp is a piece of data that represents information that already exists at a particular point in time. It is mainly proposed to provide an electronic evidence for users to prove the generation time of some data of users, ensuring the freshness of information. We employ Ellipse Curve Cryptography (ECC) in our protocol, which require a small amount of computation, faster processing speed, and less storage space and transmission bandwidth. We carry out BAN logic proof for our proposed protocol. We also perform performance comparisons and efficiency analyses. The result shows that our improved protocols have higher securing with little more computation cost.

2 Review of Amin *et al.*'s Scheme

We review of Amin *et al.*'s scheme. All notations that have been used, are described in Table 1.

Table 1: Notations

Symbol	Definition
U	User
M_s/S	Medical Server
ID	Identity of U
PW	Password of U
x	Secret key
r, R	A random number
P	A point on the elliptic curve
$P\ x$	The value of on x-axis
A	The adversary
SC	The smart card
$E_k(c)/D_k(\cdot)$	Symmetric key encryption/decryption by key k
$h(\cdot)$	One-way hash function
$\oplus$	Bitwise XOR operation
$\|\|$	Concatenation operation
T	The current time of system
SK	Session-key
ΔT	The maximum time interval for transmission delay
$H(\cdot)$	Bio-hash function
B	Biological characteristics

2.1 Registration Phase

he registration phase of Amin *et al.*' scheme is shown in Figure 1.

Step 1: U_i/smartcard chooses ID_i, PW_i, T_i, computes $A_i = h(ID||PW_i)$, $F_i = H(T_i)$, sends messages $\{ID_i, A_i, F_i\}$ to server.

Step 2: Server computes $W = h(ID_s||x||ID_i)$, $B_i = h(ID_i||A_i) \oplus W$, $CID_i = ENCx(ID_i||Ran)$, embeds messages $F_i, CID_i, A_i, B_i, h(\cdot), H(\cdot)$ in smartcard, delivers smartcard to U_i.

2.2 Login and Authentication Phase

The login and authentication phase of Amin *et al.*' scheme is shown in Figure 2.

2.3 Password Change Phase

The password change phase of Amin *et al.*' scheme is shown in Figure 3.

3 Weaknesses of Amin *et al.*'s Protocol

3.1 Weakness 1: Privileged Insider Attack

The U_i sends $< ID_i, h(ID||PW_i), H(T_i) >$ to S. A privileged insider user of medical server S being an attacker named A, who knows $< ID_i, h(ID||PW_i), H(T_i) >$. A with knowing $< ID_i, h(ID||PW_i), H(T_i) >$ can acquire PW_i as follow:

Step 1: Guesses a PW_i*.

Step 2: Computes $A_i* = h(ID_i||PW_i*)$.

Step 3: If A_i^* is equal to A_i, so $PW_i^* = PW_i$, otherwise A guesses another PW_i^* and computes A_i^* until $A_i^* = A_i$.

3.2 Weakness 2: Replay Attack

Let's say A listens message $< C_2, CID_i, C_4 >$. Then A sends same message $< C_2', CID_i', C_4' >$ to S. S computes all the following calculations without realizing that the message is a duplicate message. $W = h(ID_s||x||ID_i)$, $r_i^* = C2 \oplus W$, $C_1^* = r_i^* \cdot P$, $C_4^* = h(ID_i^*||r_i^*||W)$. S checks C_4^* is equal to received C_4' or not. Since C_4^* is equal to C_4', so S believes $< C_2', CID_i', C_4' >$ is not sent by an illegal user. Then the attacker A is authenticated. The attacker forwards the old eavesdropped message $< C_2, CID_i, C_4 >$ to S by retransmission and old login message. Because S has no way to tell when the message is delivered.

3.3 Weakness 3: Stolen Smart Card Attack

We suppose that an attacker A has stolen the smart card. A can extract the message $\langle F_i, CID_i, A_i, B_i, h(\cdot), H(\cdot)\rangle$. Then A computes $A_i = h(ID^*||PW_i^*)$, where A selects ID^* and PW_i^* respectively. If the equation is equal, A obtains the correct identity and password of the legitimate user. Otherwise, A chooses another identity and password until he/she finds the correct answer.

3.4 Weakness 4: User Impersonation Attack

The attacker be an illegal user with IDA, he will masquerade as any user. Firstly, A manipulates the smartcard to generate $< C_{2A}, CID_A, C_{4A} >$ in the name of U, where $C_{2A} = r_u \oplus W$, $CID_A = ENCx(ID_u||Ran)$, $C_{4A} = h(ID_u||r_u||W)$. After that, the smartcard sends $< C_{2A}, CID_A, C_{4A} >$ to S over the public channel. S cannot distinguish between a fresh message and old message. The telecare server accepts the attacker A as a legal user with identity ID_u.

4 Proposed Protocol

4.1 Registration Phase

The registration phase of the proposed scheme is shown in Figure 4.

4.2 Login and Authentication Phase

The login and authentication phase of the proposed scheme is shown in Figure 5.

Step 1: U_i inserts the smart card and inputs messages $\{ID_i, PW_i, T_i\}$, verifies whether $F_i^* = H(T_i) = F_i$, $A_i^* = h(PW_i||r) = A_i$, $RID_i^* = h(ID_i||r) = RID_i$ hold, if these equations are true, U_i generates random number r_i, computes $C_1 =_r i \cdot P$, $W = B_i \oplus h(RID_i||A_i)$, $C_2 = r_i \oplus W$ $C_4 = h(RID_i||r_i||W||T_1)$, sends messages $\{C_2, CID_i, C_4, T_1\}$ to server.

Step 2: Server checks $|T_s - T_1| \leq \Delta T$, extracts RID_i from CID_i, computes $W = h(ID_s||x||RID_i)$, $r_i^* = C_2 \oplus W$, $C_i^* = r_i^* \cdot P$, $C_4^* = h(RID_i||r_i^*||W||T_1)$, verifies whether $C_4^* = C_4$ holds, if the equation is true, server generates random number r_j, computes $D_1 = r_j \cdot P$, $SK = r_j \cdot C_1^*$, $G_1 = D_1 + C_1^*$, $L_i = h(RID_i^*||h_1(D_1)||W||T_2)$, $CID_i' = ENCx(RID_i||Ran')$, sends messages $\{L_i, G_1, CID_i', T_2\}$ to U_i.

Step 3: U_i checks $|T_c - T_2| \leq \Delta T$, computes $D_1^* = G_1 - C_1^*$, $L_i^* = h(RID_i||h_1(D_1^*)||W||T_2)$, $SK = r_i \cdot D_1^* = r_i \cdot r_j \cdot P$, verifies whether $L_i^* = L_i$ holds, if the equations is true, computes $Z_i = h(RID_i||SK)$, replay CID_i with CID_i', sends messages $\{Z_i\}$ to server.

User Ui /Smartcard		Server S
$Chooses < ID_i,\ PW_i,\ T_i >$		
$ChomputesA_1 = h(ID\|\|PW_i)$		
$F_1 = H(T_i)$	$\xrightarrow{<ID_i,\ A_i,\ F_I>}$	$Computes\ W = h(ID_s\|\|x\|\|ID_i)$
		$B_i = h(ID_i\|\|A_i) \oplus W$
		$CID_i = ENC_x(ID_i\|\|Ran)$
		$Embeds < F_i,\ CID_i,\ A_i,\ B_i,\ h(\cdot),\ H(\cdot) > in\ SC$
	$\xleftarrow{delivers\ SC\ to\ U_i}$	

Figure 1: Registration phase

User U_i /Smartcard		Server S
$Insert\ the\ smart\ card\ and$		
$inputs\ < ID_i,\ PW_i,\ T_i >$		
$comprtex\ F_i^* = H(T_i) = F_i$		
$A_i^* = h(ID_i\|\|PW_i) = A_i$		
$generates\ random\ number\ r_i$		
$C_1 = r_i \cdot P$		
$W = B_i \oplus\ h(ID_i\|\|A_i)$		
$C_2 = r_i \oplus W$		
$C_4 = h(ID_i\|\|r_i\|\|W)$	$\xrightarrow{\{C_2\ CID_i, C_4, T_i,\} to\ S}$	$S\ extroctsID_i from\ CID_i$
		$S\ computesW = h(ID_i\|\|x\|\|ID_i)$
		$r_i^* = C_2 \oplus W,\ C_1^* = r_i^* \cdot P$
		$C_4^* = h(ID_i\|\|r_i^*\|\|W)$
		$Checks\ \ C_4^* = C_4$
		$Generates\ random\ number\ \ r_j$
		$D_1 = r_j \cdot P,\ SK = r_j \cdot C_1^*$
		$G_1 = D_1 + C_1^*$
		$L_i = h(ID_i^*\|\|h_1(D_1)\|\|W)$
		$CID_i' = ENG_x(RID_i\|\|Ran')$
	$\xleftarrow{\{L_i,\ G_1 MCID_i'\}}$	
$U_i\ computex\ D_i^* = G_1 - C_1^*$		
$L_i^* = h(ID_i\|\|h_1(D_1)\|\|W)$		
$SK = r_i \cdot D_1^* = r_i \cdot r_j \cdot P$		
$Checks\ L_1^* = L_i$		
$Computes\ Z_i = h(RID_i\|\|SK)$		
$Re\ places\ old\ CID_i\ with\ new\ CID_I'\ in\ SC$		
	$\xrightarrow{\{Z_i\}}$	$S\ computesZ_i^* = h(ID_i\|\|SK)$
		$checks\ \ Z_i = Z_i$

Figure 2: Login and authentication phase

U_i /Smartcard		Server S
$U_i\ inputs\ < ID_i,\ PW_i >$		
$SC\ computes\ F_i^* = H(T_i) = F_i$		
$A_i^* = h(ID_i\|\|PW_i) = A_i$		
$inputs\ new\ PW_i^{new}$		
$A_i^{new} = h(ID_i\|\|PW_i^{new})$		
$B_i^{new} = h(ID_i\|\|A_i^{new}) \oplus W$		
$replaces\ < A_i,\ B_i >\ widh < A_i^{new},\ B_i^{new} >$	$\xrightarrow{Secure\ \ chanmel}$	$\xrightarrow{insecre\ \ charmel}$

Figure 3: Password change phase

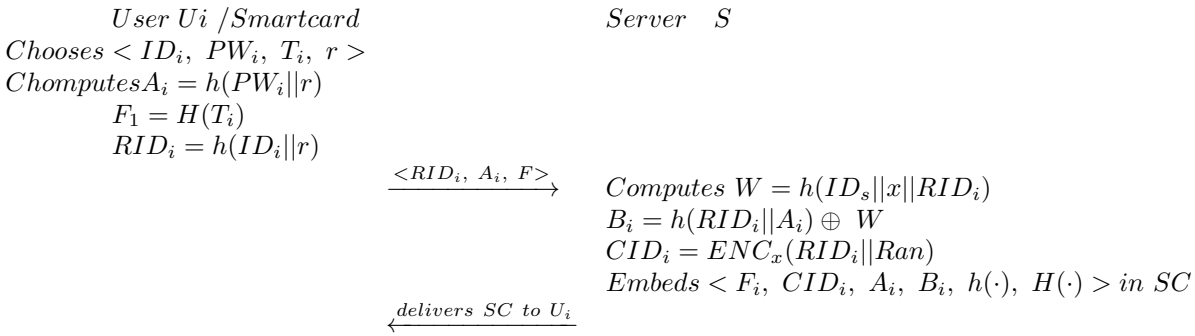


Figure 4: Registration phase

Step 4: Server computes $Z_i^* = h(RID_i||SK)$, verifies whether$Z_i^* = Z_i$holds.

4.3 Password Change Phase

The password change phase of the proposed scheme is shown in Figure 6.

5 Security Analysis of the Proposed Scheme

1) Privileged insider attack. Once the user sends $< RID_i, A_i, F_i >$ securely to S. The attacker gets all available information from the server and guesses user password. But ID_i, PW_i, T_i and r are never sent in plaintext. In addition, r is a random nonce.

2) Replay attack. We add the timestamp to the original scheme. We assume that A listens on the login message $< C_2, CID_i, C_4, T_1 >$ that U_i sends to S. Because the timestamp mechanism means is not the latest. The server will check $|T_s - T_1| \leq \Delta T$. Even if the attacker logs in at the same time as the user, he/she cannot compute $W = (ID_s||x||RID_i^*)$, $_i^* = C_2 \oplus W$ and pass the test of $C_4^* = h(RID_i * ||r_i^*||W||T_1)$.

3) Stolen smart card attack. We assume that A has stolen SC. A can extract the message $< F_i, CID_i, A_i, B_i, h(\cdot), H(\cdot) > inSC$. r is a random nonce. The A should compute $RIDi = h(ID_i||r)$, $Ai = h(PW_i||r)$. The agreement succeeded in fending off the A 's attack.

4) User impersonation attack. Because the timestamp mechanism indicates that every session message between the two is not delayed and the test of $C_4^* = h(RID_i^*||r_i^*||W||T_1)$, where $W = h(ID_s||x||RID_i^*)$ and $r_i^* = C_2 \oplus W$, is not an easy question. The attacker cannot be a malicious user with and she/he can masquerade as any user.

6 Security Analysis Using BAN Logic

In this section, we use BAN logic to perform a formal security analysis of the proposed protocol.

Goals: We use the BAN logic structure to prove that our proposed scheme can achieve mutual authentication.

Goal 1: $User| \equiv (User \overset{SK}{\longleftrightarrow} S)$.

Goal 2: $S| \equiv (User \overset{SK}{\longleftrightarrow} S)$.

The arrangement of proposed scheme to idealized form is as follows.

Message 1: $User \rightarrow S : \{User \overset{SK}{\longleftrightarrow} S, T_c\} r_j \cdot C_1^*$.

Message 2: $S \rightarrow User : \{User \overset{SK}{\longleftrightarrow} S, T_2\} r_i \cdot D_1^*$.

Assumptions: We make the following assumptions to analyze our proposed scheme.

H1: $User| \equiv (User \overset{r_i \cdot D_1^*}{\longleftrightarrow} S)$.

H2: $S| \equiv (User \overset{r_j \cdot C_1^*}{\longleftrightarrow} S)$.

H3: $User| \equiv \#(T_2)$.

H4: $S| \equiv \#(T_c)$.

H5: $User| \equiv S \Rightarrow (User \overset{SK}{\longleftrightarrow} S)$.

H6: $S| \equiv User \Rightarrow (User \overset{SK}{\longleftrightarrow} S)$.

Based on the above assumptions and the rules of BAN logic, we analyze the idealized form of the proposed scheme and the main steps of proof.

From Message 1, we have:

$S \triangleleft \{User \overset{SK}{\longleftrightarrow} S, T_c\} r_j \cdot C_1^*$.

From H2 and message-meaning rule, we have:

$S| \equiv User| \sim (User \overset{SK}{\longleftrightarrow} S, T_c)$.

From H4 and freshness rules, we have:

$S| \equiv \#(User \overset{SK}{\longleftrightarrow} S, T_c)$.

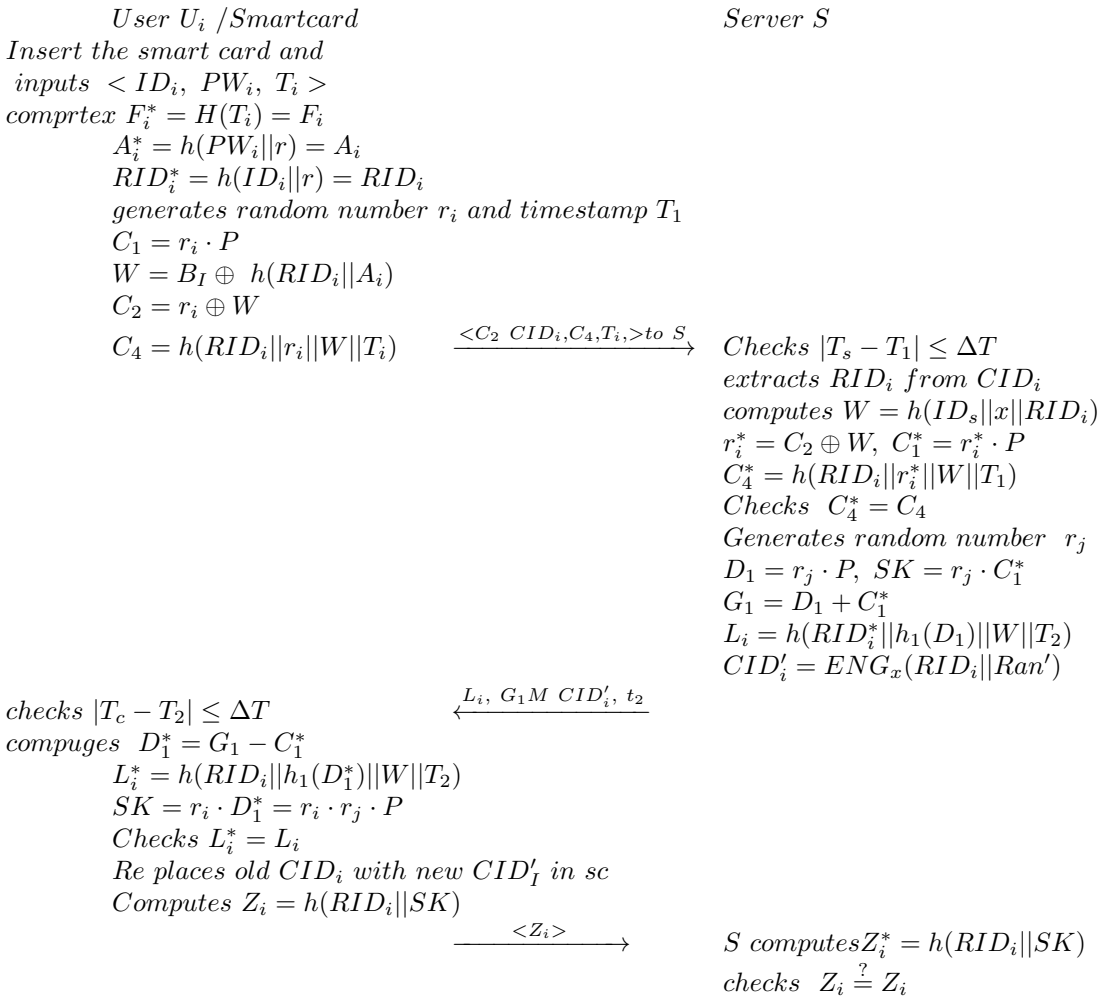


Figure 5: Login and authentication phase

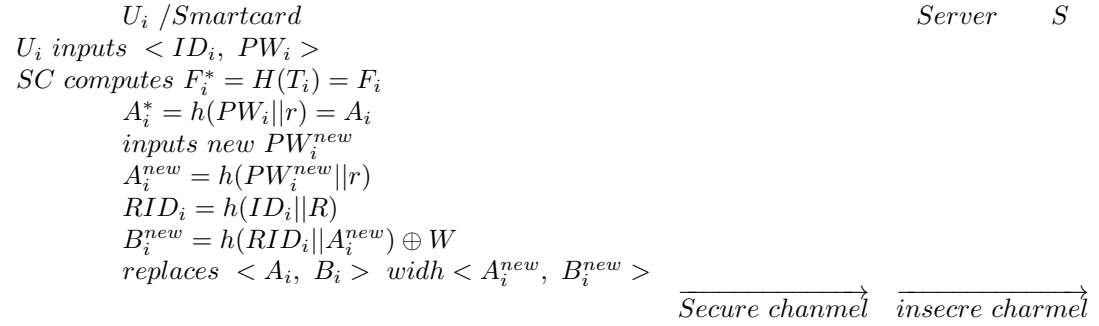


Figure 6: Password change phase

From $S| \equiv User| \sim (User \overset{SK}{\longleftrightarrow} S, T_c)$ and nonce verification rule, we have:

$S| \equiv User| \equiv (User \overset{SK}{\longleftrightarrow} S, T_c)$.

From message judgment rule, we have:

$S| \equiv User| \equiv (User \overset{SK}{\longleftrightarrow} S)$.

From H6 and message judgment rule, we have:

$S| \equiv (User \overset{SK}{\longleftrightarrow} S)$. (Goal 2)

From Message 2, we have:

$User \triangleleft \{User \overset{SK}{\longleftrightarrow} S, T_2\} r_i \cdot D_1^*$.

From H1 and message-meaning rule, we have:

$User| \equiv S| \sim (User \overset{SK}{\longleftrightarrow} S, T_2)$.

From H3 and freshness rules, we have:

$User| \equiv \#(User \overset{SK}{\longleftrightarrow} S, T_2)$.

From $User| \equiv S| \sim (User \overset{SK}{\longleftrightarrow} S, T_2)$ and nonce verification rule, we have:

$User| \equiv S| \equiv (User \overset{SK}{\longleftrightarrow} S, T_2)$.

From message judgment rule, we have:
$User| \equiv S| \equiv (User \overset{SK}{\longleftrightarrow} S)$.
From H5 and message judgment rule, we have:
$User| \equiv (User \overset{SK}{\longleftrightarrow} S)$. (Goal 1)

7 Performance Comparison and Efficiency Analysis

According to the Tables 1. and 3, the proposed agreement adds a small amount of computing and provides more security.

In Table 2, F1: Privileged insider attack; F2: Replay attack; F3: Stolen smart card attack; F4: User impersonation attack; F5: User untraceability; F6: Offline password guessing attack; F7: Session key disclosure attack; F8: Server not knowing password; F9: Forward secrecy; F10: User anonymity; and F11: Mutual authentication.

Table 2: Performance comparison

Perfomance	Amin et al. [3]	Amin et al. [4]	Lu *et al.* [35]	Ours
F1	No	No	Yes	Yes
F1	No	No	Yes	Yes
F2	No	No	Yes	Yes
F3	No	No	Yes	Yes
F4	No	Yes	No	Yes
F5	Yes	No	Yes	Yes
F6	Yes	No	Yes	Yes
F7	Yes	No	Yes	Yes
F8	Yes	No	Yes	Yes
F9	Yes	No	Yes	Yes
F10	Yes	Yes	No	Yes
F11	Yes	Yes	Yes	Yes

In Table 3, T_h =Time to compute a one-way hash function; T_{fun}=Time to compute a symmetric encryption or decryption function [2]; T_{mul}=Time complexity of a point multiplication operation on elliptic.

8 Conclusions

In this paper, we analyse Amin *et al.*'s authentication protocols and find that there were privileged internal attacks, replay attacks, stolen smart card attacks and user impersonation attacks on their protocols. In our view, Amin *et al.*'s protocol has a good framework, so we propose an improved authentication protocol based on their protocol, and use the BAN logic structure to prove that our proposed scheme can achieve mutual authentication. And we make performance comparison and efficiency analysis for the proposed protocol in Table 2 and Table 3. It can be seen that our protocol is not adding much computation, but greatly improving security.

Acknowledgments

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

References

[1] D. S. AbdElminaam, "Improving the security of cloud computing by building new hybrid cryptography algorithms," *International Journal of Electronics and Information Engineering*, vol. 8, no. 1, pp. 40-48, 2018.

[2] G. R. Alavalapati, G. Reddy, A. K. Das, E. J. Yoon, and K. Y. YOO, "A secure anonymous authentication protocol for mobile services on elliptic curve cryptography," *IEEE Access*, vol. 4, pp. 4394-4407, 2016.

[3] R. Amin, G. P. Biswas, "A secure three-factor user authentication and key agreement protocol for tmis with user anonymity," *Journal of Medical Systems*, vol. 39, no. 8, pp. 1-19, 2015.

[4] R. Amin, and G. P. Biswas, "An improved rsa based user authentication and session key agreement protocol usable in tmis," *Journal of Medical Systems*, vol. 39, no. 8, pp. 1-14, 2015.

[5] H. Arshad, M. Nikooghadam, "Three-factor anonymous authentication and key agreement scheme for telecare medicine information systems," *Journal of Medical Systems*, vol. 38, no. 12, pp. 1-12, 2014.

[6] M. U. Aslam, A. Derhab, *et al.*, "A survey and taxonomy of the authentication schemes in telecare medicine information systems," *Journal of Network and Computer Applications*, vol. 87, pp. 1-19, 2017.

[7] S. A. Chaudhry, H. Naqvi, and M. K. Khan, "An enhanced lightweight anonymous biometric based authentication scheme for TMIS," *Multimedia Tools and Applications*, vol. 77, no. 5, pp. 5503-5524, 2018.

[8] A. K. Das, A. Goswami, "A secure and efficient uniqueness-and-anonymity-preserving remote user authentication scheme for connected health care," *Journal of Medical Systems*. vol. 37, no. 3, pp. 1-16, 2013.

[9] Y. K. Ever, "Secure-anonymous user authentication scheme for e-healthcare application using wireless medical sensor networks," *IEEE Systems Journal*, vol. 13, no. 1, pp. 456-467, 2019.

[10] S. Challa, A. K. Das, V. Odelu *et al.*, "An efficient ECC-based provably secure three-factor user authentication and key agreement protocol for wireless healthcare sensor networks," *Computers & Electrical Engineering*, vol. 69, pp. 534-554, 2018.

[11] P. Chandrakar and H. Om, "An efficient two-factor remote user authentication and session key agreement scheme using Rabin cryptosystem," *Arabian Journal for Science and Engineering*, vol. 43, no. 2, pp. 661-673, 2018.

[12] W. Feifei, X. Guoai, and G. Lize, "A secure and efficient ECC-based anonymous authentication

Table 3: Comparison regarding computation costs

	Amin *et al.* [3]	Amin *et al.* [4]	Lu *et al.* [35]	Ours
User	$4T_h + T_{mul}$	$7T_h$	$5T_h + 2T_{mul}$	$8T_h + 2T_{mul}$
Server	$7T_h + 4T_{mul} + 2T_{fun}$	$4T_h$	$6T_h + 2T_{mul}$	$6T_h + 3T_{mul} + 2T_{fun}$
Total	$11T_h + 5T_{mul} + 2T_{fun}$	$11T_h$	$11T_h + 4T_{mul}$	$14T_h + 5T_{mul} + 2T_{fun}$

protocol," *Security and Communication Networks*, vol. 2019, no. 1, pp. 1-13, 2019.

[13] Z. Z. Guo, "Cryptanalysis of a certificateless conditional privacy-preserving authentication scheme for wireless body area networks," *International Journal of Electronics and Information Engineering*, vol. 11, no. 1, pp. 1-8, 2019.

[14] O. Hamdi, M. A. Chalouf, D. Ouattara, F. Krief, "eHealth: Survey on research projects, comparative study of telemonitoring architectures and main issues," *Journal of Network and Computer Applications*, vol. 46, pp. 100-112, 2014.

[15] M. S. Hwang, Li-Hua Li, "A New Remote User Authentication Scheme Using Smart Cards", *IEEE Transactions on Consumer Electronics*, vol. 46, no. 1, pp. 28–30, Feb. 2000.

[16] M. S. Hwang, J. W. Lo, S. C. Lin, "An efficient user identification scheme based on ID-based cryptosystem", *Computer Standards & Interfaces*, vol. 26, no. 6, pp. 565–569, 2004.

[17] T. Hwang, Y. Chen, and C. J. Laih, "Non-interactive password authentications without password tables," in *Conference Proceedings of IEEE Region 10 Conference on Computer and Communication Systems (TENCON'90)*, pp. 429-431, 1990.

[18] M. Jiaqing, H. Zhongwang, L. Yuhua, "Cryptanalysis and security improvement of two authentication schemes for healthcare systems using wireless medical sensor networks, *Security and Communication Networks*, vol. 2020, pp. 1-11, 2020.

[19] A. T. B. Jin, D. N. C. Ling, A. Goh., "Biohashing: Two factor authentication featuring fingerprint data and tokenised random number," *Pattern Recognition*, vol. 37, no. 11, pp. 2245-2255, 2004.

[20] L. Lamport, "Password authentication with insecure communication," *Communications of the ACM*, vol. 24, no. 11, pp. 770-772, 1981.

[21] C. C. Lee, C. H. Liu, M. S. Hwang, "Guessing attacks on strong-password authentication protocol", *International Journal of Network Security*, vol. 15, no. 1, pp. 64–67, 2013.

[22] C. H. Ling, C. C. Lee, C. C. Yang, and M. S. Hwang, "A secure and efficient one-time password authentication scheme for WSN", *International Journal of Network Security*, vol. 19, no. 2, pp. 177-181, Mar. 2017.

[23] C. H. Liu and Y. F. Chung, "Secure user authentication scheme for wireless healthcare sensor networks," *Computers & Electrical Engineering*, vol. 59, pp. 250-261, 2017.

[24] L. Liu, Z. Z. Guo, *et al.*, "An improvement of one anonymous identity-based encryption scheme," *International Journal of Electronics and Information Engineering*, vol. 9, no. 1, pp. 11-21, 2018.

[25] W. R. Liu, X. He, Z. Y. Ji, "An improved authentication protocol for telecare medical information system," *International Journal of Electronics and Information Engineering*, vol. 12, no. 4, pp. 170–181, 2020.

[26] T. Micha, C. Tomas, S. Nathaniel, "Survey of authentication and authorization for the Internet of things," *Security and Communication Networks*, vol. 2018, pp. 1-17.

[27] M. Mohammad, A. Safiyyeh, "A survey and taxonomy of the authentication schemes in telecare medicine information systems," *Journal of Network and Computer Applications*, vol. 87, no. 2017, pp. 1-19.

[28] W. Ping, L. Bin Lin, S. Hong, *et al.*, "Revisiting anonymous two-factor authentication schemes for IoT-enabled devices in cloud computing environments," *Security and Communication Networks*, vol. 2019, no. 2, pp. 1-13, 2019.

[29] N. Ravanbakhsh, M. Nazari, "An efficient improvement remote user mutual authentication and session key agreement scheme for E-health care systems," *Multimedia Tools and Applications*, vol. 77, pp. 55-88, 2016.

[30] S. Shan, "An efficient certificateless signcryption scheme without random oracles," *International Journal of Electronics and Information Engineering*, vol. 11, no. 1, pp. 9-15, 2019.

[31] D. R. Shiao, Y. F. Chang, and S. H. Yu, "A uniqueness-and anonymity- preserving remote user authentication scheme for connected health care," *Journal of Medical Systems*, vol. 37, no. 2, pp. 1-9, 2013.

[32] C. Shouqi, L. Wanrong, C. Liling, *et al.*, "An improved anonymous authentication protocol for location-based service," *IEEE Access*, vol. 7, pp. 114203-114212, 2019.

[33] B. Weixin, G. Prosanta, C. Yongqiang, L. Qingde, "Bio-AKA: An efficient fingerprint based two factor user authentication and key agreement scheme," *Journal of Network and Computer Applications*, vol. 109, pp. 45-55, 2020.

[34] L. Xu, F. Wu, "Cryptanalysis and improvement of a user authentication scheme preserving uniqueness and anonymity for connected health care," *Journal of Medical Systems*, vol. 39, no. 2, pp. 1-9, 2015.

[35] L. Yanrong, L. Lixiang, P. Haipeng, Y. Yixian, "An enhanced biometric-based authentication scheme for telecare medicine information systems using elliptic curve cryptosystem," *Journal of Medical Systems*, vol. 39, no. 3, pp. 1-8, 2015.

Biography

Liu Wanrong biography. received her bachelor's degree in electrical engineering and automation from Luoyang Institute of Technology in 2018. Now, she is a student at the College of Engineering Science and Technology, Shanghai Ocean University. Her main research is communication security and Internet of things technology.

He Xin biography. received his bachelor's degree in mechanical engineering from Anhui Polytechnic University in 2018. Now, he is a student at the College of Engineering Science and Technology, Shanghai Ocean University. He main research is Internet of things technology.

Ji Zhiyong biography. received his bachelor's degree from Nanjing University of Aeronautics and Astronautics in 2012. He received his MS degree Jiangsu University in 2017. He is the master's supervisor of mechanical engineering of Shanghai Ocean University. He is also the medical equipment senior engineer and deputy director of Shanghai Sixth People's Hospital East. His research directions include the development and application of wearable medical devices based on the Internet of things and the information security of the medical Internet of things.

An Improved Certificateless Signature Scheme for IoT-based Mobile Payment

Fen Yan, Linggen Xing, and Zhenchao Zhang
(Corresponding author: Fen Yan)

School of Information Engineering, Yangzhou University
198 West Huayang Street, Chahe Town, Hanjiang, Yangzhou, Jiangsu, China
Email: yanfen@yzu.edu.cn

(Received June 16, 2020; Revised and Accepted Apr. 24, 2021; First Online Aug. 17, 2021)

Abstract

With the popularization of the Internet of Things (IoT), mobile payments are widely used in our daily life. However, the design of secure and efficient signature schemes for mobile payment is still a topic studied by researchers. In 2018, Yeh *et al.* [24] designed a certificateless signature (CLS) scheme for IoT-based mobile payment and claimed their scheme is secure. However, we point out that Yeh *et al.*'s scheme is unable to resist the public key replacement attack. To solve the problem, an improved scheme is proposed in this paper. And we implement the security verification for this scheme under the random oracle model. Furthermore, the performance evaluation shows the efficiency of our scheme is comparable to related CLS schemes.

Keywords: Certificateless Signature (CLS); Internet of Things (IoT); Mobile Payment; Public Key Replacement Attack

1 Introduction

Mobile payment plays an increasingly important role in our life. It requires an application installed on a mobile device. Meanwhile, mobile payment platforms provide online payment services. Mobile-Payment System is mainly used for online shopping, paying utilities and transferring money. In recent years, Internet of Things (IoT) [2] devices have become a new mobile payment carrier. Users can purchase items easily with IoT devices by its contactless payment method. In addition, users can use applications on the wearable device to make payments since near field communication (NFC) makes contactless payments possible. Most of NFC applications are based on Android platform, and its open source features cause security challenges. To protect privacy and prevent malicious attacks, Feng *et al.* [5] proposed a lightweight protocol about NFC mobile sensors payment authentication, which reach mutual authentication between Reader-Tag, Back-end Database-Reader and Back-end Database-Tag. However, the growing number of online payments involving users' personal information raise the risk of user data being obtained by hackers. The risk is particularly exposed when mobile users transfer the payment data in a public network. The conceptional security services in a mobile payment system include authentication, access control, confidentiality, integrity, non-repudiation, and availability [20]. Digital signature is able to ensure data security and privacy because of its unforgeability and non-repudiation. Therefore, it is necessary to design a signature scheme for mobile payment. Furthermore, due to the limited computational resource of IoT devices, the signature scheme designed for the transaction should be lightweight.

Traditional public key infrastructure (PKI) security is based on the certificate issued by a trusted certificate authority. However, It is known that certificate management is resource-intensive and costly. To eliminate the certificate management, Shamir [18] introduced the idea of identity-based cryptosystem (IBC) where the user's identity information is used as its public key. In IBC, the private key is generated by a key generation center (KGC), which causes the key escrow problem. Certificateless signcryption [1] is an important multi-function cryptography primitive which solves the key escrow problem. In a certificateless signature (CLS) scheme, KGC generates the partial private key, and a secret value is generated by the user. Huang *et al.* [9] proved that the CLS scheme [1] can't resist the public key replacement attack. In the security model [9], adversaries are divided into two types: type I and type II. A type I adversary can perform a public key replacement attack. That is, the type I adversary can replace the public key of any user with any value. And the type II adversary knows the master secret key but cannot replace the user's public key. Li *et al.* [13] proposed the first CLS scheme with bilinear pairing. However, the bilinear pairing operation highly occupies the computational resource. He *et al.* [8] proposed a CLS scheme without bilinear pairings and the efficiency of the pairing-free scheme is greatly increased. Its low

resource consumption makes it suitable for application in IoT devices. However, the scheme [8] is vulnerable to a type II adversary [21, 22].

Liu *et al.* [14] proposed the first certificateless signcryption (CLSC) scheme in the standard model. However, in 2013, Miao *et al.* [15] claimed that the scheme [14] is not secure. Other CLSC schemes [3, 19] were proposed later. In 2018, Pakniat *et al.* [16] presented a CLS scheme and claimed their scheme is secure in the standard model, but Zhang *et al.* [25] demonstrated that the proposal is not robust against super type I adversary. Ji [11] *et al.* designed an efficient and certificateless conditional privacy-preserving authentication scheme. In their scheme [11], the trusted third parties can extract the real identity of users when necessary. However, the scheme is vulnerable to the modification attack, denial of service attack and impersonation attack [7]. Jia *et al.* [12] proposed an efficient provably-secure CLS scheme for IoT devices. Later, their scheme was found to be vulnerable to the type II adversary [26].

In 2018, Yeh *et al.* [24] presented a CLS scheme for Android-based mobile payment which does not exceed the inherent requirements of the resource limitations of the device. The performance evaluation given by the paper shows that the scheme is efficient and the scheme is existentially unforgeable against the super Type I and super Type II adversaries. However, we noticed this scheme [24] is insecure.

1.1 Our Contributions

In this paper, our contributions can be summarized as follows. We review Yeh *et al.*'s scheme [24] and point out the flaw of the scheme. We simulate a public key replacement attack against their scheme successfully, which indicates the scheme is not secure since a super type I adversary can easily forge a signature. In order to enhance the security of Yeh *et al.*'s scheme, we propose an improved CLS scheme for mobile payment and prove its unforgeability against super adversaries. We also give a specific transaction scheme for mobile payment. In addition, a comparison of our scheme with related schemes [12, 16, 24] is also presented.

1.2 Organization

This article is organized as follows. Section 2 introduces the definition of the elliptic curve, the definition of the CLS scheme and the security model. After that, we review Yeh's scheme [24] in Section 3 and point out the security issue. In Section 4, we present our improved certificateless scheme. Then we introduce the scheme for the transaction in Section 5. Section 6 analyzes the security of the scheme. The performance evaluation is presented in Section 7. Section 8 concludes the paper.

2 Preliminaries

2.1 Elliptic Curve Cryptography

Elliptic curve cryptography (ECC) is a public key encryption algorithm. ECC can provide the same or higher level of security than RSA and use shorter keys [10]. Let p be a λ-bit prime number. The symbol E/F_p denotes an elliptic curve E over a prime finite field F_p. The equation of E is established as $y^2 \equiv (x^3 + ax + b) \mod p$, where a, $b \in F_p$ and the discriminant $\Delta = (4a^3 + 27b^2) \mod p \neq 0$ are satisfied. The point at infinity on E/F_p represents O. G is the set of all points on the curve. It is defined as $G = \{(x, y) : x, y \in F_p$, and $(x, y) \in E/F_p\} \cup \{O\}$. G is an additive cyclic group under the point addition operation. Let n be the order of group G. Then P is called a generator of group G if n is the smallest number such that $nP = O$. The scalar multiplication in group G is denoted as $tP = P + P + ... + P$ (t times) where $t \in Z_q^*$.

Elliptic curve cryptography relies on the widely recognized difficulty in solving the elliptic curve discrete logarithm problem. The computational assumption used in the presented protocol is described as follows.

Elliptic curve discrete logarithm problem. Let G be an additive cyclic group of order n. Let P be a generator of the group G. They satisfy the formula $Q = aP$. Given the point $Q \in G$, finding an integer $a \in Z_q^*$ from P and Q is computationally difficult for any polynomial time-bounded algorithm.

2.2 Certificateless Signature Scheme

There are three entities in the CLS scheme: The KGC, the signer and the verifier. And a CLS scheme contains seven phases:

1) *Setup*: The algorithm is run by the KGC. With a security parameter λ, the algorithm returns a master secret key s and pubic system parameters $Params$.

2) $PartialPrivateKeyExtract$: The KGC executes this algorithm for each registered user. With the master secret key s , public parameters $Params$ and the user's identity ID, the algorithm extracts the user's partial private key s_{ID}. Then the KGC sends s_{ID} to the user through a secure channel.

3) $SetSecretValue$: With $Params$ and the user's identity ID, the algorithm outputs a secret value x_{ID}.

4) $SetPrivateKey$: With $Params$, s_ID and the secret value x_ID, the algorithm returns the private key SK_{ID}, that is, the user sets $SK_{ID} = (s_{ID}, x_{ID})$.

5) $SetPublicKey$: The algorithm is executed by each user. With $Params$ and x_{ID}, the algorithm outputs the public key PK_{ID}.

6) *Sign*: The input includes a message m, the public parameters $Params$, the signer's identity ID and the

private key SK_{ID}. The signer calls this algorithm to generate a signature σ.

7) $Verify$: The input includes the signature σ, the public parameters $Params$, the signer's identity ID, the public key PK_{ID} and the message m. The verifier executes this algorithm and acquires a $True$ or $False$ to indicate the validity of the signature σ.

2.3 Security Model of Certificateless Signature Scheme

In CLS, there are two types of adversaries: type I adversaries and type II adversaries. The type I adversary can replace the user's public key at will. That is, the type I adversary can change the user's public key to any value it wishes, but the adversary cannot obtain the master secret key s of the system. A type II adversary emulates a malicious KGC, which means it can obtain the value of the master secret key s, but cannot replace the user's public key. At the same time, adversaries are divided into three levels: Normal, strong and super. A normal adversary only has the ability to obtain a valid signature. A strong adversary is able to acquire a valid signature when the adversary is aware of the secret value x. A super adversary is able to get a valid signature even if the adversary is unaware of the secret value x. We only discuss the super type I and type II adversaries. The following queries can be executed by the adversaries:

1) $ExtractPublicKey(ID_i)$: This query allows an adversary $\mathcal{A}$ to obtain the public key PK_i of the corresponding user i. Here ID_i denotes the identity of the user i.

2) $ReplacePublicKey(ID_i, PK_i, PK_i^*)$: An adversary $\mathcal{A}$ can replace the public key PK_i of the user i with a new value PK_i^* through this query.

3) $ExtractSecretValue(ID_i)$: The adversary $\mathcal{A}$ could get the secret value of the user i through the query. It returns a null if $ReplacePublicKey(ID_i, PK_i, PK_i^*)$ has been queried.

4) $ExtractPartialPrivateKey(ID_i)$: The adversary $\mathcal{A}$ could get the partial private key $D_i = (s_i, R_i)$ of the user i through the query.

5) $SuperSign(ID_i, m)$: The adversary $\mathcal{A}$ could get a signature σ on message m through the query. The signature σ could be verified as $VALID$. This query can still give a valid signature even if the public key of the user i is replaced. Because of the super adversary's capabilities, this query does not require $\mathcal{A}$ to provide a secret value corresponding to the replaced public key PK_i^*.

The model defines two games: **Game 1** and **Game 2**. As described below, **Game 1** simulates the public key replacement attack by a super type I adversary $\mathcal{A}_{\mathcal{I}}$ and **Game 2** simulates the malicious KGC attack of a super type II adversary $\mathcal{A}_{\mathcal{II}}$.

Game 1. The game is performed between a challenger $\mathcal{C}$ and a super Type I adversary $\mathcal{A}_{\mathcal{I}}$.

- **Phase 1.** The challenger $\mathcal{C}$ runs the Setup algorithm to generate a master secret key s and public system parameters $Params$. Then $\mathcal{C}$ maintains s secret and sends $Params$ to the adversary $\mathcal{A}_{\mathcal{I}}$.
- **Phase 2.** In this phase, the adversary $\mathcal{A}_{\mathcal{I}}$ can adaptively access the queries $ExtractPublicKey(ID_i)$, $ReplacePublicKey(ID_i, PK_i, PK_i^*)$, $ExtractSecretValue(ID_i)$, $ExtractPartialPrivateKey(ID_i)$, and $SuperSign(ID_i, m)$.
- **Phase 3.** $\mathcal{A}_{\mathcal{I}}$ submits a signature σ on m^*. When the following conditions are satisfied, $\mathcal{A}_{\mathcal{I}}$ wins the game:
  1) $\mathcal{A}_{\mathcal{I}}$ has never queried the oracle $ExtractPartialPrivateKey(ID_i)$.
  2) $\mathcal{A}_{\mathcal{I}}$ has never queried the oracle $SuperSign(ID_i, m^*)$.
  3) The result of Verify$(m^*, \sigma, Params, ID_i, PK_i)$ is $True$.

Definition 1. *The proposed certificateless signature scheme is existentially unforgeable against a super type I adversary in polynomial time if the success probability $Succ_{A_I}$ is negligible when $\mathcal{A}_{\mathcal{I}}$ wins in **Game 1**.*

Game 2. This game is performed between a challenger $\mathcal{C}$ and a super Type II adversary $\mathcal{A}_{\mathcal{II}}$.

- **Phase 1.** The challenger $\mathcal{C}$ runs the $Setup$ algorithm to generate a master secret key s and public system parameters $Params$. Then $\mathcal{C}$ sends s and $Params$ to the adversary $\mathcal{A}_{\mathcal{II}}$.
- **Phase 2.** In this phase, the queries (i.e., $ExtractPublicKey(ID_i)$, $ReplacePublicKey(ID_i, PK_i, PK_i^*)$, $ExtractSecretValue(ID_i)$, $ExtractPartialPrivateKey(ID_i)$, and $SuperSign(ID_i, m)$) can be adaptively accessed by the adversary $\mathcal{A}_{\mathcal{II}}$.
- **Phase 3.** $\mathcal{A}_{\mathcal{II}}$ submits a signature σ on m^*. When the following conditions are satisfied, $\mathcal{A}_{\mathcal{II}}$ wins the game:
  1) $\mathcal{A}_{\mathcal{II}}$ has never queried $ExtractSecretValue(ID_i)$ and $ReplacePublicKey(ID_i, PK_i, PK_i^*)$.
  2) $\mathcal{A}_{\mathcal{II}}$ has never queried the oracle $SuperSign(ID_i, m^*)$.
  3) The result of Verify$(m^*, \sigma, Params, ID_i, PK_i)$ is $True$.

Definition 2. *The proposed certificateless signature scheme is existentially unforgeable against a super type II adversary in polynomial time if the success probability* $Succ_{A_{II}}$ *is negligible when* $\mathcal{A}_{\mathcal{II}}$ *wins in* ***Game 2****.*

3 Revisiting Yeh *et al.*'s Scheme

Here we review Yeh *et al.*'s scheme and present its vulnerability.

3.1 Review of Yeh *et al.*'s Scheme

Yeh's scheme consists of six phases. The details are as follows.

1) *Setup*: Given a security parameter k, KGC generates a group G of elliptic curve points with prime order q and determines a generator P of G. Then, KGC chooses a master key $s \in Z_q^*$ and a secure hash function $H : \{0,1\}^* \times G \rightarrow Z_q^*$. Next, KGC calculates a master public key $PK_{KGC} = sP$. Finally, KGC publishes $Params = \{G, P, PK_{KGC}, H\}$ and keeps s secure.

2) *PartialPrivateKeyExtract*: Given $Params$, s, and the identity ID_i of the user i, KGC generates a random number $r_i \in Z_q^*$, and calculates $R_i = r_iP$, $h_i = H(ID_i, R_i, PK_{KGC})$ and $s_i = r_iID_i + h_is \mod q$. Then, KGC returns a partial private key $D_i = (s_i, R_i)$ to the user i. The user i checks the validity of D_i. D_i will be verified as $VALID$ if the equation $s_iP = R_iID_i + h_iPK_{KGC}$ is satisfied.

3) *SetSecretValue*: The user i chooses a random number $x_i \in Z_q^*$ as user's own secret value.

4) *SetPublicKey*: Given *params* and x_i, the user i computes $PK_i = x_iP$ as user's public key.

5) *Sign*: Given *params*, D_i, x_i and a message m, the user i generates a signature for m. The user i chooses a random number $t_i \in Z_q^*$. Then the user i computes $k_i = H(m, T_i, PK_i, h_i)$, $T_i = t_iP$ and $\tau_i = t_i + k_i(x_i + s_i) \mod q$. The user i finally outputs $\sigma_i = (R_i, T_i, \tau_i)$ as the signature of the message m.

6) *Verify*: Given *params*, ID_i, PK_i, m, and $\sigma_i = (R_i, T_i, \tau_i)$, the verifier examines the validity of σ_i. The verifier computes $h_i = H(ID_i, R_i, PK_{KGC})$ and $k_i = H(m, T_i, PK_i, h_i)$. The signature σ_i is verified as $VALID$ if $\tau_iP = T_i + k_i(PK_i + ID_iR_i + h_iPK_{KGC})$.

3.2 Vulnerability of Yeh *et al.*'s Scheme

Yeh *et al.* [24] claimed that their CLS scheme is unforgeable against super type I and type II adversaries. However, we have found Yeh *et al.*'s scheme is unable to resist public key replacement attacks. Here we present the process of the public key replacement attack. Let $\mathcal{A}$ and $\mathcal{C}$ represent a super type I adversary and a challenger.

1) $\mathcal{A}$ issues *CreateUser* query with input ID_i and receives the user i's public key PK_{ID_i} as output.

2) $\mathcal{A}$ randomly chooses $r' \in Z_q^*$, and calculates $R_i' = r'P$, $h_i' = (ID_i, R_i', PK_{KGC})$.

3) $\mathcal{A}$ randomly chooses $a, t' \in Z_q^*$, and calculates $PK_i' = aP - h_i'PK_{KGC}$, $T = t'P$. Then $\mathcal{A}$ issues *ReplacePublicKey* query to change the public key of ID^* to PK_i'.

4) $\mathcal{A}$ calculates $k' = H(m^*, T, PK_i', h_i')$.

5) $\mathcal{A}$ calculates $\tau' = t' + k'(a + ID_ir')$.

6) $\mathcal{A}$ outputs the forgery signature $\sigma' = (R_i', T, \tau')$ and submits it to $\mathcal{C}$.

ExtractSecretValue query, *ExtractPartialPrivateKey* query and *Sign* query have never been issued. The forged signature $\sigma' = (R_i', T, \tau')$ on message m^* will be verified as $VALID$: $\tau'P = (t' + k'(a + ID_ir'))P = t'P + k'(aP + ID_ir'P) = T + k'(aP - h_i'PK_{KGC} + ID_iR_i' + h_i'PK_{KGC}) = T + k'(PK_i' + ID_iR_i' + h_i'PK_{KGC})$

That means $\mathcal{A}$ could successfully forge a signature. Therefore, this scheme is not secure against super type I adversaries.

4 Our Improved Scheme

To prevent the public key replacement attacks, our improved scheme is proposed. The improved scheme is shown as below.

1) *Setup*: The KGC generates the master secret key $s \rightarrow Z_q^*$ and public parameters $Params$ based on the input security parameter λ. The KGC publishes $Params = (G, P, PK_{KGC}, H)$ where P is a generator on an elliptic additive group G of order q over the finite field F_p. And here $PK_{KGC} = sP$. The master key s maintained as secret. Besides, H denotes three one-way hash functions: $H_1 : \{0,1\}^* \times G \times G \rightarrow Z_q^*$. $H_2 : \{0,1\}^* \times G \rightarrow Z_q^*$. $H_3 : \{0,1\}^* \times G \times G \times G \rightarrow Z_q^*$.

2) *PartialPrivateKeyExtract*: The KGC randomly chooses $r_i \in Z_q^*$ for the user i with the identity of ID_i, and computes $R_i = r_iP$, $h_i = H_1(ID_i, R_i, PK_{KGC})$, $s_i = (s + r_ih_i) \mod q$.

3) *SetSecretValue*: The user randomly chooses a $x_i \in Z_q^*$ as the secret value.

4) *SetPublicKey*: The public key PK_i represents (R_i, P_i) where $P_i = x_iP$.

5) *Sign*: With the message m, $Params$, signer's ID_i , x_i, s_i and R_i, the signer randomly chooses $t \in Z_q^*$ and calculates $T = tP$, $k_1 = H_2(m, T)$, $k_2 = H_3(m, P_i, T, PK_{KGC})$. Then signer computes $\tau = x_i + k_1 t + k_2 s_i \mod q$. Finally, the signer generates the signature $\sigma = (T, \tau)$.

6) *Verify*: With the message m, $Params$, signer's ID_i, PK_i and $\sigma = (T, \tau)$, the verifier calculates $h_i = H_1(ID_i, R_i, PK_{KGC})$, $k_1 = H_2(m, T)$ and $k_2 = H_3(m, P_i, T, PK_{KGC})$. Then the verifier examines if the equation $\tau P = P_i + k_1 T + k_2(PK_{KGC} + h_i R_i)$ is satisfied. If it is satisfied, σ will be verified as $VALID$. Otherwise, it returns $INVALID$.

The correctness of the equation is presented as follows: $\tau P = (x_i + k_1 t + k_2 s_i)P = x_i P + k_1 tP + k_2 s_i P = P_i + k_1 T + k_2(s + r_i h_i)P = P_i + k_1 T + k_2(PK_{KGC} + h_i R_i)$

5 Transaction Scheme for Mobile Payment

In actual mobile transactions, both customers and merchants both need to signup with the mobile payment platforms and provide their bank accounts to these acquirers and the acquirers handle the transactions in turn [23]. In this section, we present a transaction scheme for mobile payment and simulate the actual mobile payment scenario as a transaction process to introduce how our CLS scheme implements in mobile payment. There are four roles in the transaction scheme: the user with an IoT device, the merchant server, the mobile payment platform and KGC. The execution process is shown below.

5.1 Initialization

With a security parameter λ, the KGC generates the master secret key $s \to Z_q^*$ and publishes the public parameters $Params = \{G, P, PK_{KGC}, H\}$ where P is a generator on an elliptic additive group G of order q over the finite field F_p. KGC is defined as a trusted security service and provide the mobile pay service. As it is designed in original scheme, $PK_{KGC} = sP$. The master key s is kept secret. Besides, H denotes three one-way hash functions: $H_1 : \{0,1\}^* \times G \times G \to Z_q^*$. $H_2 : \{0,1\}^* \times G \to Z_q^*$. $H_3 : \{0,1\}^* \times G \times G \times G \to Z_q^*$.

Meanwhile, the user i randomly chooses a number $x_i \in Z_q^*$ as the secret value and $P_i = x_i P$. The mobile payment platform chooses a number $x_{MPP} \in Z_q^*$ and $P_{MPP} = x_{MPP} P$.

5.2 Transaction Process

For the convenience of description and discussion, we divide the transaction process into 12 steps here. The steps are shown in Figure 1, Figure 2, and Figure 3. We introduce each step in detail.

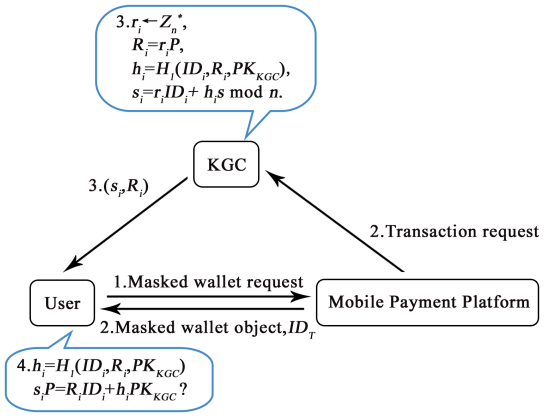


Figure 1: Steps 1-4 of the transaction process

1) A user i makes a new transaction by clicking on an application. When the application receives the request, it will send a masked wallet request to the mobile payment platform.

2) After receiving the masked wallet request, the mobile payment platform sends a request to the KGC. And the mobile payment platform sends a masked wallet object including the transaction ID_T and the purchase information to the user i.

3) Given $Params$, s, the identity ID_i of the user i, and a unique identity for this transaction ID_T , the KGC generates a random number $r_i \in Z_q^*$, and calculates $R_i = r_i P$, $h_i = H_1(ID_i, R_i, PK_{KGC})$ and $s_i = r_i ID_i + h_i s \mod q$. The KGC sends (s_1, R_i) to the user i. Here (s_1, R_i) is the partial private key of user i, which needs to be transmitted through a secure channel.

4) After receiving (s_i, R_i), the user i calculates $h_i = H_1(ID_i, R_i, PK_{KGC})$ and examines if the equation $s_i P = R_i ID_i + h_i PK_{KGC}$ is satisfied. If the equation is satisfied, the transaction process will continue. Otherwise, the user's application screen will show the error message. Then the user i verifies the purchase information. When the accepted information is verified to be correct, the application will show a confirmation page for the user itself to confirm. The user i click on the confirm order button to continue.

5) The application establishes a service connection to the mobile payment platform and sends a full wallet request to the mobile payment platform. Then, the user i sends a verification request to the KGC.

6) With $Params$, s, the identity of the mobile payment platform ID_{MPP} and ID_T, the KGC creates a random number $r_{MPP} \in Z_q^*$, and computes $R_{MPP} = r_{MPP} P$, $h_{MPP} = H_1(ID_{MPP}, ID_T, R_{MPP}, PK_{KGC})$, and $s_{MPP} = r_{MPP} ID_{MPP} + h_{MPP} s \mod q$. Next,the KGC

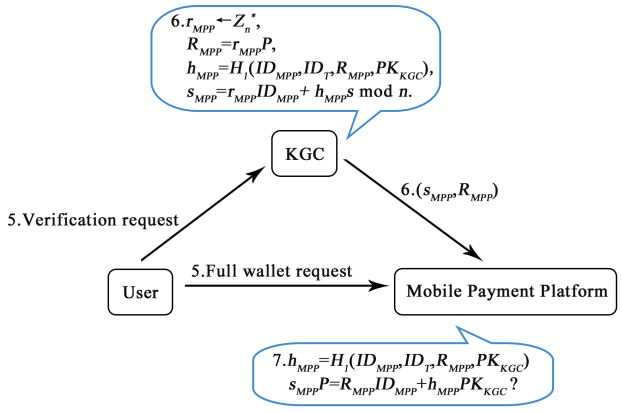


Figure 2: Steps 5-7 of the Transaction Process

sends (s_{MPP}, R_{MPP}) to the mobile payment platform. (s_{MPP}, R_{MPP}) is the partial private key of the mobile payment platform.

7) After receiving (s_{MPP}, R_{MPP}), the mobile payment platform calculates $h_{MPP} = H_1(ID_{MPP}, ID_T, R_{MPP}, PK_{KGC})$ and checks if $s_{MPP}P = R_{MPP}ID_{MPP} + h_{MPP}PK_{KGC}$ is satisfied. If the validation fails, an error message will be returned and the transaction will be closed.

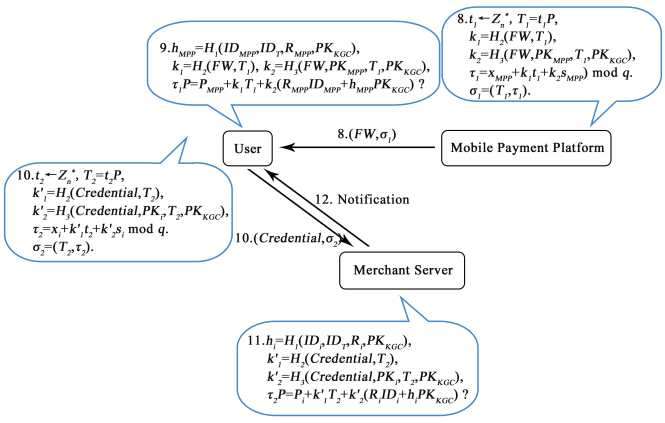


Figure 3: Steps 8-12 of the transaction process

8) If the partial private key of the mobile payment platform is verified as $VALID$, the mobile payment platform will create a $FULLWALLET(FW)$. FW contains all the details of the mobile payment credentials for the transaction. With $Params$, ID_{MPP}, ID_i, ID_T, PK_{MPP}, (s_{MPP}, R_{MPP}), x_{MPP}, and FW, the mobile payment platform calculates $T_1 = t_1P$ with a random number $t_1 \in Z_q^*$, $k_1 = H_2(FW, T_1)$ and $k_2 = H_3(FW, PK_{MPP}, T_1, PK_{KGC})$. Then the mobile payment platform computes $\tau_1 = x_{MPP} + k_1t_1 + k_2s_{MPP} \mod q$. The mobile payment platform generates the signature $\sigma_1 = (T_1, \tau_1)$. Finally, the mobile payment platform returns (FW, σ_1) to the user i.

9) After receiving (FW, σ_1), the user i calculates $k_1 = H_2(FW, T_1)$, $k_2 = H_3(FW, PK_{MPP}, T_1, PK_{KGC})$, $h_{MPP} = (ID_{MPP}, ID_T, R_{MPP}, PK_{KGC})$, and checks if the equation $\tau_1 P = P_{MPP} + k_1T_1 + k_2(R_{MPP}ID_{MPP} + h_{MPP}PK_{KGC})$ is satisfied. If it is satisfied, the transaction will continue.

10) The application forwards the mobile payment credentials in the FW to the merchant server in order to process the payment. The user i performs the following computations and then sends $(Credential, \sigma_2)$ to the merchant server. That is, given $Params$, ID_i, ID_T, PK_i, (s_i, R_i), x_i, and $Credential$, the application computes $T_2 = t_2P$, $k_1^{'} = H_2(Credential, T_2)$, $k_2^{'} = H_3(Credential, PK_i, T_2, PK_{KGC})$ and $\tau_2 = x_i + k_1^{'}t_2 + k_2^{'}s_i \mod q$, where $h_i = H_1(ID_i, ID_T, R_i, PK_{KGC})$ and $t_2 \in Z_q^*$ is a random number. The signature is $\sigma_2 = (T_2, \tau_2)$.

11) Once the merchant server receives $(Credential, \sigma_2)$, the server verifies it's validity. That is, given $Params$, ID_i, ID_T , PK_i, $Credential$, $\sigma_2 = (T_2, \tau_2)$, the merchant server computes $h_i = H_1(ID_i, ID_T, R_i, PK_{KGC})$, $k_1^{'} = H_2(Credential, T_2)$ and $k_2^{'} = H_3(Credential, PK_i, T_2, PK_{KGC})$. Then, the server examines if the equation $\tau_2 P = P_i + k_1^{'}T_2 + k_2^{'}(R_iID_i + h_iPK_{KGC})$ is satisfied. If it is satisfied, the transaction will continue.

12) The merchant server sends a notification of the transaction to the user i. The application in mobile device will display the notification of the payment.

6 Security Analysis

Our improved scheme is analyzed to be provably secure against the super type I and type II adversaries based on the intractability of elliptic curve discrete logarithm problem (ECDLP) under the random oracle model.

Lemma 1. *According to the security model, if a super type I adversary $\mathcal{A}_\mathcal{I}$ could succeed in **Game 1** with non-negligible probability β in a polynomial time, $\mathcal{C}_\mathcal{I}$ could solve the ECDLP with the possibility:*

$$\beta^{'} \geq (1 - q_{H_1}/q)^{q_{cu}}(1 - 1/q_{cu})^{q_{ep}}(1/q_{cu}) \\ (1 - q_{H_2}/q)(1 - q_{H_3}/q)\beta.$$

In the equation above, q_{H_1}, q_{H_2}, q_{H_2}, q_{cu} and q_{ep} represents the number of H_1 query, H_2 query, H_3 query, $CreateUser$ query and $ExtractPartialPrivateKey$ query respectively.

Proof. Assuming $\mathcal{A}_\mathcal{I}$ could succeed in **Game 1** with possibility β, the challenger $\mathcal{C}_\mathcal{I}$ is required to solve the problem in which $Q = sP$, where P is a generator of

group G over an elliptic curve with an order q. $\mathcal{C}_{\mathcal{I}}$ sets $PK_{KGC} = Q$ and need to compute s according to the attacker's forged signature. In the following processes, $\mathcal{C}_{\mathcal{I}}$ maintains four lists L_1, L_2 , L_3 and L_u which are initially empty, to record the information about H_1 query, H_2 query, H_3 query and $CreateUser$ query.

- **Phase 1.** $\mathcal{C}_{\mathcal{I}}$ randomly chooses an ID^* as the target identity, and sets public parameters $Params = (G, P, PK_{KGC} = Q)$ and sends $Params$ to $\mathcal{A}_{\mathcal{I}}$.
- **Phase 2.** $\mathcal{A}_{\mathcal{I}}$ can issue the following queries in polynomial times.
  - H_1 query: When $\mathcal{A}_{\mathcal{I}}$ issues the H_1 query with (ID_i, R_i, PK_{KGC}), $\mathcal{C}_{\mathcal{I}}$ searches list L_1 and returns the record if the information of (ID_i, R_i, PK_{KGC}) exists. Otherwise, $\mathcal{C}_{\mathcal{I}}$ queries $CreateUser(ID_i)$ and extracts h_i from the returned parameters and sends it to $\mathcal{A}_{\mathcal{I}}$.
  - H_2 query: When $\mathcal{A}_{\mathcal{I}}$ issues the H_2 query with (m, T), $\mathcal{C}_{\mathcal{I}}$ searches the list L_2 and returns the record if (m, T) exists. Otherwise, $\mathcal{C}_{\mathcal{I}}$ randomly chooses $v \in Z_q^*$, calculates $V = vP$, which sets $H_2(m, T) = v$. $\mathcal{C}_{\mathcal{I}}$ sends v to $\mathcal{A}_{\mathcal{I}}$ at last.
  - H_3 query: When $\mathcal{A}_{\mathcal{I}}$ issues the H_2 query with (m, P_i, T, PK_{KGC}), $\mathcal{C}_{\mathcal{I}}$ searches the list L_3 and returns the record if (m, P_i, T, PK_{KGC}) exists. Otherwise, $\mathcal{C}_{\mathcal{I}}$ randomly chooses $v \in Z_q^*$, calculates $V = vP$, which sets $H_3(m, P_i, T, PK_{KGC}) = v$. $\mathcal{C}_{\mathcal{I}}$ sends v to $\mathcal{A}_{\mathcal{I}}$ at last.
  - $CreateUser(ID)$. When $\mathcal{A}_{\mathcal{I}}$ issues the $CreateUser$ query with ID_i, $\mathcal{C}_{\mathcal{I}}$ searches the list L_u and returns the PK_i if there exists a record of ID_i. Otherwise, $\mathcal{C}_{\mathcal{I}}$ will complete the following steps to create a record and add it to L_u. If $ID_i \neq ID^*$, $\mathcal{C}_{\mathcal{I}}$ randomly chooses s_i, h_i, x_i, and computes $R_i = {h_i}^{-1}(s_iP - PK_{KGC})$, $P_i = x_iP$. If $ID = ID^*$, $\mathcal{C}_{\mathcal{I}}$ randomly chooses r_i, h_i, x_i and sets $R_i = r_iP$, $P_i = x_iP$, $s_i = null$. In addition, if there exists a record $(ID_i, R_i, H_1(ID_i, R_i, PK_{KGC}))$ but $H_1(ID_i, R_i, PK_{KGC}) \neq h_i$, $\mathcal{C}_{\mathcal{I}}$ aborts the game. Otherwise, $\mathcal{C}_{\mathcal{I}}$ returns PK_i to $\mathcal{A}_{\mathcal{I}}$ and adds the record $(ID_i, s_i, x_i, R_i, P_i)$ and (ID_i, R_i, h_i) to the lists L_u and L_1 respectively.
  - $ReplacePublicKey(ID_i, x_i^{'}, PK_i^{'})$. $\mathcal{C}_{\mathcal{I}}$ will replace the user's public key with $(x_i^{'}, PK_i^{'})$. We assume that the $CreateUser$ query has been executed with the identity of ID. And given the ability of a super type I adversary, $\mathcal{A}_{\mathcal{I}}$ is unnecessary to provide the value of x_i, which means x_i can be null.
  - $ExtractSecretValue(ID_i)$. $\mathcal{C}_{\mathcal{I}}$ searches the list L_u and returns x_i to $\mathcal{A}_{\mathcal{I}}$ if there exists a record $(ID_i, s_i, x_i, R_i, P_i)$ on ID. Otherwise, $\mathcal{C}_{\mathcal{I}}$ issues $CreateUser$ query on ID_i and returns x_i to $\mathcal{A}_{\mathcal{I}}$. In addition , $\mathcal{C}_{\mathcal{I}}$ maintains idle if $ReplacePublicKey$ has been queried on ID_i where x_i is null.
  - $ExtractPartialPrivateKey(ID_i)$. If $ID_i \neq ID^*$, $\mathcal{C}_{\mathcal{I}}$ searches the list L_u and returns s_i. If $ID_i = ID^*$, $\mathcal{C}_{\mathcal{I}}$ aborts the game. We assume that $CreateUser$ on ID_i has been executed.
  - $SuperSign(ID_i, m)$. $\mathcal{C}_{\mathcal{I}}$ searches for the record $(ID_i, s_i, x_i, R_i, P_i)$, (ID_i, R_i, PK_{KGC}), (m, T) and (m, P_i, T, PK_{KGC}) in the lists L_u, L_1, L_2 and L_3. If $ID = ID^*$ or $x_i = null($ $ReplacePublicKey$ has been queried on ID^*), $\mathcal{C}_{\mathcal{I}}$ randomly chooses $T \in G$, $\tau \in Z_q^*$ and sends (T, τ) to $\mathcal{A}_{\mathcal{I}}$. Otherwise, $\mathcal{C}_{\mathcal{I}}$ randomly chooses $t, k_1, k_2 \in Z_q^*$, and calculates $T = tP$,$\tau = x_i + k_1t + k_2s_i \mod q$. Then, $\mathcal{C}_{\mathcal{I}}$ adds (m, T, k_1) and $(m, P_i, T, PK_{KGC}, k_2)$ to L_2 and L_3 respectively. Finally, $\mathcal{C}_{\mathcal{I}}$ returns (T, τ) to $\mathcal{A}_{\mathcal{I}}$. In that case, the signature will be verified because the equation $\tau P = P_i + k_1T + k_2(PK_{KGC} + h_iR_i)$ is satisfied.
- **Phase 3.** $\mathcal{A}_{\mathcal{I}}$ gives a forgery signature (T^*, τ^*). Then $\mathcal{C}_{\mathcal{I}}$ checks if the identity of this signature is ID^*. If not, $\mathcal{C}_{\mathcal{I}}$ ends the game. Otherwise, $\mathcal{C}_{\mathcal{I}}$ searches the lists L_u, L_1, L_2 and L_3 for the records $(ID_i, s_i, x_i, R_i, P_i)$, $h_i = (ID_i, R_i, PK_{KGC})$, $k_1(m, T)$, $k_2 = (m, P_i, T, PK_{KGC})$. If there is no records of h_i, k_1 and k_2 in the lists, $\mathcal{C}_{\mathcal{I}}$ aborts the game. Next, if the signature can be authenticated, the equation $\tau^* = x_i^* + k_1^*t^* + k_2^*(s + r_i^*h_i^*) \mod q$ is satisfied. In this equation, there are only three unknown values i.e. x_i^*, s, and t^*. In addition, x_i may not be provided according to the $ReplacePublicKey$ query and s is the value $\mathcal{C}_{\mathcal{I}}$ need to compute to solve the ECDLP problem. According to the principle of forking lemma [17], $\mathcal{C}_{\mathcal{I}}$ needs to repeat the same steps and provide different values of h_i, k_1 and k_2, then compute three different signatures as follows.
  1) $\tau^* = x_i^* + k_1^*t^* + k_2^*(s + r_i^*h_i^*)$;
  2) $\tau^{*'} = x_i^* + k_1^{'}t^* + k_2^{'}(s + r_i^*h_i*')$;
  3) $\tau^{*''} = x_i^* + k_1^{''}t^* + k_2^{''}(s + r_i^*h_i^{''})$;

Therefore, $\mathcal{C}_{\mathcal{I}}$ acquires the value of s for the ECDLP problem. Let the probability of $\mathcal{C}_{\mathcal{I}}$ solving the ECDLP problem be $Pr[succ]$. The equation denotes the probability that $pr[succ] = pr[E_1 \wedge E_2]$ where E_1 means the **Game 1** successfully completed all the steps without being terminated and E_2 means the signature forged by $\mathcal{A}_{\mathcal{I}}$ with ID^* identity is verified. Suppose that $\mathcal{A}_{\mathcal{I}}$ can forge a valid signature with probability β, we can computes: $pr[succ] = pr[E_1 \wedge E_2] = pr[E_1]pr[E_2E_1] = pr[E_1]\beta$. E_1 requires these conditions corresponding to the respective probabilities: The probability of $(1 - q_{H_1}/q)^{q_{cu}}$ represents that there exists no collisions in the $CreateUser$ query. The probability of $(1 - 1/q_{cu})^{q_{ep}}$ represents that

$\mathcal{A}_{\mathcal{I}}$ doesn't query the partial private key of ID^*. The probability of $(1/q_{cu})$ represents that $\mathcal{A}_{\mathcal{I}}$ sends the signature where $ID = ID^*$. The probability of $(1 - q_{H_2}/q)$ denotes that the values of $H_2(m, T)$ about the forged signature sent by $\mathcal{A}_{\mathcal{I}}$ can be found in the list L_2 in **Phase 3**. The probability of $(1 - q_{H_3}/q)$ denotes that the values of $H_3(m, P_i, T, PK_{KGC})$ about the forged signature sent by $\mathcal{A}_{\mathcal{I}}$ can be found in the list L_3 in **Phase 3**.

Therefore, we have $\beta^{'} \geq (1 - q_{H_1}/q)^{q_{cu}}(1 - 1/q_{cu})^{q_{ep}}$ $(1/q_{cu})(1 - q_{H_2}/q)(1 - q_{H_3}/q)\beta$. □

Lemma 2. *Assuming the probability that $\mathcal{A}_{\mathcal{II}}$ generates a legal signature and wins **Game 2** in polynomial time is β, the probability that $\mathcal{C}_{\mathcal{II}}$ is able to solve the ECDLP problem is $\beta^{''}$:*

$$\beta^{''} \geq (1 - q_{H_1}/q)^{q_{cu}}(1 - 1/q_{cu})^{q_{rp}}(1 - 1/q_{cu})^{q_{es}} (1/q_{cu})(1 - q_{H_2}/q)(1 - q_{H_3}/q)\beta.$$

q_{H_1}, q_{H_2}, q_{H_3}, q_{cu}, q_{es} and q_{rp} represents the number of H_1 query, H_2 query, H_3 query, $CreateUser$ query, $ExtratSecretValue$ query and $ReplacePublicKey$ query.

Proof. Assuming that $\mathcal{A}_{\mathcal{II}}$ can win in **Game 2** between $\mathcal{A}_{\mathcal{II}}$ and $\mathcal{C}_{\mathcal{II}}$ in a polynomial time, for a given G where $Q = sP$ and s is unknown, $\mathcal{C}_{\mathcal{II}}$ could get the value of s based on the signature given by $\mathcal{A}_{\mathcal{II}}$. $\mathcal{C}_{\mathcal{II}}$ maintains L_u, L_1, L_2 and L_3 initially empty for $CreateUser$ query, H_1 query, H_2 query and H_3 query as in **Game 2**.

- **Phase 1.** $\mathcal{C}_{\mathcal{II}}$ randomly chooses an ID^* and a number $s \in Z_q^*$, calculates $PK_{KGC} = sP$, sets $Params = (G, P, PK_{KGC})$, then sends $Params$ and s to $\mathcal{A}_{\mathcal{II}}$.
- **Phase 2.** $\mathcal{A}_{\mathcal{II}}$ can issue any of the following queries in polynomial times.
  - H_1, H_2 and H_3 queries are as same as in **Game 1**.
  - $CreateUser(ID_i)$. If there exists a record of ID_i, $\mathcal{C}_{\mathcal{II}}$ returns the public key PK_i. Otherwise, if $ID \neq ID^*$, $\mathcal{C}_{\mathcal{II}}$ randomly chooses r_i, x_i and $h_i \in Z_q^*$, and calculates $R_i = r_iP$, $s_i = s + h_ir_i$, $P_i = x_iP$. If $ID_i = ID^*$, $\mathcal{C}_{\mathcal{II}}$ randomly chooses r_i, h_i, and calculates $R_i = r_iP$, $s_i = s + h_ir_i$, $P_i = Q$, $x_i = null$. Finally, $\mathcal{C}_{\mathcal{II}}$ returns $PK_i = (R_i, P_i)$ to $\mathcal{A}_{\mathcal{II}}$ and adds $(ID_i, r_i, s_i, x_i, R_i, P_i)$ and (ID_i, R_i, h_i) to the lists L_u and L_1 respectively.
  - $ReplacePublicKey(ID_i, x_i^{'}, PK_i^{'})$: If $ID = ID^*$, $\mathcal{C}_{\mathcal{II}}$ aborts the game. Otherwise, $\mathcal{C}_{\mathcal{II}}$ replaces ID_i's public key with given $PK_i^{'}$ even if $x_i^{'}$ is null, and updates the list L_u.
  - $ExtractSecretValue(ID_i)$. Note that $\mathcal{A}_{\mathcal{II}}$ is not allowed to access the secret value of ID^*, and $\mathcal{C}_{\mathcal{II}}$ will abort the game if $ID_i = ID^*$. Otherwise, $\mathcal{C}_{\mathcal{II}}$ searches the record of ID_i and returns x_i to $\mathcal{A}_{\mathcal{II}}$. However $\mathcal{C}_{\mathcal{II}}$ may outputs a null if the PK_i has been replaced with a null x_i.
  - $ExtractPartialPrivateKey(ID_i)$. $\mathcal{C}_{\mathcal{II}}$ returns s_i from the list L_u if there exists a record of ID_i.
  - $SuperSign(ID_i, m)$. $\mathcal{C}_{\mathcal{II}}$ searches for the records of (ID_i) in the list L_u, L_1, L_2 and L_3. If $x_i = null$, which means $ID_i = ID^*$ or the PK_i has been replaced with $x_i^{'} = null$, $\mathcal{C}_{\mathcal{II}}$ randomly chooses $T \in G$, $\tau \in Z_q^*$ and sends (T, τ) to $\mathcal{A}_{\mathcal{II}}$. Otherwise, $\mathcal{C}_{\mathcal{II}}$ randomly chooses $t, k_1, k_2 \in Z_q^*$, calculates $T = tP$, $\tau = x_i + k_1t + k_2(s + r_ih_i) \mod q$. Eventually the signature (T, τ) generated by $SuperSign$ query will be sent to $\mathcal{A}_{\mathcal{II}}$ and the record (m, T, k_1) and $(m, P_i, T, PK_{KGC}, k_2)$ will be added to L_2 and L_3.
- **Phase 3.** $\mathcal{C}_{\mathcal{II}}$ submits the forged signature (T^*, τ^*) to $\mathcal{C}_{\mathcal{II}}$ at this stage. If the ID of this signature is not ID^*, the game will be terminated. Otherwise, $\mathcal{C}_{\mathcal{II}}$ will check the records of ID^* in lists L_u, L_1, L_2 and L_3. Then $\mathcal{C}_{\mathcal{II}}$ searches if $k_1 = (m, T)$ and $k_2 = (m, P_i, T, PK_{KGC})$ exist. If not, $\mathcal{C}_{\mathcal{II}}$ aborts the game.

From the signature given by $\mathcal{A}_{\mathcal{II}}$, we can get $\tau^* = x_i^* + k_1^*t^* + k_2^*(s + r_i^*h_i^*) \mod q$. And according to the forking lemma [17], $\mathcal{C}_{\mathcal{II}}$ can obtain another signature submitted by $\mathcal{A}_{\mathcal{II}}$, which satisfies the equation: $\tau^{'*} = x_i^* + k_1^{'*}t^* + k_2^{'*}(s + r_i^{'*}h_i^{'*}) \mod q$.

From these two equations extracted from the forged signatures, the values of x_i^* and t^* can be calculated. $\mathcal{C}_{\mathcal{II}}$ obtains the value of x_i^*, which means $\mathcal{C}_{\mathcal{II}}$ solves the ECDLP problem.

$\mathcal{A}_{\mathcal{II}}$ successfully generates the right signature with the following conditions:

No collision of the hash function is happened in the $CreateUser$ query. The probability is $(1 - q_{H_1}/q)^{q_{cu}}$. $\mathcal{A}_{\mathcal{II}}$ has not queried $ReplacePublicKey$ with ID^*. The probability is $(1 - q_{H_1}/q)^{q_{cu}}$. $\mathcal{A}_{\mathcal{II}}$ has not queried $ExtractSecretValue$ with ID^*. The probability is $(1 - 1/q_{cu})^{q_{rp}}$. The submitted signature must satisfies $ID = ID^*$. The probability is $(1/q_{cu})$. The probability that $k_1 = H_2(m, T)$ will be found in **Phase 3** is $(1 - q_{H_2}/q)$. The probability that $k_2 = H_3(m, P_i, T, PK_{KGC})$ will be found in **Phase 3** is $(1 - q_{H_3}/q)$. In conclusion, if $\mathcal{A}_{\mathcal{II}}$ can complete **Game 2** with a non-negligible probability β, the probability that $\mathcal{C}_{\mathcal{II}}$ can solve the ECDLP problem is:

$$\beta^{''} \geq (1 - q_{H_1}/q)^{q_{cu}}(1 - 1/q_{cu})^{q_{rp}}(1 - 1/q_{cu})^{q_{es}} (1/q_{cu})(1 - q_{H_2}/q)(1 - q_{H_3}/q)\beta.$$

□

Theorem 1. *Our improved scheme is said to be existentially unforgeable against adaptively chosen message and*

Table 1: Performance evaluation and security

	Jia *et al.* [12]	**Pakniat *et al.* [16]**	**Yeh *et al.*'s [24]**	**Our scheme**
Sign	$2T_h + T_{sm} + 2T_m + T_i$	$4T_h + T_{sm} + 2T_m + 2T_i$	$T_h + T_{sm} + T_m$	$2T_h + T_{sm} + 2T_m$
Verify	$2T_h + 4T_{sm} + 2T_a$	$5T_h + 3T_{sm} + 3T_a$	$2T_h + 4T_{sm} + 3T_a$	$3T_h + 4T_{sm} + 3T_a$
Total	$4T_h + 5T_{sm} + 2T_m + 2T_a$ $+T_i$(0.764ms)	$9T_h + 4T_{sm} + 2T_m$ $+3T_a + 2T_i$(0.781ms)	$4T_h + 5T_{sm} + 1T_m$ $+3T_a$(0.701ms)	$5T_h + 5T_{sm} + 2T_m$ $+3T_a$(0.732ms)
Size of Signature	$\|G\| + \|Z_q^*\|$ (1184 bits)	$\|G\| + \|Z_q^*\|$ (1184 bits)	$\|G\| + \|G\| + \|Z_q^*\|$ (2208 bits)	$\|G\| + \|Z_q^*\|$ (1184 bits)
Public key replacement	Yes	Yes	Yes	No

identity attacks if for any polynomial-time super adversary $\mathcal{A}$, the advantage $Succ_A(\lambda)$ is negligible assuming the ECDLP assumption holds.

7 Performance

Here we compare the performance of our improved scheme and other schemes [12, 16, 24]. In order to ensure the security level, we choose a widely accepted parameter size since 160-bit private key in the ECC has the same level of security as the 1024-bit RSA private key [6]. Let G be an additive cyclic group of order q over an elliptic curve E/F_p, where p is a prime number of 512 bits and q is a prime number of 160 bits. P is a generator of G. The program is performed on the Windows 10 system 64 bit with an intel(R) Core(TM) i3-8100 CPU @ 3.60GHz and 8.00GB Random Access Memory(RAM). The relative operations are implemented using Pairing Based Cryptography 2.00(JPBC) [4].

A comparison of the performance in terms of computational cost, communication overhead and the security, of our proposed scheme and the schemes [12, 16, 24] is presented in Table 1. The symbols used in Table 1 are listed below:

T_h: The execution time of performing a one-way hash function.

T_{sm}: The execution time of performing an ECC-based scalar multiplication.

T_m: The execution time of performing a general multiplication operation.

T_a: The execution time of performing a general addition operation.

T_i: The execution time of modular inverse in Z_q^*.

$|G|$: The size of the point in group G.

$|Z_q^*|$: The size of the number in Z_q^*.

The execution time required for each algorithm and time consumption on the total time are shown in Table 1. It shows the efficiency of our scheme is slightly higher than that of scheme [12] and scheme [16]. However, our scheme require $5T_h+5T_{sm}+2T_m+3T_a$ (0.732ms) while Yeh *et al.*'s scheme [24] requires $4T_h + 5T_{sm} + 1T_m + 3T_a$ (0.701ms).

At the same time, the comparison of communication costs is also noteworthy. The most critical factor affecting communication overhead is the size of the signature. Table 1 also shows the difference in the size of the signature between our CLS scheme and the schemes [12,16,24]. We can find that the signature size of our scheme is the same as that of scheme [12] and C [16]. Yeh *et al.*'s scheme [24] needs $|G| + |G| + |Z_q^*|$(2208 bits). It can be seen that our scheme has advantages in reducing communication overhead.

In terms of security, the schemes [12, 16, 24] are unable to prevent public key replacement attacks. But our improved scheme is unforgeable against super Type I adversaries [25, 26].

Comparing to the schemes [12, 16, 24], the efficiency of our scheme is competitive and the security level is enhanced.

8 Conclusions

In this paper, we analyzed in detail that Yeh *et al.*'s scheme [24] cannot resist the public key replacement attack. In other words, a type I adversary can easily forge a valid signature on any message and cheat the verifier. Then we proposed an improved scheme and proved its security against super type I and type II adversaries. The transaction scheme used for the actual mobile payment applications is also presented. The evaluation shows the performance of our proposed scheme is comparable to related schemes [12, 16, 24].

Research and discussion of group key agreement for IoT devices would be our future work.

References

[1] S. S. Al-Riyami and K. G. Paterson, "Certificateless public key cryptography," in *International Conference on the Theory and Application of Cryptology and Information Security*, pp. 452–473, 2003.

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] Z. Caixue, "Certificateless signcryption scheme without random oracles," *Chinese Journal of Electronics*, vol. 27, no. 5, pp. 1002–1008, 2018.

[4] A. D. Caro and V. Iovino, "JPBC: Java pairing based cryptography," in *IEEE Symposium on Computers and Communications (ISCC'11)*, pp. 850–855, 2011.

[5] T. H. Feng, M. S. Hwang, and L. W. Syu, "An authentication protocol for lightweight NFC mobile sensors payment," *Informatica*, vol. 27, no. 4, pp. 723–732, 2016.

[6] P. Gallagher, "Digital signature standard (DSS)," *Federal Information Processing Standards Publications*, pp. 186–3, 2013. (`https://csrc.nist.gov/csrc/media/publications/fips/186/3/archive/2009-06-25/documents/fips_186-3.pdf`)

[7] Z. Guo, "Cryptanalysis of a certificateless conditional privacy-preserving authentication scheme for wireless body area networks," *International Journal of Electronics and Information Engineering*, vol. 11, no. 1, pp. 1–8, 2019.

[8] D. He, J. Chen, and R. Zhang, "An efficient and provably-secure certificateless signature scheme without bilinear pairings," *International Journal of Communication Systems*, vol. 25, no. 11, pp. 1432–1442, 2012.

[9] X. Huang, W. Susilo, Y. Mu, and F. Zhang, "On the security of certificateless signature schemes from asiacrypt 2003," in *International Conference on Cryptology and Network Security*, pp. 13–25, 2005.

[10] M. S. Hwang, S. F. Tzeng, C. S. Tsai, "Generalization of proxy signature based on elliptic curves", *Computer Standards & Interfaces*, vol. 26, no. 2, pp. 73–84, 2004.

[11] S. Ji, Z. Gui, T. Zhou, H. Yan, and J. Shen, "An efficient and certificateless conditional privacy-preserving authentication scheme for wireless body area networks big data services," *IEEE Access*, vol. 6, pp. 69603–69611, 2018.

[12] X. Jia, D. He, Q. Liu, and K. K. R. Choo, "An efficient provably-secure certificateless signature scheme for internet-of-things deployment," *Ad Hoc Networks*, vol. 71, pp. 78–87, 2018.

[13] X. X. Li, K. F. Chen, and L. Sun, "Certificateless signature and proxy signature schemes from bilinear pairings," *Lithuanian Mathematical Journal*, vol. 45, no. 1, pp. 76–83, 2005.

[14] Z. Liu, Y. Hu, X. Zhang, and H. Ma, "Certificateless signcryption scheme in the standard model," *Information Sciences*, vol. 180, no. 3, pp. 452–464, 2010.

[15] S. Miao, F. Zhang, S. Li, and Y. Mu, "On security of a certificateless signcryption scheme," *Information Sciences*, vol. 232, pp. 475–481, 2013.

[16] N. Pakniat and B. A. Vanda, "Cryptanalysis and improvement of a pairing-free certificateless signature scheme," in *The 15th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC'18)*, pp. 1–5, 2018.

[17] D. Pointcheval and J. Stern, "Security arguments for digital signatures and blind signatures," *Journal of cryptology*, vol. 13, no. 3, pp. 361–396, 2000.

[18] A. Shamir, "Identity-based cryptosystems and signature schemes," in *Workshop on the theory and application of cryptographic techniques*, pp. 47–53, 1984.

[19] S. Shan, "An efficient certificateless signcryption scheme without random oracles," *International Journal of Electronics and Information Engineering*, vol. 11, no. 1, pp. 9–15, 2019.

[20] W. Stallings and M. P. Tahiliani, *Cryptography and Network Security: Principles and Practice*, 2014.

[21] M. Tian and L. Huang, "Cryptanalysis of a certificateless signature scheme without pairings," *International Journal of Communication Systems*, vol. 26, no. 11, pp. 1375–1381, 2013.

[22] J. L. Tsai, N. W. Lo, and T. C. Wu, "Weaknesses and improvements of an efficient certificateless signature scheme without using bilinear pairings," *International Journal of Communication Systems*, vol. 27, no. 7, pp. 1083–1090, 2014.

[23] Y. Wang, C. Hahn, and K. Sutrave, "Mobile payment security, threats, and challenges," in *The Second International Conference on Mobile and Secure Services*, pp. 1–5, 2016.

[24] K. H. Yeh, "A secure transaction scheme with certificateless cryptographic primitives for iot-based mobile payments," *IEEE Systems Journal*, vol. 12, no. 2, pp. 2027–2038, 2018.

[25] Z. C. Zhang, Y. L Liu, X. C. Yin, and K. K. Huang, ""analysis and improvement of certificateless signature schemes," *Journal of Cryptologic Research*, vol. 7, no. 3, pp. 389–403, 2020.

[26] Z. Zhang, Y. Liu, X. Yin, and X. Li, "A new pairing-free certificateless signature scheme for internet of things," in *International Conference on Science of Cyber Security*, pp. 371–379, 2019.

Biography

Fen Yan completed her Ph.D degree in department of computer science and technology in 2009 at Nanjing University. Her research is focused on computer network and information security.

Linggen Xing received the B.Eng. degree at Yangzhou University in 2018. He is currently a master candidate at the college of information engineering, Yangzhou University. He has research interests in wireless networks and cryptography.

Zhenchao Zhang is currently a master candidate at the college of information engineering, Yangzhou University. His main research interests include information security and cryptography.

Fast Scalar Multiplication Algorithm Based on Co_Z Operation and Conjugate Point Addition

Shuang-Gen Liu, Ying Zhang, and Shi-Yao Chen
(Corresponding author: Shuang-Gen Liu)

School of Cyberspace Security, Xi'an University of Posts and Telecommunications
Xi'an 710121, China
Email: liusgxupt@163.com

(Received July 4, 2020; Revised and Accepted Apr. 24, 2021; First Online Aug. 17, 2021)

Abstract

This paper aims to study the elliptic curve over a prime field in Jacobian projective coordinate to improve the efficiency of the elliptic curve scalar multiplication algorithm. There are three main contributions: Firstly, based on the Co_Z operation, the operation of $3P$ and $5P$ have been improved, which costs decreased by 18.17%, 13.5% respectively compared with existing formulas. Secondly, by combining the Co_Z operation with Conjugate point addition, the composite formula of $3P\pm Q$ and $5P\pm Q$ has been proposed, which can be adopted to accelerate most scalar multiplication, the costs of $3P+Q$ and $5P+Q$ decreased by 10.2%, 11.5% respectively, the additional costs for getting $3P-Q$ from $3P+Q$ is only $1M+1S$. The same is true for $5P-Q$. Thirdly, we proposed a new encoding way based on quinary. Specifically, this paper describes two efficient scalar multiplication algorithms against SPA. Compared with the Highest-weight Symbolic Ternary Form (HSTF) scalar multiplication with Anti-Simple Power Attack (SPA), BJMontgomery algorithm, Symbolic Ternary Form (STF) scalar multiplication algorithm, the efficiency of the first algorithm is increased by 19.3%, 10%, 19% respectively, and that of the second algorithm efficiency is increased by 27.8%, 19.5%, 27.5% respectively.

Keywords: Co_Z operation; Conjugate Point Addition; Elliptic Curve Cryptography; New Symmetry Ternary Form; Scalar Multiplication

1 Introduction

1.1 Background

Elliptic Curve Cryptography(ECC) [13, 25], a public key cryptosystem, which introduced by Koblitz and Miller independently since 1985, has received widely attention due to its shorter key. ECC is based on a one-way function (elliptic curve discrete logarithm), which is much more complicated than the discrete logarithm used by RSA [30]. And this one-way function is more difficult than RSA. Therefore, ECC has higher security and is well suited for use in embedded mobile environments with limited resources. ECC provides the same level of security with a shorter secret key than other cryptosystems, for example, 160-bit ECC has the same security strength as 1024-bit RSA [23], so there are various cryptosystems based on ECC. Encryption, decryption, signature, and verification operations are all related to scalar multiplication of points on an elliptic curve. The computational performance of the entire cryptosystem depends on the computational speed of scalar multiplication on an elliptic curve, which is the multiplication of the base point P on a given elliptic curve with a private key k, which is a positive integer, to get the public key $kP = P + P + ... + P$ (k times).

There are two main method to improve the efficiency of scalar multiplication: The first method is encoding k with different ways either Binary form, Non-adjacent form (NAF) [10, 12], or Balanced symmetric ternary [5], they can reduce the number of addition and doubling operations of scalar multiplication; The second method is reducing the computation of point addition and doubling operation. In general, field inversion, field multiplication, squartion and cube operation are denoted as I, M, S, C respectively, the quantity relationship among them is $S/M = 0.8$, $C = 1.37M$, $I/M>8$. It is obvious that field inversion is the most time-consuming operation, it always be eliminated by exchanging the coordinate, the Jacobian coordinate is the most commonly used [4].

In recent years, ECC has been studied by many researchers and has made great progress. In 2007, Co_Z operation was first introduced by Meloni [24], which is computing the point addition of two point with the same Z-coordinate, then used to develop a specific exponentiation algorithm, based on Zeckendorf representation(k is an integer and $(F_i)_{i\geq 0}$ is the Fibonacci sequence, then k can be uniquely written as $k = \sum_{i=2}^{l} d_i F_i$, with $d_i \in \{0, 1\}$ and $d_i d_{i+1} = 0$, so can be denoted as $k = (d_{l-1}, ...d_2)_z$, where, the limit of i is $l-1$, and the z stands the sequence is Zeckendorf representation); Then, Goundar *et al.* [7] de-

rived the formula of $P \pm Q$ on Weierstrass elliptic curve, by combining the conjugate point addition with Co_Z operation and used for Montgomery ladder algorithm; In 2008, the rule of $dP + Q$ over prime fields was proposed by Longa *et al.* [21], in which $3P + Q$ costs $16M + 9S$, combined with wNAF, the efficiency get rapidly improved;

In 2011, Li *et al.* [15] proposed the Co_Z point addition on Hessian Curve and used these operations in precomputation to improved the efficiency of scalar multiplication algorithm. In 2016, Lai *et al.* [16] proposed Co_Z point addition, double point addition, Conjugate point addition operations with only (X, Y) coordinates on Hessian curve and used these operations to improved the traditional Montgomery ladder algorithm and zero-free and signed binary scalar multiplication algorithm. In 2017, Yu *et al.* [33] proposed the Co_Z Montgomery algorithm over $GF(3^m)$, by adding the skill of not calculating the y-coordinate in the middle of the loop. At the same year, Liu *et al.* [28] proposed new Symbolic Ternary Form and the $3P$ operation, and in this literature, the $3P$ operation costs $9M+7S$. In 2018, Xu *et al.* [26] proposed $3P$ and $5P$ formula based on pseudo 4D projective coordinate, which costs $7M + 7S$, $11M + 12S$, respectively. In 2019, Liu *et al.* [19] proposed the $3P$ formula based on Co_Z operation over GF(2^m), which costs $12M + 3S$.

1.2 Simple Power Analysis

Side Channel Attack (SCA) [18] is a method of attacking the cryptographic algorithm of the chip by using the information that the cryptographic chip inadvertently leaks in the operation process. SCA is mainly divided into four types: Simple Power Attack(SPA) [31, 32], Differential Power Attack (DPA) [9, 11], Refined Power Analysis (RPA) and Zero-value Point Attack (ZPA) [6]. These energy attacks have posed a serious threat to the security of mobile devices. SPA, as one of the SCA, has became a huge threat to the ECC algorithm. And the scalar multiplication algorithm is most vulnerable to this attack. In this attack, the adversary can guess the side channel information by analyzing the energy consumption during the operation of the cryptographic algorithm.

The principle of SPA recovery the secret key is to judge the running process of the encryption device on time and to perform a certain analysis of the energy curve. Due to its simplicity and high successful rate, SPA has became one of the most effective side channel attacking methods.

SPA can be avoided by the following methods: Adopt Side-Channel Atomicity [22, 32]; Use a uniform formula of point addition and point doubling, such as Edwards curve [1]; Use the randomization method; Use point addition and point doubling in each loop of the algorithm to make the energy consumption of each loop same, such as Montgomery algorithm [27, 29]. In this paper, our algorithm mainly implements the same operation in each loop to resist SPA.

1.3 Contribution

In this paper, we proposed two secure and fast scalar multiplication algorithm based on Co_Z operation and Conjugate point addition over prime fields. Firstly, the operation of $3P$ and $5P$ have been improved by using Co_Z operation. Then, the formula of $3P \pm Q$ and $5P \pm Q$ has been proposed based on Co_Z operation and Conjugate point addition. Finally, we proposed a new encoding way based on quinary. By applying the newly proposed formula to the new proposed scalar multiplication algorithm, the efficiency of first algorithm improved 19.3%, 10%, 19%, compared with Algorithm 3.4 [17], BJMontgomery [29], STF Algorithm [17], respectively. And that of the second Algorithm improved 27.8%, 19.5%, 27.5% respectively. Further, the efficiency of second Algorithm is 10.5% higher than the first Algorithm.

1.4 Organization

This paper is organized as follows. The second part introduces the basic knowledge of elliptic curves. In the third part, the operation of computing $3P$ and $5P$ is improved and the formula of $3P \pm Q$ and $5P \pm Q$ based on Co_Z point addition and Conjugate point addition are proposed, and a new encoding way is proposed. And these formula are used in the new proposed scalar multiplication algorithm to improve the efficiency of ECC. In the fourth part we analyze the performance of the proposed algorithm. At last part, we do a conclusion for this paper.

2 Preliminaries

2.1 Basic Knowledge

Definition 1. *An elliptic curve E over a prime field F_p is given by the Weierstrass equation:*

$$E : y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6. \tag{1}$$

Where, a_1, a_2, a_3, a_4, $a_6 \in F_p$, Δ is the discriminant of E, it can be defined as:

$$\begin{cases} \Delta = -{d_2}^2 d_8 - 8{d_4}^3 - 27{d_6}^2 + 9d_2d_4d_6 \\ d_2 = {a_1}^2 + 4a_2 \\ d_4 = 2a_4 + a_1a_3 \\ d_6 = {a_3}^2 + 4a_6 \\ d_8 = {a_1}^2a_6 + 4a_2a_6 - a_1a_3a_4 + a_2{a_3}^2 - {a_4}^2 \end{cases} \tag{2}$$

and $\Delta \neq 0$, which ensures the curve is smoothly, that is to say, there is only one tangent at any point on the curve.

If char p is not equal to 2 or 3, Equation (1) can be simplified as:

$$y^2 = x^3 + ax + b, \tag{3}$$

where $a, b \in F_p$, and Δ = -16$(4a^3 + 27b^2)$. According to chord-and- tangent, the third point on the curve can be

obtained by adding two more point on curve, thus the set rational point forming an Abelian group. The point at infinity denoted as O. ECC is constructed by this group [8].

If $P = (x_1, y_1)$ on E and $Q = (x_2, y_2)$ on E, $P \neq \pm Q$, then $P + Q = (x_3, y_3)$ is given by:

$$\begin{cases} x_3 = \lambda^2 - x_1 - x_2 \\ y_3 = \lambda(x_1 - x_3) - y_1 \\ \lambda = \frac{y_2 - y_1}{x_2 - x_1} \end{cases} \quad (4)$$

If $P = (x_1, y_1)$ on E, $P \neq -P$, then $2P = (x_3, y_3)$ is given by:

$$\begin{cases} x_3 = \lambda^2 - 2x_1 \\ y_3 = \lambda(x_1 - x_3) - y_1 \\ \lambda = \frac{3{x_1}^2 + a}{2y_1} \end{cases} \quad (5)$$

From Equation (4) and Equation (5), it can be deduced that under the affine coordinate, the point addition costs $1I+2M+1S$; The point doubling costs $1I+2M+2S$. It is well known that field inversion is the most time-consuming operation, so the projective coordinate can be adopted to eliminate the field inversion to decrease the costs of point addition and point doubling.

2.2 Formula of Point Addition and Point Doubling under the Jacobian Coordinate

In order to eliminate field inversion, $E(F_p)$ can be put on the projective coordinate, so the efficiency can be measured by field squartion and field multiplication. The correspondence between (x, y) and (X, Y, Z) is $x = X/Z^2$, $y = Y/Z^3$. The point at infinity, O, is the only point with a Z-coordinate equal to 0, denoted as $O = (1 : 1 : 0)$. Under Jacobian coordinate, $(\lambda^2 X{:}\lambda^3 Y{:}\lambda Z)$ is equal to $(X : Y : Z)$ for all $\lambda \in F_p$ and $\lambda \neq 0$. The Weierstrass equation under Jacobian coordinate is defined as Equation (6):

$$E : Y^2 = X^3 + aXZ^4 + bZ^6, \quad (6)$$

where, $a, b \in F_p$.

2.2.1 Point Addition

Let $P = (X_1, Y_1, Z_1)$, $Q = (X_2, Y_2, Z_2)$, $P + Q = (X_3, Y_3, Z_3)$, $P \neq Q$, then the point addition formula shown as Equation (7):

$$\begin{cases} X_3 = R^2 + G - 2V \\ Y_3 = R(V - X_3) - 2K_1 G \\ Z_3 = ((Z_1 + Z_2)^2 - I_1 - I_2)H \end{cases} \quad (7)$$

Where,

$$\begin{aligned} R &= 2(K_1 - K_2), \\ G &= FH, \\ V &= U_1 F, \\ K_1 &= Y_1 J_1, \\ K_2 &= Y_2 J_2, \\ F &= (2H)^2, \\ H &= U_1 - U_2, \\ U_1 &= X_1 I_2, \\ U_2 &= X_2 I_1, \\ J_1 &= I_1 Z_1, \\ J_2 &= I_2 Z_2, \\ I_1 &= Z_1^2, \\ I_2 &= Z_2^2. \end{aligned}$$

2.2.2 Doubling

Let $P = (X_1, Y_1, Z_1)$, and $P = Q$, then $2P = (X_3, Y_3, Z_3)$, the point doubling formula is:

$$\begin{cases} X_3 = M^2 - 2D_1 \\ Y_3 = M(D_1 - X_3) - 8L \\ Z_3 = (Y_1 + Z_1)^2 - E - N \end{cases} \quad (8)$$

where, $M = 3B_1 + aN^2$, $D_1 = 2((X_1 + E)^2 - B_1 - L)$, $L = E^2$, $B_1 = X_1^2$, $E = Y_1^2$, $N = Z_1^2$.

Therefore, under Jacobian coordinate, the point addition costs $11M + 5S$ and the point doubling costs $1M + 8S + 1c$, where c denote the cost of a multiplication by curve parameter a.

Also, it is important to note that it has suggested that the parameter a be fixed at -3 for efficiency purposes. In fact, most curves recommended by public-key standards use $a = -3$, which has been shown to not impose significant restrictions to the cryptosystem. In this case, the cost of point doubling is reduced to only $3M + 5S$ [7]. when $Z_1 = 1$ the doubling costs drop to $1M + 5S$, at this time, $(4X_1 Y_1^2, 8Y_1^4, 2Y_1) \sim P$, denoted as $P^{(1)}$.

3 Co_Z point Addition and Conjugate Point Addition

3.1 Co_Z point Addition

Co_Z operation, introduced by Meloni [24], which means giving two point with the same Z coordinate, that is let $P = (X_1, Y_1, Z)$, $Q = (X_2, Y_2, Z)$, then $P + Q = (X_3, Y_3, Z_3)$, by using this operation, the point addition formula becomes as Equation (9):

$$\begin{cases} X_3 = D - {W_1}' - {W_2}' \\ Y_3 = (Y_1 - Y_2)({W_1}' - X_3) - A_1 \\ Z_3 = Z(X_1 - X_2) \end{cases} \quad (9)$$

where, $A_1 = Y_1(W_1{}' - W_2{}')$, $W_1{}' = X_1T$, $W_2{}' = X_2T$, $T = (X_1 - X_2)^2$, $D = (Y_1 - Y_2)^2$. The key observation in Equation (9) is that the computation of $R = P + Q$ yields for free an equivalent representation for input point P with its Z-coordinate equal to that of output point R, namely $(X_1(X_1 - X_2)^2 : Y_1(X_1 - X_2)^3 : Z_3) \sim P$.

From Equation (9), the unified Z-coordinate transformation has been calculated in the Co_Z point addition operation, so no extra calculations are needed. At this time, the point addition costs $5M + 2S$. If Z-coordinate of P and Q are equal to 1, the corresponding point addition costs $4M + 2S$.

3.2 Tripling and Quintuple on Co_Z Operation

3.2.1 Tripling Operation

From above conclusion, when $Z_1 = 1$, $P^{(1)}$and $2P$ have the same Z-coordinate, so using the Co_Z operation between them can get the formula of $3P$. let $P = (X_1, Y_1, 1)$, $P^{(1)} = (X_1^{(1)}, Y_1^{(1)}, Z_1^{(1)}) = (4X_1Y_1^2, 8Y_1^4, 2Y_1)$, and $3P = 2P + P^{(1)} = (X_4, Y_4, Z_4)$, then,

$$\begin{cases} X_4 = (Y_1^{(1)} - Y_3)^2 - (X_1^{(1)} - X_3)^2 \\ Y_4 = (Y_1^{(1)} - Y_3)(X_3(X_1^{(1)} - X_3)^2 - X_4) \\ \qquad - Y_3(X_1^{(1)} - X_3)^3 \\ Z_4 = Z_3(X_1^{(1)} - X_3) \end{cases} \tag{10}$$

And it is presented in Algorithm 1. According to the costs of point doubling and point addition, it is readily seen that Algorithm 1 costs $7S + 6M$.

Algorithm 1 Tripling on Co_Z

1: **Input:** $P^{(1)} = (X_1^{(1)}, Y_1^{(1)}, Z_1^{(1)}) = (4X_1Y_1^2, 8Y_1^4, 2Y_1)$, $2P = (X_3, Y_3, Z_3)$
2: **Output:** $3P = 2P + P^{(1)} = (X_4, Y_4, Z_4)$
3: $A \leftarrow (X_1^{(1)} - X_3)$
4: $B \leftarrow (Y_1^{(1)} - Y_3)$
5: $X_4 \leftarrow B^2 - A^3$
6: $Y_4 \leftarrow B(X_3A^2 - X_4) - Y_3A^3$
7: $Z_4 \leftarrow Z_3A$
8: **Return** (X_4, Y_4, Z_4)
9: End

3.2.2 Quintupling Operation

Similarly, when $Z_1 = 1$, let $4P = 2(2P) = (X_5, Y_5, Z_5)$, we can get that,

$$\begin{cases} X_5 = {M_1}^2 - 2D_2 \\ Y_5 = M_1(D_2 - X_5) - 8L_1 \\ Z_5 = (Y_3 + Z_3)^2 - E_1 - N_1 \end{cases} \tag{11}$$

where, $M_1 = 3B_2 + aN_1, D_2 = 2((X_3 + E_1)^2 - B_2 - L_1), L_1 = {E_1}^2, B_2 = {X_3}^2, E_1 = {Y_3}^2, N_1 = {Z_3}^2$. Then, using Co_Z point addition between $4P$ and P can get the formula of $5P$, At this time, $P \sim P^{(2)} = (X_1^{(2)}, Y_1^{(2)}, Z_1^{(2)}) = (X_1(2Y_3Z_3)^2, Y_1(2Y_3Z_3)^3, 2Y_3Z_3)$, then,

$$\begin{cases} X_6 = (Y_1^{(2)} - Y_5)^2 - (X_1^{(2)} - X_5)^3 \\ Y_6 = (Y_1^{(2)} - Y_5)(X_1^{(2)}(X_1^{(2)} - X_5) - X_6) \\ \qquad - Y_1^{(2)}(X_1^{(2)} - X_5)^3 \\ Z_6 = Z(X_1^{(2)} - X_5) \end{cases} \tag{12}$$

The details shown as Algorithm 2. According to the costs of point doubling and point addition, it costs $9M + 12S$.

Algorithm 2 Quintupling on Co_Z

1: **Input:** $P^{(2)} = (X_1^{(2)}, Y_1^{(2)}, Z_1^{(2)}) = (X_1(2Y_3Z_3)^2, Y_1(2Y_3Z_3)^3, 2Y_3Z_3)$, $4P = (X_5, Y_5, Z_5)$
2: **Output:** $5P = 4P + P^{(2)} = (X_6, Y_6, Z_6)$
3: $A_1 \leftarrow (X_1^{(2)} - X_5)$
4: $B_1 \leftarrow (Y_1^{(2)} - Y_5)$
5: $X_6 \leftarrow {B_1}^2 - {A_1}^3 - 2X_5{A_1}^2$
6: $Y_6 \leftarrow B_1(X_5{A_1}^2 - X_6) - Y_5{A_1}^3$
7: $Z_6 \leftarrow Z_5A_1$
8: **Return** (X_6, Y_6, Z_6)
9: End

3.3 Conjugate Point Addition

Conjugate point addition, proposed by the literature Goundar *et al.* [7], which would be used between nP and Q to get the formula of $nP \pm Q(n = 3, 5)$ in this paper.

3.3.1 $3P \pm Q$ Operation Based on Co_Z Operation

The necessary condition for using Co_Z operation between $3P$ and Q is that they must agree to the Z coordinate. According to the formula of $3P$, then let $Q = (X_2\lambda^2, Y_2\lambda^3, \lambda)$, with $\lambda = Z_4$, in this way, we can ensure $3P$ and Q have the same Z-coordinate. Then $3P + Q = (X_4', Y_4', Z_4')$, $3P - Q = (\overline{X_4'}, \overline{Y_4'}, Z_4')$, the detailed algorithm is given hereafter.

From Algorithm 3, the calculation of $3P + Q$ costs $10S + 15M$, where W_1, W_2 and A_2 have computed during the course of $3P + Q$, the additional cost for getting $3P - Q$ from $3P + Q$ is thus of only $1M + 1S$. Hence, the total cost for the $3P - Q$ is $11S + 16M$.

3.3.2 $5P \pm Q$ Operation based on Co_Z Operation

Similarly, according to the formula of $5P$, then let $Q = (X_2\lambda_1^2, Y_2\lambda_1^3, \lambda_1)$, with $\lambda_1 = Z_5$, in this way, we can ensure $5P$ and Q have the same Z-coordinate. Then $5P + Q = (X_6', Y_6', Z_6')$, $5P - Q = (\overline{X_6'}, \overline{Y_6'}, Z_6')$, the detailed algorithm is given hereafter.

Algorithm 3 $3P \pm Q$ on Conjugate and Co_Z Operation(TC-ZADD)

Input: $3P = (X_4, Y_4, Z_4)$, $Q = (X_2{Z_4}^2, Y_2{Z_4}^3, Z_4)$
Output: $3P + Q = (X_4', Y_4', Z_4')$, $3P - Q = (\overline{X_4}', \overline{Y_4'}, Z_4')$
$C_1 \leftarrow (X_4 - X_2Z_4)^2$
$W_1 \leftarrow X_4C_1$; $W_2 \leftarrow X_2C_1Z_4$
$H_1 \leftarrow (Y_4 - Y_2{Z_4}^3)^2$, $A_2 \leftarrow Y_4(W_1 - W_2)$
$X_4' \leftarrow H_1 - W_1 - W_2$
$Y_4' \leftarrow (Y_3 - Y_2{Z_4}^3)(W_1{X_4}') - A_2$
$Z_4' \leftarrow Z_4(X_4 - X_2{Z_4}^2)$
$\overline{H_1} \leftarrow (Y_4 + Y_2{Z_4}^3)^2$
$\overline{X_4'} \leftarrow \overline{H_1} - W_1 - W_2$
$\overline{Y_4'} = (Y_1 + Y_2)(W_1 - \overline{X_4'}) - A_2$
Return $({X_4}', {Y_4}', {Z_4}')$, $(\overline{X_4'}, \overline{Y_4'}, Z_4')$
End

Algorithm 4 $5P \pm Q$ on Conjugate and Co_Z Operation(QC-ZADD)

Input: $5P = (X_6, Y_6, Z_6)$, $Q = (X_2{Z_6}^2, Y_2{Z_6}^3, Z_6)$
Output: $5P + Q = (X_6', Y_6', Z_6')$, $5P - Q = (\overline{X_6}', \overline{Y_6'}, Z_6')$
$C_2 \leftarrow (X_6 - X_2Z_6)^2$
$W_1' \leftarrow X_6C_2$; $W_2' \leftarrow X_2C_2Z_6$
$H_2 \leftarrow (Y_6 - Y_2{Z_6}^3)^2$, $A_2' \leftarrow Y_6(W_1' - W_2')$
$X_4' \leftarrow H_2 - W_1' - W_2'$
$Y_4' \leftarrow (Y_3 - Y_2{Z_6}^3)(W_1'{X_6}') - A_2'$
$Z_4' \leftarrow Z_6(X_6 - X_2{Z_6}^2)$
$\overline{H_2} \leftarrow (Y_6 + Y_2{Z_6}^3)^2$
$\overline{X_6'} \leftarrow \overline{H_2} - W_1' - W_2'$
$\overline{Y_6'} = (Y_1 + Y_2)(W_1' - \overline{X_6'}) - A_2'$
Return $({X_6}', {Y_6}', {Z_6}')$, $(\overline{X_6'}, \overline{Y_6'}, Z_6')$
End

From Algorithm 4, the calculation of $5P + Q$ costs $18M + 15S$, where W_1', W_2' and A_2' have computed during the course of $5P + Q$, the additional cost for getting $5P - Q$ from $5P + Q$ is thus of only $1M + 1S$. Hence, the total cost for the $5P - Q$ is $16S + 19M$.

3.4 Scalar Multiplication Algorithm

3.4.1 New Symbolic Ternary Form Scalar Multiplication Algorithm

The scalar k often encoded as symmetric ternary form to improve the efficiency of scalar multiplication operation, because this encoding length is shorter than the binary length. To further improve efficiency, new symbolic ternary form has been proposed by Wang *et al.* [28], in which $k_i \in \{-2, -1, 0, 1, 2\}$. The specific algorithm shown as Algorithm 5.

In this encoding form, non-zero digits can be reduced to half. The specific coding principle is that from the lowest bits, if $k_{i+1} = 1$ and $k_i = -1$, then set $k_{i+1} = 0$ and $k_i = 2$, if $k_{i+1} = -1$ and $k_i = 1$, then set $k_{i+1} = 0$ and $k_i = -2$, the sequence of New Symbolic Ternary Form encoding can be deduced. After that, we perform a scalar multiplication algorithm based on the sequence and perform different operations based on different values, the specific algorithm is shown in Algorithm 6 [28].

Algorithm 5 New Symbolic Ternary Form Encoding Algorithm

Input: $k = \sum_{i=0}^{n-1}(k_i 3^i), k_i \in \{-1, 0, 1\}$
Output: $k = \sum_{i=0}^{n-1}(k_i 3^i), k_i \in \{-2, -1, 0, 1, 2\}$
for $i = 0$ to n-1 **do**
  if $k_{i+1} = 1, k_i = -1$ **then**
    $k_{i+1} = 0, k_i = 2$
  else if $k_{i+1} = -1, k_i = 1$ **then**
    $k_{i+1} = 0, k_i = -2$
  end if
  $i++$
end for
Return k
End

Algorithm 6 Scalar Multiplication Algorithm

Input: $P(X, Y, Z), k_i \in \{-2, -1, 0, 1, 2\}$
Output: $Q = kP$
$i = 0$, $P_1 = O$, $Q = O$
while $i \leqslant n - 1$ **do**
  if $k_i = 1$ **then**
    $Q = Q + P, P = 3P$;
  else if $k_i = -1$ **then**
    $Q = Q - P, P = 3P$;
  else if $k_i = 2$ **then**
    $P_1 = 2P, Q = Q + P_1, P = 3P$;
  else if $k_i = -2$ **then**
    $P_1 = 2P, Q = Q - P_1, P = 3P$
  else
    $P = 3P$
  end if
  $i++$
end while
Return $Q = (X_q, Y_q, Z_q)$
End

Algorithm 6 costs $nT + \frac{n}{6}D + \frac{n}{2}A$, where T, D, A represente tripling, doubling and point addition respectively. And it perform different operations based on different k, therefore the attacker can deduce the side channel information to recover the private k. Depend on these conclusions, we propose a new secure and fast algorithm to against SPA by combining the algorithms in [17] and [14]. It is shown as Algorithm 7, and then applying the formula of $3P \pm Q$ in Algorithm 7, the efficiency of scalar multiplication algorithm can get improved.

According to statistical rules, the probability of non-zero coefficient in this encoding scheme is about $\frac{1}{2}$, and positive and negative are about $\frac{1}{4}$ respectively. When the sequence length is n, Algorithm 7 costs $(\frac{3n}{4}TA + \frac{n}{4}TD) + D$

Algorithm 7 New Proposed Algorithm with Anti-SPA

Input: $P(X,Y,Z), k_i \in \{-2,-1,0,1,2\}$
Output: $Q = kP$
$Q = O$, $Q_1 = Q_{-1} = P$, $Q_2 = Q_{-2} = 2P$
for $i = n-1$ to 0 **do**
  $Q = TC - ZADD(Q, Q_{k_i})$
end for
Return $Q = (X_q, Y_q, Z_q)$
End

totally, where TA, TD and D represents tripling-add operation, tripling-subtraction operation and doubling operation respectively. In the process of Algorithm 7, $k_i \in \{-2,-1,-0,1,2\}$, if $k_i \geq 0$, $3P+Q$ operation would be executed, if $k_i < 0$, the algorithm would execute $3P-Q$ operation. In the loop of Algorithm 7, the specific operation is independent of specific location of scalar k, so, the attacker can not deduce the specific information according to Side channel information leakage. That is to say, because the costs among the $3P+Q$, $3P-Q$ are different, the attacker can only distinguish between different operations, but the specific scalar value cannot be deduced. Therefore, the attacker cannot guess the side channel information. Hence, Algorithm 7 can resist SPA. Due to the limition of hardware condition, the algorithm security can only be analysed at the theoretical leval, unable to implement on hardware.

3.4.2 Scalar Multiplication Algorithm Based on New Encoding Form

In this section, a new from encoding based on quinary is proposed. This code length about $l/(\log_2 5)$, l represents the Binary length. Compared with other encodings, this encoding length is shorter, in which, $k_i \in \{-2,-1,0,1,2\}$. The specific algorithm shown as Algorithm 8.

Algorithm 8 New Encoding Algorithm

Input: k
Output: $k = (k_{n-1}, k_{n-2}, ..., k_1, k_0)_5, k_i \in \{-2,-1,0,1,2\}$
$i = 0$;
while $k > 0$ **do**
  if $k mod 5 = 4$ **then**
    $k_i = -1, k = \lceil k/5 \rceil, i++$;
  else if $k mod 5 = 3$ **then**
    $k_i = -2, k = \lceil k/5 \rceil, i++$;
  else if $k mod 5 = 2$ **then**
    $k_i = 2, k = \lfloor k/5 \rfloor, i++$;
  else if $k mod 5 = 1$ **then**
    $k_i = 1, k = \lfloor k/5 \rfloor, i++$;
  else
    $k_i = 0, k = k/5, i++$;
  end if
end while
Return $(k_{n-1}, k_{n-2}, ..., k_1, k_0)_5$
End

The probability of each element in this encoding is 1/5. The scalar multiplication based on the encoding shown as Algorithm 9.

Algorithm 9 scalar multiplication with anti-SPA based on new encoding

Input: $P(X,Y,Z), k_i \in \{-2,-1,0,1,2\}$
Output: $Q = kP$
$Q = O$, $Q_1 = Q_{-1} = P$, $Q_2 = Q_{-2} = 2P$
for $i = n-1$ to 0 **do**
  $Q = QC - ZADD(Q, Q_{k_i})$
end for
Return $Q = (X_q, Y_q, Z_q)$
End

The cost of Algorithm 9 about $\frac{3}{5}FA + \frac{2}{5}FD + D$ totally, where FA, FD, and D represent $5P+Q, 5P-Q$ and Double operation respectively. In the process of Algorithm 9, $k_i \in \{-2,-1,0,1,2\}$, if $k_i \geq 0$, $5P+Q$ would be executed, if $k_i < 0$, the algorithm would executed $5P-Q$, else, $5P$ operation would be executed. It is clearly that the process of Algorithm 9 is independent of specific location of scalar k. So, attacker can not deduce the specific information. Therefore, Algorithm 9 can resist SPA. Due to the limition of hardware condition, the algorithm security can only be analysed at the theoretical leval, unable to implement on hardware.

4 Performance Analysis

In this section, we would analyze the efficiency of the improved scalar multiplication algorithm.

The bottom formulas $3P+Q$, $3P$, $5P+Q$ and $5P$ are calculated based on Conjugate point addition and Co_Z operation. Firstly, the improved formulas are compared with the other articles, as shown in Table 1 and Table 2 respectively. when adopt $I/M = 8$, the costs of $3P+Q$ and $5P+Q$ decreased by 10.2%, 11.5% respectively and the costs of $3P$ and $5P$ decreased by 18.17%, 13.5% respectively.

Table 1: Comparisons of computation cost of $3P+Q$ and $3P$ in different literatures

Literature	$3P+Q$	$3P$
Liter [21]	$16M+9S$	-
Liter [26]	-	$7M+7S$
Liter [3]	$2I+4S+9M$	-
Liter [17]	-	$1I+4S+7M$
Ours	$10S+15M$	$7S+6M$

The computation cost of different scalar multiplication algorithms shown in Table 3, where $n = 101bits$, $I/M = 8$. Compared with Algorithm 3.4 [17], BJMontgomery [29], STF Algorithm [17], the cost of Algorithm 7 decreased by 19.3%, 10%, 19% respectively, and that of Algorithm 9 decreased by 27.8%, 19.5%, 27.5%

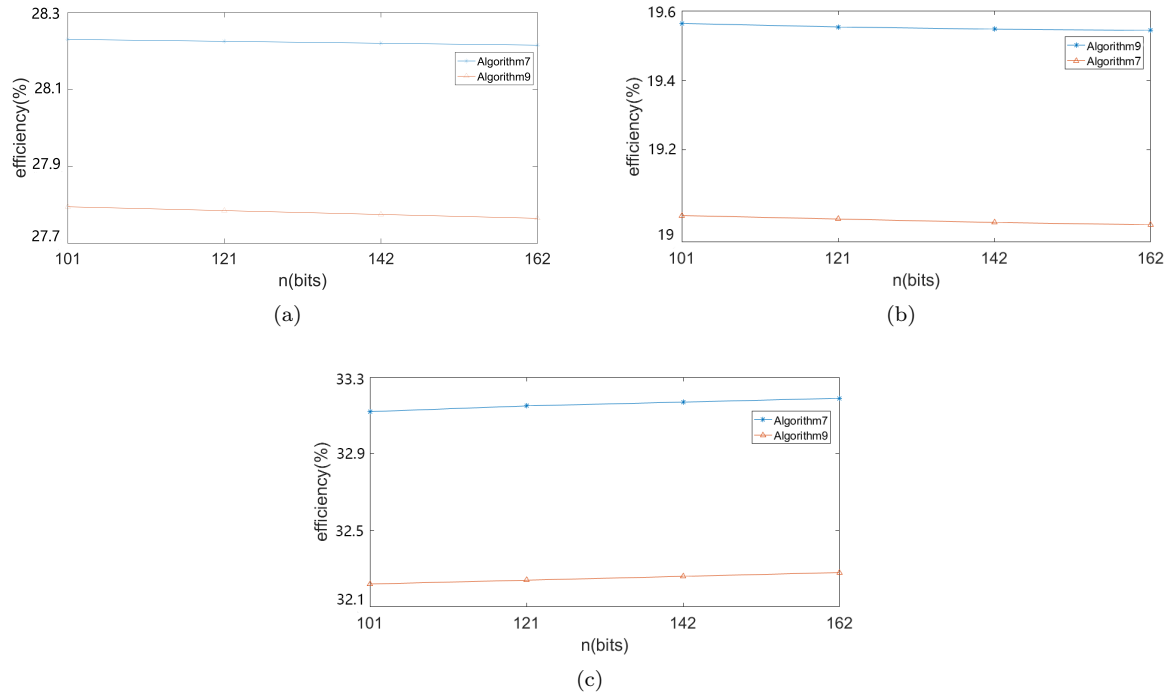


Figure 1: Different efficiency growth under different bit values compared with different Algorithm. (a), (b), (c) represent Algorithm 3.4 [17], BJMontgomery [29], STF Algorithm [17], respectively

Table 2: Comparisons of computation cost of $5P + Q$ and $5P$ in different literatures

Literature	$5P + Q$	$5P$
Liter [21]	$26M + 13S$	-
Liter [26]	-	$11M + 12S$
Liter [20]	$17M + 18S$	-
Liter [2]	-	$12M + 13S$
Ours	$18M + 15S$	$9M + 12S$

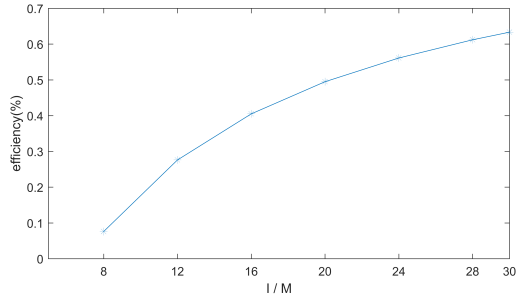


Figure 2: Increased efficiency for HSTF algorithm [17]

respectively. Further, Algorithm 9 costs 10.5% less than Algorithm 7. The difference between Algorithm 6 and Algorithm 7 is shown in Table 4, it is obvious that Algorithm 7 is not only faster than Algorithm 6 but also Anti-SPA. Table 5 shows the costs of Anti-SPA HSTF [17] scalar multiplication algorithm with different bottom operations, after adopting the new formulas, the efficiency is improved by 7.6%.

Then, we will use a more intuitive way to display the efficiency analysis. Let $\alpha = 8$, which is the ratio of field inversion and field multiplication. To get a better display effect, the three efficiency curves are put in different graphs. Because of the large difference in efficiency, let the efficiency of Algorithm 7 add 9%. When adopted different bits, the efficiency can be deduced from Equation (13):

$$efficiency = 1 - \frac{\#I_1 + \#M_1}{\#I_2 + \#M_2} = 1 - \frac{\alpha + d}{m_1\alpha + n_1} \quad (13)$$

where, $\#I_1 + \#M_1$ represents the costs of our new proposed Algorithm, $\#I_2 + \#M_2$ represents the costs of algorithm in other Literatures, and d, m_1, n_1, represent constants.

From Figure 1, It is clearly that the efficiency increase value of Algorithm 7 and Algorithm 9 relative to Algorithm 3.4 [17] and BIMontgomery [29] decreases with the increase of bit value, while the efficiency increase value relative to STF Algorithm [17] increase with the increase of bit value. Therefore, changing the bit value n plays an important role in improve efficiency. And the bit value

Table 3: Computation cost of different scalar multiplication algorithms

Algorithm	Total cost	$n = 101bits$
Algorithm 3.4 [17]	$n(1I + 1S + 2M) + n(1I + 4S + 7M) + (1I + 2S + 2M)$	$2940.6M$
BJMontgomery [29]	$16.4n \log_2 3 + 13.5$	$2637.5M$
STF Algorithm [17]	$n(I + 4S + 7M) + n(I + S + 2M)$	$2929M$
Proposed Algorithm 7	$\frac{3}{4}n(10S + 15M) + \frac{1}{4}n(11S + 16M) + (1M + 5S)$	$2373M$
Proposed Algorithm 9	$\frac{3}{5}n/\log_3 5(18M + 15S) + \frac{2}{5}n/\log_3 5(19M + 16S) + (1M + 5S)$	$2123M$

Table 4: Comparison of the calculation costs of the algorithms

Algorithm	Total cost	Anti-SPA	$n = 101bits$
Algorithm 6	$n(9M + 7S) + \frac{1}{6}n(4M + 6S) + \frac{1}{2}n(12M + 4S)$	no	$2390M$
Proposed Algorithm 7	$\frac{3}{4}n(10S + 15M) + \frac{1}{4}n(11S + 16M) + (1M + 5S)$	yes	$2373M$

Table 5: Computation cost of HSTF algorithm [17] with different operations

HSTF Algorithm [17]	Tripling	Addition	Double	Total Costs($n = 162bits$)
Using new formulas	$7S + 6M$	$11M + 5S$	$1M + 5S$	$4362M$
Original	$1I + 4S + 7M$	$1I + 1S + 2M$	$1I + 2S + 2M$	$4720M$

has the greatest influence on the efficiency improvement of Algorithm 7 and Algorithm 9 relative to STF Algorithm [17].

Similarly, the efficiency increase of HSTF are shown as Figure 2, α is defined as the ratio of field inversion and field multiplication, the value of α ranges from 8 to 30. The efficiency can be deduced from following equation, which derive from Equation (13):

$$efficiency = 1 - \frac{4362}{326\alpha + 2112.4} \tag{14}$$

It can be obtained from Figure 2, with the increase of α, the efficiency can be further improved. when $\alpha = 30$, the efficiency can reach to 63.3%.

5 Conclusions

In this paper, we aim to improve the scalar multiplication over $GF(p^m)$ (p>3). For bottom operation, we have proposed $3P \pm Q$, $3P$, $5P \pm Q$ and $5P$ by adopting Conjugate point addition and Co_Z operation. And the formula of $3P$, $5P$, $5P + Q$ and $3P + Q$ have been improved to make the number of field multiplications and field square calculations reduced in the operation process, the costs of proposed $3P + Q$ and $3P$ decreased by 10.2%, 18.17% respectively than before, the costs of $5P - Q$ and $5P$ decreased by 11.5%, 13.5% averagely. We also proposed a new encoding way based on quinary, the length can be reduced to $l/log_2 5$, where l represents the binary length. And the efficiency of proposed Algorithm 7 increased by 19.3%, 10%, 19% respectively, compared with Algorithm 3.4 [17], BJMontgomery [29], STF Algorithm [17], and that of Algorithm 9 is increased by 27.8%, 19.5%, 27.5% respectively. Further, The efficiency of Algorithm 9 is 10.5% higher than Algorithm 7. Then, when using the improved operation in the HSTF Algorithm [17], the computational efficiency is increased by 7.6% compared with original.

Next, if Algorithm 8 can be optimized to increase the proportion of 0 in the sequence, the efficiency of scalar multiplication would get improved.

Due to the limitation of hardware conditions, the scalar multiplication algorithm cannot be implemented in hardware, so the efficiency and security analysis can only be a theoretical analysis.

In the future, if the suitable curve can be choose to combine with the new formula, the efficiency would get a higher improvement, and the scalar multiplication would be more suitable for use in specific hardware environments.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (61872058), the Key Research and Development Program of Shaanxi (Program No.2021NY-211). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] M. Boudabra and A. Nitaj, "A new public key cryptosystem based on Edwards curves," *Journal of Applied Mathematics and Computing*, no. 1–2, pp. 1–20, 2019.

[2] S. M. Cho, S. G. Gwak, C. H. Kim, and S. Hong, "Faster elliptic curve arithmetic for triple-base chain by reordering sequences of field operations," *Multi-*

media Tools and Applications, vol. 75, no. 22, pp. 1–13, 2016.

[3] M. Ciet, M. Joye, K. Lauter, and P. L. Montgomery, "Trading inversions for multiplications in elliptic curve cryptography," *Designs Codes and Cryptography*, vol. 39, no. 2, pp. 189–206, 2006.

[4] D. V. Chudnovsky, G. V. Chudnovsky, "Sequences of numbers generated by addition in formal groups and new primality and factorization tests," *Applied Mathematics*, vol. 7, pp. 385–434, 1986.

[5] W. Y. Deng and X. H. Miao, "Application of balanced ternary in elliptic curve scalar multiplication," *Computer Engineering*, vol. 38, no. 5, pp. 152–154, 2012.

[6] L. Goubin, "A refined power-analysis attack on elliptic curve cryptosystems," in *International Workshop on Public Key Cryptography*, vol. 2567, pp. 199–211, Dec 2002.

[7] R. R. Goundar, M. Joye, and A. Miyaji, "Co_Z addition formulea and binary ladders on elliptic curves," *Lecture Notes in Computer Science*, vol. 6225, pp. 65–79, 2010.

[8] D. Hankerson, A. Menezes, and S. V. Springer, *Guide to Elliptic Curve Cryptography*, 2004. (`https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.394.3037&rep=rep1&type=pdf`)

[9] S. Hou, Y. Zhou, H. Liu, and N. Zhu, "Improved DPA attack on rotating s-boxes masking scheme," in *IEEE 9th International Conference on Communication Software and Networks (ICCSN'17)*, Dec. 2017. DOI: 10.1109/ICCSN.2017.8230283.

[10] G. U. Jianguang, "Scalar multiplication algorithm resisting power analysis attacks using NAF with threshold window," *Computer Engineering*, no. 8, pp. 296–299, 2019.

[11] I. Kabin, Z. Dyka, D. Klann, and P. Langendoerfer, "Horizontal DPA attacks against ECC: Impact of implemented field multiplication formula," in *The 14th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS'19)*, pp. 1–6, 2019.

[12] D. Khleborodov, "Fast elliptic curve point multiplication based on window non-adjacent form method," *Applied Mathematics and Computation*, vol. 334, pp. 41–59, 2018.

[13] N. Koblitz, "Elliptic curve cryptosystems," *Mathematics of Computation*, vol. 48, no. 177, pp. 203–209, 1987.

[14] L. Li, "Research on ternary algorithm in elliptic curve operation," *Network Security Technology and Application*, no. 11, pp. 94–96, 2015.

[15] L. Li and Z. Tao, "Co_Z addition operation of Hessian curve," in *The Seventh International Conference on Computational Intelligence and Security*, Jan. 2012. DOI: 10.1109/CIS.2011.206.

[16] Z. Lai and Z. Zhang, "Scalar multiplication on Hessian curves based on Co_Z operations," *Bulletin of Science and Technology*, vol. 32, pp. 28–33, Feb. 2016.

[17] S. Liu, H. Yao, and A. W. Xu, "SPA resistant scalar multiplication based on addition and tripling indistinguishable on elliptic curve cryptosystem," in *The 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC'16)*, Mar. 2016. DOI: 10.1109/3PGCIC.2015.20.

[18] K. Liao, X. Cui, N. Liao, T. Wang, D. Yu, X. Cui, "High-performance noninvasive side-channel attack resistant ECC coprocessor for $GF(2)^m$," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 1, pp. 727–738, 2017.

[19] S. Liu, Y. Ding, R. Shi, and S. Lu, "Co_Z point addition algorithm on elliptic curve over characteristic two," *Journal of Wuhan University (Science Edition)*, vol. 65, pp. 207–212, Apr. 2019.

[20] P. Longa and A. Miri, "Fast and flexible elliptic curve point arithmetic over prime fields," *IEEE Transactions on Computers*, vol. 57, no. 3, pp. 289–302, 2008.

[21] P. Longa and A. Miri, "New composite operations and precomputation scheme for elliptic curve cryptosystems over prime fields," in *Public Key Cryptography (PKC'08)*, pp. 229-247, 2008.

[22] C. Y. Lu, S. M. Jen, and C. S. Laih, "A general framework of side-channel atomicity for elliptic curve scalar multiplication," *IEEE Transactions on Computers*, vol. 62, no. 3, pp. 428–438, 2013.

[23] F. Mallouli, A. Hellal, N. S. Saeed, and F. A. Alzahrani, "A survey on cryptography: Comparative study between RSA vs ECC algorithms, and RSA vs El-gamal algorithms," in *IEEE International Conference on Cyber Security and Cloud Computing / IEEE International Conference on Edge Computing and Scalable Cloud*, pp. 173–176, 2019.

[24] N. Meloni, "New point addition formulae for ECC applications," in *International Workshop on the Arithmetic of Finite Fields*, vol. 4547, pp. 189–201, 2007.

[25] V. S. Miller, "Use of elliptic curves in cryptography," in *Conference on the Theory and Application of Cryptographic Techniques*, vol. 218, pp. 417–426, 1985.

[26] X. U. Ming and L. Shi, "Pseudo 4D projective coordinate-based multi-base scalar multiplication," *Journal on Communications*, vol. 38, no. 5, pp. 74–84, 2018.

[27] D. B. Roy and D. Mukhopadhyay, "High-speed implementation of ECC scalar multiplication in GF(p) for generic montgomery curves," *IEEE Transactions on Very Large Scale Integration Systems*, no. 99, pp. 1–14, 2019.

[28] Y. Xi Wang, C. R. Zhang, B. H. Zhang, and Z. Zhu, "Efficient scalar multiplication of ECC based on composite operations over prime fields," *Application Research of Computers*, vol. 30, no. 11, pp. 3385–3387, 2013.

[29] Y. U. Wei, Wang Kunpeng, L. I. Bao, and Tian Song, "Montgomery algorithm over a prime field," *Chinese Journal of Electronics*, vol. 28, no. 1, pp. 43–48, 2019.

[30] Y. Wu and X. Wu, "Implementation of efficient method of RSA key-pair generation algorithm," in *IEEE International Symposium on Consumer Electronics*, pp. 72–73, 2017.

[31] F. Xiang and S. Li, "A high-speed and SPA-resistant implementation of ECC point multiplication over GF(p)," in *IEEE Trustcom/BigDataSE/ICESS*, 2017. DOI: 10.1109/Trustcom/BigDataSE/ICESS.2017.245.

[32] L. I. Yang, Jin Lin Wang, Xue Wen Zeng, and Y. E. Xiao-Zhou, "A segmented montgomery scalar multiplication algorithm with resistance to simple power analysis SPA attacks," *Computer Engineering and Science*, vol. 39, no. 1, pp. 92–102, 2017.

[33] W. Yu, B. Li, P. Wang, W. Li, and S. Tian, "Co_Z montgomery algorithm on elliptic curve scalar multiplication," *Journal of Computer*, vol. 40, no. 5, pp. 1121–1133, 2017.

Biography

Shuang-Gen Liu, born in 1979. He received the PH.D. in cryptography form Xidian University in 2008. He is currently an associate professor with the school of cyber security, Xi'an University of Posts and Telecommunications, Xi'an, China. He is a member of the China Computer Federation, and a member of the Chinese Association for Cryptologic Research. His recent research interests include crptography and information security.

Ying Zhang, born in 1997. A graduate student of Xi'an University of posts and telecommunications. She is mainly engaged in the research of elliptic curve cryptosystem.

Shi-Yao Chen, born in 1999. An undergraduate student of Xi'an University of posts and telecommunications.She is mainly engaged in the research of elliptic curve cryptosystem.

A Hyperchaotic Encrypted Speech Perceptual Hashing Retrieval Algorithm Based on 2D-Gabor Transform

Yi-bo Huang[1], Shi-hong Wang[1], Yong Wang[1], Yuan Zhang[1], and Qiu-yu Zhang[2]
(Corresponding author: Yi-bo Huang)
College of Physics and Electronic Engineering, Northwest Normal University[1]
967 Anning E Rd, Anning District, Lanzhou 730070, Gansu, China
School of Computers and Communication, Lanzhou University of Technology[2]
Lanzhou 730050, China
Email: huang_yibo@nwnu.edu.cn
(Received July 15, 2020; Revised and Accepted Apr. 24, 2021; First Online Aug. 17, 2021)

Abstract

A hyperchaotic encrypted speech perceptual hashing retrieval algorithm based on 2D-Gabor transform and PCA dimension reduction has been proposed in this paper. The proposed algorithm first uses 2D-Gabor transform to extract speech features. Then use PCA to reduce the dimension of the extracted feature. Finally, the proposed algorithm uses the extracted feature to construct a hash sequence, then uploads the hash sequence to the cloud to establish a hash sequence table. At the same time, use the four-dimensional hyperchaotic encryption method to encrypt the spech, and then upload it to the cloud and establishes a phonetic table with a one-to-one correspondence with the hash sequence table. When the user needs to retrieval speech, compare the generated hash sequence of target speech with the hash sequence table. After the matching is completed, the speech corresponding to the matching result is returned to the user. Thus, the proposed retrieval method can achieve successful matching without downloading and decrypting speech. Experimental results show the proposed algorithm in this paper improves the algorithm's accuracy compared with the previous retrieval algorithm and has high discrimination and nice robustness.

Keywords: 2D-Gabor Feature Extraction; Encrypted Speech Retrieval; Four-Dimensional Hyperchaotic System; Perceptual Hashing; Principal Components Analysis

1 Introduction

With the continuous development of multimedia technology, people put higher requirements in speech storage and retrieval, so how to effectively retrieval target speech from massive speech in the cloud has become a challenging tasks [3, 6]. However, the third-party cloud services are not a place you can rest assured. Therefore, it is necessary to strengthen the security of speech during transmission [9,22,24] under the premise of continuously improving retrieval accuracy and speed. However, there is not much research now about how to reteieval in encrypted speech. Therefore, how to retrieve encrypted speech has become an important field in speech retrieval research.

The current mainstream speech retrieval scheme is content-based encryption speech retrieval technology [2, 5, 12]. The content-based speech retrieval scheme realizes the retrieval of speech through the physical characteristics of speech, such as speech amplitude, speech spectrum and other characteristics, thereby greatly improving the discrimination and robustness of speech features. The perceptual hash sequence calculated by this way can have a high accuracy during retrieval. The content-based speech retrieval solution can also retrieve encrypted speech without downloading and decrypting. Therefore, this scheme not only ensures the security of speech data but also improves the efficiency and accuracy of retrieval.

Content-based encrypted speech retrieval technology mainly includes three aspects: Speech feature extraction, Speech encryption technology, Speech retrieval technology. The main speech feature extraction method now includes speech fingerprints [15, 17] and perceptual hash [1, 4], *etc.* Speech encryption technology mainly includes chaotic map encryption [14], DNA encoding encryption [11] and Haar Transform and Permutation encryption [13], *etc.* Speech retrieval technology mainly includes feature matching [20], example speech search [8], *etc.*

In 2013, Wang *et al.* [18] proposed a retrieval method for encrypted speech using perceptual hashing. It encrypted speech with Chus's chaotic circuit and piecewise linear(PWL), and used the speech zero-crossing rate to

extract speech features for retrieval. This scheme has good robustness and high retrieval speed, but due to poor discrimination, the retrieval accuracy is low. He *et al.* [6] proposed a perceptual hashing based on syllable-level to encrypt speech. Although this scheme has nice retrieval speed and improves the retrieval efficiency, but the characteristics of the algorithm are too simple, the discrimination is not high. Zhang *et al.* [23] proposed a perceptual hashing based on short-time zero-crossing rate to retrieve encrypted speech. Although it has good retrieval accuracy and robustness, the algorithm has low discrimination and the encryption effect is not enough. Zhang *et al.* [21] proposed a perceptual hashing based on IFFT and measurement matrix to retrieve encrypted speech. Although this scheme has better encryption and discrimination, the robustness and retrieval accuracy of this algorithm are not high. Zhang *et al.* [19] proposed a perceptual hashing based on Chirp-Z and second feature extraction to retrieve encrypted speech. Although this algorithm has good discrimination and retrieval efficiency, the robustness of the algorithm is still not high, and the accuracy of retrieval is not enough.

These studies show that there are still many problems and shortcomings in the current encrypted speech retrieval scheme, such as discrimination and security are not enough. In order to solve these problems, we propose a hyperchaotic encrypted speech perceptual hashing retrieval algorithm based on 2D-Gabor transform and PCA dimension reduction. The proposed algorithm first use Gabor transform to extract speech feature, then use PCA on extracted feature to reduce the dimensional. At the same time, the algorithm use four-dimensional hyperchaotic to encrypt the speech files. After that, the algorithm operate the dimensional-reduced feature to generate perceptual hash sequence, and store them in the hash sequence table. In the retrieval process, the algorithm first extract the feature of the speech to be retrieved and generate a perceptual hash, then match it with the hash sequence table, finally return the matching result to the user.

The main contributions of our approach can be summarized as follow:

1) The Gabor feature extraction used in this paper can effectively generate a hash sequence that can well represent the feature information of speech, and has good discrimination and nice robustness;
2) This paper uses PCA to reduce a number of feature datas, which can greatly improve the efficiency of retrieval;
3) This paper uses four-dimensional hyperchaotic encryption for speech encryption, which not only has a large key space and can not easily brute-forced, but also greatly reduces the correlation between each frame of speech;
4) This paper uses Minimum code distance to tamper detection,which can not only detect that the speech has been tampered, but also accurately locate the tampered location.

The remaining part of this apper is organized as follows. Section 2 introduces the related theory, including Gabor Feature Extraction, Four-dimensional hyperchaotic map system. Section 3 describes in detail the specific implementation process of the proposed algorithm in this paper. Section 4 gives the experimental results and performance analysis as compared with other related algorithms. Finally, we conclude our paper and give the future perspectives in Section 5.

2 Related Theory

2.1 Gabor Feature Extraction

Fourier transform is a mathematical analysis method widely used in the field of signal processing, however it is mainly used to analyze stationary signal, but the characteristics of the signal in the local area cannot be processed well. Gabor transform [10] was proposed to solve this problem:

Generally for any $f(t) \in L^2(R)$, the Gabor transform defined as Equation (1):

$$G_f(b,w) = \int_{-\infty}^{+\infty} f(t)e^{-jwt}s(t-b)dt. \tag{1}$$

When the Window function $s(x)$ is Gaussian function, thus $s_a(x) = \frac{1}{2\sqrt{\pi a}} \exp\left(-\frac{x^2}{4a}\right)$, $a > 0$. So there is one-dimensional Gabor core function as Equation (2):

$$g(a,b,w,t) = e^{jwt}s_a(t-b). \tag{2}$$

Although one-dimensional Gabor transform have many improvements when dealing with local features compared with Fourier transform, but the one-dimensional Gabor transform cannot completely describe the characteristics of the signal. In order to better describe the signal characteristics, 2D-Gabor [7] was proposed to solve this problem. Expanding the one-dimensional Gabor core function into two-dimensional space, we can get the 2D-Gabor core function as Equation (3):

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right) * \exp\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \tag{3}$$

The real part as Equation (4):

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right) * \cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{4}$$

And the imaginary part as Equation (5):

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right) * \sin\left(2\pi\frac{x'}{\lambda}+\psi\right) \quad (5)$$

Among them $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$. And λ is the wavelength of sine function, θ is the direction of the core function, φ is the phase shift, σ is the Gaussian standard deviation, and γ is the aspect ratio of the two directions: x, y (That is, the ellipticity of the Gabor function is specified).

Gabor Feature Extraction is to process the signal S(x,y), but if we directly operate the obtained data, because the dimension is too high, which is not conducive to subsequent processing. So we generally block the signal first, for example: Take 16 equal divisions in the horizontal and vertical directions respectively, divide the signal into 16 × 16 sub-signal blocks, then calculate the energy corresponding to each block as Equation (6):

$$e(k) = \sum_{i=1}^{16}\sum_{j=1}^{16}|a(k)|^2; k = 1,2,\cdots,64 \quad (6)$$

Finally, we can get the frequency energy matrix E.

2.2 Four-Dimensional Hyperchaotic System

A currently accepted definition of Chaotic is Li-Yorke chaotic [16], the defination is as follow:

If there is a closed interval I, and $f(x)$ is a continuous self-mapping on I, if the following conditions are met, it is considered chaotic:

1) Continuous self-mapping function $f(x)$ is unbounded for any period;

2) In a closed interval I there has an uncountable subset S, and meet the following conditions:

   a. For any x and y, and $x \in S$, $y \in S$, there is: $\liminf_{n\to\infty}|f^n(x) - f^n(y)| = 0$;

   b. For any x and y, and $x \in S$, $y \in S$, and satisfy $x \neq y$ there is: $\limsup_{n\to\infty}|f^n(x) - f^n(y)| > 0$;

   c. For any x and y, and$x \in S$, $y \in S$, among them y is any period of $f(x)$, there is: $\limsup_{n\to\infty}|f^n(x) - f^n(y)| < 0$.

However, the dynamic equations of low-dimensional chaotic system are too simple, and the key sensitivity is not high enough. Therefore, this paper proposes a new four-dimensional hyperchaotic system as Equation (7):

$$\begin{cases} x_1(n+1) = \frac{2\alpha_1\sin(\beta_1 x_1(n))}{\gamma_1\sin(x_4(n))^2+w_1} \\ x_2(n+1) = \frac{2\alpha_2\sin(\beta_2 x_2(n))}{\gamma_2\sin(x_1(n))^2+w_2} \\ x_3(n+1) = \frac{2\alpha_3\sin(\beta_3 x_3(n))}{\gamma_3\sin(x_2(n))^2+w_3} \\ x_4(n+1) = \frac{2\alpha_4\sin(\beta_4 x_4(n))}{\gamma_4\sin(x_3(n))^2+w_4} \end{cases} \quad (7)$$

Among them $x(n)$ represent the chaotic sequence, $\alpha_i, \beta_i, \gamma_i, w_i$ are the parameters of chaotic sequence, and satisfy $\alpha_i, \beta_i, \gamma_i, w_i \neq 0$, $i = 1,2,3,4$.

The Lyapunov exponent is a quantitative description of chaotic system, which reflects the overall effect of the movement trajectories generated by the nonlinear mapping being close to or separated from each other, and describes the sensitivity of the system to the initial value when the parameters change in the chaotic motion system and the local instability in changing process. Therefore, having positive Lyapunov exponent can be used as the basis for discriminating chaotic system. The formula of Lyapunov exponent as Equation (8):

$$\lambda = \lim_{n\to\infty}\frac{1}{n}\sum_{n=0}^{n-1}\ln\left|\frac{df(x_n,\mu)}{dx}\right| \quad (8)$$

For a low-dimensional chaotic systems, at least there have one positive Lyapunov exponent, but for a high-dimensional hyperchaotic system, there must be at least two positive Lyapunov exponents, so the behavior is more complicated than general chaotic system, making it more difficult to predict.

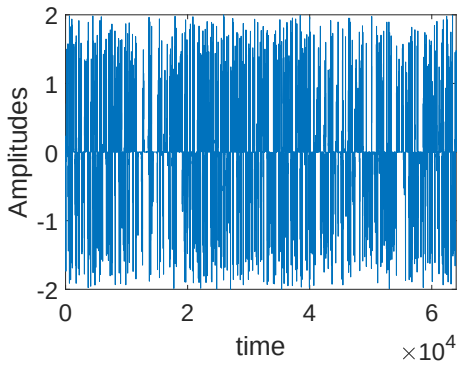


Figure 1: The waveform of ecryption system

According to Equation (8), Equation (7) have two positive Lyapunov exponents, $\lambda_1 = 2.7161$ and $\lambda_2 = 0.2445$, so this hyperchaotic system has sufficiently complex chaotic property. The time-domain waveform of hyperchaotic system as Figure 1. It can be clearly seen that the time-domain waveform of this hyperchaotic system is sufficiently complex to hide the waveform of target signal in the hyperchaotic waveform.

3 The Proposed Algorithm

3.1 System Model

The system model of the scheme mainly includes three parts: Sever terminal, Client terminal and Speech retrieval. As shown in Figure 2, we first generate perceptual hashing of all original speech files on the server side, store all the perceptual hashing sequences in the hash sequence table, then store them in the cloud. At the same time,

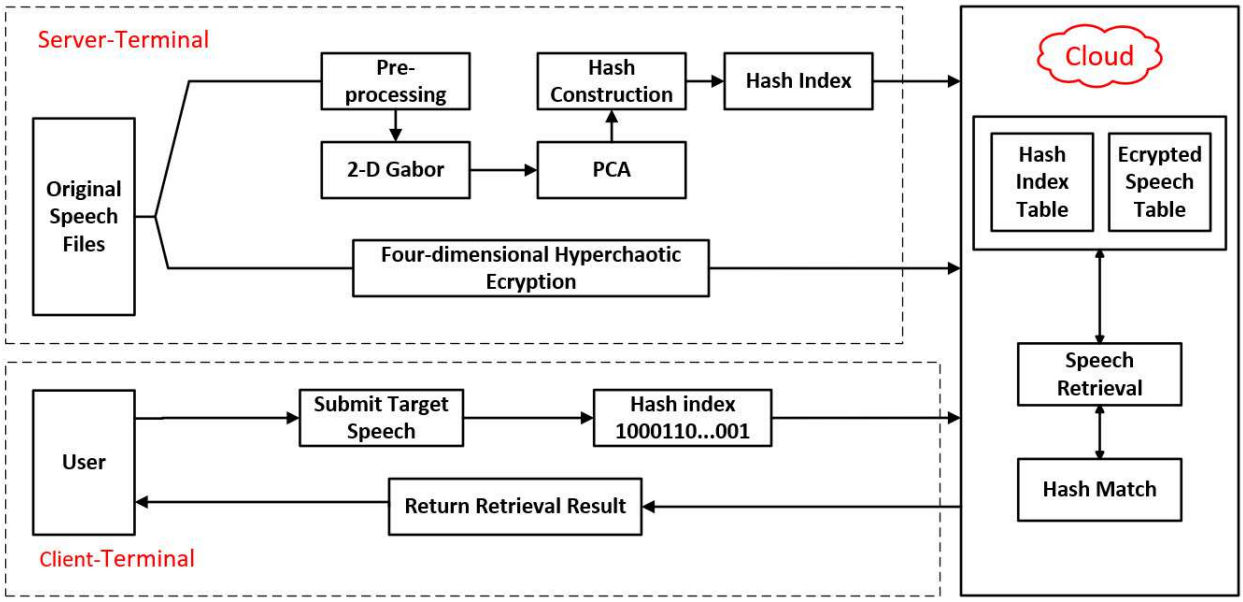


Figure 2: The flow chart of the proposed speech retrieval algorithm

we encrypt all the speech files, upload to the cloud, and realize the one-by-one mapping relationship with the corresponding speech hash sequence. When the user needs to retrieve the speech, first submit the speech to be retrieved to the client, then the system will process the retrieved speech to generate a hash sequence and upload it to the cloud. In the cloud, the system will match the hash sequence generated by the speech to be retrieved with the hash sequence table. If the same sequence is matched, the corresponding encrypted speech is found, and the result of a successful match is returned. If not found, the result of the failed match is returned to the user.

3.2 Speech Encryption Process

Assume the size of the speech to be encrypted is $1 \times m$ (The original speech has been processed into mono), the encryption steps are as follow:

1) Given the initial conditions and system parameters, repeatedly iterate Equation (7) N times, remember the result of the $N-th$ time is $X_N = [x_1(N), x_2(N), x_3(N), x_4(N)]$ (This step is to keep the initial conditions random enough to minimize human factors);

2) Use X_N as the initial conditions, then iterate m times, get the chaotic sequence: $\{H(k)|k = 1, 2, \cdots, m\}$, and satisfy $H(k) = x_1(k)$;

3) We arrange the chaotic sequence in ascending order $\boldsymbol{H} = \{h_1, h_2, h_3, \ldots, h_n\}$ to get a new sequence $\boldsymbol{K} = \{k_1, k_2, k_3, \ldots, k_j\}$, $j = 1, 2, 3, \ldots, M$;

4) Chaotic sequence $\{H(k)|k = 1, 2, \cdots, m\}$ Meet the mapping relationship with the encrypted speech E: $E(j) = H(i)$, scrambling the original speech signal according to the mapping relationship.

In the algorithm proposed in this paper, given system parameters, initial conditions as follow:

$$\text{keys} = \left\{ \begin{array}{c} \alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4 \\ \gamma_1, \gamma_2, \gamma_3, \gamma_4, w_1, w_2, w_3, w_4 \\ x_1(1), x_2(1), x_3(1), x_4(1), N \end{array} \right\}$$

Upload the encrypted speech after the above operation to the cloud.

3.3 Feature Extraction and Hash Sequence Construction

The steps of feature extraction and generating hash sequence algorithm are as follows:

1) Pre-processing: Pre-emphasis the input signal $s(t)$ to get $s(t)'$, Pre-emphasis can increase the features of the speech signal's high-frequency components. Then, the processed signal is framed, $s(t)'$ is divided into m frame, and get $f_i = \{f_i(n)|n = 1, 2, \ldots, L/m, i = 1, 2, \ldots, m\}$. L is the length of speech, m is the total number of frames, and $f_i(n)$ is the $n-th$ frame;

2) Feature extraction: According to Equation (3) to process $f(n)$ with 2D-Gabor transform, then use Equation (3) to get feature vector $\{V(k)|k = 1, 2, \cdots, m\}$;

3) Dimensional reduction: We use PCA to reduce the feature vector $\{V(k)|k = 1, 2, \cdots, m\}$ dimensional to get vector H;

4) Hash sequence generation: We use Equation (9) to get the hashing sequence:

$$\boldsymbol{h}(i) = \begin{cases} 1, & \text{if } H(i+1) > H(i) \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

where $\boldsymbol{h}(1) = 0$, thus we get the hash sequence $\boldsymbol{h} = \{h(i)|i = 1, 2, \ldots, m\}$.

4 Experimental Results and Analysis

4.1 Experimental Environment and Main Parameter Settings

We conducted a series of experiments to evaluate our approach using speech samples from the standard Texas Instrument and Massachusetts Institute of Technology (TIMIT) and Text to Speech (TTS) speech library as the test speech. The library consists of 1200 speech clips stored as 16 bits 16kHz mono recordings. The operating experimental hardware platform is Intel(R) core(TM) i5-7500 CPU @ 3.40 GHz, with memories of 4G. The operating system is window 7. And the simulation platform is MATLAB R2018b.

4.2 Performance Analysis of Perceptual Hashing

In this section, we use discrimination and robustness for evaluate the performance of the extracted speech perceptual hashing. Whether to extract the perceptual hashing sequence with good performance is the important part of speech retrieval. At the same time, we also analyzed Encryption effect and retrieval effect.

4.2.1 Discrimination Analysis

Discrimination is one of the important indicators for evaluating the speech hash sequence. The discrimination of perceptual hash is used to determine the degree of similarity between two speeches. We determine the similarity between two speechs by calculating the hamming distance between two speech hash sequences(Also named bit error rate, BER), the calculating formula of the normalized Hamming distance $D(H_{\mathrm{x}}, H_{\mathrm{q}})$ is shown in Equation (10):

$$\begin{aligned} D(H_x, H_q) &= \frac{1}{N}\sum_{p=1}^{N}(|H_x(p) - H_q(p)|) \\ &= \frac{1}{N}\sum_{p=1}^{N} H_x(p) \oplus H_q(p). \end{aligned} \tag{10}$$

Where H_{q} is the perceptual hashing sequence of query speech, H_{x} is perceptual hashing sequence in the hash sequence table, N is the length of perceptual hashing value and $p = 1, 2, \ldots, N$, setting T as the similarity threshold.

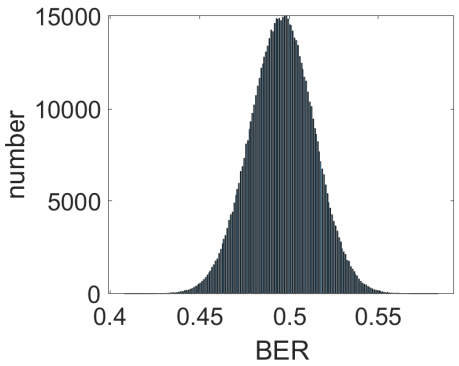


Figure 3: Statistics histogram of 1200 speech clips matching results

If $D(H_{\mathrm{x}}, H_{\mathrm{q}}) < T$, then the two corresponding hash sequences match successfully, otherwise the match is wrong, and the accuracy of retrieval is related to the threshold. The statistic histogram of BERs of the matching results is shown in Figure 3.

As shown in Figure 3, The hash BER of different speech contents basically conforms to the normal distribution. That shows the perceptual hash sequence algorithm proposed in this paper has a nice randomness and collision resistance perforamnce. The probability of the BER normal distribution is shown in Figure 4.

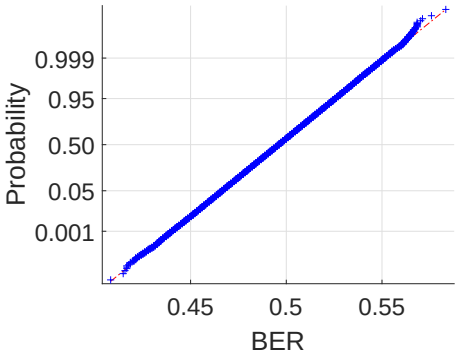


Figure 4: Probability distributions of 1200 speech clips matching results

As can be seen in Figure 4, the probability distributions of BER values of different speech basically conforms to the standard normal distribution. This indicates that the Hamming distance between different speechs is approximately normal distribution.

In order to better quantify the discrimination of the proposed algorithm, the False Accept Rate (FAR) and False Reject Rate (FRR) are mentioned. Their calculation formula are as Equation (11) and Equation (12):

$$FAR(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{11}$$

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{12}$$

where τ is the similarity threshold, μ is the expected value, σ is the standard deviation. FAR and FRR are used to evaluate the discrimination and the robustness of the algorithm. The lower FAR means better discrimination, and the lower FRR means better robustness. According to the De Moivre-Laplace central limit theorem, the Hamming distance(also named BER) approximately obey the normal distribution ($\mu = p, \sigma = \sqrt{p(1-p)/N}$, Nis the number of bits in hashing sequence, p represents the probability of 0 or 1). In this paper, the length of the hash sequence of the speech clips is N = 1068. According to the De Moivre-Laplace central limit theorem, the mean value of normal distribution is $\mu = 0.5$, the variance is $\sigma = 0.0153$. The mean value of the experimental is $\mu_0 = 0.4960$, the variance is $\sigma_0 = 0.0179$. It can be seen from Figure 5, that the values of μ and σ measured in this paper are very close to the theoretical value. This shows that the hash sequence generated by this algorithm has high randomness and collision resistance, so as to ensure that each speech has its own unique hash sequence. In Table 1, we compare the FAR value under different

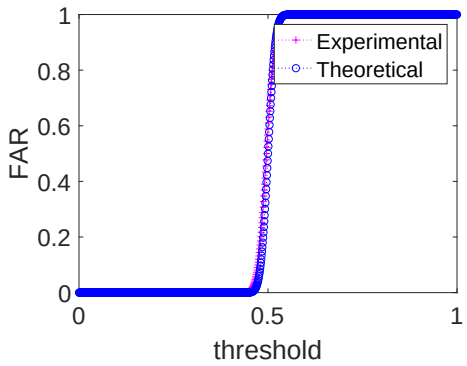


Figure 5: The FAR curve of hashing sequence

thersholds of the proposed algorithm with those existing algorithms [19–21, 23].

It can be seen from Table 1, the smaller the matching threshold τ is, the smaller the FAR value is. In the proposed algorithm, when the threshold $\tau = 0.16$ is set, about 1.3 of each 10^{80} speech clips are false accepted. This indicates that the algorithm proposed in this paper has high discrimination. When $\tau = 0.16$, about 1.8 of each 10^{-33} speech clips in [20] false accepted, about 2.3 of each 10^{-22} speech clips in [23] false accepted, about 3.3 of each 10^{-29} speech clips in [21] false accepted, about 2.5 of each 10^{-29} speech clips in [19] false accepted.

All in all, compare with the proposed algorithm in [19–21, 23], the proposed algorithm in this paper have lower FAR. What this means is that compare with the proposed algorithm in [19–21, 23], the proposed algorithm have higher discrimination.

4.2.2 Robustness Analysis

Robustness is to judge the same speech under different Content Preserving Operation (CPO) the degree of change of the speech perceptual hash sequence. The lower the robustness value, the less the extracted perceptual hash under different CPOs will be affected.

Table 2 introduces the different CPOs and their operations in proposed algorithm. Under seven kinds of CPOs, 1200 speech clips paired to compare the BER, the FRR-FAR curve is obtained in Figure 6. As can be seen from

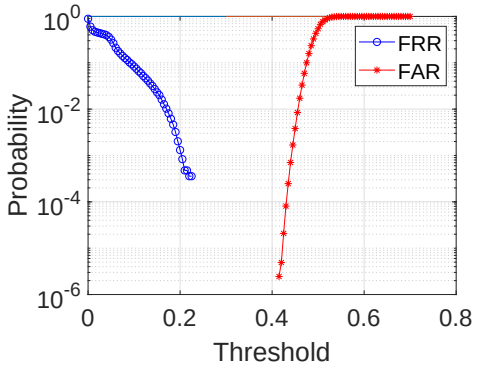


Figure 6: The FRR-FAR curve

Figure 6, there is no cross section, and the interval is still very large, indicating that the algorithm proposed in this paper has a large space to accurately determine different speech content, even after the CPOs. Obviously, when the matching threshold can be setting between 0.20 and 0.41, a good retrieval effect will be obtained. At the same time, it also shows that the algorithm proposed in this paper has both high discrimination and robustness.

Table 3 shows the BER mean value, max value and the variance of the BER value. It can be obtained from Table 3: The average BER obtained by the algorithm proposed in this paper does not exceed 0.1, and the maximum BER does not exceed 0.2. It shows that the algorithm in this paper can maintain good robustness under different CPOs. And if we do not consider the case of narrowband noise with SNR=30db, the average BER obtained in this paper does not exceed 0.05, and the maximum BER does not exceed 0.08. It shows that the algorithm in this paper can maintain nice robustness even under different CPOs. The mean BER comparison results of this algorithm is then compared with the algorithms in [19–21, 23] and the result is shown in Table 4.

As can be seen from Table 4, the average BER value of the algorithm proposed in this paper is smaller than the algorithm proposed in the [20] regardless any CPOs, which means that the algorithm in this paper is more robust than the algorithm in [20]. Our result is equal to or better than the algorithm proposed in the [19, 21, 23], indicating that the algorithm proposed in this paper is at least as good as the [19, 21, 23]. But from the previous part of this paper, we can see that the discrimination of the algorithm in this paper is much better than [19,21,23].

Table 1: Comparison of FAR values

τ	**Proposed method**	**[20]**	**[23]**	**[21]**	**[19]**
0.02	6.0115×10^{-160}	1.4486×10^{-66}	7.9324×10^{-43}	2.1743×10^{-56}	1.9366×10^{-56}
0.04	6.1132×10^{-147}	6.2274×10^{-61}	1.7718×10^{-39}	6.0859×10^{-52}	5.2700×10^{-52}
0.06	1.7184×10^{-134}	3.3854×10^{-56}	2.8523×10^{-36}	1.1039×10^{-47}	9.3200×10^{-48}
0.08	1.3354×10^{-122}	7.6358×10^{-52}	3.3141×10^{-33}	1.2978×10^{-43}	1.0713×10^{-43}
0.10	2.8699×10^{-111}	9.9134×10^{-47}	2.7782×10^{-30}	0.8909×10^{-40}	8.0059×10^{-40}
0.12	1.7058×10^{-100}	2.3020×10^{-42}	1.6830×10^{-27}	4.8877×10^{-36}	3.8902×10^{-36}
0.14	2.8053×10^{-90}	3.6436×10^{-38}	7.3382×10^{-25}	1.5665×10^{-32}	1.2295×10^{-32}
0.16	1.2768×10^{-80}	1.7923×10^{-33}	2.3147×10^{-22}	3.2571×10^{-29}	2.5281×10^{-29}

Table 2: Content preserving operation

CPO	**Operation method**	**Abbreviation**
Re-sampling	8-16kbps	R
Amplitude increase	3 db for amplitude increase	$A\uparrow$
Amplitude decrease	3 db for amplitude decrease	$A\downarrow$
Narrowband Noise 1	SNR=30db	N1
Narrowband Noise 2	SNR=50db	N2
MP3 compression	128kbps	M
Echo addition	Attenuation 50%	E

Table 3: The BER value after CPO

CPO	**Mean**	**Max**	**Variance**
R	0.0267	0.0774	4.0292×10^{-4}
$A\uparrow$	0.0010	0.0128	2.2262×10^{-6}
$A\downarrow$	0.0021	0.0088	2.3241×10^{-6}
N1	0.0941	0.2049	1.4000×10^{-3}
N2	0.0219	0.0705	2.3855×10^{-4}
M	0.0031	0.0091	3.6840×10^{-6}
E	0.0473	0.0617	2.0654×10^{-5}

It can be seen that the algorithm proposed in this paper balances the discrimination and robustness of the perceptual hash sequence, on the premise of greatly improving the discrimination of the algorithm, the robustness of the algorithm does not lost. It shows that the perceptual hash sequence in this paper will not be greatly lost in different speech environments, so it can meet the needs of speech retrieval.

4.3 Encryption Performance Analysis

We encrypted and decrypted the speech signal during transmission using the Four-dimensional hyperchaotic map system described previously. First, given the key:

$$\text{keys} = \left\{ \begin{array}{c} 1,2,3,4,2,2,4,1 \\ 3,4,2,2,4,2,2,1 \\ 3,2,5,4,500 \end{array} \right\}$$

Then we randomly select a speech from the speech library to encrypt and decrypt it, and analyze the effectiveness of the encryption algorithm proposed in this paper. Figure 7 shows the speech waveform before encryption, Figure 8 shows the the speech waveform after encryption and Figure 9 shows the speech waveform after decryption.

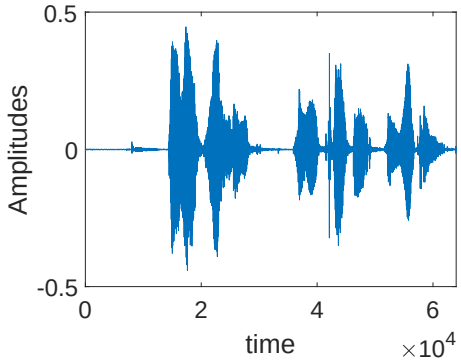


Figure 7: Original speech waveform

As we can seen from Figure 8 and Figure 9, encrypted speech has no regularity at all, and can't see any features of the original speech at all and the decrypted speech is basically the same as the original speech. For the effectiveness of an encryption algorithm, correlation analysis is also a very important criterion. For a speech clip, use Equation (13) to calculate the correlation coefficient be-

Table 4: The BER mean value of different algorithm

CPO	Proposed method	[20]	[23]	[21]	[19]
R	0.0267	0.0304	0.0033	-	0.0283
$A\uparrow$	0.0010	0.0925	0.0160	0.0052	0.0054
$A\downarrow$	0.0021	0.0089	0.0038	0.0015	0.0014
$N1$	0.0941	-	0.0248	0.0424	-
$N2$	0.0219	0.0416	-	0.0032	0.0267
M	0.0031	-	0.0090	0.2028	0.1928
E	0.0473	0.2375	-	0.1467	0.1505

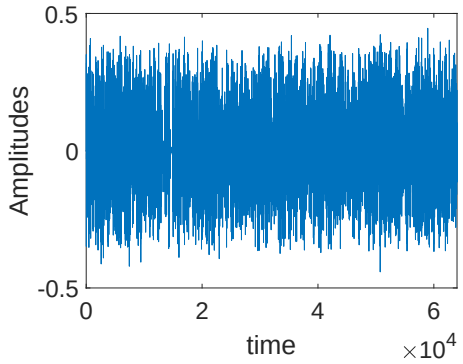


Figure 8: Encryption speech waveform

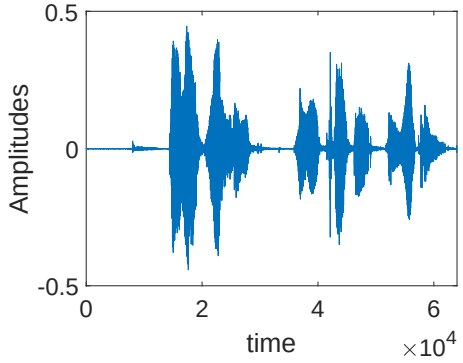


Figure 9: Decryption speech waveform

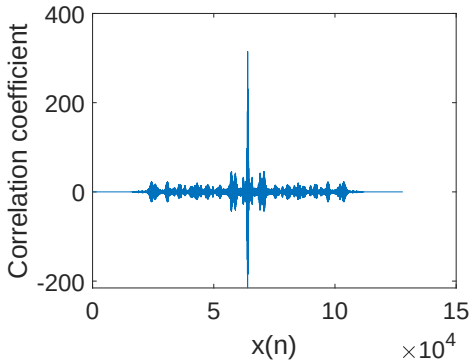


Figure 10: The correlation coefficient before encryption

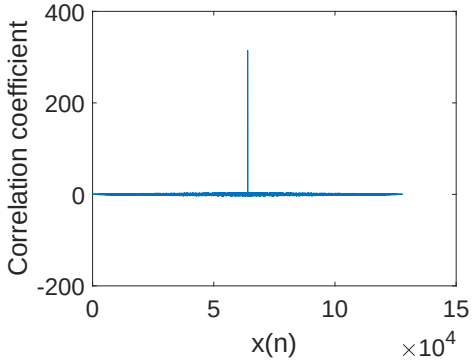


Figure 11: The correlation coefficient after encryption

tween adjacent sample points of speech.

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^*, & m \geq 0 \\ \hat{R}_{yx}^*(-m), & m < 0 \end{cases} \tag{13}$$

According to Equation (13), we can get the speech correlation coefficient before encryption and after encryption as Figure 10 and Figure 11. As can be seen from Figure 10 and Figure 11, the encrypted speech correlation coefficient is basically zero, indicating that the encryption effect is nice (At zero point, the speech x is exactly same as itself, and the maximum peak of the correlation coefficient appears). The original speech has a clear correlation, but the encrypted speech can't see the obvious correlation at all. This shows that the encryption algorithm proposed in this paper confuses the relevant characteristics of the original speech, so it also proves that the algorithm in this paper has high security.

In order to measure the disorder of our encryption algorithm, the position number before scrambling and the change of position number after scrambling are used to describe in this paper.

If the position number before and after the scrambling of a speech segment has not changed, there have Equation (14):

$$\Delta_i = l(i) - l'(i) = 0 \tag{14}$$

If the position number before and after the scrambling of a speech segment changes, then, there have Equa-

tion (15):

$$\Delta_i = l(i) - l'(i) \neq 0 \tag{15}$$

where, Δ_i represents the position difference.$l(i)$ represents the $i-th$ position number of the original hash sequence.$l'(i)$ represents the $i-th$ position number after scrambling. It can be seen from Figure 12 that Δ_i

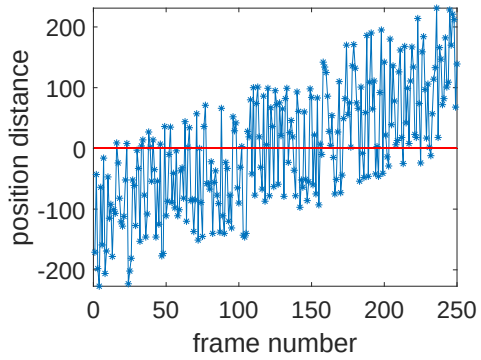


Figure 12: The intersection of Δ and the line $y = 0$

and the line $y = 0$ has few intersections, which proves that our encryption algorithm has good disorder.

4.4 Tamper Detection and Location

Aiming at the low-resolution tamper detection capability and low BER of speech segments, this paper proposes a tamper detection and localization algorithm based on minimum code distance (MCD) of Hamming code. In the detection process, for the original speech $x(n)$ and the original speech $x'(n)$ after the malicious attack, the hash sequence $h(n)$ and $h'(n)$ are obtained through the hash template. Then the MCD of the Hamming code between each frame of the two sequences is calculated. Finally, determine whether the speech has been attacked or not. defined as Equation (16):

$$MCD(i) = \begin{cases} 1, & h(i) \neq h'(i) \\ 0, & h(i) = h'(i) \end{cases} \tag{16}$$

Where, $MCD(i)$ is the MCD of the Hamming code of the $i-th$ frame, and its matrix form is as follows:

$$MCD(i) = \left[\begin{array}{cccc} MCD(1) & MCD(2) & \ldots & MCD(i) \end{array} \right]$$

As shown in Figure 13,Figure 14, the blue represents speech and red represents an area where speech content is tampered. It can be seen from Figure 15 that the algorithm can effectively detection and localization tamper. Which proves that the algorithm has good tamper detection and location capabilities for small-scale malicious attacks.

4.5 Retrieval Performance Analysis

As can be seen from above paper, according to FAR-FRR curve, we know that it is most appropriate to set the

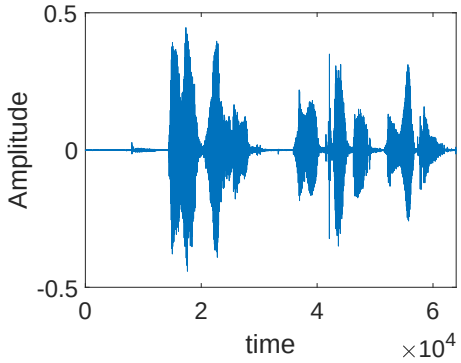


Figure 13: Original speech

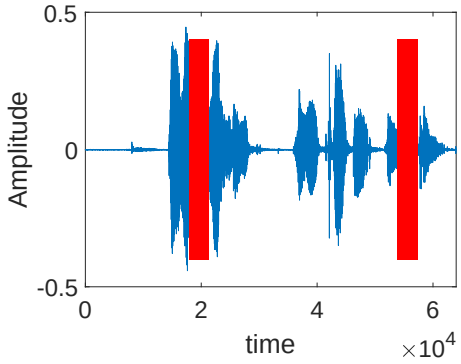


Figure 14: Tampered speech

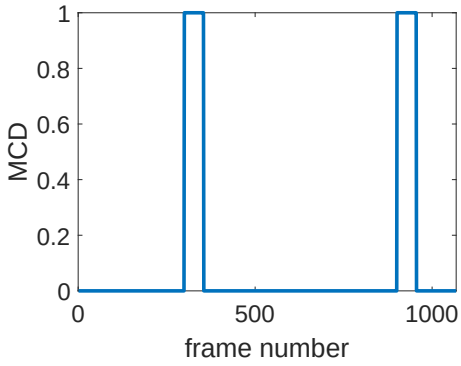


Figure 15: The minimum code distance of Hamming code

threshold between 0.25 and 0.41, so in the next experiment, we will set the threshold T=0.3, so if $D(H_x, H_q) < T$, the retrieval is successful. In order to test whether the retrieval system we proposed in this paper can accurately retrieve speech clips, we randomly select a speech clip for retrieval. The retrieval result is show in Figure 16. As show in Figure 16, our chosen threshold T=0.3 is suitable. In speech hash sequence table, except the speech to be retrieved, there is no BER value less than 0.3, which basically floats around 0.5. So the retrieval algorithm proposed in this paper is successful. The criteria for judging the performance of speech retrieval are Recall ratio (R)

Table 5: Comparison of the recall ratio after CPO

CPO	The recall ratio (%)				
	Proposed method	**[20]**	**[23]**	**[21]**	**[19]**
R	100	100	100	100	100
$A\uparrow$	100	100	100	100	100
$A\downarrow$	100	100	100	100	100
N1	100	-	100	-	-
N2	100	100	-	-	100
M	100	-	100	100	100
E	100	100	-	-	100

Table 6: Comparison of the precision ratio after CPO

CPO	The Precision ratio (%)				
	Proposed method	**[20]**	**[23]**	**[21]**	**[19]**
R	100	100	100	100	100
$A\uparrow$	100	100	100	100	100
$A\downarrow$	100	100	100	100	100
N1	100	-	100	-	-
N2	100	100	-	-	100
M	100	-	100	97	98
E	100	96	-	-	99

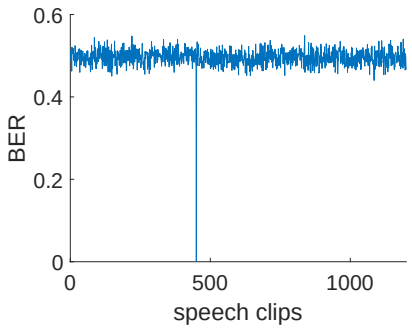


Figure 16: The matching result

and Precision ratio (P), the combination of these two criteria can describe the success rate of retrieval. The R is to describe the ability of retrieve relevant content, as show in Equation (17):

$$R = \frac{a}{a+b} \times 100\% \tag{17}$$

Where a represents the number of retrieved successfully, and b represents the number of retrieval failures. The P is to describe the ability of accurately retrieved relevant content, as show in Equation (18):

$$P = \frac{a}{a+d} \times 100\% \tag{18}$$

Where d is the number of errors retrieved.

Next, we use R and P to compare the performance of the retrieval algorithm proposed in this paper with those existing algorithms [19–21, 23]. As shown in Table 5 and Table 6, regardless of any CPOs, the recall ratio and precision ratio of the retrieval algorithm proposed in this paper can be maintained at 100% retrieval success rate, indicating that the algorithm proposed in this paper is very effective. But in [19–21, 23], due to various reasons, there will be more or less some retrieval failures.

Table 7: Efficiency comparison of different algorithms

Algorithm	**Average retrieval time**
Proposed algorithm	0.1209 s
[20]	0.1467 s
[21]	0.0649 s
[19]	0.0582 s

The efficiency of the retrieval algorithm is also an important indicator to judge the performance of the algorithm. In order to test the efficiency of our proposed retrieval algorithm, we perform a circular search on all the speechs in the speech library to ensure that each speech was retrieved once, and calculate each speech average retrieval time. The retrieval time comparison between the retrieval algorithm proposed in this paper and the algorithm in [19–21] is shown in Table 7.

As can be seen from Table 7, although the algorithm proposed in this paper has significantly improved the discrimination and robustness of the perceptual hash compared to the [19–21], but the feature extraction algorithm is too complicated, which leads to the retrieval efficiency

reduce.

5 Conclusions

In this paper, a hyperchaotic encrypted speech perceptual hashing retrieval algorithm based on 2D-Gabor transform and PCA dimension reduction has proposed. Compared with the existing algorithms, the feature extraction method of this algorithm can greatly improve the discrimination of the algorithm. Moreover, the four-dimensional hyperchaotic system reduces the correlation of speech, thereby greatly improving the security of speech in the transmission process. The experimental results show the advantages of our algorithm:

1) The generated perceptual hashing sequence has high discrimination, which can greatly improve the accuracy of retrieval.

2) The generated perceptual hashing sequence has nice robustness, which can adapt to multiple CPOs.

3) The key space of the four-dimensional hyperchaotic encryption system proposed in this paper is very large, so it can resist brute force cracking, and there is no obvious correlation between the speech that before and after encrypt, so it has high security.

The proposed algorithm has high security, retrieval accuracy and retrieval efficiency, it may be applied to cloud search, mobile speech assistant, *etc.*

The shortcoming of the proposed algorithm is when the perceptual hashing sequence is in the narrowband noise with SNR=30db, the robustness is poor and the efficiency of retrieval is not high. Therefore, improving the robustness of the perceptual hashing algorithm and further reducing the complexity of the algorithm is the next research goal.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61862041), Youth Science and Technology Found of Gansu Province of China (No.1606RJYA274).

References

[1] R. Biswas, R. A. Vasco-Carofilis, E. F. Fernandez, F. J. Martino, and P. B. Medina, "Perceptual hashing applied to tor domains recognition," *Computer Vision and Pattern Recognition*, 2020. arXiv:2005.10090.

[2] M. Blaß and R. Bader, "Content-based music retrieval and visualization system for ethnomusicological music archives," in *Computational Phonogram Archiving*, pp. 145–173, 2019.

[3] D. Dash, P. Ferrari, S. Malik, and J. Wang, "Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP'18)*, pp. 489–493, 2018.

[4] L. Du, A. T. S. Ho, and R. Cong, "Perceptual hashing for image authentication: A survey," *Signal Processing: Image Communication*, vol. 81, pp. 115713, 2020.

[5] L. Fan, "Audio example recognition and retrieval based on geometric incremental learning support vector machine system," *IEEE Access*, vol. 8, pp. 78630–78638, 2020.

[6] S. He and H. Zhao, "A retrieval algorithm of encrypted speech based on syllable-level perceptual hashing," *Computer Science and Information Systems*, vol. 14, no. 3, pp. 703–718, 2017.

[7] P. Huang, T. Mao, Q. Yu, Y. Cao, J. Yu, G. Zhang, and D. Hou, "Classification of water contamination developed by 2-d gabor wavelet analysis and support vector machine based on fluorescence spectroscopy," *Optics express*, vol. 27, no. 4, pp. 5461–5477, 2019.

[8] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*, pp. 7120–7124, 2019.

[9] S. Kassim, O. Megherbi, H. Hamiche, S. Djennoune, and M. Bettayeb, "Speech encryption based on the synchronization of fractional-order chaotic maps," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT'19)*, pp. 1–6, 2019.

[10] J. Luo, G. Liu, Z. Huang, and S. S. Law, "Mode shape identification based on gabor transform and singular value decomposition under uncorrelated colored noise excitation," *Mechanical Systems and Signal Processing*, vol. 128, pp. 446–462, 2019.

[11] P. K. Naskar, S. Paul, D. Nandy, and A. Chaudhuri, "DNA encoding and channel shuffling for secured encryption of audio data," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 25019–25042, 2019.

[12] G. Qian, "A music retrieval approach based on hidden markov model," in *The 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA'19)*, pp. 721–725, 2019.

[13] A. Saleh and S. B. Sadhkan, "A proposed speech scrambling based on haar transform and permutation," in *The 2nd International Conference on Engineering Technology and its Applications (IICETA'19)*, pp. 31–36, 2019.

[14] P. Sathiyamurthi and S. Ramakrishnan, "Speech encryption algorithm using FFT and 3D-lorenz–logistic chaotic map," *Multimedia Tools and Applications*, vol. 79, pp. 17817-17835, 2020.

[15] J. S. Seo, J. Kim, and H. Kim, "Audio fingerprint matching based on a power weight," *The Journal of the Acoustical Society of Korea*, vol. 38, no. 6, pp. 716–723, 2019.

[16] Y. Wang, E. Chen, and X. Zhou, "Mean li-yorke chaos for random dynamical systems," *Journal of Differential Equations*, vol. 267, no. 4, pp. 2239–2260, 2019.

[17] M. Wang, K. Li, L. Luo, X. Song, Z. Zhou, and H. Qin, "An subarea localization algorithm based on combination features using representative audio fingerprint," in *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA'19)*, pp. 374–380, 2019.

[18] H. Wang, L. Zhou, W. Zhang, and S. Liu, "Watermarking-based perceptual hashing search over encrypted speech," in *International Workshop on Digital Watermarking*, pp. 423–434, 2013.

[19] Q. y. Zhang, Z. X. Ge, Y. J. Hu, J. Bai, and Y. B. Huang, "An encrypted speech retrieval algorithm based on chirp-z transform and perceptual hashing second feature extraction," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6337–6361, 2020.

[20] Q. Y. Zhang, Z. X. Ge, and S. B. Qiao, "An efficient retrieval method of encrypted speech based on frequency band variance," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 9, no. 6, pp. 1452–1463, 2018.

[21] Q. Zhang, Z. Ge, L. Zhou, and Y. Zhang, "An efficient retrieval algorithm of encrypted speech based on inverse fast fourier transform and measurement matrix," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, no. 3, pp. 1719–1736, 2019.

[22] Q. Zhang, Y. Li, Y. Hu, and X. Zhao, "An encrypted speech retrieval method based on deep perceptual hashing and cnn-bilstm," *IEEE Access*, vol. 8, pp. 148556–148569, 2020.

[23] Q. Y. Zhang, L. Zhou, T. Zhang, and D. H. Zhang, "A retrieval algorithm of encrypted speech based on short-term cross-correlation and perceptual hashing," *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 17825–17846, 2019.

[24] Q. Zhang, D. Zhang, and L. Zhou, "An encrypted speech authentication method based on uniform sub-band spectrum variance and perceptual hashing," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 5, pp. 2467–2482, 2020.

Biography

Yi-bo Huang received the PhD degree from Lanzhou University of Technology in 2015, and now working as a Associate Professor in the college of physics and electronic engineering in Northwest Normal University. His research interests include multimedia information processing, information security and speech recognition.

Shi-hong Wang received the BS degree in Liren college of Yanshan University, Hebei,China, in 2018. His research interests include speech signal processing and application,

multimedia retrieval techniques.

Yong Wang received the BS degree in Henan Institute of Science and Technology, Henan, China, in 2017. His research interests include speech signal processing and application, multimedia authentication techniques.

Yuan Zhang received the B.S. degree in Wuhan Institute of Technology, Hubei, China, in 2017. His research interests include speech signal processing and application, and multimedia authentication.

Qiu-yu Zhang Researcher/PhD supervisor, graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is vice dean of Gansu manufacturing information engineering research center, a CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, image understanding and recognition, multimedia communication technology.

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.