CHAPTER 1: THE PROBLEM

This chapter discuss about the underlying background of the research, followed with overview of several text steganography methods prior that were known before Noiseless Steganography paradigm and Noiseless Steganography itself.

1.1 Rationale

In the era of communication, people need to share information to intended recipients. On the other hand, confidential information is vulnerable to eavesdropping by malicious parties. Therefore, efforts to secure information become necessary.

Protecting the information can be done using steganography or cryptography. Steganography is the practice of concealing a file, message, image, or video within another file, message, image, or video such that its presence cannot be detected, except by its intended recipients. Using steganography, the intended secret message does not attract attention to itself as an object of scrutiny. Likewise, Cryptography scrambles the message to conceal the information in contains [1]. Steganography is used when it is desirable to hide the message, instead of scramble it [1]. When it is necessary, both techniques can be combined to add multiple layers of security [2].

However, there are some cases where steganography has the advantage over cryptography. There are some occasions where cryptography cannot be used due to policies of companies or governments that either limit the strength of cryptosystems or completely ban them [1].

There are several steganography approaches based on the cover media, such as video [3], image [4] and text [5], [6], [7]. Among these steganography techniques, Text steganography is considered a very challenging task due to lack of redundancy in text as compared to image or audio [8].

Text steganography can be classified into three basic categories [9]: format-based, random and statistical generation, and linguistic method.

Format-based method embeds secret message by modifying physical text formatting of an existing text. This includes insertion of spaces, text sizing, or misspellings throughout the text [9]. While this method can pass human inspections, it cannot pass computer system, since computer system will detect the difference in the formatting. Some word processing programs can even correct the insertion of spaces between sentences. The SNOW (Steganographic Nature Of Whitespace) program introduced by Matthew Kwan

[9] is the example of the implementation of format-based method which concealing secret messages by appending tabs and spaces on the end of lines. This program also incorporates encryption to prevent an attacker to extract the secret message without knowing the password.

Random and statistical generation method embeds secret message by generating text-cover based on statistical properties of characters sequences, word sequences. This method can also use word-length and letter and letter frequency to create text that appears to be statistically normal, but actually does not have any lexical value. Random and statistical generation can be implemented using mimic functions [9].

Linguistic method considers the linguistic properties of generated and modified text. One of the approaches of linguistic method is the synonym based method, as implemented in NICETEXT and SCRAMBLE [9]. NICETEXT transforms secret message by choosing corresponding plaintext in a type of category which is specified by the writing style sources. This writing style sources can be according to examples or by Context-Free Grammars. The reverse process, or SCRAMBLE ignores the style, and relies on a simple substitution mechanism to recover the secret message. On the contrary of format based method which was discussed earlier, linguistic method is more vulnerable to human inspection rather than computer examinations.

In all of the three classifications of text steganography above, noise is introduced in different ways. In format-base method, noise is introduced in the alteration of existing text, which can raise suspicion and detection regardless of whether or not the secret message is revealed. In random and statistical generation, noise is introduced by meaningless and semantically incoherent cover-text. While in Linguistic method, noise is introduced by incorrect syntax, grammar or the use of synonym that does not make sense. Hence, those steganography methods can be considered as “noisy”, as they either generate noise or conceal message in noise the involvement of noise makes the steganography method vulnerable to attacks [5].

In 2009, Desoky introduced Noiseless Steganography or Nostega paradigm [5], [6], [7]. Nostega describes a paradigm for designing steganography system, which does not introduce noise to its cover, nor exploit noise as stego-carrier. There are several Nostega-based methodologies, such as Listega or List-based steganography, Graphstega or Graph steganography, Chestega or Chess steganography, Edustega or Education-centric steganography, etc.

Listega camouflages secret message by manipulating a noiseless list of legitimate items. Listega establishes a covert channel among communicating parties to avoid suspicions in transmission of generated covers. This is achieved by employing justifiable reasons based on the common practice of using textual lists of items. Listega's simple but powerful method, with lists for the cover of Listega can be found in any companies, makes the author chose Listega as the previous method to be enhanced in this research.

1.2 Theoretical Framework

Listega takes advantage of textual list to camouflage data by exploiting textual lists of itemized data, e. g. book titles, CD titles, computer parts, etc. The scenario below illustrates the implementation of Listega:

Bob and Alice are on a spy mission. Both of them run an online business to buy and sell items such as book, CD's and computer parts. Before they went on a mission, they agreed to set rules for communication covertly by concealing secret messages in the list of items by manipulating a list of items to embed data. Since they have online business, their exchanges of information in the form of list of items will not look suspicious. Furthermore, Alice is not the only recipient of Bob's list, and vice versa. Other customers can also receive their list.

In above illustration, prior to exchanging secret messages, Bob and Alice have determined the domain to establish the covert communication channel, which was an online business. They then can embed secret message into the items of their online store. They generate legitimate list of the items as the list-cover, and make a communication protocol which regulates how a send and a recipient would communication covertly, including the decoder scheme to decode the secret message.

To enhance the capability of Listega, this research proposes the two enhancements, which are the enhancement of encoding scheme and the enhancement by making it more adaptive to the various list-covers in the real world implementations.

1.3 Conceptual Framework/Paradigm

As mentioned in previous section, Listega can be enhanced by elaborating the encoding scheme and make it more adaptive to various covers. To make the enhancements, this research observes the embedding capacity which is defined as the characters of list-cover needed to convey the characters of secret message, measured in percentage.

Another variable observed in this research is the successful embedding ratio, which is defined as the percentage of successful embedding out of all embedding done.

1.4 Statement of the Problem

Based on the theoretical and conceptual framework, there are several problems that can be enhanced from Listega. First, Listega uses 4-bit slices to encode message into first characters of list-cover items. Since each character is in 8-bit ASCII representation, Listega requires 2 rows of list-cover item to encode a character of message, which yields low embedding capacity of 1.32% up to 3.87%. Second, the use of latin square to map the characters of secret messages into the first letter of list items can lead to embedding failure. This failure can occur when an index latin square requires song titles or employee names that begin with, for example Q, X, Z, which are less common than those that begin with A, I, E in Indonesian language. Third, the previous method only use embedded items which can lead to suspicion and less resilient to noise.

1.5 Hypothesis

The List Steganography Based on Syllable Patterns as the proposed method exploits the use of syllable pattern in order to improve the embedding capacity of Listega. Syllable patterns are grouped based on the set of consonants (C) and vowels (V) which form a syllable patterns. The use of the syllable pattern as a method of steganography is a novelty, and since syllables are not as immediately observable as characters, their implementation as a method of steganography is one of the contributions of this research. Using the List Steganography Based on Syllable Patterns, to embed one character, one row of list-cover is needed, compares to two rows of list-cover in original Listega. This is due to direct mapping between character and syllable pattern in this method, as opposed to mapping half character (4-bit slice) to first letter of list item in original Listega. Therefore, the other contributions of List Steganography Based on Syllable Patterns is that it can be expected to have embedding capacity twice as much as original Listega, with the same or better performance.

1.6 Assumption

This research assume employee birthday list as list-cover in its steganographic system. Therefore it is assumed that the list can be divided to several period or interval of time, according to the birth date of the employees. The sender and the recipient are assumed to have the same data and set of rules have been agreed before the system is run.

1.7 Scope and Delimitation

The List Steganography Based on Syllable Patterns exploits syllable pattern to encode message, while using list-cover to conceal data. For implementing this method, a company employee birthday case is used. The method uses a list of a company's employee birthday list, which has around 17,500 entries. The method uses two first syllables of the employee name, which is one of the columns of employee list. The names of the employees are syllabified according to the rules from Indonesian grammar. The syllable patterns which resulted from the syllabification process are then used to embed the secret message.

The secret message is in Indonesian Language, while the characters used in the secret message are the upper-case alphabets without accent marks (A-Z) and SPACE. In the implementation using unfiltered list-cover, beside its function as regular characters, the character A is used to mark the beginning of a message. Three SPACE's are used to mark the end of a message.

1.8 Importance of the Study

This research opens new possibilities in noiseless steganography in general, and list-based steganography method in particular, by exploring the use of syllables to conceal secret messages. The employee birthday list is just an implementation case of the List Steganography Based on Syllable Patterns. It is possible to implement the method with various lists in organizations as an option of information security enhancement without attacting any suspicion.