A Modified Simple Substitution Cipher With Unbounded Unicity Distance

Bruce Kallick
Curmudgeon Associates
Winnetka, IL 60093
curmudgeon@rudegnu.com

Abstract. The classic simple substitution cipher is modified by randomly inserting key-defined noise characters into the ciphertext in encryption which are ignored in decryption. Interestingly, this yields a finite-key cipher system with unbounded unicity.

INTRODUCTION

The notion of unicity distance was introduced by Shannon [2] as

> [A measure of] how much intercepted material is required to obtain a solution to a secrecy system. [...] In general we may say that if a proposed system and key solves a cryptogram for a length of material considerably greater than the unicity distance the solution is trustworthy. If the material is of the same order or shorter than the unicity distance the solution is highly suspicious.

and is frequently defined in the current literature as, for example [1]

> The minimum amount of ciphertext (number of characters) required to allow a computationally unlimited adversary to recover the unique encryption key.

Shannon [2] wrote

> It appears from this analysis that with ordinary languages and the usual types of ciphers (not codes) this "unicity distance" is approximately H(K) / D. Here H(K) is a number measuring the "size" of the key space. If all keys are *a priori* equally likely H(K) is the logarithm of the number of possible keys. D is the redundancy of the language and measures the amount of "statistical constraint" imposed by the language. In simple substitution with random key H(K) is $\log_{10}26!$ or about 20 and D (in decimal digits per letter) is about .7 for English. Thus unicity occurs at about 30 letters.

Correcting the error in calculating $\log_{10}26!$ (which is actually about 26.6), unicity should occur at about 38 characters for a simple substitution with 26! possible keys. Using a different value for D (3.2 bits or .96 decimal digits) Menezes [1] calculated the unicity of such a cipher as 28 characters.

A CLOAKED SUBSTITUTION CIPHER

In this paper we consider a cipher system which operates on plaintext strings over the 26-character alphabet $\mathcal{A}$ comprising the 26 uppercase letters A-Z, and generates ciphertext strings over the 52-character alphabet $\mathcal{A}'$ comprising the 52 uppercase and lowercase letters A-Z and a-z as follows:

Some permutation of 26 of the 52 characters of $\mathcal{A}'$ is chosen as a secret shared 26-character key K, say for example,

Q j u f G C t w b U z S N L q H A g V D O o a n s I

which is used as the basis of a simple substitution cipher which can be displayed as a substitution table, ξ_K

m:	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
ξ_K(m):	Q	j	u	f	G	C	t	w	b	U	z	S	N	L	q	H	A	g	V	D	O	o	a	n	s	I

Then a plaintext message M is encrypted into a ciphertext cryptogram C using the key K by means of the following

Simple Substitution Encryption algorithm: C ⇐ E(K,M)
Replace each character m of M by ξ_K(m).

and a ciphertext cryptogram C is decrypted into a plaintext message M using the key K by means of the following

Simple Substitution Decryption algorithm: **M ⇐ D(K,C)**
Replace each character c of C by $\xi_{K^{-1}}$(c).

The 26 characters of $\mathcal{A}'$ used by a key are said to be that key's ***signal characters***, and the 26 unused characters are its ***noise characters***, so each key K divides $\mathcal{A}'$ into a subset S_K of signal characters and a complementary subset $\mathcal{N}_K$ of noise characters.

For the example key above,

S_K = {A C D G H I L N O Q S U V a b f g j n o q s t u w z}

$\mathcal{N}_K$ = {B E F J K M P R T W X Y Z c d e h i k l m p r v x y}

Now consider the following modification of the simple substitution cipher:

Cloaked Substitution Encryption algorithm: **$C \Leftarrow E^*(K,M)$**

1. $C' \Leftarrow E(K,M)$, i.e., encrypt M with the simple substitution algorithm E using K.
2. Randomly intersperse noise characters into C' producing a string C as follows:

```
i←0   j←0
REPEAT
        Flip a coin
        IF heads
                Randomly choose a noise character n∈ℵK
                increment i
                Ci←n
        ELSE
                increment j
                IF  j ≤ | C′ |
                        increment i
                        Ci← C′j
                ENDIF
        ENDIF
UNTIL  j > | C′ |
```

Cloaked Substitution Decryption algorithm: **$M \Leftarrow D^*(K,C)$**

1. Drop all noise characters from C, leaving a string C' of signal characters.
2. $M \Leftarrow D(K, C')$, i.e., decrypt C' with the simple substitution algorithm D using K.

Considering that there are $52! / 26!^2$ possible ways that 26 characters can be selected from 52, for the cloaked substitution cipher there are $26! \times (52! / 26!^2) \approx 2 \times 10^{41}$ possible keys, and H(K) / D is about 59 using Shannon's value for D, or 43 using Menezes's.

But in fact the unicity distance for this cipher is actually unbounded, as can be shown by the following argument.

Theorem: **The unicity of the cloaked substitution cipher is unbounded.**

Proof:

Two ciphertext strings S and T can be randomly interleaved by the following

Random Interleaving algorithm: **$R \Leftarrow I(S,T)$**

Given two strings S and T, the string R is constructed as follows:

```
i←1   j←1   k←1
WHILE i ≤ |S| + |T|
    Flip a coin
    IF heads
        IF j ≤ |S|
                R_i←S_j
                increment i
                increment j
        ENDIF
    ELSE
        IF k ≤ |T|
                R_i←T_k
                increment i
                increment k
        ENDIF
    ENDIF
ENDWHILE
```

Let us say that two keys K1 and K2 are ***complementary*** in case for each key, $S_{K1} = \mathcal{N}_{K2}$ and $S_{K2} = \mathcal{N}_{K1}$; i.e., for each key, the subset of $\mathcal{A}'$ comprising its signal characters is the subset comprising the other's noise characters. (So for any given key there will be 26! complementary keys.)

If for two such complementary keys, C1 <= E(K1,M1) and C2 <= E(K2,M2) for two messages M1 and M2, and these two cryptograms are randomly interleaved as $C \Leftarrow I(C1,C2)$, then $M1 \Leftarrow D^*(K1, C)$ and $M2 \Leftarrow D^*(K2, C)$.

That is, C is one of the infinitely many possible cloaked substitution encryptions of M1 using K1, and at the same time it's one of the infinitely many possible cloaked substitution encryptions of M2 using K2, so the cloaked substitution cipher has an unbounded unicity distance (since the length of C can be made arbitrarily large).

Q.E.D.

As a whimsical illustration, take as K1 the example key given above and as K2 take the complementary key

W X k v E d c i Z R r m B e K T J x P M F y h l Y p

with substitution table, ξ_K

m:	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
$\xi_{K2}(m)$:	W	X	k	v	E	d	c	i	Z	R	r	m	B	e	K	T	J	x	P	M	F	y	h	l	Y	p

Then the 365-character ciphertext

bMiZaPZQLfGgPMiEGdKfxESPMTqxLZBGSsQEyWVQmuSqMOfDwQiEDBCSqQDVqFLxwbtBFxZwecqTGZgeEoQSPWevMiEGV
iEBQLfwbmKSkSrVaPXwEGLWxQSvSEvhQZMiBKDPqLPuGbWVQaeQvZuegqcaWxBfEeMQwqPcVxEEDqeZCtqSfeGLfQvCZC
qfPbSMZVekjMZeGMiVEbfMGDhwZGSmZQzcGjiMPGLGMWQDewDvmZrwEGvDgxGFGZvPVCKSdODEmvhZMDiGgyKbZLkEPtQ
PLWvWffQevTLxubKTiEMLZktbLPDMwGjgGGIWGevmZrEiWxTExPiKWxhZMiXEWxvPMiWMxEPMKeMiEZxXKPKBP

is decrypted, using K1, as (with spaces added for legibility):

I WANDERED LONELY AS A CLOUD THAT FLOATS ON HIGH OER VALES AND HILLS WHEN ALL AT ONCE I SAW A CROWD A HOST OF GOLDEN DAFFODILS BESIDE THE LAKE BENEATH THE TREES FLUTTERING AND DANCING IN THE BREEZE

but, using K2, is decrypted as:

THIS IS THE FOREST PRIMEVAL THE MURMURING PINES AND THE HEMLOCKS BEARDED WITH MOSS AND IN GARMENTS GREEN INDISTINCT IN THE TWILIGHT STAND LIKE DRUIDS OF ELD WITH VOICES SAD AND PROPHETIC STAND LIKE HARPERS HOAR WITH BEARDS THAT REST ON THEIR BOSOMS

RELATED WORK

Relatively little has been published regarding finite-key ciphers that have unbounded unicity distance. The author is only aware of work by Massey and Ingemarsson [3] [4] and Maurer [5].

POSSIBLE DIRECTIONS FOR FURTHER RESEARCH

It would be of interest to explore what effect interspersing random noise characters into the ciphertext will have on the various hill-climbing based attacks routinely used against substitution ciphers. One promising direction to follow would be to use the previous ploy, bisecting the plaintext, encrypting each half with complementary keys, and randomly interleaving the two encryptions. This could be expected to have two desirable consequences: 1) using one encrypted part of the plaintext as the noise characters to be interleaved with the other encrypted part instead of randomly generating them will prevent their being uniformly distributed, which would clearly

present an exploitable weakness; 2) the fact of there being two distinct solutions may interfere with some hill-climbers converging in certain cases.

CONCLUSION

A slight modification of the classic simple substitution cipher is introduced, in which key-defined noise characters are randomly introduced into the ciphertext in encryption and then ignored in decryption.

For this modified cipher, called a Cloaked Substitution Cipher, calculating the ratio of the number of bits required to express the key divided by the redundancy of English in bits per character, the unicity distance should be 43. Yet, it is shown that for any two plaintext messages M1 and M2 there are keys K1 and K2 and a ciphertext C such that C is decrypted with K1 to M1, and decrypted with K2 to M2.

Inspired by Gil Hayward's [6] recalling his devising a set of wheel patterns to test the Tunny (Lorenz SZ40) cipher machines by encrypting "Now is the time for all good men to come to the aid of the party" (the standard test sentence used by telegraph people) into the opening lines of Wordsworth's lyric poem "I Wandered Lonely as a Cloud," the idea is illustrated by devising two keys one of which encrypts the Wordsworth text and the other encrypts the opening lines of Longfellow's most famous work "Evangeline," each into the same ciphertext.

References

1. A. Menezes, P. van Oorschot, S. Vanstone, *Handbook of Applied Cryptography* (1996), Chapter 7, available at http://cacr.uwaterloo.ca/hac/about/chap7.pdf.
2. C. E. Shannon, Communication theory of secrecy systems. *Bell System Technical Journal* 28 (1949) 659-715, available at http://netlab.cs.ucla.edu/wiki/files/shannon1949.pdf.
3. J.L. Massey, I. Ingemarsson. The Rip van Winkle cipher - a simple and provably computationally secure cipher with a finite key. Proc. IEEE Int. Symp. Information Theory (Abstracts), page 146, 1985.
4. Rip van Winkle cipher. https://en.wikipedia.org/wiki/Rip_van_Winkle_cipher
5. Ueli M Maurer, Conditionally Perfect Secrecy and Provably Secure Randomized Cipher. *Journal of Cryptology* vol. 5, no. 1, pp 53-66 1992, available at ftp://ftp.inf.ethz.ch/pub/crypto/publications/Maurer92b.ps
6. https://www.telegraph.co.uk/news/obituaries/technology-obituaries/8837249/Gil-Hayward.html