Cryptography Basics

Cryptography

The idea of an **encryption system**, or **cipher**, is to code a message which is to be passed through a public channel so that anyone reading the encrypted message cannot recover the original message without some extra information which is not publicly available. Such secret information is called a **key**. We require an ecryption function E, a decryption function D, and a key K. Often the encryption scheme/cipher involves an algorithm to make use of these elements.

A message m (the **plaintext**) is transformed into its **ciphertext** $c = E(m, K)$, and decryption recovers the original message, so $m = D(c, K)$. By convention, we tend to write plaintext messages in lower case and ciphertext in upper case, though this is not always the case.

We will look at some basic classical ciphers which make use of modular arithmetic or permutations, motivating the principles behind cryptography. We will later move on to more modern systems of encryption which involve more advanced mathematical ideas and techniques.

Transposition Ciphers

The key in a **transposition cipher** is simply a permutation where the ciphertext is just an anagram of the plaintext. That is, the letters in the plaintext message are swapped around according to some rule in order to generate the ciphertext. For example, a very simple transposition cipher is to just write everything backwards. So the plaintext '`mathematics`' would becomes '`SCITAMEHTAM`'. This might be fairly secure against small children, but it probably wouldn't take anybody else that long to spot what is going on.

With longer text and a complex enough permutation, this may prove difficult to unscramble, but it is also an unwieldy form of communication.

Rail Fence Ciphers

Rail ciphers, also called zigzag ciphers, are a type of transposition cipher which are very much like a riffle shuffle that one can do with a pack of playing cards. In a riffle shuffle the deck of cards is cut into two halves, the Top and the Bottom. Then the top card of the Top remains the top card of the shuffled deck, but the second card of the shuffled deck is the top card of the Bottom. The cards are

then alternately added to the new deck, one from the Top, then the Bottom, then the Top again, etc.

The reason that the cipher is called a rail cipher is essentially because we can write the original plaintext out on two 'rails'.

m		t		e		a		i		a	
	a		h		m		t		c		l

We then generate the ciphertext by reading the top rail first, followed by the second rail. For the plaintext '`mathematical`', this would generate the ciphertext '`MAATTIHCEAML`'.

We can also increase the number of rails in the cipher. For example, with three rails, the word '`mathematical`' becomes '`MEIHAHMTCLTAA`'.

Rail ciphers are quite similar to another type of transposition cipher, called a Scytale cipher.

Substitution Ciphers

Whereas a transposition cipher keeps the same letters as the original messages, a **substitution cipher** replaces each letter in the original message. Many substitution ciphers are **monoalphabetic** ciphers, meaning that each letter in the plaintext corresponds to a letter in the ciphertext. Examples of these below are the Caesar cipher and the affine cipher. Other substitution ciphers, such as the Vigenère cipher, are **polyalphabetic** ciphers, meaning that each plaintext letter may be encrypted as multiple different letters in the ciphertext.

Caesar Cipher

A simple (and insecure) example of an encryption system simply permutes the letters of a message in English. The key is a permutation, often denoted π (not the number). If, for example, π is the permutation that shifts each letter in the alphabet along by one (e.g., $a \mapsto B$, $b \mapsto C$, ..., $z \mapsto A$), then

$$E(\text{'the cat sat on the mat'}, \pi) = \text{'UIF DBU TBU PO UIF NBU'}$$

and decryption consists simply by applying the permutation in the opposite direction (shifting back one letter).

A cipher of this type, where each letter is moved on a fixed number of places in the alphabet, is called a **Caesar cipher**. Julius Caesar is reputed to have

m				e				i			
	a		h		m		t		c		l
		t				a				a	

used this method, moving letters three places. We can make the description of this cipher slightly more mathematical using modular arithmetic.

There is a one-to-one correspondence between the sets

$$\{a, b, c, \ldots, x, y, z\}$$

and

$$\{0, 1, 2, 3, \ldots, 24, 25, 25\},$$

where a corresponds to 0, b corresponds to 1, and so on up to z corresponding to 25. We can use the table below for reference.

a	b	c	d	e	f	g	h	i	j	k	l	m
0	1	2	3	4	5	6	7	8	9	10	11	12
n	o	p	q	r	s	t	u	v	w	x	y	z
13	14	15	16	17	18	19	20	21	22	23	24	25

The Caesar cipher then relies on a shift number k, which acts as the key. This shift k is an element of the set $\{0, 1, 2, 3, 4, \ldots, 25\}$. If α represents the corresponding number from the table above of a single letter in our message, then the encryption function is given by

$$E(\alpha, k) = \alpha + k \mod 26.$$

The use of mod 26 is what allows the shift to 'wrap around' when it gets to `Z`. The first example above had a shift value of $k = 1$, while Julius Caesar's version of the cipher had a shift value of $k = 3$. Obviously, a trivial shift value of $k = 0$ won't be very effective!

The decryption key works in much the same way.

$$D(\alpha, k) = \alpha - k \mod 26.$$

The problem with this sort of cipher is that it is very insecure. If somebody intercepts a large amount of ciphertext encrypted by such a cipher and suspects that this cipher was indeed used, then it doesn't take long to write a program or even create a spreadsheet in order to check and display all possible shifts to reveal the original message.

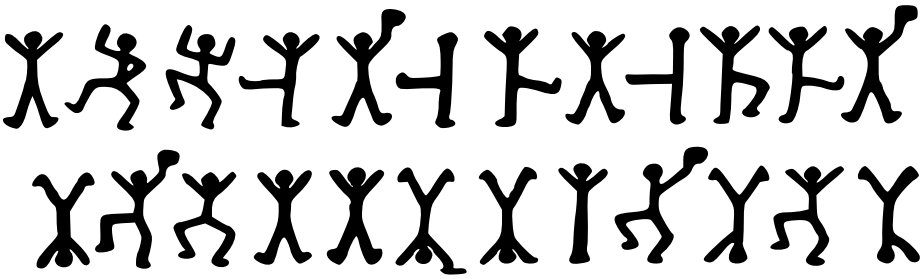
It can even be decrypted quickly by hand by making use of frequency analysis (see the exercises). Currently `e` is the most frequent letter in the English language. Finding the most frequent letter in the ciphertext may then let you guess that this corresponds to `e` in the plaintext and so easily discover the shift value.

Frequency Analysis

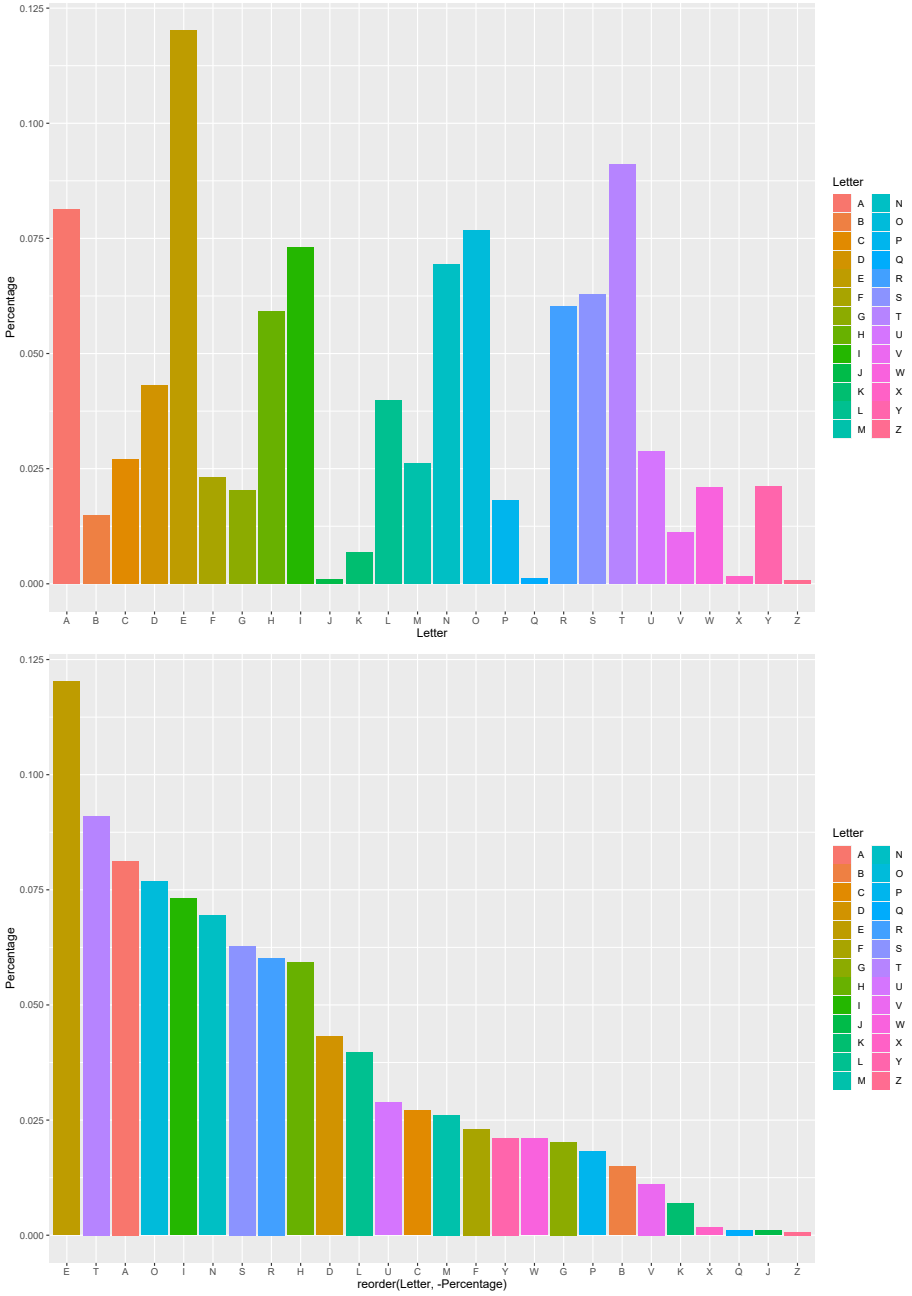
Simple substitution ciphers can be fairly easy to break using the method of **frequency analysis**. The letters in the alphabet can be ranked by how often they appear in written texts. For example, in English, the most common letter is `E`, followed usually by `A` or `T`. The frequencies are not static since more texts are always being written and the use of language changes over time and between disciplines.

However, we can still use this principle of letter frequency to analyse messages which we suspect have been encrypted using a substitution cipher. For example, if the most common letter of the message we have intercepted is `Q`, then it is likely that this corresponds to an `E` in the original message. Similarly for the second most common letter, and so on. The frequencies will not always match up exactly, since the intercepted message may not be very long. For this reason, this method usually works more reliably on longer texts. Another weakness is in the groupings of letters - double letters will still be double letters, and common digraphs (double letter combinations) such as `CH`, `TH`, and `SH` will start to become apparent.

If the encipher is not too savvy, they might even leave punctuation in the text, including spaces. This has the effect of allowing the codebreaker to identify single letter words, such as A or I, and common two-letter words such as `IN`, `AS`, `OR`, and `ON`.



Some substitution ciphers substitute different symbols for the original letters, in order to further disguise the content of the message. Sometimes even whole words were replaced by single symbols. Two famous examples of this are the messages sent between Mary Queen of Scots and her approval of the assassination plot of her cousin, Queen Elizabeth I - further examples of messages sent by Mary have been discovered much more recently. A fictional example is found in Sir Arthur Conan Doyle's Sherlock Holmes story 'The Adventure of the Dancing Men', as seen above.



Affine Encryption

The Caesar cipher is in fact a special case of an **Affine cipher**. An Affine cipher works modulo n, where n is the number of characters in the plaintext alphabet. Then another number, a, is chosen so that a and n have highest common factor of 1. I.e., $\gcd(a, n) = 1$. A final number, $0 \leq b < n$ is also chosen. Then a letter

m_i of the plaintext message is encrypted using the encryption function

$$E(m_i) = (am_i + b) \mod n.$$

Decryption is achieved using the decryption function

$$D(c_i) = a^{-1}(c_i - b) \mod n.$$

As an example, we will consider our alphabet to be the letters `a` up to `z`, considered as the numbers 0 up to 25. Choosing key values of $a = 5$ and $b = 12$, the table below shows how letters would be encrypted using the encryption function

$$E(m_i) = (5m_i + 12) \mod n.$$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
12	17	22	1	6	11	16	21	0	5	10	15	20	25	4	9	14	19	24	3	8	13
M	R	W	B	G	L	Q	V	A	F	K	P	U	Z	E	J	O	T	Y	D	I	N

For example, the plaintext '`claudeelwoodshannon`' would be encrypted to the ciphertext '`WPMIBGGPSEEBYVMZZEZ`'.

Vigenère Cipher

Another cipher based on shifts using modular arithmetic is the **Vigenère cipher**, a form of which was originally conceived by Leon Battista Alberti in 1467 and which remained unbroken for over three centuries. Due to its notoriety as an incredibly difficult cipher to break, it gained the moniker of *le chiffrage indéchiffrable*: the indecipherable cipher.

When implemented by hand, this relies on the large table below, or a cipher disk.

	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 **keyword** is used in this cipher. For example, suppose that our message is `maths is great` and our keyword is `euler`. Then we line up our message and the keyword as below to generate the ciphertext. For example, we look up the `m` row and the `e` column to get `Q`. Similarly, we then look up the `a` row and the `u` column to get `U`. And so on.

Letter position, i	1	2	3	4	5	6	7	8	9	10	11	12
Plaintext, m_i	m	a	t	h	s	i	s	g	r	e	a	t
Key, k_i	e	u	l	e	r	e	u	l	e	r	e	u
Ciphertext, c_i	Q	U	E	L	J	M	M	R	V	V	E	N

The encryption method seems a bit more complicated but can be described using modular arithmetic again. Essentially a different Caesar shift is being applied to each letter of the message, depending on what letter of they keyword it matches up with. So, encryption is achieved by performing the addition

$$c_i = m_i + k_i \mod 26,$$

where m_i is the ith letter of the message and k_i is the ith letter of the keyword repeated over and over again. For example, with our message above: $m_7 = 18$

Letter position, i	1	2	3	4	5	6	7	8	9	10	11	12
Ciphertext, c_i	Q	U	E	L	J	M	M	R	V	V	E	N
Key, k_i	e	u	l	e	r	e	u	l	e	r	e	u
Plaintext, m_i	m	a	t	h	s	i	s	g	r	e	a	t

(s), $k_7 = 20$ (u), and so

$$c_7 = m_7 + k_7 = 18 + 20 = 38 = 12 \mod 26.$$

So $c_7 = 12$ which corresponds to M, as above.

Decryption works in a similar fashion, so

$$m_i = c_i - k_i \mod 26.$$

Essentially we use the key again to work backwards.

Frequency Analysis and Vigenère

The Vigenère cipher is much more resistant to techniques such as frequency analysis. However, it is not impossible as demonstrated by Friedrich Kasiski and possibly even earlier by Charles Babbage.

Kasiski's and Babbage's attacks on the Vigenère cipher relied on having a long enough ciphertext. If the keyword length was known, then the long message would allow frequency analysis to be applied in groupings. For example, the keyword above is '`euler`' and has five letters, so every fifth letter is encrypted with the same key. Frequency analysis would be applied to the letters which appeared first, sixth, eleventh, and so on. Similarly we would look at the grouping for the second, seventh, twelfth, etc. And so on for the third, fourth, and fifth groupings.

Kasiski Attacks

The main form of attack on a Vigenère cipher is known as a 'Kasiski attack'. This is a combination of looking at patterns in the ciphertext to determine a key length, and then applying frequency analysis on groupings of letters, as described above.

Kasiski's idea was to use repeated patterns in the ciphertext, along with the length of the gap between them, to determine the key length. For example, the ciphertext passage below has some highlighted patterns which are repeated, such as HYQ, XAIA, and FLX.

```
ZSHRSNAYEHVRHIUIZZQZXHWEFLXPOJFCXEFJ
AJMLSEURXXSVZXAGSEFYKCHYMXMLWJISKPRN
MWUIWESATXQYQHDISEXCTRRTXSLIZPNCBRHV
XPBKSEOILKFVMXXVHYMRFEBJMRWCSKMWFSFK
MPTWVZESPRHYMXTWAVZFNWWVPXAIAJQPOIGR
NSNXHYQMKZOIUSNWQFZGXVBJFLXCKVDILGFL
FMGMGVPEGHGKGHBIRGQVAEDJMPFSGKMWGEFI
```

```
AAECOJMQTRKZFLTQWTDSLGCGQQBKVKEGKYHZ
ZMLIHYQXKEBJUIGXQIQEMYFVEXAEHJIEKQOE
PQNPHZBPRMBRPVHTCWIEMIFNUXAMBWURBXST
AQIPOTQRVCAVZAXRHKAEGHTIASOIFKTMLKZF
NITFCLFXAIWIXMMXZVMJYEWIEWXVSEQMGXVV
UVTWGLDEGGSFRXAIWIQQIMFVAZXVARFXXVWK
UWISGJUFEIHYMXMLSZZJNWCIUENRRVDXAIAZ
OVHWQFBIWSHYQWTQSEASGIURHITXVFGKAXHF
FLXSZUQVPSFCPWHJGGMGXEGJAYKGSJAJAYAR
ZHTRUVDSKXVFGKAXCWFLXQCEXCMSRZEQBWGK
TIBHSRAJEMTVGTHRHYQQTWWDBSLWWSXIHVWD
BVHFOSXIBXWJOYKMCLEXHVSTMPEWCDQSYXVV
YIGXOCTEUMHJAJMLCJQHXTOIFIWHOPEEMQCJ
FXXVFVEXKMOCYIGJOEOMXHHYQVXQWXTXUICK
TIKQSEGTHRARDWIIFYMTLMBWQVBSFKAXAIAJ
QPOIGRZHKIOUKXHASCOSFIODUWLMCEMVRIBK
QVIVWJQCXXOTDSLWHYQKNPTFRWIEQVYMGHGK
TEMEFVFSHYFDURWWOJAYKWOIQXHXVFEIHJHY
QFXEGKEXAEHGQVBWVZZXXPZVOXLZOJFEGHQF
APTRRLZWRQDRFLXXWTDIZEFUQHMLWJQEKXVN
UXAIBMUSNWSPQWTRRJXSPPMRZHLYFVXCWVSN
        FLXMFGXEGWOXMMGWHLE
```

The sequence `XAIA` is repeated in the ciphertext and the two sequences are 288 letters apart. This length is called the **interval**. If we break this interval into its prime factors we can start to determine possible key lengths:

$$288 = 2 \times 2 \times 2 \times 2 \times 2 \times 3 \times 3.$$

The key length could be 2, 3, 4, 6, 8, 9, 12, 18, 24, 32, 36, 48, 72, 96, 133, or 288, since these are all of the possible numbers we can make by selecting some of the factors of this interval. This narrows things down, but that's still a lot of options, so we repeat the process.

The sequence `FLX` is repeated in the ciphertext and the two sequences are 60 letters apart. We break this interval into its prime factors:

$$60 = 2 \times 2 \times 3 \times 5.$$

So the key length could be 2, 3, 4, 5, 6, 10, 12, 15, 30, or 60, and some of these options match what we had before.

Repeating this again for the sequence `HYQ` we find that the interval is 108. We break this interval into its prime factors:

$$108 = 2 \times 2 \times 3 \times 3 \times 3.$$

This gives possible key lengths of 2, 3, 4, 6, 9, 12, 18, 27, 36, 54, or 108. Again, some of these options again match what we had before.

Carrying on in this way, we can create a table of sequences, intervals, and possible key lengths.

It is possible that some repeated sequences are just due to coincidence, but with longer sequences, this becomes less likely. If we analyse the factors in the

sequence	interval	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
XAIA	288	X	X	X		X		X	X			X				X
FLX	60	X	X	X	X	X				X		X			X	
HYQ	108	X	X	X		X			X			X				
XVF	54	X	X	X		X			X			X				
YMX	96	X	X	X		X		X				X				X
KAX	54	X	X	X		X			X			X				
FXXV	258	X	X			X										

table, the obvious candidates for key lengths are 2, 3, 4, and 6. Since key lengths of 2, 3 or 4 are quite short, we can rule them out for now. (We may have to come back to them.) The key length is likely to be 6 and as we found out in the lecture, the actual keyword was `meteor`.