Spring 2012
Chris Christensen

Simple Substitution Ciphers

The art of writing secret messages – intelligible to those who are in possession of the key and unintelligible to all others – has been studied for centuries. The usefulness of such messages, especially in time of war, is obvious; on the other hand, their solution may be a matter of great importance to those from whom the key is concealed. But the romance connected with the subject, the not uncommon desire to discover a secret, and the implied challenge to the ingenuity of all from who it is hidden have attracted to the subject the attention of many to whom its utility is a matter of indifference.

Abraham Sinkov
In *Mathematical Recreations & Essays*
By W.W. Rouse Ball and H.S.M. Coxeter, c. 1938

We begin our study of cryptology from the romantic point of view – the point of view of someone who has the "not uncommon desire to discover a secret" and someone who takes up the "implied challenged to the ingenuity" that is tossed down by secret writing. The material in this section will help you do the *Quiptoquip* in the morning newspaper, and it is excellent preparation for an appearance on the gameshow *Wheel of Fortune*. (And, it will prepare you for our future work.)

A simple substitution cipher is a method of concealment that replaces each letter of a plaintext message with another letter (or, perhaps, a symbol). Here is the key to a simple substitution cipher:

Plaintext letters: `abcdefghijklmnopqrstuvwxyz`
Ciphertext letters: `EKMFLGDQVZNTOWYHXUSPAIBRCJ`

The key gives the correspondence between a plaintext letter and its replacement ciphertext letter. (It is traditional to use small letters for plaintext and capital letters, or small capital letters, for ciphertext. We will not use small capital letters for ciphertext so that plaintext and ciphertext letters will line up vertically.) Using this key, every plaintext letter `a` would be replaced by ciphertext `E`, every plaintext letter `e` by `L`, etc. The plaintext message `simple substitution cipher` would become `SVOHTL SAKSPVPAPVYW MVHQLU`.

The key above was generated by randomly drawing slips of paper with letters of the alphabet written on them from a bag that had been thoroughly shaken to mix up the slips. The first letter drawn E became the substitution for a, the second letter drawn K became the substitution for b, etc.

Encryption (or enciphering) is the process of using the key to produce ciphertext from plaintext. **Decryption** (or deciphering) is the process of using the key to produce plaintext from ciphertext.

To encrypt a message requires knowing two things: the method of encryption (in our case, simple substitution) and the key (in our case, the letter substitutions). Notice that if we believed that our messages were no longer secure, we could leave the method unchanged (simple substitution) but change the key (use different letter substitutions).

Here is a message to decrypt. It has been encrypted with a simple substitution cipher with key:

Plaintext letters: abcdefghijklmnopqrstuvwxyz
Ciphertext letters: HUFRCOGMTZXLKPNWYVABQSIEDJ

BMC XTP MHBM PNBC NO HLL BMHB BMCD TPBCPR,
UD TPBCVFCBTNP IMTFM BMCD RVCHK PNB NO.

Decrypt the message. Knowing the key, this should not be a problem. Although it might be useful to have the ciphertext letters in alphabetical order for decryption, the key is the same for encryption and decryption.

Plaintext letters: steyxdgawzmlhofnudvibrpkqj
Ciphertext letters: ABCDEFGHIJKLMNOPQRSTUVWXYZ

But, how would a person solve the message not knowing the key? Solving the message not knowing the key is called **cryptanalysis**. Cryptanalysts take up the "implied challenged to the ingenuity" that is tossed down by secret writing, and they find, when successful, satisfaction of their "not uncommon desire to discover a secret."

Brute Force

If the cryptanalyst knew that the method of encryption were simple substitution cipher, then the cryptanalyst could try all possible keys to solve the message. Or, maybe not! How many keys are possible? How long would it take to try them all?

When constructing a key for a simple substitution cipher, there are 26 choices of letters to substitute for a, then 25 remaining letters that can be substituted for b, then 24 remaining letters that can be substituted for c, etc. This results in

$$26 \times 25 \times 24 \times 23 \times ... \times 3 \times 2 \times 1 = 26!$$

possible keys. That's a lot of keys; in fact, there are

403,291,461,126,605,635,584,000,000

keys.

Now, not all of these would make good choices for a key. One of the choices is plaintext, and others keep many plaintext letters unchanged. If many common plaintext letters remained unchanged, it would not be much of a challenge to cryptanalyze the ciphertext message.

The security of cryptosystems often depends on forcing the cryptanalyst into doing a brute force attack – forcing the crypanalyst to try all possible keys – and "having a large keyspace" – having too many possible keys to making trying them all practical.

> *Cardano [an Italian mathematician, 1501 – 1576] heads a long line of cryptographers in erroneously placing cryptographic faith in large numbers – a line that stretches right down to today. … Cryptanalysts do not solve [simple substitution ciphers] – or any cipher for that matter – by testing one key after another. … If the cryptanalyst tried one of these [403,291,461,126,605,635,584,000,000 possibilities] every second, he [or she] would need*

$$\frac{403{,}291{,}461{,}126{,}605{,}635{,}584{,}000{,}000}{60\times 60\times 24\times 365} \approx 1.2788\times 10^{19} \text{ years]} \ldots$$

> *to run through them all. Yet most[simple substitution ciphers] are solved in a matter of minutes.* David Kahn, *The Codebreakers: The comprehensive history of secret communication from ancient times to the internet*, Scribner, 1996.

Ok, so it is not a good idea to try to solve one of these by brute force. Would a computer do better? Yes, a computer would do better. Computers now provide an alternative to hand checking of possible keys, but even checking 1000 or 10,000 keys per second wouldn't make a significant dent in the time required to check all possibilities. Brute force attack is just not a good attack. It is certainly not an elegant way of cryptanalysis.

Discovering Patterns

How are simple substitution ciphers attacked? By finding patterns. Every language has rules so that the language "makes sense." There are rules for punctuation, there are rules for combining letters, there is word length, … . These rules create patterns in messages that can be exploited by cryptanalysts. Usually cryptograms that appear in newspapers preserve word length and punctuation, and they preserve letter frequencies. For example, `e` is the most frequent letter in plaintext English. If we used the key

```
abcdefghijklmnopqrstuvwxyz
EKMFLGDQVZNTOWYHXUSPAIBRCJ
```

we would expect that the most frequent ciphertext letter would be L. Now, it might not be, but it is likely that the most frequent ciphertext letter corresponds to one of the most frequent letters `e`, `t`, `a`, `o`, `i`, `n`, or `s`. An attack on ciphertext that uses letter frequencies is called **frequency analysis**. Using letter frequencies and other patterns, cryptanalysts are usually able to quickly solve simple substitution ciphers.

Cryptanalysis

Here is a cryptogram that was taken from a local newspaper.

D RNXHT VHRVCK VKKXOW FYVF V OVFY

GENBWKKNE'K PWEC BVPNEDFW TWKKWEF DK GD.

This puzzle is called a *Cryptoquip*. The method used for encrypting it was simple substitution. It obeys the traditional rule for such puzzles that no letter is encrypted as itself. This is very useful information. For example, in this message we know that PWEC cannot be the ciphertext for when.

If you did this puzzle daily, you would become familiar with the puzzler's writing style. You would know that the plaintext message is a humorous statement. Information about the writing style of the sender or the nature of the plaintext message is often available to cryptanalysts. Use it.

Often cryptogram puzzles give a clue – typically one plaintext/ciphertext correspondence is given. We will attack this message without a clue.

Even though this puzzle might not require all the effort that we will spend on it, we will try to establish a pattern by collecting a great deal of information prior to starting the cryptanalysis.

On the next page is that form that we will use to gather information from the ciphertext.

CIPHERTEXT

Most frequent English letters: `etaoins`
Ciphertext frequencies

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

1-letter English words: `a i`
1-letter ciphertext words

Most frequently doubled letters in English: `setflmo`
Doubled letters in ciphertext:

Most frequent 2-letter words in English: `an, at, as, he, be, in, is, it, on, or, to, of, do, go, no, so, my`
2-letter ciphertext words:

Most frequent 3-letter words in English: `the, and, for, was, his, not, but, you, are, her`
3-letter ciphertext words:

Most frequent initial letters in English: `tasoi`
Initial letters of ciphertext strings:

Most frequent final letters in English: `esdnt`
Final letters of ciphertext strings:

Plaintext letters used: `abcdefghijklmnopqrstuvwxyz`

Here is the information that was gathered about the ciphertext

D RNXHT VHRVCK VKKXOW FYVF V OVFY

GENBWKKNE'K PWEC BVPNEDFW TWKKWEF DK GD.

Most frequent English letters: etaoins

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

The five most frequent letters appear above in **bold**.

1-letter English words: a i

One-letter words: D, V

Most frequently doubled letters in English: setflmo

Doubled letters: K, K, K

Most frequent 2-letter words in English: an, at, as, he, be, in, is, it, on, or, to, of, do, go, no, so, my

Two-letter words: DK, GD

Most frequent 3-letter words in English: the, and, for, was, his, not, but, you, are, her

Three-letter words:

Most frequent initial letters in English: tasoi

V

Initial letters: R V F O P B T D G

Most frequent final letters in English: esdnt

K W F

Final letters: T K W F Y C D

Here's a cryptanalysis of the message.

We begin with the one-letter words D and V. V is more frequent than D; so, it is likely that V is a and D is i. Put those in place above the letters of the ciphertext.

Usually we would hunt for a three-letter word that could be the, but there are no three-letter words in this *Cryptoquip*.

Notice the 'K. This suggests that K could be s. Because K is doubled and K appears often as a final letter, there is additional information suggesting that K is s. Put that in place. Additional confirmation that our choice is correct comes from noting that DK becomes is.

Notice ass_ _ _ with the final letter being high frequency. This suggests that X is u and O is m and W is e. Put those in place.

Notice FYaF. F is a high frequency initial and final letter. This is likely to be that. Put those letters in place.

We have now identified all the high frequency ciphertext letters other than E.

Notice math _ _ _ _ e s s _ _ 's. Doesn't math professor's just leap out? Put those letters in place.

We still do not seem to have any contradictions.

Everything comes together quickly now:

f a _ o r i t e suggests that P = v.

v e r _ suggests that C = y.

_ e s s e r t suggests that T = d.

_ o u _ d suggests that H = l.

a l _ a y s suggests that R = w.

Done! Funny?

We have the plaintext message, and we have much of the key:

```
abcdefghijklmnopqrstuvwxyz
V  TWB YD  HO NG  EKFXPR C
```

We have the ciphertext letters that correspond to almost all of the most frequent plaintext letters. Given another message encrypted with the same key, we could probably make sense of it, and after several additional messages, we could probably complete the key.

Ciphertext Attack

The type of attack that we made on the cryptogram is called a **ciphertext attack**. What was available to us was only a ciphertext message. We will discuss other types of attack later.

Definitions

Mathematicians make careful definitions so that they can use words to precisely describe situations.

Initially, we will be a little vague and will be satisfied with getting an idea what words mean.

For example, it's hard to define what we mean by a key.

Here are some words that describe a simple substitution cipher. A simple substitution cipher is a *monographic* cipher. **Monographic** means that a single ciphertext letter replaces a single plaintext letter. A simple substitution cipher is also a *monoalphabetic* cipher. **Monoalphabetic** means that a *single* alphabet is used to replace plaintext characters with ciphertext characters. The ciphertext alphabet is just a rearrangement of the plaintext alphabet. It might not be obvious yet what a polyalphabetic cipher would look like, but we will encounter several of those later.

Self-Reciprocal Keys

A key is said to self-reciprocal if it "swaps letters." For example, if plaintext a becomes ciphertext R, then plaintext r becomes ciphertext A.

If $a \rightarrow R$, then $r \rightarrow A$; or $a \leftrightarrow r$.

The Old Testament of The Holy Scriptures contains several examples of a self-reciprocal simple substitution cipher called *atbash*. In Jeremiah 25: 26 and 51: 41, SHESHACH is substituted for *Babel* ("Babylon"); and in Jeremiah 51:1 LEB KAMAI ("heart of my enemy") is substituted for *Kashdim* ("Chaldeans").

> *Both transformations resulted from the application of a traditional substitution of letters called "atbash," in which the last letter of the Hebrew alphabet replaces the first, and vice versa; the next-to-last replaces the second, and vice versa; and so on. …* David Kahn, *The Codebreakers.*

Using the English alphabet and this scheme, we would have the following key:

```
abcdefghijklmnopqrstuvwxyz
ZYXWVUTSRQPONMLKJIHGFEDCBA
```

Notice that the key is self-reciprocal – if W substitutes for d, then D substitutes for w, etc.

There is no need for separate encryption and decryption keys; the same key works equally well for both processes.

Difficulties with the Key

One problem with the key that is always faced by the cryptographer is how to distribute the key to people who are authorized to use it; there must be some way for the sender and receiver to agree on and exchange the key.

Traditionally this has been done face to face by means of a trusted courier. This is the **key distribution** problem.

Consider a key for a simple substitution cipher. If the ciphertext alphabet is randomly arranged, it is unlikely that the sender and receiver can memorize the plaintext-ciphertext letter correspondence; so, it is likely that the key must be written. That leaves open the possibility that an unauthorized person might steal the key and then be able to break all messages enciphered by it. This is the key security, or **key management**, problem.

Finally, it is often assumed by the receiver that any received message that has been properly encrypted comes from an authorized sender. But, if the key has been stolen, an unauthorized sender might be sending messages. Various prearranged authentication schemes have been used; e.g., the sender and receiver might agree that the sender will in each message make three "mistakes" of substituting an `x` for a `t`. This is the **authentication** problem.

Exercises

First, a few general words about exercises.

1. The plaintext of the ciphertext messages that occur in the exercises are often taken from cryptology or history texts. No attempt has been made to attribute these statements.

2. One tool that is of value to cryptanalysts is knowing something about the nature of the plaintext message. Sometimes this information comes from having seen other messages written by the sender. You will, while solving ciphertexts, recognize some interests of the sender. Use that information.

3. No errors have intentionally been made while encrypting messages. However, errors are common during encryption, and errors are present in some of the ciphertexts you will encounter in these notes. Cryptanalysts must deal encryption errors; so, errors that have occurred have not been corrected.

1. Construct a key for a simple substitution cipher. Explain how you chose your plaintext-ciphertext letter correspondence. Then use your key to encrypt the message

> During World War I, much of the early success at sea enjoyed by the Allied Powers stemmed directly from the Russian recovery of German radiotelegraphic codebooks from the cruiser Magdeburg, which had run aground in the Baltic Sea.

2. In your key, did any letter substitute for itself? How many simple substitution keys have no letter substituting for itself?

3. During World War II the Japanese Navy used a cipher that randomly scrambled the six vowels (a, e, i, o, u, y) among themselves and the 20 consonants among themselves. How many such keys are there?

4. Decrypt the following message that was encrypted with a simple substitution cipher and the following key:

Plaintext	abcdefghijklmnopqrstuvwxyz
Ciphertext	ZRGHLWPXYCQIDUKFJNEASTMBVO

Ciphertext message:

NKVZI WIZPE MZTL QYUPE ZRKTL

The plaintext message is a mnemonic to remember the turnover positions on the first five Enigma rotors.

Would you have found a “decryption key” useful?

5. Write a key for a simple substitution cipher that is self-reciprocal.

6. Decrypt the following message that was encrypted with a simple substitution cipher and the following key:

Plaintext	abcdefghijklmnopqrstuvwxyz
Ciphertext	YNFROTMKPHELQWBDJXZAUSVCGI

Ciphertext message:

PWMBR VOAXU ZAYLL BAKOX ZVOQB WPABX

The plaintext message is an “intercept operator’s motto.”

Word length was not given. Did that create a problem for decryption?

Would you have found a “decryption key” useful?

7. Before attacking some cryptograms, here are some warm-up exercises to get you thinking "the right way:"

7a. R and Q appear as one-letter words in a cryptogram. R has frequency 7, and Q has frequency 3. Which ciphertext letter is likely to be the substitute for a, and which is likely to be the substitute for *i*? Explain your reasoning.

7b. The two-letter word KC appears in a cryptogram. K has frequency 5, and C has frequency 2. One of K and C is a consonant, and the other is a vowel. Which is likely to be which? Explain your reasoning.

7c. BKKL appears in a cryptogram. What are some possible plaintext words for this four-letter encrypted word?

7d. EOQE'D appears in a cryptogram. What plaintext letters are likely to correspond to ciphertext D? Explain your reasoning.

7e. LWW appears in a cryptogram. What are some possible plaintext words for this three-letter encrypted word?

7f. Y'P appears in a cryptogram. Which plaintext letter is likely to be substituted by Y, and which by P?

7g. HOWX'JW appears in a cryptogram. W is the highest frequency letter. What is the likely plaintext of this string?

7h. AVBA appears in a cryptogram. A = t. B has high frequency. What is the likely plaintext of this string?

7i. "RYK." appears at the end of a sentence in a cryptogram. R and K are each high frequency letters. Is it likely that R = t, Y = h, and K = e? Explain your reasoning.

7j. PVDY APE appears in a cryptogram. V = h, D = e, and A = t. What is the likely plaintext?

7k. DAXOIKDD appears in a cryptogram. D is a high frequency letter. What is the likely plaintext word?

7l. F UIJMDL appears in a cryptogram. F = i. Is it likely that UIJMDL substitutes for a verb or a noun?

7m. S ZCA appears in a cryptogram. S = a. What can you say about the plaintext of ZCA?

7n. KC appears twice in a cryptogram, and BKKL appears once. Any guesses as to the plaintext for K?

7o. UUJLQUJYN appears in a cryptogram. U is a high frequency letter. Any guesses as to the plaintext?

7p. PYJJHHO appears in a cryptogram. Any guesses as to the plaintext?

7q. MPN and MP'N both appear in a cryptogram. What plaintext letters correspond to M, P, and N?

7r. QWAPSNUDNB appears in a cryptogram. Is it possible that the plaintext word ends in ing?

8 Cryptanalyze the following ciphertext message that was encrypted with a simple substitution cipher. (A letter may substitute for itself.)

```
WOR  YRSZGN  ASDEPRS  ZGYVRJDSY  MGP  BGSW  LT
WOR  JGFWEA  TFRRW
```

Write as much of the key as you are able to determine.

Can you extend the key? Why or why not?

In a few sentences describe the process you used to cryptanalyze the message.

9 The following two ciphertexts were encrypted using a simple substitution cipher and the same key. Cryptanalyze them.

Ciphertext message number one:

```
GEQ  QPKWVC  JCD  CP  QYOQTTOPG  VCOEKPQ
```

Ciphertext message number two:

```
GEQ  WQNVCP  DQOBNKGA  DXQOKCTKDGD  DCJ  PZ
JCA  GECG  KG  OZBTH  FQ  FNZLQP
```

Write as much of the key as you are able to determine.

Can you extend the key? Why or why not?

In a few sentences describe the process you used to cryptanalyze the message.

10 Here is a plaintext message: `alan turing was a prodigy`. Sometimes in an effort to increase security a message is encrypted more than once. We will encrypt this message once using a simple substitution cipher with one key and then re-encrypt it using a simple substitution cipher with a second key.

The first key:

Plaintext	`abcdefghijklmnopqrstuvwxyz`
Ciphertext	`KPFHIGLDEXCVTOUBJQZMRNAYSW`

Plaintext message:	`alan turing was a prodigy`
Ciphertext #1	`KVKO MRQEOL AKZ K BQUHELS`

The second key:

Plaintext	`abcdefghijklmnopqrstuvwxyz`
Ciphertext	`XFCZIJRNATWSQOLUBMVPYHKGDE`

Ciphertext #1	`KVKO MRQEOL AKZ K BQUHELS`
Ciphertext #2	`WHWL QMBILS XWE W FBYNISV`

The ciphertext message that is transmitted is `WHWL QMBILS XWE W FBYNISV.`

The result of re-encrypting the message is called a composition of the two ciphers. What kind of cipher results from composing these two simple substitution ciphers? (There aren't many choices are there?)

What is the key for the composed ciphers?

Has re-encryption increased the security of the cipher?

11 The following ciphertext message is a patristocrat; i.e., word length and punctuation are not given. Try to cryptanalyze the message. Letter contact is an important tool in solving patristocrats (see the appendix). The message is a *Cryptoquip*. This is hard!

```
JPFXB    XYXFQ    EBCLP    FCERR    VBNPR

EBEWF    YSGLV    FSBXS    EPGNW    BSFBX

BCCRJ    FNLBQ    F
```

In a few sentences describe the process you used to cryptanalyze the message.

David Kahn notes that Giovanni Battista Porta (1535 – 1615), one of the outstanding cryptologists of the European Renaissance, "gave the first published description in Europe of how to solve a monoalphabetic cipher with no word divisions or with false word divisions, at a time when cryptanalysts often depended on the presence of word divisions." Kahn continues by quoting Porta's advice on work techniques:

> *There is required the most complete concentration, the most perfect diligence, so that the mind, free from all distracting thoughts, and with everything else put aside, may devote itself entirely to the single task of carrying the whole understanding to a successful conclusion. Still, if the task sometimes requires unusual concentration and expenditure of time, this concentration should not go on uninterrupted; the brain should not be racked over-anxiously. For excessive pains and prolonged mental effort bring on brain-fag, so that the mind is afterwards less fit for these things and accomplishes nothing. … This has often been my experience at such times as I came upon particularly involved ciphers, in the working-out of these. For after spending the whole day in this task (scarcely seven or eight hours seemed to me go have gone by), I hardly thought it was more than one or two o'clock, so I was not aware of the approach of evening except through the shadows and the failing of light.*

12 The following ciphertext is also a patristocrat, but it might be easier to cryptanalyze than the ciphertext in exercise eleven because for this message the sender has inserted the plaintext letter `x` to separate plaintext words. What is likely to be the most frequent plaintext letter?

Here is the ciphertext:

```
JAGJS  LZRCS  ROWAR  OSRTD  OWUTL  TOVLZ  JROWT
LTOGN  POAKA  ZFAJF  T
```

In a few sentences describe the process you used to cryptanalyze the message.

13 The following message was encrypted with a simple substitution cipher. It has been partially cryptanalyzed. The letters `e, t, a, o, i, n, s` have been inserted into the ciphertext. Complete the cryptanalysis.

```
tMe ZaWanese VeWLaFeR tMe VeR KaFMine IitM a
KaFMine to Ue XnoIn as WQVWLe
```

In a few sentences describe the process you used to cryptanalyze the message.