An Extended Method for Transmitting Secret Messages in Textual Documents Based on Paragraph Resizing

Benjamin Aziz[1][a], Estabraq Makiyah[1][b] and Aysha Bukhelli[2][c]
[1]*School of Creative and Digital Technologies, Buckinghamshire New University, High Wycombe, U.K.*
[2]*Office of the Prime Minister, Bahrain*

Keywords: Formal Methods, Information Hiding, Lexical Steganography, Text Steganography, Linguistic Steganography.

Abstract: This short paper presents an extended method for the embedding of secret messages in text documents based on the readjustment of paragraph sizes in a document. The new method improves on an existing method in literature proposed by the authors previously, by introducing the idea of choice functions, which allows for any two paragraphs in a document to be compared. This new method provides for greater flexibility when performing text steganography. The paper also defines a modified algorithm, based on the Diffie-Hellman protocol, for establishing an agreement between two communicating parties on the choice of paragraphs to compare prior to the commencement of the communication session. Finally, the paper demonstrates the applicability of the extended method by means of a few examples.

1 INTRODUCTION

Steganography, as an art of embedding secret content in cover media, has taken up in recent decades as an exciting and important field of research part of the wider field of data security and protection research, as a result of the plethora of digital content format and content generation technologies. And although the majority of the research in the area of steganography is concerned with rich media content, such as images (Cheddad et al., 2010), audio (Jayaram et al., 2011), video (Liu et al., 2019) and even virtual reality (Wilson, 2019), text steganography remains an active area of research encompassing several methods for the embedding of secret content.

We can classify all steganographic methods in one of three categories:

- *Alteration of Cover Media*. In this method, the cover medium is altered to embed the secret content. In image-based steganography, the resulting *stego*-object is expected to be visually indistinguishable from the original cover object and the resulting image will still look the same as the original cover to the normal human eye. Moreover, in the absence of a mechanism for determining the originality of the cover image (e.g. by digitally signing it), it is also impossible to determine which set of pixels represents the original cover and which other set represents the stego-object. Despite the fact that the majority of literature so far on textual steganography uses this method, alteration of text remains an insecure method as a result of the fact that such modifications can easily be detected with a suitable level of precise document comparison with the original. This is evident in the lack of discussion of the security of this method in literature on textual steganography.
- *Generation of New Media*. In this method, we generate a new cover altogether with characteristics that match the embedding of our secret message. For example, this would be similar to the capturing of an image using a digital camera, and where that image has already characteristics matching our secret message. Another example would be to generate (e.g. using generative AI methods) some new text excerpt, again that encodes the secret message within.
- *Search for a Suitable Media*. In this method, we search among the currently available media for a cover medium that matches the embedding of our secret message. In this case, we neither alter existing cover nor generate a new one, but simply match with an existing one, drawn

[a] https://orcid.org/0000-0001-5089-2025
[b] https://orcid.org/0000-0002-6432-8596
[c] https://orcid.org/0000-0001-7578-977X

Aziz, B., Makiyah, E. and Bukhelli, A.
An Extended Method for Transmitting Secret Messages in Textual Documents Based on Paragraph Resizing.
DOI: 10.5220/0012716600003767

In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*, pages 389-396
ISBN: 978-989-758-709-2; ISSN: 2184-7711

from a database of such media, that exhibits the right characteristics for our secret message. This *matching* cover becomes itself the stego-object for the communication.

In what follows, we assume that the attacker (the warden) has no access to the original text document, as our method is not immune to visual analysis.

In Section 2, we discuss a few works in literature related to our paper. In Section 3, we review the theoretical background underlying the method proposed by (Aziz et al., 2022), which forms the basis for our extended method. In Section 4, we introduce our new extended method for embedding secret content into textual documents. We also define in this section the extraction process reversing our embedding method. In Section 5, we demonstrate our extended method through a simple example. In Section 6, we give a sketch of how the new method can be implemented based on a previous mechanism for sharing secrets that was defined in (Aziz, 2021). Finally, in Section 7, we conclude the paper with future research directions.

2 RELATED WORK

Recently, there have been several attempts to design format-based textual steganographic methods for different languages like English, Persian and Arabic. Format-based steganography is when the physical features of text symbols are used to conceal a message. The features are altered in such a manner that the human eye cannot detect them (Baawi et al., 2018). For example, lines within the text are moved up and down to conceal bits of secret data. Similarly, words are moved left or right or up and down. In some cases, white spaces in-between words or in-between paragraphs or lines, are used to hide such data. In feature-based encoding, the physical features of the words themselves are altered in order to conceal information. This is reliant on actual symbols in the language being used. Naharuddin et al. (Naharuddin et al., 2018) suggested a method that maps the secret message's bits onto a cover text using the American Standard Code for Information Interchange (ASCII) characters, comprising punctuation, spaces and symbols. The secret text is initially embedded using a one-time pad and transformed into a stego-object. Then, each character is transformed into 7-bit binary numbers. The embedding procedure is carried out by mapping one bit of the secret text onto the first bit of the stego-object character comprising the same quantity of bits. Each bit position for the bit of the secret text is documented as a stego key, which is placed on the bit of the stego-object. The stego key functions as a key to extract the secret text embedded in the stego-object.

Maher (Maher, 1995), on the other hand, designed a text data-hiding programme called TEXTO. TEXTO was designed to transform PGP ASCII-armored data into English sentences. This method is convenient for exchanging binary data, especially embedded data. The secret data here are replaced by English words, meaning TEXTO works like a simple substitution cipher, which results in reducing suspicion over the produced encoded text. Chapman and Davida (Chapman and Davida, 1997) introduced a steganographic scheme that consists of two functions: NICETEXT and SCRAMBLE. NICETEXT transforms a secret message text into a text that looks like natural language, by parsing the cover text and extracting syntactical patterns, i.e Part-of-Speech (PoS) tags. SCRAMBLE does the opposite; it parses sole words from the generated text and recreates the encoded text by using codes from the dictionary table. Later, Chapman (Chapman et al., 2001) expanded this approach by using sentence models and large dictionaries of words classified by PoS tags. By using the "extensible contextual template" approach, combined with a synonym-based replacement strategy, the generated text seems more realistic than it is with the original NICETEXT function.

Attallah et al. (Atallah et al., 2001) proposed a watermarking scheme for natural language text by embedding small portions of the watermark bit strings in the syntactic structure of a number of selected sentences in the text. This scheme is suitable for extended meaning texts, such as reports, manuals and so on, which need integrity protection rather than secrecy protection as is the case with critical texts that governments and industries produce in abundance. Later, Attallah (Atallah et al., 2002) also presented a semantically-based scheme, which improves the information-hiding capacity of any texts through two techniques: first by modifying the granularity of meaning of individual sentences, and second, by dividing the number of sentences affected by the watermark, which makes it possible to watermark short texts too. On the other hand, Moerland proposed in (Moerland, 2003) a text steganography technique, which is based on using selected characters from words. For instance, the first letter of all the paragraphs can be used to conceal the secret message. By placing all selected characters together, the complete secret message can be extracted.

In (Bergmair, 2004), Bergmair presented the linguistic problem of word-sense ambiguity and demonstrated its relevance to current computer security applications in the context of Human Interactive Proofs

(HIPs). HIPs enable a machine to automatically determine whether the machine is interacting with another machine or with a human. In doing so, Bergmair used the linguistic anomaly, which states that a word can have different meanings dependent on the context it is used in. In the same line of semantic interpretation of text, Chand and Orgun (Chand and Orgun, 2006) developed a linguistically robust embedding application called LUNABEL, which converts a message into semantically mundane text. LUNABEL uses word replacement with substitution classes based on traditional word replacement features, as well as features like semantic criteria and frequency statistics. LUNABEL creates text, which preserves the syntactic structure and semantic context of the original cover text. In the same manner, Liang (Liang and Iranmanesh, 2016) proposed a method by adding five white-space characters to random positions in a line using a key to correlate to the characters required for embedding secret information. This method is advantageous as randomly spread white-spaces can encode a message differently using different keys. The white-spaces contained in the secret text regulate the embedding process.

When it comes to non-English text steganography, we find also plenty of literature, specially in relation to languages such as Persian and Arabic. For example, Shirali-Shaherza (Shirali-Shahreza and Shirali-Shahreza, 2006) created an application based on hiding binary values in Arabic or Persian scripts using a feature-coding method. This method depends on the points inherited in the Arabic or Persian alphabet. The points' location within the pointed letters hide information as follows: First, the hidden information is translated into a binary value. Then, the cover text is scanned, and whenever a pointed letter is detected, the location of the point may be affected if hidden binary value is one or zero. The location of the point is slightly shifted up if the hidden bit value is one. Otherwise, the location remains unchanged. In (Shirali-Shahreza, 2008), Shirali-Shahreza also proposed a text steganography technique based on different spellings of words in British and American English. Some words in both dialects have different spellings for the same word; such as 'colour' and 'color'. Furthermore, in (Shirali-Shahreza and Shirali-Shahreza, 2008), Shirali-Shahrezaand and Shirali-Shahreza evolved the previous technique to cater for the different terms for the same word in British and American English dialects, and substituting the text to hide secret data. For example, the term 'elevator' in American English is referred to as 'lift' in British English, and substituing one for the other facilitates the hiding of one bit.

Baawi et al. (Baawi et al., 2020) suggested a technique to enhance the embedding capacity for format-based text steganography using the font as well as other text characteristics for encoding secret information. This technique uses similar symbols for several codes, known as Set of High-Frequency Letters (SHFLs). The embedding process is based on replacing English letters with codes that share similar shapes. One pass encodes two bits, where 00, 01, 10 and 11 conceal glyph1, glyph2, glyph3 and glyph4, respectively (see Table 1 for an example of four letters, 'e', 't', 'a' and 'o'). The steganographic capacity of the table can be enhanced, and the technique is based on lower-case SHFL. This two-bit technique outperforms the standard text steganography because it enhances the embedding capacity of the stego-text.

Table 1: Selected letters in SHFL for the hiding process (Baawi et al., 2020).

Letters	**ASCII Code**		**Unicode**	
	S = 00	**S = 01**	**S = 10**	**S = 11**
e	0065	0023	0026	002A
t	0074	003C	003D	003E
a	0061	005B	005D	005E
o	006F	007B	007C	007D

Most of the new space insertion techniques are inspired from the Kashida technique developed by Taha et al. in (Taha et al., 2020), who suggested a format-based text steganographic technique intended for the Arabic language and based on the Kashida and Unicode text (including zero-width non-joiner (discrete), zero-width joiner, little space and zero-width space along with traditional spaces). The cover text can be used to conceal one bit of the secret text in each letter by transforming the letters using their position (i.e. end, middle or beginning of a word, or an isolated word). The shapes of the letters are corrected using a software that changes typographic sequences depending on letter positioning (i.e. isolated, end, middle or beginning). Kashida, typed as '_', represents a character in the Arabic langauge that extends a letter but does not change the word meaning.

3 THEORETICAL BACKGROUND

We give a recap here of the theory of the new textual embedding method that was presented in (Aziz et al., 2022) for completeness. In (Aziz et al., 2022), the authors presented a model of a textual document $S \in \mathcal{S}$ that assumed the document S consisted of a number of paragraphs such that $S = (P_1, \ldots, P_n)$, where $\mathcal{S}$ is the set of all possible such documents. $P, P', \ldots \in \mathcal{P}$ is the set of every possible sound paragraph written in English, hence $S \subseteq \mathcal{P}$, or in other words, a document is a finite sequence of such paragraphs. In addition to this, a paragraph P was assumed to consist of a finite number of (possibly repeating) characters, defined by the function:

$$ch\_of : \mathcal{P} \rightarrow \wp(\mathcal{C})$$

where $\mathcal{C}$ is the multi-set of all possible characters in English and $\wp(\mathcal{C})$ is the power-set generated from $\mathcal{C}$ and preserving the multiplicity of each character in the multi-set, as defined by Axiom V of (Blizard, 1988). Hence:

$$ch\_of(P) = \{|c : c \in P|\}$$

Where we assume that a paragraph has at least one sentence defined by a punctuation mark, that is a '!' or a '?' or a '.'. Additionally, the model of (Aziz et al., 2022) assumes the condition that no paragraphs have empty number of characters, i.e.:

$$\forall P \in \mathcal{P} : |ch\_of(P)| > 0$$

A similar condition applies to documents, where no documents are assumed to be empty, or in other words, $S = (\ )$ does not exist. Moreover, a stronger condition is required to hold if a document is to be used as a *cover* document, that is

$$\forall S \in \mathcal{S} : is\_cover(S) \Leftrightarrow |S| > 1$$

where $is\_cover : \mathcal{S} \rightarrow \mathbb{B}$ is a predicate that asserts whether a document can be used as a cover for embedding secret messages or not. This last condition assumes that, unless a document contains two or more paragraphs, it is neither suitable for the method of (Aziz et al., 2022) nor to our extended method proposed later.

The main mechanism, which was used in (Aziz et al., 2022) for embedding secret messages in text documents was the $R : \mathcal{P} \times \mathcal{P} \rightarrow \{0,1\}$ function, which essentially compares two paragraphs and returns a 0 or a 1 depending on the result of the comparison. It is possible to define R in any way one prefers, however, this was defined in (Aziz et al., 2022) as being a size comparison function on the number of characters in a paragraph:

$$R(P_l, P_r) = \begin{cases} 0 & \textit{if } |ch\_of(P_l)| \leq |ch\_of(P_r)| \\ 1 & \textit{otherwise} \end{cases}$$

which will be the same definition we will be using throughout the rest of this paper. Note that we refer to the parameters of R as the left paragraph (P_l) and the right paragraph (P_r), and the choice of instantiating these depends on the specific case that uses R (discussed in the next section).

4 THE EXTENDED EMBEDDING METHOD

Our extended embedding process builds on the same approach used in (Aziz et al., 2022) for embedding secret messages M in text documents. However, we introduce a different method by which M is built:

$$M = [R(c_{i_1}, c_{i_2}), \ldots, R(c_{i_{k-1}}, c_{i_k})] \quad \textit{where, } i_1, i_2, \ldots, i_{k-1}, i_k \in \mathbb{N}^+$$

which gives us a $(k-1)$-long secret message. This definition introduces an n-wise choice function, c_i:

$$c_i((p_1, \ldots, p_n)) = p_i$$

A choice function chooses the i^{th} element in a sequence. Every c_i is a partial function whenever $i < 1$ or $i > n$, for an n-long sequence.

In our embedding method, we shall call the sequence of pairs of such choice functions, ρ, such that:

$$\rho = ((c_{i_1}, c_{i_2}), \ldots, (c_{i_{k-1}}, c_{i_k}))$$

where there is a requirement that:

$$1 \leq i_1, i_2, \ldots, i_{k-1}, i_k \leq |S|$$

for the document S on which the choice functions are applied.

In order to extract a message, the message receiver will need to have agreed on the definition of R above with the sender of that message, beforehand. Such definition of R could in real terms vary, giving therefore rise to different variations of this method, each with its own definition of R. However, here we stick with one definition of R. With this in mind, the extraction logic can be defined as follows:

$$\mathbf{Y}\ \omega\ (P_1, \ldots, P_n)\ \rho\ [\,] = \omega\ (\mathbf{Y}\ \omega)\ (P_1, \ldots, P_n)\ \rho\ [\,]$$

where $\mathbf{Y}$ is Curry's fixed-point combinator as defined in (Curry and Feys, 1958, p.178), $(P_1, \ldots, P_n)$ is the text document received by the receiver, ρ is as sequence of pairs of choice functions as defined in the previous section, $[\,]$ is an empty list, which will be filled with the bits of the secret message during the extraction, and finally, ω is defined as follows:

$$\omega = \lambda f.\lambda s.\lambda r.\lambda \ell.\ \textit{if } r = [\,] \textit{ then } \ell \textit{ else } f\ s\ (r \backslash fst(r)) \\ (\ell : R(fst(fst(r))(s), snd(fst(r))(s)))$$

which is a λ expression (Church, 1932) that embeds the definition of R. Here, $fst : \mathcal{S} \rightharpoonup \mathcal{P}$ is a partial function that returns the first paragraph element in a sequence, $snd : \mathcal{S} \rightharpoonup \mathcal{P}$ is a partial function that returns the second paragraph element in a sequence and $\backslash : \mathcal{S} \times \mathcal{P} \rightharpoonup \mathcal{S}$ is a partial function that takes a sequence and a paragraph, and returns the sequence resulting from the removal of that paragraph from the input sequence. Finally, $(\ell : n) = \ell'$ is an operation that joins an element n to the tail of an existing list ℓ such that n becomes the last element of the new list ℓ'. In our case, $n = R(\mathit{fst}(\mathit{fst}(r))(s), \mathit{snd}(\mathit{fst}(r))(s))$, which informally, is the bit resulting from the application of the R function to two paragraphs, $\mathit{fst}(\mathit{fst}(r))(s)$ and $\mathit{snd}(\mathit{fst}(r))(s)$. The former is chosen based on the *first* choice function included in the first element of the received ρ (or r), and the latter is chosen based on the *second* choice function included in that element of ρ (or r). Both *fst* and *snd* are partial functions since they are not defined over anything other than pairs (basically, our pairs of choice functions in a ρ element). $\setminus$, on the other hand, is partial since the element being removed from a sequence may not be a member of that sequence.

5 EXAMPLE

As a simple example, let us consider the 5-paragraph excerpt in Figure 1 taken from Jules Verne's "Journey to the Centre of the Earth". In its current form, this excerpt naturally encodes the message $M = [1,0,1,1]$, given a sequential comparison of subsequent paragraphs, i.e. $\rho = ((c_1,c_2),(c_2,c_3),(c_3,c_4),(c_4,c_5))$.

5.1 Modes of Selection

In the most general case, there are no conditions on how ρ is selected. In fact, there are no conditions on the size of the transmitted message M. However, we only consider below interesting choices of ρ, where the size of the transmitted message is at least as large as the number of paragraphs in the cover text minus one. We do not consider cases of smaller messages, since these can be transmitted in text documents with fewer number of paragraphs, therefore, they do not represent any new cases.

5.1.1 $|M| = |S|-1$: Base Case

This case corresponds to the original embedding algorithm proposed in (Aziz et al., 2022). Here, the definition of ρ is restricted by the following format:

$$\rho = ((c_1,c_2),\ldots,(c_{n-1},c_n))$$

for a text document of n number of paragraphs, which represents the comparison of sequential paragraphs' sizes. In our example of Figure 1, if we wanted to embed the secret message $M_1 = [0,0,1,0]$, we will have to modify the excerpt such that $R(P_1,P_2) = 0$, $R(P_2,P_3) = 0$, $R(P_3,P_4) = 1$ and $R(P_4,P_5) = 0$. An example of a modified excerpt embedding this message is that of Figure 2, which we refer to as S_{Fig2}.

When extracting this message, we simply apply the following fixed-point calculation:

Y $\omega\, S_{Fig2}\ ((c_1,c_2),(c_2,c_3),(c_3,c_4),(c_4,c_5))\ [\,] = [0,0,1,0]$

5.1.2 $|M| = |S|-1$: Random Selection Case

In this more generic case, the choice of ρ is only bounded by one condition: that the size of the embedded message must be one fewer than the number of paragraphs in the cover text document (i.e. similar to the size of the message in the base case). Otherwise, the definition of ρ may allow for any two paragraphs to be compared depending on the pre-communication agreement made by the two communicating entities. We will discuss in slightly more detail in Section 6 how such pre-communication agreement can be established securely. For now, we assume that both entities know which two paragraphs need to be compared, for each of the bits in the secret message.

As an example, let us assume that $\rho = ((c_3,c_1),(c_5,c_2),(c_4,c_5),(c_2,c_3))$ and using the modified excerpt of Figure 2, this will allow us to embed the following message:

$$M = [1,0,0,0]$$

In this case, in order to extract the secret message, we apply the following fixed-point calculation:

Y $\omega\, S_{Fig2}\ ((c_3,c_1),(c_5,c_2),(c_4,c_5),(c_2,c_3))\ [\,] = [1,0,0,0]$

5.1.3 $|M| > |S|-1$: Finite but Unbounded Case

In this last case, the number of message bits is finite but unbounded, meaning it can be any number of bits, as long as the two communicating entities have a pre-agreed length of the secret message. For example, if we assume that $|M| = 8$, then we may agree that the following definition of ρ consisting of eight pairs of choice functions, would be used as our embedding method:

$$\rho = ((c_1,c_3),(c_4,c_5),(c_5,c_1),(c_2,c_4),(c_1,c_2),\\ (c_5,c_3),(c_2,c_3),(c_1,c_4))$$

Consequently, to embed the following message:

$$M = [1,0,0,1,0,1,1,1]$$

we will need to modify the original excerpt of Figure 1 to a new excerpt that matches this message and

```
P1: Really, what was the good of making such a fuss about an old quarto volume,
the back and sides of which seemed bound in coarse calf|a yellowish old book,
with a faded tassel dangling from it?

P2: However, the professor’s vocabulary of adjectives was not yet exhausted.

P3: "Look!" he said, asking himself questions, and answering them in the same breath;
"is it handsome enough? Yes; it is first-rate. And what binding! Does it open easily?
Yes, it lies open at any page, no matter where. And does it close well? Yes;
for binding and leaves seem in one completely. Not a single breakage in this back after 700 years
of existence! Ah! this is binding that Bozerian, Closs, and Purgold might have been proud of!"

P4: All the while he was speaking, my uncle kept opening and shutting the old book.
I could not do less than ask him about the contents, though I did not feel the least interest
in the subject.

P5: "And what is the title of this wonderful volume?" I asked.
```

Figure 1: A five-paragraph excerpt from Jules Verne’s ”Journey to the Centre of the Earth”.

```
P1: Really, what was the good of making such a fuss about an old quarto volume,
the back and sides of which seemed bound in coarse calf|a yellowish old book,
with a faded tassel dangling from it?

P2: However, the professor’s vocabulary of adjectives was not yet exhausted.
"Look!" he said, asking himself questions, and answering them in the same breath;
"is it handsome enough? Yes; it is first-rate.

P3: And what binding! Does it open easily?
Yes, it lies open at any page, no matter where. And does it close well? Yes; for binding and
leaves seem in one completely. Not a single breakage in this back after 700 years of existence!
Ah! this is binding that Bozerian, Closs, and Purgold might have been proud of!"

P4: All the while he was speaking, my uncle kept opening and shutting the old book.

P5: I could not do less than ask him about the contents, though I did not feel the least interest
in the subject. "And what is the title of this wonderful volume?" I asked.
```

Figure 2: The modified excerpt, S_{Fig2}.

```
P1: Really, what was the good of making such a fuss about an old quarto volume,
the back and sides of which seemed bound in coarse calf|a yellowish old book,
with a faded tassel dangling from it? However, the professor’s vocabulary of adjectives
was not yet exhausted.

P2: "Look!" he said, asking himself questions, and answering them in the same breath;
"is it handsome enough? Yes; it is first-rate. And what binding! Does it open easily?
Yes, it lies open at any page, no matter where. And does it close well? Yes;
for binding and leaves seem in one completely. Not a single breakage in this back after 700
years of existence!

P3:  Ah! this is binding that Bozerian, Closs, and Purgold might have been proud of!"

P4: All the while he was speaking, my uncle kept opening and shutting the old book.

P5: I could not do less than ask him about the contents, though I did not feel the least interest
in the subject. "And what is the title of this wonderful volume?" I asked.
```

Figure 3: The modified excerpt, S_{Fig3}.

the choice of ρ above. An example of such (suitably) modified excerpt could be that shown in Figure 3. We call this second modified excerpt, S_{Fig3}.

Now, in order to extract the secret message for this case, we again apply our fixed-point calculation:

$$\mathbf{Y}\ \omega\ S_{Fig3}\ ((c_1,c_3),(c_4,c_5),(c_5,c_1),(c_2,c_4),(c_1,c_2),(c_5,c_3),(c_2,c_3),(c_1,c_4))\ [\,] = [1,0,0,1,0,1,1,1]$$

6 IMPLEMENTATION USING SHARED SECRET KEYS

A new and important usage of Diffie-Hellman shared secret keys (Diffie and Hellman, 1976) was demonstrated in (Aziz, 2021) as a means of agreement between two entities on the semantic interpretation of the exchanged messages. This approach opens the possibility, in our case, to achieve a sort of separation of concerns, between defining ρ and sharing a secret key that would allow both communicating entities to agree on ρ, before the commencement of the communication session.

Assuming ι is an index function defined as:

$$\iota : \mathbb{N} \rightarrow ((\Xi \times \Xi) \times \ldots \times (\Xi \times \Xi))$$

where $\Xi = \{c, c', \ldots\}$ is the set of all possible choice functions, then we can simply retrieve a specific definition of ρ by applying:

$$\iota(K_{AB}) = \rho$$

K_{AB} is the Diffie-Hellman key pre-agreed between A and B. A generic definition of ι was given in (Aziz, 2021) as an indexing function that can be used to retrieve a semantic domain (ρ in our case).

We can also modify the algorithm given in (Aziz, 2021) for agreeing a semantic domain by introducing a new albeit slightly modified algorithm as shown in Algorithm 1 below. The net outcome of calling the ALICE procedure will be that $\rho_B = \rho_A$.

As a result, both ALICE and BOB will end up agreeing on the same ρ value to use in the modification of textual documents in the subsequent communication session(s).

```
procedure ALICE
    Choose some $a \in \mathbb{N}$
    Compute $K = \text{BOB}(c_A)^a \ mod \ p$
    ▷ where $c_A = g^a \ mod \ p$
    Set in internal state $\rho_B =: \iota(K)$
    return
end procedure

procedure BOB($c_A$)
    Choose some $\rho \in (\Xi \times \Xi) \times \ldots \times (\Xi \times \Xi)$
    Choose a specific $b \in \mathbb{N}$ such that
    $\iota({c_A}^b \ mod \ p) = \rho$
    Set in internal state that $\rho_A = \rho$
    return $c_B$
    ▷ where $c_B = g^b \ mod \ p$
end procedure
```

Algorithm 1: A modified DH-based ρ Agreement Algorithm [p, g, ι are global parameters, ρ_A, ρ_B are local state variables].

7 CONCLUSION

This paper has presented an overview of recent procedures for hiding secret messages in text documents through steganography. Both format-based techniques modifying text attributes and linguistic methods producing credible language covers were reviewed. However, limitations were recognised with the capacity and security of the existing approaches.

To address these concerns, a novel extended embedding method is proposed based on encoding messages by adjusting the size of comparative paragraph. The choice of the paragraphs being compared is allowed to be flexible, unlike the original sequential comparison method proposed in (Aziz et al., 2022). The theory underlying our approach is formally expressed using the λ-calculus, and method for agreeing the choice of paragraph functions is also proposed by an algorithm based on the Diffie-Hellman protocol (Diffie and Hellman, 1976). We hypthosise that this this new technique will provide increased hiding capacity and resilience against statistical attacks by exploiting structural properties of the textual document.

Future work will be focused on exploring a number of research directions: First, we intend to explore substitute definitions of the paragraph comparison function R to define optimal configurations of the cover text. Second, it would be interesting to define a search algorithm that searches for text documents with the most suitable features, such that both the hiding capacity as well as the security of the stego-object are maximised. Extending the approach to other media formats, such as images, video and audio would also be possible, by adopting a block comparison approach, where parts of the image or video file are compared against other parts as a means of hiding a bit of information. Finally, we propose that recent generative artificial intelligence methods may be useful in generating natural language cover documents with suitable characteristics, or even generating the actual stego-object documents.

REFERENCES

Atallah, M. J., Raskin, V., Crogan, M., Hempelmann, C., Kerschbaum, F., Mohamed, D., and Naik, S. (2001). Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Proceedings of the 4th International Workshop on Information Hiding*, IHW '01, page 185–199, Berlin, Heidelberg. Springer-Verlag.

Atallah, M. J., Raskin, V., Hempelmann, C. F., Karahan, M., Sion, R., Topkara, U., and Triezenberg, K. E. (2002). Natural language watermarking and tamperproofing. In *International workshop on information hiding*, pages 196–212. Springer.

Aziz, B. (2021). A note on the problem of semantic interpretation agreement in steganographic communications. *Journal of Internet Services and Information Security*, 11(3):47–57.

Aziz, B., Bukhelli, A., Khusainov, R., and Mohasseb, A. (2022). A novel method for embedding and extracting secret messages in textual documents based on paragraph resizing. In *Proceedings of the 19th International Conference on Security and Cryptography - Volume 1: SECRYPT,*, pages 714–719. INSTICC, SciTePress.

Baawi, S. S., Mokhtar, M. R., and Sulaiman, R. (2018). A comparative study on the advancement of text steganography techniques in digital media. *ARPN Journal of Engineering and Applied Sciences*, 13(5):1854–1863.

Baawi, S. S., Nasrawi, D. A., and Abdulameer, L. T. (2020). Improvement of "text steganography based on unicode of characters in multilingual" by custom font with special properties. In *IOP Conference Series: Materials Science and Engineering*, volume 870.

Bergmair, R. (2004). Towards linguistic steganography: A systematic investigation of approaches, systems, and issues. *Final year thesis, B. Sc.(Hons.) in Computer Studies, The University of Derby*.

Blizard, W. D. (1988). Multiset theory. *Notre Dame Journal of Formal Logic*, 30(1):36 – 66.

Chand, V. and Orgun, C. (2006). Exploiting linguistic features in lexical steganography: Design and proof-of-concept implementation. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 6, pages 126b–126b.

Chapman, M. and Davida, G. (1997). Hiding the hidden: A software system for concealing ciphertext as innocuous text. In *International Conference on Information and Communications Security*, pages 335–345. Springer.

Chapman, M., Davida, G. I., and Rennhard, M. (2001). A practical and effective approach to large-scale automated linguistic steganography. In *International Conference on Information Security*, pages 156–165. Springer.

Cheddad, A., Condell, J., Curran, K., and Mc Kevitt, P. (2010). Digital image steganography: Survey and analysis of current methods. *Signal processing*, 90(3):727–752.

Church, A. (1932). A set of postulates for the foundation of logic. *Annals of Mathematics*, 33(2):346–366.

Curry, H. B. and Feys, R. (1958). *Combinatory Logic*. Number v. 1 in Combinatory Logic. North-Holland Publishing Company.

Diffie, W. and Hellman, M. E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory*, IT-22(6):644–654.

Jayaram, P., Ranganatha, H., and Anupama, H. (2011). Information hiding using audio steganography–a survey. *The International Journal of Multimedia & Its Applications (IJMA) Vol*, 3:86–96.

Liang, O. W. and Iranmanesh, V. (2016). Information hiding using whitespace technique in Microsoft word. In *Proceedings of the 2016 International Conference on Virtual Systems and Multimedia, VSMM 2016*. Institute of Electrical and Electronics Engineers Inc.

Liu, Y., Liu, S., Wang, Y., Zhao, H., and Liu, S. (2019). Video steganography: A review. *Neurocomputing*, 335:238–250.

Maher, K. (1995). Texto. *URL: ftp://ftp.funet.fi/pub/crypt/steganography/texto. tar. gz.*

Moerland, T. (2003). Steganography and steganalysis. *Leiden Institute of Advanced Computing Science*.

Naharuddin, A., Wibawa, A. D., and Sumpeno, S. (2018). A high capacity and imperceptible text steganography using binary digit mapping on ascii characters. In *Proceeding - 2018 International Seminar on Intelligent Technology and Its Application, ISITIA 2018*, pages 287–292. Institute of Electrical and Electronics Engineers Inc.

Shirali-Shahreza, M. (2008). Text steganography by changing words spelling. In *2008 10th International Conference on Advanced Communication Technology*, volume 3, pages 1912–1913. IEEE.

Shirali-Shahreza, M. H. and Shirali-Shahreza, M. (2006). A new approach to persian/arabic text steganography. In *5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering,Software Architecture and Reuse (ICIS-COMSAR'06)*, volume 2006, pages 310–315.

Shirali-Shahreza, M. H. and Shirali-Shahreza, M. (2008). A new synonym text steganography. In *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 1524–1526. IEEE.

Taha, A., Hammad, A. S., and Selim, M. M. (2020). A high capacity algorithm for information hiding in arabic text. *J. King Saud Univ. Comput. Inf. Sci.*, 32:658–665.

Wilson, S. (2019). Unreal steganography: Using a vr application as a steganography carrier. https://www.forensicfocus.com/stable/wp-content/uploads/2019/07/dissertation.pdf.