Syntactic Steganography for Document Leak Identification

Hardik Rajpal

October 23, 2023

1 Introduction

This project seeks to implement the principles of syntactic steganography for identification of the sources of leaks of confidential documents from a protected server. The security model is based on the following assumptions:

1. Each protected document is stored on the server in .DOCX format.
2. Each user has an identification number, known to the server.
3. The only way to access the documents is to request them from the server after authentication.

In addition to steganography, additional techniques are employed to futher complicate the task of leaking the documents, such as:

1. Ghostscript

2 Reading Notes from This Paper

1. Abstract
   - Syntactic bank-based linguistic steganography approach for information security.
   - Uses the Stanford parser to obtain a syntax tree from sentences.
   - Use Shannon-Fano coding to compress input messages (user identification in the project's case) in as few bits as possible.
   - Syntax transformation task: searches the syntax set of the given sentence within the syntax bank and transforms it into a syntax that can represent the secret in the sentence.
   - The resulting text is still "innocent-looking," and the transformation doesn't affect the semantics.
   - Additionally use a HMAC to improve robustness of the stego text.
2. Introduction
   - There are three dimensions in a stego system:
     (a) Payload Capacity: the ratio of hidden information to cover information.
     (b) Robustness: the ability of the system to resist against changes in the cover object.
     (c) Imperceptibility: the potential of the generated stego object to remain indistinguishable from other objects in the same category.
   - The three dimensions often contradict each other.
   - Concealing any data in a text file is the most difficult kind of steganography due to the lack of redundant information in a text file.
   - Text steganography is broadly classified into two categories:
     (a) Linguistic Approach: the art of using written natural language to conceal secret messages.

- Further divided into semantic and syntactic methods, with the syntactic method being divided into line-shift, word-shift, open-space and feature encoding.

(b) Format-based Approach: uses physical formatting of text as a place in which to hide information.

3. Linguistic Steganography
   - The changes made to embed information in the cover text **do not result in an ungrammatical or unnatural text.**
   - Most methods of this type use either lexical (semantic) (ex. synonym substitution) or syntactic transformations (of the grammatical style of the original sentences) or a combination of the two.
4. Syntax of Language
   - Set of rules that the language uses to combine words to create sentences.
   - Parts of speech combine into phrases.
   - Clauses can be broken down into phrases.
   - Sentences can have one or more independent or dependent clauses.
5. Proposed Approach
   - The cover text (document in our case) is parsed by the parser to obtain a syntax tree, while the secret message (user identifer in our case) is compressed by the Shannon-Fano algorithm.
   - How do we get the syntax set of a sentence?

3 References

1. CoreNLP Setup
2. parse tree
3. Combat text selection
4. Combat text selection still
5. read this on paraphrasing