CS G513 : Network Security

SECOND SEMESTER 2021-22

# An Approach for Text Steganography Based on Markov Chains

SUBMITTED BY

Sahil Sanjiv Rasaikar - 2021H1030026G
Akshat Saxena - 2021H1030017G
Gaurav Saxena - 2021H1030056G
Shashank Pal - 2021H1030064G

**INDEX……………………………………………………**

# Introduction:

Steganography is an area of research which deals with the transfer of message through a channel which is readily available for the third party to be read,However it is not suspected that some relevant information is traveling in the channel.Steganography deals with stealth information or hiding information in plain sight such that the information hidden is not suspicious.Steganographic techniques deal with hiding information in images,audios,videos and other media.

Types of Steganography:

1. Text Steganography
2. Image Steganography
3. Video Steganography
4. Audio Steganography
5. Network Steganography

In this Report and for our project work we will solely focus on the approaches used for Text Steganography.

Text Steganography is the process of hiding relevant information in plain text and in plain sight such that only the receiver can make out that there is information in the text.It involves changing the format of the text ,changing words within a text,generating random text,or using cfg to generate readable texts.Various methods used to stealth data are:

- Format based text
- Random and statistical generation
- Linguistic method

For Eg SNOW which is a steganography tool uses tabs and spaces at the end of each line to hide information,There are also other techniques start from base text(The Covertext),and modify it to some secondary text by using several techniques like switching words to near synonyms ,or by changing the sentences grammatically to sound and mean something very different than the original text.

Or Some methods modify sentences such that they look like orthographical or typographical errors.

Sometimes Texts are generated using Grammar models. This kind of system Generates Texts that make sense at grammatical level and not at semantic level.In this report we discuss one such method which uses the theory and implementational qualities of probability and statistics defined by the Markov Model, and explore a way to modify data in this way.

## Motivation:

Traditionally for Text transfer usually preferred method is Encryption ,However Encryption though is secure it creates suspicious scenario in the channel ,Suppose a user is sending some information through a public channel to a receiver,The admin is an integral part of the network and thus can see that a ciphertext is being transmitted in the network,Inturn to which he will block the access for the User,However if steganography is used then the people who want to communicate can use various steganographic techniques to generate text which can be sent over the network without any suspicion being created and thus the two original senders can use the network for communication without the network Admin getting noticed about such knowledge transfer.As suggested in the Introduction a large number of steganography techniques existes for Image,Video and Text data.Considering Text Steganography we see a large number of algorithms focusing on Techniques such as raw text generation using Context Free Grammars,Grammar manipulation,Word replacement methods etc.Through this project we wanted to bring forward something different from the usual paradigm ,something from the domain of probability theory and statistics,This Project focuses on Markov chain implementation without simplification of the model.Furthermore in future this method can be combined with other methods which are language based Steganographic systems to make better Steganographic text generation models,and compare them with other Steganographic models.

# Literature Survey:

There are Many Text Steganography methods which hide Texts and are summarized in Linguistic Steganography survey, which suggests steganographic text generated by tool such as spammimic or as Suggested in Applying Statistical Methods to Text Steganography we can use T-lex method which uses word replacement.Or we could use grammar generation rules to generate texts.Nice Text also shows a way to transform text from one form to another,using some custom styles,some custom Context Free Grammars and dictionaries.Similar to this project,an approach is used in Text Steganography using Markov Chain and BinText Steganography Based on Markov State Transfering Probability.However to avoid complex calculations these papers assume equal probability of all state transitions.This is the reason which can affect text transitions heavily,Suppose we are considering a phrase, ‘The’ and ‘Naturally’ are two most frequently used words in phrases ,However the word ‘The’ is more likely to come over Naturally,This dependency is not preserved due to simplification of the model.Similar Models on markov chain use such specifications which require simplification of chain of markov models by making probabilities of every state equal or by replacing them by other ones by using an encoding algorithm.This project presents a way to preserve the probabilities in the markov model to higher level of accuracy.

# Objectives:

- Our main objective is to design and implement two functions: encode and decode. One will be used to create text (based on a literature) called as stegotext out of the input that it receives and the other will help in getting back the original plain text. These functions will be implemented as different modules to be called upon during the execution of the program.

- Generate a markov chain model which will be used to compute probabilities for phrases. These probabilities will be used further for the generation of stegotext. Each n-gram text will be represented as a node whose current probability will depend on the previous state.

- Our final objective will be to deploy our model as a webapp where the user can enter the text in real-time and a corresponding stegotext will be generated. This will be used to demonstrate the use of our model but the model can also be used as a backend of a network.

# Methodology:

The proposed assignment is based on the use of markov chains to encode messages. Markov chain has two fundamental properties defined as follows:
1)limited horizon property
2)Time Invariant property

Limited Horizon Property:
Limited Horizon Property says that the markov chain doesn't have memory of any states beyond the last one. This property gives us the benefit of not maintaining the excessive calculations at each step, we can simply check the probability, and make judgment.
Time Invariant Property:
Whereas Time Invariant property means that the conditional probabilities for all states do not depend on the position on the sequence. This property ensures that if a state s0 has probability p1 at time instance t0 then the probability for the same state s0 at some time t0+k will be equal to p1 As the probabilities of each state in Markov chain models do not get updated with time and thus stay constant.This property also ensures decoding of the message will always give back the original the original plain text irrespective of the time when it is decrypted.Hence can be used to encode text even for longer durations.

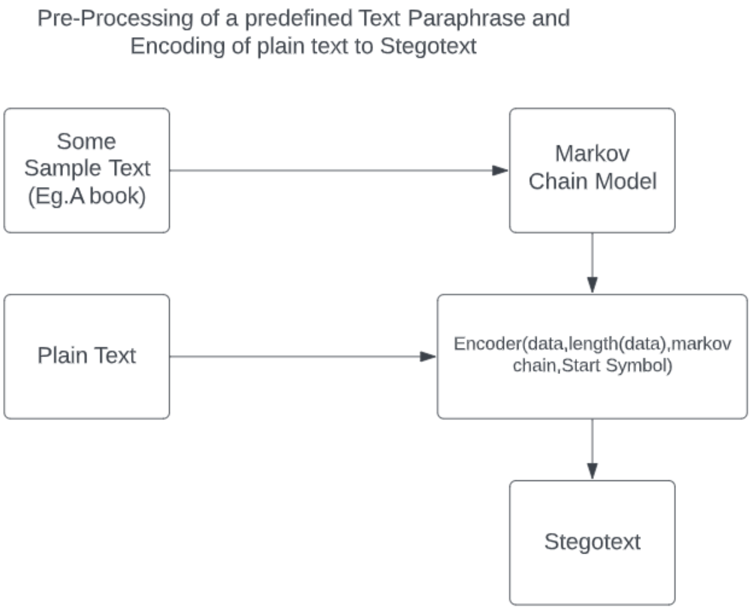
The assignment works in three phases:
1)Pre-processing
2)Encoding
3)Decoding

Pre-processing phase:In this phase we use a predetermined text data.This source text has to be predetermined between sender and receiver.This text can be taken from Novels,Text Conversations,Newspapers,Textbooks etc.Once the data source is finalized then we create the Unique markov chain for that predetermined text.

Encoding:

In the encoding process first we decide a n bit number which is going to be the size of the longest input data.After this we create a continuous sequence of numbers from 0 to n representing the subranges.Once these are finalised and generated we will start moving from start state in markov chain based on the probability of the next states.As we are moving down the chain we are also dividing out original range into subranges of numbers as per the probability distribution and this process continues till our subrange contains only a single number. Now the state path we have followed from the start to the end state will represent the words used to encode that n bit data. In this way we are going to encode all the input characters one by one and their respective state path will give us the encoded data for those characters.

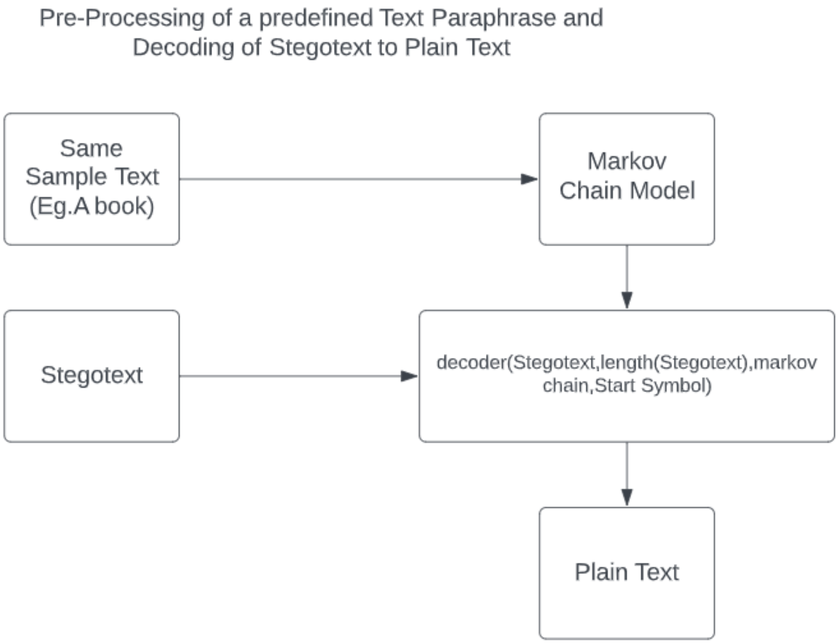
The flowchart for Encoding is as follows:


Pre-Processing of a predefined Text Paraphrase and Encoding of plain text to Stegotext
Some Sample Text (Eg.A book)
Markov Chain Model
Plain Text
Encoder(data,length(data),markov chain,Start Symbol)
Stegotext

Decoding:

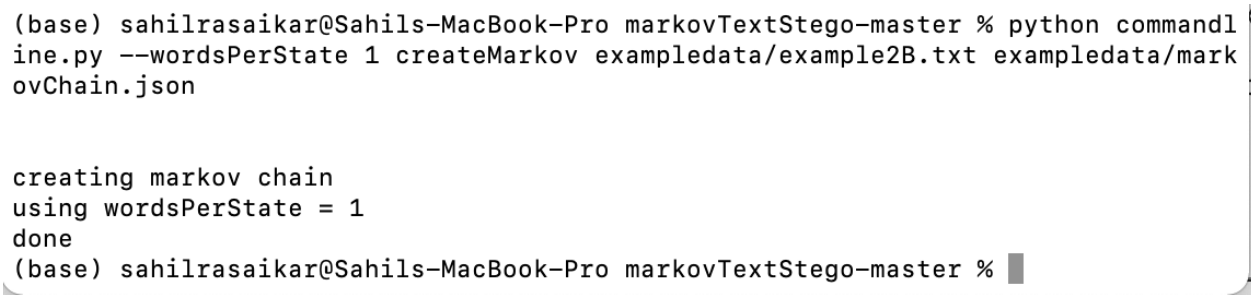
As from the encoding method we know for sure that the encoded messages are formed due to a chain of words which are strung together using the makrov chain upto a state where the subranges for the characters converge to a single value.Now if we take the encoded message and start from the initial state and go further in the chain using the set of words in the encoded text we reach a point where there is no link for from the current state to the state for the next word.This is where we know that the particular chain has ended at that particular state. Using this we can go to the start of the chain and determine the character for which this chain was developed.Similarly for every character a chain can be traced and the original character can be determined from the encoded text message.Thus repeated determinism of characters will lead to the generation of original characters and thus the original plain text is obtained as a result.

The flowchart for Decoding is as follows:


Pre-Processing of a predefined Text Paraphrase and Decoding of Stegotext to Plain Text
Same Sample Text (Eg.A book)
Markov Chain Model
Stegotext
decoder(Stegotext,length(Stegotext),markov chain,Start Symbol)
Plain Text

# Results:

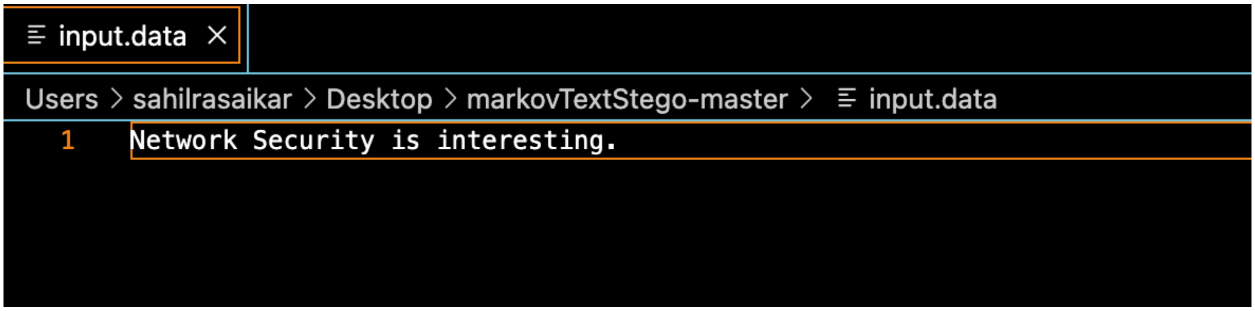
- **Creating a markov chain for some example text**



```
(base) sahilrasaikar@Sahils-MacBook-Pro markovTextStego-master % python commandl
ine.py --wordsPerState 1 createMarkov exampledata/example2B.txt exampledata/mark
ovChain.json


creating markov chain
using wordsPerState = 1
done
(base) sahilrasaikar@Sahils-MacBook-Pro markovTextStego-master %
```

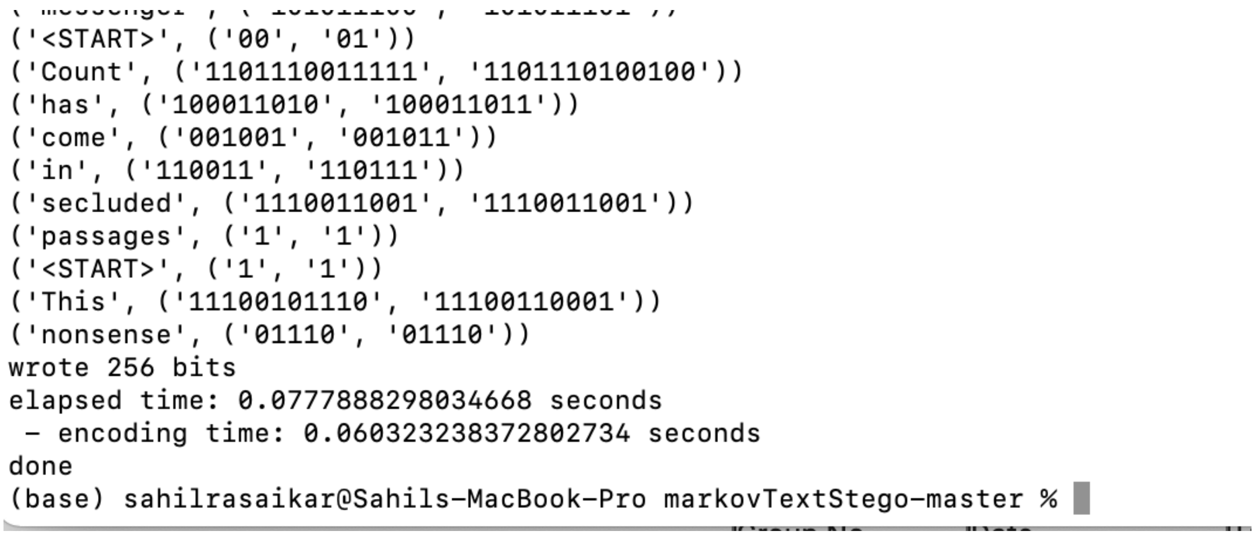
- **Some Input text**



input.data
Users > sahilrasaikar > Desktop > markovTextStego-master > input.data

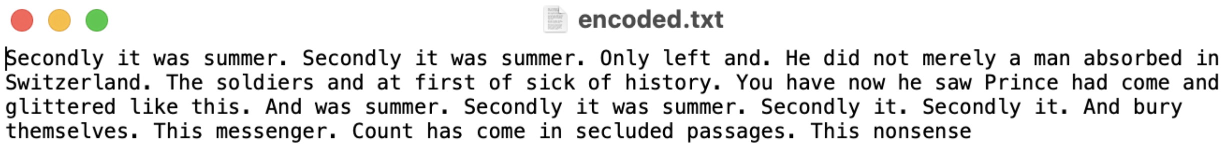
```
1    Network Security is interesting.
```

- **Encoding the Input text**



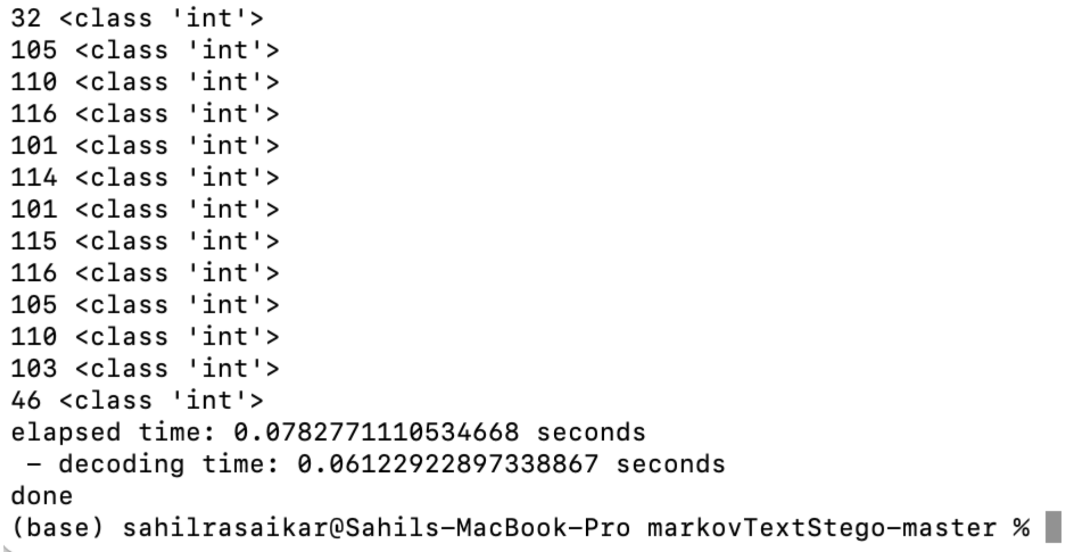
```
('messenger', ('101011100', '101011101'))
('<START>', ('00', '01'))
('Count', ('1101110011111', '1101110100100'))
('has', ('100011010', '100011011'))
('come', ('001001', '001011'))
('in', ('110011', '110111'))
('secluded', ('1110011001', '1110011001'))
('passages', ('1', '1'))
('<START>', ('1', '1'))
('This', ('11100101110', '11100110001'))
('nonsense', ('01110', '01110'))
wrote 256 bits
elapsed time: 0.0777888298034668 seconds
 - encoding time: 0.060323238372802734 seconds
done
(base) sahilrasaikar@Sahils-MacBook-Pro markovTextStego-master %
```

- **Encoded text**



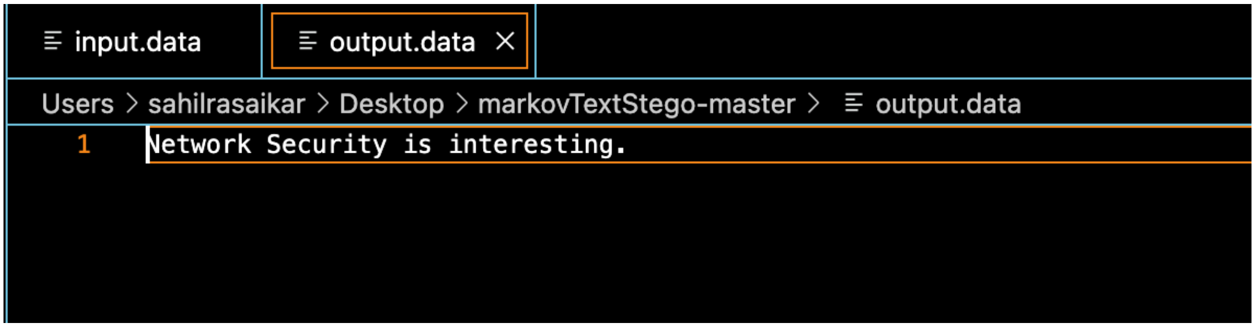
```
Secondly it was summer. Secondly it was summer. Only left and. He did not merely a man absorbed in
Switzerland. The soldiers and at first of sick of history. You have now he saw Prince had come and
glittered like this. And was summer. Secondly it was summer. Secondly it. Secondly it. And bury
themselves. This messenger. Count has come in secluded passages. This nonsense
```

- **Decoding Encoded text**



```
32 <class 'int'>
105 <class 'int'>
110 <class 'int'>
116 <class 'int'>
101 <class 'int'>
114 <class 'int'>
101 <class 'int'>
115 <class 'int'>
116 <class 'int'>
105 <class 'int'>
110 <class 'int'>
103 <class 'int'>
46 <class 'int'>
elapsed time: 0.0782771110534668 seconds
 - decoding time: 0.06122922897338867 seconds
done
(base) sahilrasaikar@Sahils-MacBook-Pro markovTextStego-master %
```

- **Decoded text**


input.data
output.data
Users > sahilrasaikar > Desktop > markovTextStego-master > output.data

```
Network Security is interesting.
```

# Conclusion:

The proposed assignment produces a way to use markov chain models to encode the original message  into an encoded text.This encoded text can be sent to the receiver via any communication channel,the receiver uses the predefined text to make a markov chain,using this markov chain the user then can decode the received encoded message.

The said assignment also expands the horizon of steganography to the field of statistics.Though not completely sensible the result is fairly more natural than some other text steganography techniques like static table replacement or synonym replacement,etc.

# References:

1. Bennett K.: Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text. CERIAS Tech Report 2004-13, Purdue University. (2004)
2. Nechta, I., Fionov, A.: Applying Statistical Methods to Text Steganography. CoRR. (2011)
3. Kwan M.: SNOW. http://www.darkside.com.au/snow/manual.html (1996)
4. Topkara M., Topkara U., Atallah M.J.: Information Hiding Through Errors, A Confusing Approach. Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents. (2007)
5. Grothoff, C., Grothoff, K., Alkhutova, L., Stutsman, R., Atallah, M.: TranslationBased Steganography. Proceedings of the 2005 Information Hiding Workshop (IH 2005). Paper 1624. (2005)
6. Dai, W., Yu, Y., Dai, Y., Deng, B.: Text Steganography System Using Markov Chain Source Model and DES Algorithm. Journal of Software, vol. 5, issue 7, pp. 785-792. (2010)
7. Dai, W., Yu, Y., Deng, B.: BinText Steganography Based on Markov State Transferring Probability. 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human (ICIS ’09). (2009)

8. Chapman, M.: Hiding the Hidden: a Software System for Concealing Ciphertext as Innocuous Text. Masters thesis, University of Wisconsin-Milwaukee (1997)
9. Manning, C.D., Sch¨utze, H.: Foundations of Statistical Natural Language Processing. The MIT Press. (1999)
10. Russell, S., Norvig, P.: Artificial Intelligence, a Modern Approach. Prentice Hall, 2nd Edition. (2002)
11. Taskiran, C.M., Topkara, U., Topkara, M, Delp, E.J.: Attacks on Lexical Natural Language Steganography Systems. SPIE International Conference on Security, Steganography, and Water-marking of Multimedia Contents. (2006)