Automating the Cracking of Simple Ciphers

by

Matthew C. Berntsen

A Proposal Submitted to the Honors Council

For Honors in Computer Science

October 13, 2004

Approved: ____________________________

Richard J. Zaccone
Thesis Advisor

____

Gary Haggard
Chair, Department of Computer Science

1 Background

Ciphers have been used throughout history to keep information secret [3]. Text ciphers work by applying some methodology and a key to transform an English plaintext into a less meaningful ciphertext. Simple text ciphers fall into one of two categories: substitution ciphers, in which each letter of plaintext is replaced by a letter of ciphertext, and transposition ciphers which seek to break up the relationships between neighboring letters. Substitution ciphers can be either monoalphabetic or polyalphabetic (Definitions for various words can be found in Appendix C).

One key point about ciphers is that one must, given the proper key, be able to decipher them. As such, they cannot be random, and hence must conform to some algorithmic methodology, which inevitably results in some form of weakness. Monoalphabetic substitution ciphers are vulnerable to frequency analysis and transposition ciphers usually rearrange letters in a pattern such that by looking for repeated combinations of letters (digrams, trigrams, etc.) one can deduce the key length. From the key length and multiple sub-groups of ciphertext, it is often a small step to determine the key.

After World War II and the advent of the computer, simple text ciphers began to fall out of popular use[3]. With computers came two things: First, the ability to process immensely more information than a human in a given amount of time, greatly improving the capabilities of brute-force and other attacks and second, data encryptions act on the bit-wise representation of data, rather than the letters themselves. Data encryptions are advantageous because they allow the same encryption algorithm to be used on any unit of data as opposed to only text, and because they are more difficult to crack.

2 Methodology

This project will automate the cracking of simple text ciphers, specifically Vigenère through a program or set of programs. This will entail intimate knowledge of the cipher(s) used and the crypt-analysis needed to determine how best to approach cracking them with a ciphertext-only attack. The project will begin by focusing on the Vigenère cipher (See Appendix A), and will broaden the scope as time allows. Both encipherment and decipherment algorithms will be implemented for all ciphers used as well as tools to analyze ciphertext so as to determine the key. Also, an analysis of different algorithms' runtimes in computing the tests necessary in cracking Vigenère will be performed.

Through analysis and careful planning, this project will systematize and thus automate the somewhat heuristic process of breaking a cipher. Vigenère can be broken with the use of two tests, the Kasiski test (Appendix B.2) and the Friedman test (Appendix B.1), which for the sake of this project will assume the English language.. The Friedman test consists of two equations which indicate the type of cipher used (monoalphabetic vs. polyalphabetic) and estimate the keylength. While the Friedman test is well defined, there are multiple ways to approach obtaining the results for the Kasiski test, which will require substantial analysis to determine the best method. The two methods that appear most obvious are: Brute-force in which each substring is blindly compared to every following substring, and Dynamic or 'intelligent,' in which the substrings to be compared are screened based on some algorithm. Dynamic algorithms are usually faster, but are inevitably more complex to implement. Analysis will show what speed is traded for simplicity. Analysis will also be conducted of the runtime of the program overall, as well as its various modules, both experimentally and mathematically. In order to determine the key in a Vigenère Cipher, there are a number of tests which are employed.

As both the Kasiski and Friedman tests estimate the keylength, their results can be combined to deduce the most likely keylength. Once one has a hypothesis for the keylength (See Appendices B.2 and B.1), the ciphertext can be arranged into columns, with one column per letter in the key. Each column will be a simple shift cipher, from which one can use frequency analysis to determine the key letter used. Once one knows the key, the ciphertext can simply be deciphered. (Please see Appendix B.3 for discussion of the shift cipher, and how it can be broken.)

A preliminary portion of this project will be to determine what programming language to work in. The languages under consideration are Java, C/C++, and Squeak/Smalltalk. This will necessitate research into the advantages and disadvantages of each language for this purpose, the result of which will be discussed in the final thesis.

3 Significance

There is little point to enciphering information if one does not wish to keep it secret. It follows that if one wishes to keep information secret that others would likely want access to that information. One could easily come up with many scenarios where one would like the information being exchanged to be out in the open or kept secret. For example, Internet shoppers want their credit card information to be secured. Would-be thieves want access to that information. Cryptanalysis can identify vulnerabilities in a cipher that help both to break existing ciphers and to create stronger ciphers that will not exhibit the same vulnerabilities.

The weaknesses of the Vigenère cipher are already known and well documented. The process of cracking Vigenère, however, traditionally involves a human. This project will build a framework that will remove the human aspect from the process of cracking Vigenère and potentially other ciphers through the use of a computer program or set of programs.

This project will be Open Source (Academic Free License [4]), with the intent for it to be used or expanded upon by others. It will be useful in the learning of cryptography and cryptanalysis, the analysis of ciphers created in those endeavors, both by academics and enthusiasts.

A The Vigenère Cipher

The Vigenère Cipher is a simple polyalphabetic cipher based on the tableau in Appendix A.1 below, in which each row of the tableau is shifted one letter from the row above. The algorithms for encipherment and decipherment are discussed in Appendix A.2.

A.1 The Vigenère Tableau

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

A.2 Encipherment and Decipherment

The cipher works by taking a keyword, “BUCKNELL”, and uses it to encode a plaintext,“HAS A NICE CAMPUS”. First, the keyword is repeated on top of the plaintext:

```
BUC K NELL BUCKNE
HAS A NICE CAMPUS
```

For each letter in the plaintext, the corresponding ciphertext letter is determined by selecting the letter in the column determind by the keyword letter column and the row determind by the plaintext. In the example above, the resulting ciphertext would be (omitting spaces): “IUUKAMNPDUOZHW”.

To decrypt the Vigenère cipher there are two methods. The first is to simply do the encoding process backwards: Repeat the keyword above the ciphertext, and select the ciphertext letter out of the column determined by the keyword. The row that the ciphertext letter appears in corresponds to the plaintext letter that was enciphered. The second method is to encode again using the inverse of the keyword, which is easier for a computer to do. This is done by taking each keyword letter’s numerical value (A=0, B=1, etc.), and subtracting it from 26, and moding by 26.

B Tests and Shift Ciphers

B.1 Friedman Test

The Friedman test relies on the concept of Index of Coincidence (IC). Simply put, the the IC is the probability that when two letters are selected random , they are the same letter. Given the number of ciphertext letters, n, where n_1 = number of A’s in the ciphertext, n_2 = number of B’s in ciphertext, etc., the IC is defined as follows:

$$IC = \sum_{i=1}^{26} \frac{n_i(n_i - 1)}{n(n-1)} \tag{1}$$

From the accepted letter frequencies, one finds that the IC for the English language is about 0.065. It follows then, that a monoalphabetic cipher will have an IC that is approximately equal to 0.065. If it is not, then the cipher is most likely polyalphabetic [2].

The Friedman test can also be used to estimate the keylength of a Vigenère ciphertext. If the ciphertext is placed in columns, with one column for each letter of the keyword, then each column is a shift cipher of the plaintext equivalent (See Appendix B.3). It follows that it will have the same IC as ordinary English. It follows then, that that IC can be written as a function of the keylength. Solving for the keylength, the equation becomes:

$$keylength \approx \frac{0.027n}{(n-1)IC - 0.038n + 0.065} \tag{2}$$

The longer the ciphertext analyzed, the better the estimated keylength value. To provide a meaninful example of the Friedman test here would take multiple pages.

B.2 Kasiski Test

The Kasiski test is used to determine the keyword length. The test finds all repeated substrings of the cipher text of length three or more. It then finds their relative distances from each other, and on the assumption that these repeated subsequences, particularly the longer ones, occur at some multiple of the keylength, pose likely key lengths. The longer the ciphertext analyzed, the better the estimated keylength value. To provide a meaninful example of the Kasiski test here would take multiple pages.

Unlike the Friedman test, the Kasiski test is not well-defined. There are many ways to go about obtaining the desired results, and an investigation will be made into the most efficient and accurate method.

B.3 Shift Ciphers

Shift ciphers assign a value to each letter (A = 0, B = 1, etc.), call this value K, and take a shift L as a key. They then assign each letter to a new one by taking (K + L) mod 26. As such, the L last letters of the alphabet are wrapped around to correspond to the beginning.

Shift ciphers are, however, very easy to crack. Given a large enough plaintext in ordinary English, the letter frequencies will be similar to the accepted values, which are commonly known. As shift ciphers do not jumble the mapping of letters, the relative frequencies of the letters can tell one what shift was used.

C Definitions

Cipher A method by which plaintext is reversibly transformed into a less intelligible ciphertext.

Ciphertext The text that results from encipherment of a plaintext.

Encryption A method by which information (text or data) is transformed such that it is reversibly less meaningful.

Key A value used by a cipher to encipher plaintext to ciphertext and decipher ciphertext to plaintext.

Keyword A key that consists only of letters.

Plaintext A text in ordinary English to be enciphered.

Mod or Modulo A mod B is defined as the remainder after A is evenly divided by B.

Monoaplhabetic A cipher in which one letter or plaintext is uniformly mapped to one and only one letter of ciphertext. [5]

Polyalphabetic A cipher in which the letters are mapped one-to-many from plaintext to ciphertext based on some algorithm.

Runtime A way of describing how the time necessary for a program to run changes in relation to the size of the input.

Shift Cipher Please see Appendix B.3.

Text Cipher A method of concealing information that works directly on the letters (text) as opposed to their data representation.

Vigenère Cipher Please see Appendix A.

References

[1] Pfleeger, Charles P. and Pfleeger, Shari L. *Security in Computing*. Third Edition. Pearson Education, 2003.

[2] Lewand, Robert E. *Cryptological Mathematics*. The Mathematical Association of America, 2000.

[3] Singh, Simon. *The Code Book: The Evolution of Secrecy from Mary, Queen of Scots to Quantum Cryptography*. Doubleday, 1999.

[4] Academic Free License v. 2.1 Accessed October, 1 2004.
http://www.opensource.org/liscenses/afl-2.1.php

[5] Oxford English Dictionary Online. Accessed October 10, 2004.
http://dictionary.oed.com