Genre-based information hiding

Russell Ogilvie and George R S Weir[1]

Department of Computer & Information Sciences
University of Strathclyde
Glasgow G1 1XH
rogilvie@cis.strath.ac.uk, george.weir@cis.strath.ac.uk[1]

Abstract. While data encryption is an effective means of keeping data private it does not conceal the presence of 'hidden' information, rather it serves as an indicator that such data is present. Concealing information and hiding the fact that information is hidden are both desirable traits of a confidential data exchange, especially if that exchange takes place across a public network such as the Internet. In the present paper, we describe an approach to textual steganography in which data is not only hidden, in virtue of its encoding, but the presence of hidden data is also concealed, through use of human-readable carrier texts. Information transmitted in this fashion remains confidential and its confidential nature is also concealed. The approach detailed addresses several shortcomings in previous work in this area. Specifically, we achieve a high rate of accuracy in message decoding and also produce carrier texts which are both coherent and plausible as human-readable plain text messages. These desirable features of textual steganography are accomplished through a system of sentence mapping and a genre-based approach to carrier text selection that produces contextually related content in the carrier messages.

Keywords: Textual steganography, genre, data hiding.

1 Introduction

The major attraction of steganography is not the goal of concealing data, since this can be achieved very effectively with current encryption techniques, but, more specifically, the goal of concealing the presence of concealed data [1]. Although such information hiding is often addressed through use of media data files as the 'carrier' for concealed data [2], there is an alternative but underdeveloped approach that seeks to conceal data within plain text. Such textual steganography is appealing but problematic when compared to transmission via image or audio carrier files. The major issues facing textual steganography are twofold:

(i) Provide an encoding mechanism that is sufficiently flexible to conceal any required message;
(ii) Create carrier texts that are plausible as 'ordinary' texts;

[1] Corresponding author.

The first criterion relates to the successful encoding and decoding of messages while the second relates to the successful concealment of hidden data. In the following, we describe several approaches that have been used previously as a basis for textual steganography before proposing an alternative technique that we have developed, based upon sentence mapping and genre-specific carrier texts.

2 Strategies for Textual Steganography

Concealing data in text requires that aspects of the text serve as markers or codes for the concealed data. This parallels the use of altered pixel brightness or colour values in images as the 'codes' for concealed data. In the textual realm we may distinguish two varieties or strategies for steganography: presentational and linguistic. The presentational approach is simpler and uses variations in features such as inter-word or inter-line spacing to represent the hidden data. This is similar in concept to image steganography since it relies upon visual aspects in the carrier presentation that will be 'inconsequential' to the human observer but discernible through appropriate software analysis. Linguistic approaches tend to be more ambitious in their selection of coding strategy and rely upon changes to the textual content as a basis for conveying the hidden message.

2.1 Presentational strategies

Presentational strategies for textual steganography include the use of features such as inter-word spacing; tabs and spaces; and line shifting. These approaches are outlined below.

Inter-word spacing.

Delina [3] describes the technique of generating a cover text depending on the length of the secret message and altering the number of spaces between the words in the carrier text. One space between a word and the next word indicated a 0 and two spaces indicated a 1. The advantage of this method is that additional spaces in documents are fairly common especially if text justification or text centring is used. Thereby, this feature is unlikely to rouse suspicion. When a small font is used for document presentation a human cannot easily tell the difference between one space and two spaces. So it is likely that the presence of a secret message will not be discerned. However, a potential down side to this approach is that to convey a single character will generally require eight bits, meaning that eight inter-word spaces would be required to represent one character. This constraint means that only short hidden messages can be transmitted, unless the carrier text is very long.

A similar approach, termed 'word shifting', is described by Kim et. al. [4]. Arguably, this form of shifting may be less easy for a human to detect because

horizontal shifting of words is common in newspapers, magazines and other documents to fill up lines of text and achieve text justification [5].

Tabs and spaces

This variation on the use of inter-word spacing was developed by Mansor et. al [6]. Their program (SNOW) takes the secret message and a carrier text as input and uses an algorithm to add extra spaces and tabs to the end of some or all of the lines in the carrier text. A tab is always added first to indicate the start of the concealed message and then sequences of spaces and tabs are added to make up the secret message. The extra spaces and tabs cannot normally be seen by readers so will not arouse suspicion to a human who has a glance at it as the original carrier text is preserved. One drawback of this technique is that extra spaces and tabs may be detectable if the carrier text was viewed in a text editor. This may arouse suspicion that confidential data is hidden within.

Line shifting

Textual steganography by means of 'line shifting' conceals data through small vertical adjustments (e.g., 1/300th of an inch) to the lines of text in the carrier message [7]. The receiving system detects these changes in vertical alignment and reconstitutes the hidden message accordingly. As with other presentational approaches, there is some risk that a human reader may notice slight variations in the alignment of the lines of text in a carrier message [8]. Furthermore, while this technique could be used in formatted electronic documents, it is best suited to printed texts [9]. Recovering the hidden information may be problematic in cases where the carrier text has been edited or retyped.

2.2 Linguistic strategies

While presentational approaches can successfully conceal data and may do so in ways that are not noticed by the average human reader, they are primarily limited by their dependence upon preservation of document format and layout. This makes them better suited to document-based data hiding than 'live' transmission systems. In contrast, linguistic approaches to textual steganography adopt strategies that depend upon changes to the meaning of the carrier message. Unlike presentational approaches, linguistic techniques require a mechanism that changes words or word combinations as a basis for concealing the data. Linguistic strategies for textual steganography include the use of features such as word spelling, synonyms and phrase structures. These approaches are outlined below.

Changing Word Spelling

In this approach, data is concealed by changing the spelling of specific words in a piece of text, for example, British and US spelling [10]. A database (the resource) is created which holds lists of words which have different spellings in the British and US and the encoding program searches through the carrier text to find words which have different British and US spellings. The system changes words in the carrier text

so that a US word denotes a 0 and a UK word denotes a 1 so that data can be hidden in the carrier text. At the receiver end the same database of words is used to search through the document in order to build up a sequence of 0 and 1s and thereby extract the hidden message. A potential downside to this approach is its sparsity of encoding. Only a small amount of data can be hidden in the carrier text because a whole word indicates only a single binary digit. Despite this limitation, this method would be very difficult to detect unless a human reader notices the mixture of American spelling and British spelling.

Synonym replacement

An attempt to use synonym replacement as the mechanism for concealing data in carrier texts was developed by Morran & Weir [11]. This required the use of part-of-speech tagging on the secret text and the carrier text in order to identify suitable terms for synonym replacement, with the aim of maintaining the meaning and sense of the carrier text. Thereby, the private message may be concealed in the cover text with a key used to identify which words have been changed and, therefore, which words conceal parts of the hidden message. In order to eliminate the risk of significant changes in meaning, this approach also adopted word-sense disambiguation. Despite the sophistication of this approach, the resultant carrier texts were not always grammatical and meaningful.

Phrase structures

Chapman et. al. [12] used part-of-speech analysis in order to identify specific phrasal forms in the carrier texts that could be replaced as a means of concealing data. This was implemented as a system called NICETEXT but, as with Morran and Weir's synonym replacement approach, NICETEXT was impaired by the limitations of part-of-speech tagging and suffered from occasional grammatical anomalies in the carrier texts [1].

3 Genre-Based Textual Steganography

With a view to addressing the main requirements of textual steganography, viz., providing a flexible encoding mechanism and creating carrier texts that are plausible as 'ordinary' texts, we combined two aspects in our prototype system. The first aspect is the use of a word to sentence mapping as a means of converting secret text to carrier text. This employed a database with a large set of words and an associated database of sentences. Words in the secret text are mapped algorithmically to specific sentences and the carrier text is a compilation of the successive sentences. This approach ensures that each sentence in the carrier text is grammatically well formed. The second aspect in our steganography system addresses the need for coherence across sentences. This is addressed through use of datasets of genre-specific sentences.

In principle, there might be any number of discrete genre-specific sentence datasets. Our prototype system presently operates with two sentence databases. The

first of these is composed of sentences drawn from a corpus of computer science literature while the second is composed of sentences drawn from the Brown corpus. Before creating a carrier text from a specific secret text, the user chooses one or other genre database. Each word of the secret text is then mapped to a sentence from the selected genre dataset. The choice of genre is also conveyed as a concealed data item in the carrier text, thereby ensuring that the decoding process operates with the appropriate sentence dataset.

In the process of formulating the carrier text, each word of the secret text is located in the word dataset. For any word not already present in the dataset, the system adopts the 'new word' procedure. In this case, the new word is added to the sender-side database and then broken down into individual characters. The word dataset also contains a set of individual characters each of which can map to an item in the sentence dataset. This permits newly encountered words to be concealed and included within the carrier text. At the receiving side, new words are identified by an in-text code, decoded from the sentences in the carrier text and automatically added to the receiver-side database.

3.1 Example codings

In the following examples, we show the secret message followed by the carrier texts from the Computer Science (CS) genre and then the Brown Corpus (Brown) genre.

Secret text 1: *I like watching films in the evening.*

CS Carrier text: *The stat_formula() interface allows derived instrumentation to be declared, creating a metric that is a function of other counters. SimpleScalar's software architecture has been improved and modularized, making the code easier to understand and modify. For tracers in real-world environments, however, especially on backbone networks where network traffic is very high, more memory will likely be needed. In GOMS terms, there must exist a method with operators to execute the command. An open distribution model has potential drawbacks, however. Direct user support creates by far the largest demand. The growth in the number of homes with always-on, high-speed broadband Internet connections has enabled hackers to spread bots widely and quickly, according to David Perry, global director of education for antivirus-software vendor Trend Micro.*

Brown carrier text: It was marked by controversy, anonymous midnight phone calls and veiled threats of violence. The United States, State Department officials explain, now is mainly interested in setting up an international inspection system which will prevent Laos from being used as a base for Communist attacks on neighboring Thailand and South Viet Nam. Everybody has a different idea on the ethics and morals of driving away neighbors, when and if. This resulted in revisions of the King James Bible in 1881-85 as the English Revised Version and in 1901 as the American Standard Version. Kansas, regarded as tops in the nation in its treatment of the mentally ill, spends $9 per day per patient. And after several correspondents went

into Pathet Lao territory and exposed the huge build-up, administration spokesmen acclaimed them for performing a great service and laid the matter before the Southeast Asia Treaty Organization. Of the millions who have served time in concentration camps in Siberia as political prisoners of the Soviet state, few emerge in the West to tell about it. Some 30 spectators remained in the court during the day and were on hand to hear the verdict read.

Secret text 2: *The queen has gone to roost.*

CS carrier text: *One is that all DR(i, k)s are distributed evenly in a given range. Both evaluations followed the same layout and question structuring. In contrast to other bots, Bobax's primary purpose is to send spam. Problems in the representation of meaning in information retrieval. Naturally, this reflects on the accuracy of the relevance decision. Expected Output: +word +word +word Actual Output: As expected Pass/Fail: Pass. Most web pages you find today are written in HTML (HyperText Markup Language).*

Brown carrier text: *Research projects as soon as possible on the causes and prevention of dependency and illegitimacy. Mr and Mrs Stevens and the bride's other uncles and aunts, the Rush C Butlers, the Homer E Robertsons, and the David Q Porters, will give the bridal dinner tonight in the Stevenses' home. We (the Chicago Association of Commerce and Industry) expect to establish closer relations with nations and their cultural activities, and it will be easy as a member of the fair staff to bring in acts, explains Mrs Geraghty. Led fight on teamsters Gladden has been an outspoken critic of the present city administration and led his union's battle against the teamsters, which began organizing city firemen in 1959. The city is not adequately compensated for the services covered by the fees, he said. Families go out to the edge of the terraces to sit on carpets around a samovar. Those three other great activities of the Persians, the bath, the teahouse, and the zur khaneh (the latter a kind of club in which a leader and a group of men in an octagonal pit move through a rite of calisthenics, dance, chanted poetry, and music), do not take place in buildings to which entrance tickets are sold, but some of them occupy splendid examples of Persian domestic architecture : long, domed, chalk-white rooms with daises of turquoise tile, their end walls cut through to the orchards and the sky by open arches.*

3.2 Issues

While every tested example of encoded text has been successfully decoded, there remain some issues with the current prototype that we aim to address in future work. In the first place, the word-to-sentence mapping results in considerable expansion from the secret message size to the carrier text size. This is an inherent consequence of the mapping of individual words to sentences but can be reduced through more careful selection of sentences for the genre datasets. Specifically, we aim to prioritize shorter sentences in order to reduce the data inflation effect.

A second issue concerns the coherence of the resultant carrier texts. Although there is a high degree of coherence in the resultant sentences, there are indications that

'noise' in the sentence datasets can detrimentally impact upon the carrier text by introducing partial sentences (e.g., computer science article titles). This signifies a need to carefully filter the content of the sentence datasets in order to ensure their sentential integrity.

4 Conclusion

The aim of this work was to establish an effective means of concealing and conveying hidden messages in plain text such that the presence of such messages would not be discernible in the carrier texts. In evaluating the encoding and decoding process, 40 different secret texts of differing lengths and compositions were tested with the system. The results showed that 100% of the secret texts fed in to the system were recovered successfully. While such testing cannot guarantee that all possible carrier text will be decoded correctly, it lends plausibility to the system's effectiveness and shows that it is able to cope with a wide range of different secret texts.

As indicated above, the plausibility of carrier texts produced by the current prototype is high but occasionally leaves room for improvement. Overall, we are led to conclude that this approach achieves the two major issues facing textual steganography:

(i) Provide an encoding mechanism that is sufficiently flexible to conceal any required message;

(ii) Create carrier texts that are plausible as 'ordinary' texts.

The system on the whole does produce plausible carrier texts containing sentences which belong to the same genre and are grammatically and syntactically correct.

References

1. Weir, G.R.S. and Morran, M.: Hiding the Hidden Message: Approaches to Textual Steganography. International Journal of Electronic Security and Digital Forensics. 3 (3),.223--233 (2010).
2. Marwaha, P.: Visual Cryptographic Steganography in Images Proceedings of the 2010 Second International Conference on Computing, Communication and Networking Technologies (ICCCNT). 1--6 (2010)
3. Delina, B.: "Information Hiding: A New Approach in Text Steganography" Proceedings of the International Conference on Applied Computer and Applied Computational Science, World Scientific and Engineering Academy and Society (WSEAS), 689--695 (2008)
4. Kim, Y Moon, K & Oh, I.: A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics" Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03) 775–-779 (2003)

5. Shirali-Shahreza, M.: A New Synonym Text Steganography. Proceedings of the 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing 1524--1526 (2008)
6. Mansor, S Din, R & Samsudin, A.: Analysis of Natural Language Steganography. International Journal of Computer Science and Security (IJCSS), Vol. 3, Issue 2, CSC Journals, Kuala Lumpur, Malaysia, 113-–125 (2010)
7. Alattar, A & Alattar, O.: Watermarking Electronic Text Documents Containing Justified Paragraphs and Irregular Line Spacing. Proceedings of SPIE, Vol. 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, June. 685—695 (2004)
8. Whitiak, D.: The Art of Steganography. SANS Institute, as part of GIAC Practical Repository, GSEC Practical (v.1.4b) (2003)
9. Shirali-Shahreza, M.: Text Steganography in SMS. Proceedings of the 2007 International Conference on Convergence Information Technology, November 21-23, 2260--2265 (2008)
10. Shirali-Shahreza, M.: Text Steganography by Changing Words Spelling. 10th International Conference on Advanced Communication Technology (ICACT 2008), 1912—1913 (2008)
11. Morran, M & Weir, G.: An Approach to Textual Steganography. Global Security, Safety, and Sustainability: Communications in Computer and Information Science, 92, 48—54 (2010)
12. Chapman, M., Davida, G.: Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text. In: Han, Y., Quing, S. (eds.) ICICS 1997. LNCS, vol. 1334, 333-–345. Springer, Heidelberg (1997)