Fall 2019
Chris Christensen
MAT/CSC 483
Section one

Monoalphabetic Substitution Ciphers (MASCs)

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.

A **monoalphabetic substitution cipher (MASC)** is a method of concealment that replaces each letter of a plaintext message with another letter consistently throughout the message[1]. Here is the key to a MASC:

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

The key gives the correspondence between a plaintext letter and its replacement ciphertext letter[2]. Using this key, every plaintext letter `a` would be replaced by ciphertext `E`, every plaintext letter `e` by `L`, etc. The plaintext

[1] The term monoalphabetic (one alphabet) is used because traditionally it is said that there is "one alphabet;" i.e., one ordering of the ciphertext letters against the plaintext letters.
[2] 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. Also, we will use a nonproportional font, like Courier New, for both plaintext and ciphertext so that all letters are the same width and, therefore, plaintext and ciphertext can be aligned vertically.

message `simple substitution cipher` would become `SVOHTL SAKSPVPAPVYW MVHQLU`:

```
simple substitution cipher
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[3] (in our case, MASC) 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 MASC 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`

The word(s) decryption (or deciphering) implying action by an authorized receiver – someone who was given the key.

[3] In modern terminology, this would be called the encryption algorithm.

But, how would a person solve the message not knowing the key? Solving the message by someone who is not authorized to know the key is called **cryptanalysis**[4]. **Cryptanalysts** (people who do cryptanalysis) 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."

People who construct ciphers are called **cryptographers**; the construction of ciphers is called **cryptography**.

Cryptology consists of two parts: cryptography and cryptanalysis.

Brute Force

If the cryptanalyst knew that the method of encryption were a MASC, 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

$$26! = 403{,}291{,}461{,}126{,}605{,}635{,}584{,}000{,}000$$

keys.

Now, not all of these would make good choices for a key. One of the 26! choices is plaintext (every letter substitutes for itself), and other choices keep many plaintext letters unchanged. If many common plaintext letters

[44] The generic term for cryptanalysis is "codebreaking." Cryptanalysts are often called "codebreakers." However, codes and ciphers are not the same. Codes are often used to preserve information, and ciphers are used to conceal it. Codes tend to substitute for words and phrases; they are linguistic objects.

remained unchanged, it would not be much of a challenge to cryptanalyze the ciphertext message.

A well-designed cipher forces the cryptanalyst into doing a **brute force attack**; i.e., trying all possible keys. The security of the cipher then depends on "having a large keyspace" – having too many possible keys to making trying all of them 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 [MASCs] – 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 MASC keys] every second, he [or she] would need*
>
> $$\frac{403{,}291{,}461{,}126{,}605{,}635{,}584{,}000{,}000}{60\times 60\times 24\times 365} \approx 1.2788\times 10^{19}$$ *years] ...*
>
> *to run through them all. Yet most [MASC] 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, of course, using a computer to try all possible keys would be faster than trying them by hand, but even checking 1000 or 10,000 keys per second wouldn't make a significant dent in the time required to check all possibilities[5].

If a cipher is properly designed, a brute force attack should not work[6].

Discovering Patterns

So, how are 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, … .

[5] 26! Takes slightly more than 88 binary digits to write in binary. So, having 26! Possible keys corresponds to 88-bit security. Today the minimum security level is usually 128 bits.
[6] Even when it does work, it is not an elegant method of cryptanalysis.

Cryptographers attempt to design ciphers that remove these patterns and make the ciphertext appear to be random. However, for all but one method of encryption[7], ghosts of these patterns remain in the ciphertext, and it is these "ghostly patterns" that can be exploited by cryptanalysts to recover the plaintext.

Frequency Analysis

MASCs "preserve letter frequencies." Ciphertext letters inherit the frequencies of the plaintext letters they replace. For example, `e` is the most frequent letter in plaintext English. If we used a MASC with 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 in English: `e`, `t`, `a`, `o`, `i`, `n`, or `s`. An attack on ciphertext that uses letter frequencies is called **frequency analysis[8]**. Using letter frequencies (and other patterns), cryptanalysts are usually able to quickly solve MASCs.

[7] Bell Labs scientist Claude Shannon (1916 – 2001) in the *Bell Labs Technical Journal* (1949) proved that the one-time pad (OTP) is the only "perfectly secure" cipher. Fortunately for cryptanalysts the OTP is difficult to implement.

[8] Frequency analysis was described in the ninth century ay Al-Kindi, an Arab polymath. The idea was known in Europe in the fifteenth century. It appears that the idea of frequency analysis developed independently (but later) in Europe.

MASC Cryptanalysis Example

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 MASC.

It obeys the traditional rule for such puzzles that no letter is encrypted as itself. This is very useful information[9]. 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 and aids in cryptanalysis.

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 can be used to gather information from the ciphertext.

[9] Cryptograms for which the method of enciphering is MASC, world length and punctuation are given, and no letter enciphers as itself, are traditionally called "aristocrats."

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`

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

Most frequent final letters in English: `esdnt`

Final letters: `T K W F Y E C W F K D`

Here's a cryptanalysis of the message given above:

We begin with the one-letter words D and V[10]. V is more frequent than D; so, it is likely that V is a and D is i[11]. 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 message. 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?

[10] a and I are the only one-letter words in English.
[11] a is a more frequent letter in English than is i.

We have the plaintext message, and we have recovered 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 recover the complete key.

Ciphertext Attack

The type of attack that was 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.

Permutations

The key for a MASC can be thought of as establishing an encryption function *E* from the "space" of plaintext characters {a, b, c, … z} onto the "space" of ciphertext characters {A, B, C, … Z}.

Consider the key

Plaintext letters: abcdefghijklmnopqrstuvwxyz
Ciphertext letters: YNROTKMCPBDVXZALEWUSFQJHGI

The encryption function *E* maps

a → Y, b → N, c → R, …, z → I.

The domain of the encryption function is the set of plaintext characters, and the range is the set of ciphertext characters. The encryption function *E* is one-to-one (i.e., no two plaintext characters are mapped to the same ciphertext character) and onto (i.e., every ciphertext character is the image of some plaintext character).

Domain of *E*: abcdefghijklmnopqrstuvwxyz
Range of *E*: YNROTKMCPBDVXZALEWUSFQJHGI

It is necessary for the encryption function to be one-to-one[12] because the encryption function must have an inverse – the decryption function *D*.

In mathematics, if the domain and the range of a function are the same finite set and the function is one-to-one from the ordered domain onto the ordered

[12] Recall that a function that is one-to-one has an inverse; in this case, that means that if we are given a ciphertext character, there is a unique plaintext character associated with it.

range, the mapping is called a **permutation** of the set. A permutation of an ordered set is just a rearrangement of its elements. We can, by assuming that the plaintext and ciphertext characters are the same set (i.e., we will not distinguish between the small characters and capital characters), regard a MASC encryption function as describing a permutation – a rearrangement -- of the letters of the alphabet.

```
YNROTKMCPBDVXZALEWUSFQJHGI
```

Key

It can be hard to define what we mean by **key** in an informal way.

For a MASC the key is the plaintext/ciphertext letter correspondences. For example:

```
abcdefghijklmnopqrstuvwxyz
EKMFLGDQVZNTOWYHXUSPAIBRCJ
```

In a more formal way, we can think of the MASC enciphering function *E* as being a two-variable function *E*(letter, key) where we can think of the key as "setting" the enciphering function so that when letters are substituted they are enciphered in the correct way.

Key Establishment

How are keys distributed to people who are authorized to use them. Traditionally this has been done face-to-face by means of a trusted courier. The same key is distributed to both the sender and the receiver. But Amazon, for example, does not want to have to send out thousands of couriers with keys to be able to engage in internet commerce with people they have never met. Since the advent of internet commerce and communication a different way is needed to distribute keys, and what have resulted are called public-key algorithms. For public-key ciphers, two keys are created – an enciphering key for the sender (which can be public and, therefore, sent over the internet) and a deciphering key for the receiver (which must be kept private).