# Cover Text Steganography: N-gram and Entropy-based Approach

Sara M. Rico-Larmer
*Kennesaw State University*, slarmer@students.kennesaw.edu

**Abstract**
Steganography is an ancient technique for hiding a secret message within ordinary looking messages or objects (*e.g.*, images), also known as cover messages. Among various techniques, hiding text data in plain text file is a challenging task due to lack of redundant information. This paper proposes two new approaches to embed a secret message in a cover text document. The two approaches are n-gram and entropy metric-based generation of stego text. We provide examples of encoding secret messages in a cover text document followed by an initial evaluation of how well stego texts look close to the plain texts. Furthermore, we also discuss several related work as well as our future work plan.

**Disciplines**
Information Security | Management Information Systems | Technology and Innovation

# INTRODUCTION

Steganography is a popular approach to hide a given secret message within an ordinary document (known as cover media). The embedded message within a cover document is known as stego document. A systematic approach is applied at the sender side to generate the stego document, while the secret message gets extracted at the receiver side. The entire approach is known as stego system.

Steganographic techniques are divided into several categories based on the employed approach [6]. For example, a substitution system would add a secret message in redundant text present in a file, whereas a statistical method would capture the statistical properties from a text file in order to encode information. The cover document can be of three types: text, image and video.

This work focuses on embedding a secret message into a text document. Generally, it is a challenging task as text documents have little or no redundant information to hide a secret message compared to images or videos. In this project, we focus on embedding a secret message in a cover document of type text. We denote it as cover text steganography.

Current text-based steganography approaches (e.g., Garg (2011), Kim (2003), Nagarhill 2014, Low et al. (1995)) assume that a cover document is huge and the secret message is smaller in size. In practice, the cover document may be much smaller. A popular steganography technique on text cover document is to change the format by adding a space or invisible characters. The generated stego text reads as misspelled words in text, variation of fonts. This would fool an ordinary reader thinking misspellings or extra spaces. Format based approach may apply various fonts sizes, gaps among lines, or paragraphs.

In this paper, we propose two simple approaches to generate stego text: N-gram and Entropy. We provide examples of the ideas.

The paper is divided as follows. Section 2 provides an overview of some related work. Section 3 introduces N-gram and Entropy-based approaches with example of stego-text generation. Section 4 concludes and provides a future work plan.

# RELATED WORK

Garg (2011) proposed to embed a secret message in HTML document. The idea is that HTML tag attributes often come with pairs (e.g., ) and the order of the pairs does not impact on the HTML document rendering process at the browser. Thus, they generate a set of tag attribute pairs and

alter their sequences to represent binary zero or one (e.g., align preceding width imply 1, where align present after width means zero). The approach works fine as long as HTML document is large enough to fit a secret message. If the secret message has a large length and HTML document does not have enough pairs of tag and attribute from the built in table, then the approach may not work well.

Kim, Moon, and Oh (2003) grouped adjacent words and generated statistics of white space. The embedding of the secret message involves modifying the statistics among adjacent words in a group. Nagarhill [3] proposed to embed a secret message through emotics by mapping alphabets and numbers with emotics available by SMS messaging application in mobile phone.

Low et al. (1995) proposes word shifting method where words are shifted horizontally and by changing the distance between the words the information is hidden. This leads to the strategy that marks a text line both vertically using line shifting and horizontally using word shifting.

Interested readers can see an extensive survey of Bennett et al. (2004) covering most known cover text-based stego text generation approaches.

# PROPOSED APPROACH

## N-GRAM ANALYSIS-BASED STEGANOGRAPHY

N-gram approach is used to generate possible close enough words for a given word. If the length of a given word is N, the approach systematically generates strings (subwords) from 1 to N-1 characters (Suen 1979). There are automatic tools available (e.g., Frequency Analysis (2016)) to generate N-gram for a given word. N-gram techniques are widely used for speech recognition, spelling correction, information extraction, etc. (N-gram (2016)).

In our proposed approach, we rely on n-gram analysis of wording to insert an embedded message. For a given word of length N (N>2), we generate words of length N-1. Then, we choose the first subword and place a space between the subword and remaining character in ordinary text to encode binary zero. For example, the word “TEST” has length 4, and it will have the following two subwords generated by a tool from (Frequency Analysis (2016)) as shown in Figure 1.

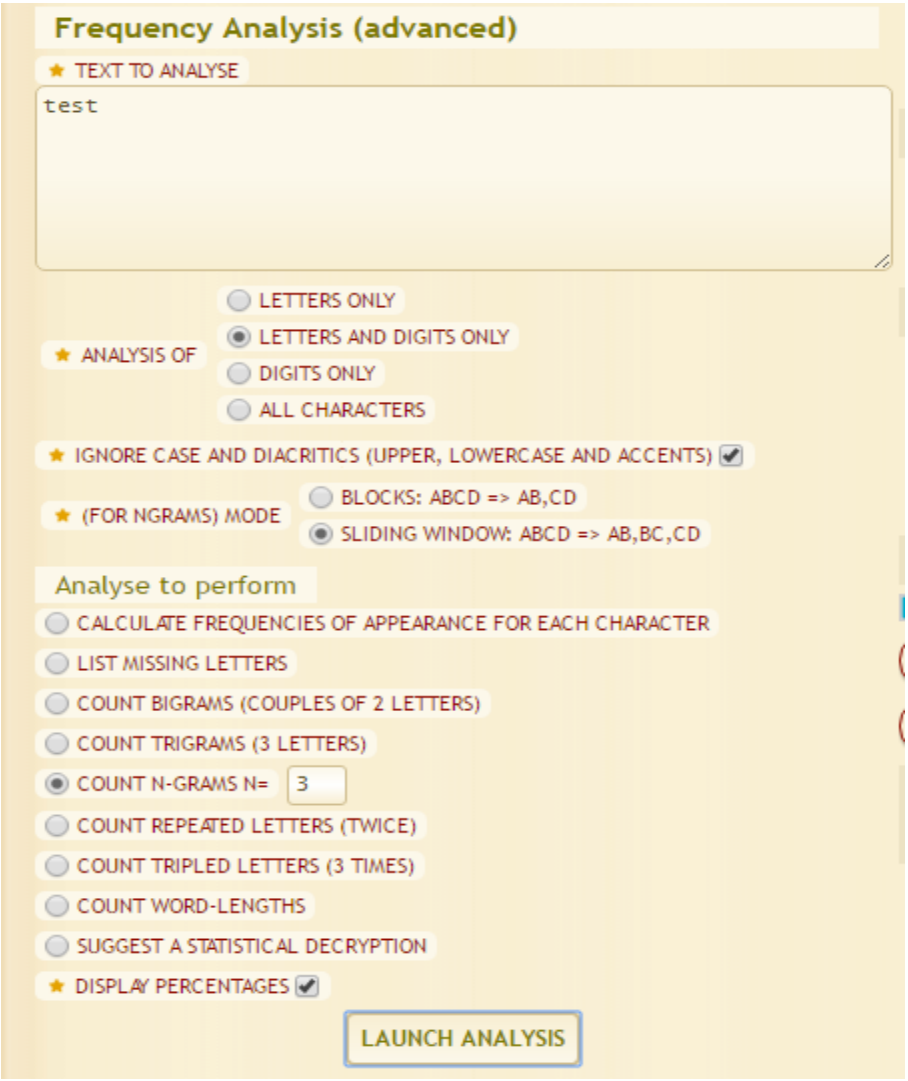


**Figure 1: Example of Trigram for the cover text "TEST"**

**Table 1: Example of secret message encoding in cover text TEST**

| Secret message | Stego text |
|---|---|
| 0 | T EST |
| 1 | TES T |

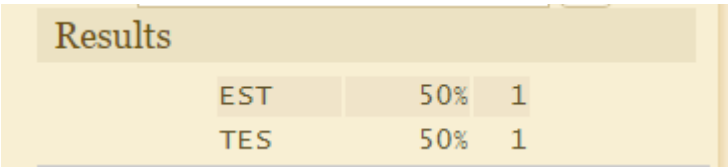


**Figure 2: Result of Trigram for the cover text "TEST"**

To encode the secret message 0, the stego text will be "T EST", where we added a space between "EST" and the beginning character "T" (first row of Table 1). We choose the last element of the generated subword (TES from Figure 1) of length N-1 to encode binary 1. Thus, a sender can encode a secret message up to the length of the number of words (as long as a character is at least 3 characters). The receiver needs to know the N-gram generator tool link to recover the secret text.

## ENTROPY BASED STEGANOGRAPHY

Entropy (*H*) is a popular metric from information theory proposed by Shannon [9]. It calculates the amount of randomness present in a message. The formula shown in Equation (i) is commonly used to compute entropy.

Let us assume that *Q* is a set of symbols (unique characters) found in a given word present in cover text. Here, $q_i$ is the occurrence of $i^{th}$ character of a given word found in cover text. Let us consider that $p(q_i)$ is the occurrence of probability of $q_i^{th}$ element. Then, the entropy of Q is

$$H(Q) = -\Sigma qi * \log_2 P(qi) \ldots \ldots (i)$$

To encode a secret message using entropy-based technique, we find two successive pair of words and compare their entropy level. To encode zero, we place the word having lower entropy at the front whereas, to encode one, we place the word having higher entropy at the front. If the entropy of two successive words are same, we skip the pair of words and move to the next.

Let us assume the cover text is "Comparing inverted files and signature files for searching a large lexicon" and secret message to be encoded is "11010". Table 2 shows each of the words (top row) of the cover text and the corresponding entropy level (bottom row). There are available open source tools to compute entropy (see for example, Shannon Entropy Calculator (2016)).

**Table 2: Words and entropy level**

| Comparing | inverted | files | and | signature | files | for | searching | a | large | lexicon |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.17 | 2.75 | 2.32 | 1.58 | 3.12 | 2.32 | 1.58 | 3.17 | 0 | 2.32 | 2.81 |

**Table 3: Stego text generation example**

| Comparing | inverted | files | and | files | signature | searching | for | a | large | lexicon |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 1 | | 0 | | 1 | | 0 | | |

Table 3 shows encoding of the secret message "11010" based comparison of entropy level for two successive word pairs in cover text. For example, the

entropy level of "Comparing" and "inverted" is 3.17 and 2.75, respectively. So, to encode "1", we keep the order as is (the word having higher entropy is already at the front). The last word ("lexicon") is an orphan and in this case to be ignored by the receiver. The sender would require mentioning the size of the secret message to the receiver.

# EVALUATION

We calculate Levenshtein Distance (LD) between plain text and stego text. The LD algorithm computes the least number of operations (substitution, insertion, and deletion) performed at character level to modify plain text to stego text [12]. We apply an online tool to measure the LD between plain cover text and stego text [11].

**Table 4: Secret message and plain text**

| Test# | Secret | Plain text | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1101011100 | Comparing | inverted | files | and | signature | files | for | searching | a | large | lexic |
| 2 | 1101011100 | Normally | the | system | probes | the | monitors | and | fills | in | the | valu |
| 3 | 1101011100 | Email | enables | you | to | communicate | with | users | on | the | local | syste |
| 4 | 1101011100 | The | keys | back | up | and | correct | the | shell | command | line | |
| 5 | 1101011100 | The | nature | of | value | differs | for | different | types | of | organizations | |
| 6 | 1101011100 | Usually | formal | approval | will | solidify | the | sponsors | of | the | project | |
| 7 | 1101011100 | The | focus | will | be | on | the | managerial | and | business | decisions | |
| 8 | 1101011100 | Typical | projects | have | six | to | ten | designers | submitting | several | designs | |
| 9 | 1101011100 | Literally | thousands | of | programmers | have | worked | on | Apache | over | the | years |
| 10 | 1101011100 | Some | critics | believe | poor | design | is | more | common | than | good | desig |

**Table 5: Stego text based on entropy and Levenshtein Distance (LD)**

| Test# | Stego text | | | | | | | | | | | LD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Comparing | inverted | files | and | signature | files | searching | for | a | large | lexicon | 9 |
| 2 | Normally | the | probes | system | the | monitors | fills | and | in | the | values | 20 |
| 3 | Enables | email | you | to | with | communicate | users | on | the | local | system | 21 |
| 4 | Keys | the | back | up | and | correct | shell | the | command | line | command | 23 |
| 5 | Nature | the | value | of | for | differs | different | types | of | organizations | | 21 |
| 6 | Formal | usually | approval | will | the | solidify | sponsors | of | the | project | | 23 |
| 7 | Focus | the | will | be | on | the | managerial | and | business | decisions | | 12 |
| 8 | Projects | typical | have | six | to | ten | submitting | designers | design | several | | 39 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | Thousands | literally | programmers | of | have | worked | Apache | on | the | over | years | 36 |
| 10 | Critics | some | believe | poor | is | design | more | common | good | than | design | 27 |

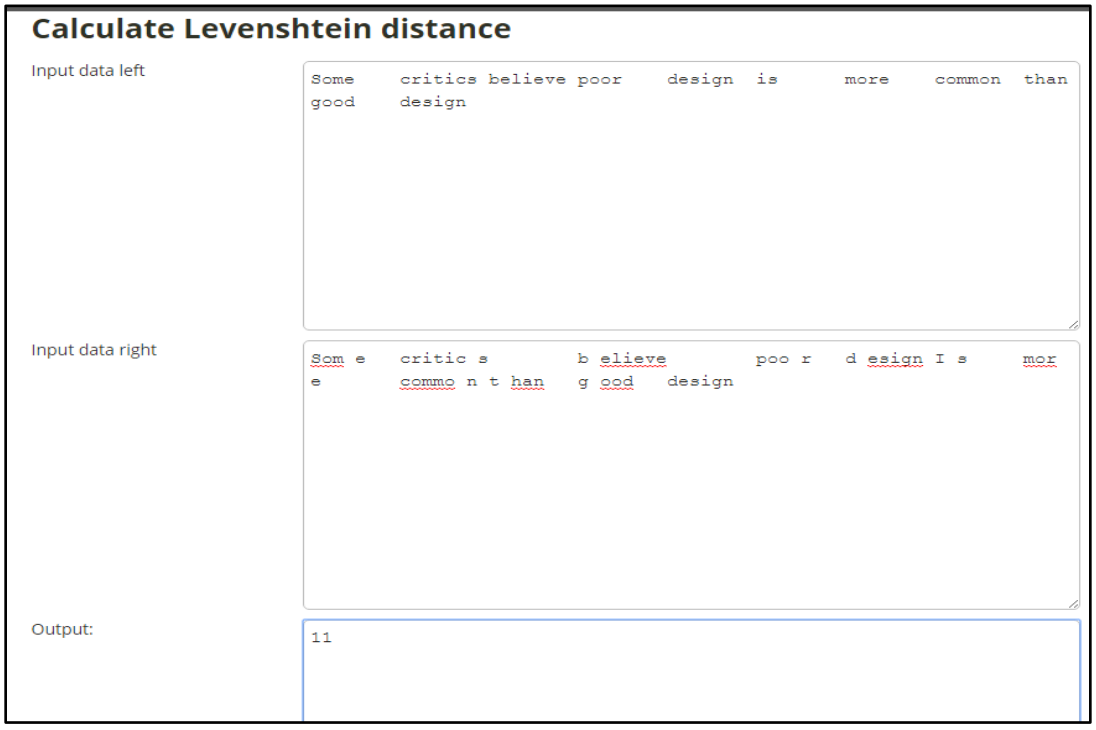


**Figure 3: Screenshot of computing LD for N-gram based steganography**

The higher the distance, the more dissimilar the two given strings are. Table 4 shows 10 original texts. We encode the secret message "1101011100" (column 2). Table 5 shows the generated stego text based on entropy. The last column of Table 5 shows LD for entropy-based stego text. Table 6 shows the generated stego text using N-gram technique. The last column of Table 6 shows LD.

**Table 6: Stego text based on N-gram and Levenshtein Distance (LD)**

| Test# | Stego text | | | | | | | | | | | LD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Comparin g | inverte d | f iles | an d | s ignature | file s | fo r | searchin g | a | l arge | lexicon | 18 |
| 2 | Normall y | th e | s ystem | probe s | t he | monitor s | an d | fill s | I n | t he | values | 12 |
| 3 | Emai l | enable s | y ou | t o | c ommunicate | wit h | user s | o n | t he | l ocal | system | 10 |
| 4 | Th e | key s | b ack | u p | a nd | correc t | th e | shel l | c ommand | l ine | | 10 |
| 5 | Th e | natur e | o f | valu e | d iffers | fo r | differen t | type s | o f | o rganizations | | 11 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | Usuall y | forma l | a pproval | wil l | s olidify | th e | sponsor s | o f | t he | p roject | | 10 |
| 7 | Th e | focu s | w ill | b e | o n | th e | manageria l | an d | b usiness | d ecisions | | 10 |
| 8 | Typica l | project s | h ave | si x | t o | te n | designer s | submittin g | severa l | design s | | 10 |
| 9 | Literall y | thousand s | o f | programmer s | h ave | worke d | o n | Apach e | o ver | t he | years | 10 |
| 10 | Som e | critic s | b elieve | poo r | d esign | I s | mor e | commo n | t han | g ood | design | 11 |

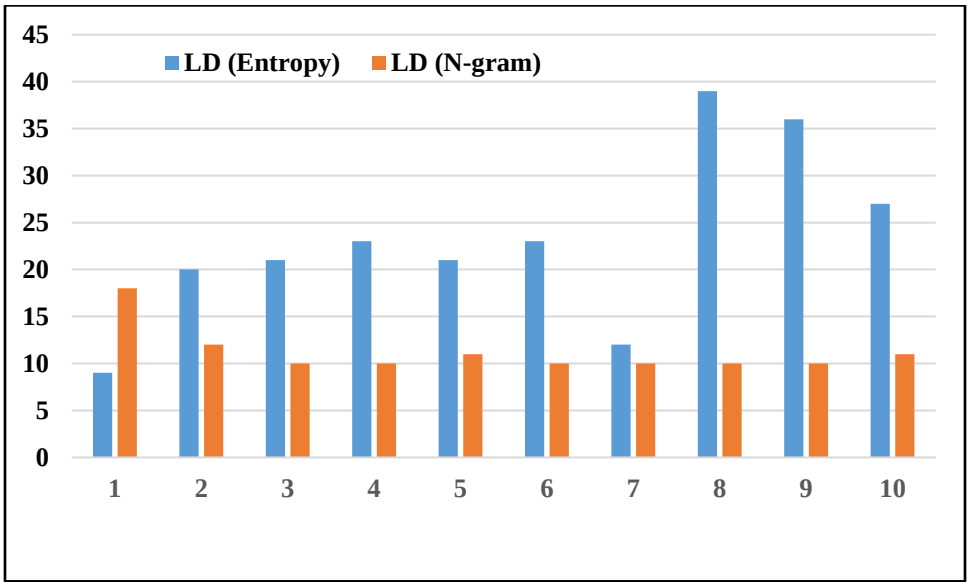


**Figure 4: Leveshtein Distance between plain text and Stego Text**

Figure 4 shows the comparison of LD between Entropy and N-gram based technique. We find the LD is lower for N-gram compared to Entropy-based technique. N-gram would be the preferred technique to generate stego text for small volume of texts. Table 7 shows the average and standard deviation for Entropy and N-gram based technique.

**Table 7: Average and standard deviation of LD for Entropy and N-gram**

| | **Entropy** | **N-gram** |
|---|---|---|
| **Avg.** | 23.1 | 11.2 |
| **Stdev.** | 9.279009 | 2.485514 |

# CONCLUSION

Steganography comprises a set of art to hide a secret message in an ordinary looking document. Though most works focused on hiding information in image or video files, text file poses the challenge of discovering redundant information to hide secret text. To overcome the limitation of some of the existing text-based steganography techniques, we propose to apply N-gram and entropy metric-based generation of stego text to hide a secret message. We show examples of hiding secret messages and performed an initial evaluation to compare between entropy and N-gram based technique. The early results indicate N-gram is better than entropy-based technique. In the future, we plan to compare with larger text files. We also plan to measure the effectiveness steganography-based approaches using other distances such as Jaro-Winker.

# REFERENCE

[1] M. Garg. (2011). A Novel Text Steganography Technique Based on Html Documents, *International Journal of Advanced Science and Technology*, Vol. 35, October, 2011, pp. 129-138. http://www.sersc.org/journals/IJAST/vol35/11.pdf

[2] Y. Kim, K. Moon, and I. Oh. (2003). A Text Watermarking Algorithm based on Word Classification and Inter word Space Statistics. *Proc. of the 7th International Conference on Document Analysis and Recognition (ICDAR)*, 2003, pp. 775-779.

[3] T. Nagarhill. (2014). A New Approach to SMS Text Steganography using Emoticons, *International Journal of Computer Applications*, pp. 1-3. Accessed from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.662.4948&rep=rep1&type=pdf

[4] S. Low, N. Maxemchuk, J. Brassil, L. O'Gorman. 1995. Document Marking and Identification Using Both Line and Word Shifting, *Proc. of 14th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, Vol.2, pp. 853-860. Accessed from http://netlab.caltech.edu/publications/InfocomID95.pdf

[5] Frequency Analysis Tool, 2016 Accessed from http://www.dcode.fr/frequency-analysis

[6] Krista Bennett. (2004). Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text, CERIAS Tech Report 2004-13, Purdue University, USA.

[7] N-gram, All our N-gram belong to you. (2006). Accessed from https://research.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

[8] C. Suen. (1979). n-Gram Statistics for Natural Language Understanding and Text Processing, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 1, Issue 2, Feb 1979, pp. 164-172. http://dl.acm.org/citation.cfm?id=2053409

[9] T. Cover. (2006). *Elements of information theory*, 2nd Edition. Wiley-Interscience.

[10] Shannon Entropy Calculator. (2016) Accessed from http://www.shannonentropy.netmark.pl/

[11] Levenshtein Distance Calculator, Accessed from http://www.unit-conversion.info/texttools/levenshtein-distance/

[12] The Levenshtein Algorithm, Accessed from http://www.levenshtein.net/