BOSS CRYPTANALYST HANDBOOK V2

TOP SECRET

Harry

HARRY

This is the official guide to codebreaking and cryptanalysis, issued by the Bureau of Security and Signals Intelligence. It should not be shared with individuals outside the organization.

HARRY

Table of Contents

SUBSTITUTION CIPHERS

CAESAR SHIFT CIPHERS

The easiest method of enciphering a text message is to replace each letter by another, shifted along the alphabet by a fixed amount. So for example every letter a may be replaced by D, and every letter b by the letter E and so on.

Applying this rule to the previous paragraph produces the text

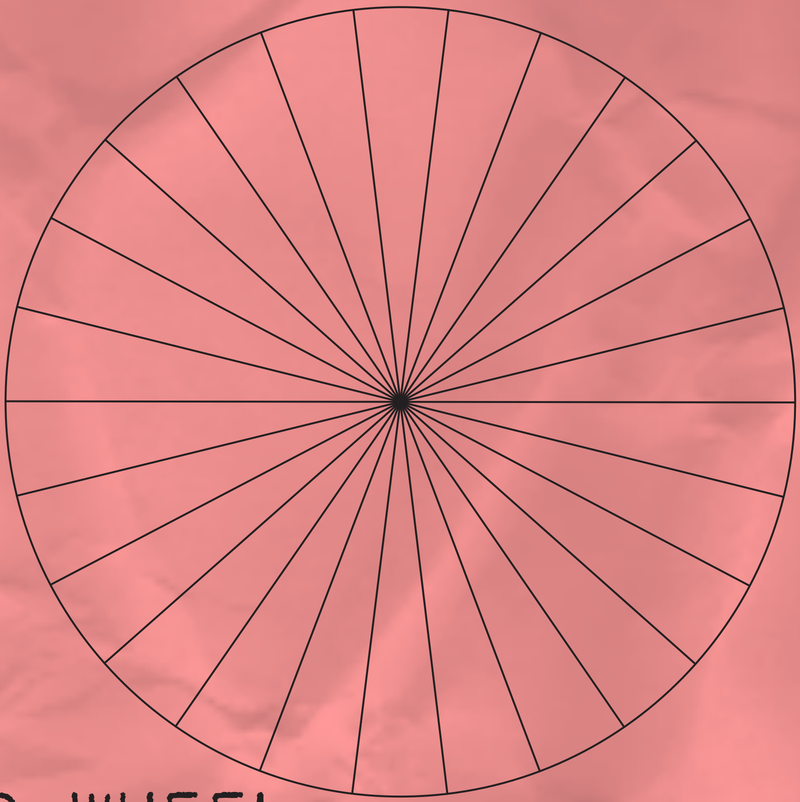
WKH HDVLHVW PHWKRG RI HQFLSKHULQJ D WHAW PHVVDJH LV WR UHSODFH HDFK OHWWHU EB DQRWKHU, VKLIWHG DORQJ WKH DOSKDEHW EB D ILAHG DPRXQW. VR IRU HADPSOH HYHUB OHWWHU D PDB EH UHSODFHG EB G, DQG HYHUB OHWWHU E EB WKH OHWWHU H DQG VR RQ.

Note the convention in these notes that ciphertext is written in capital letters, while plaintext is usually lowercase.

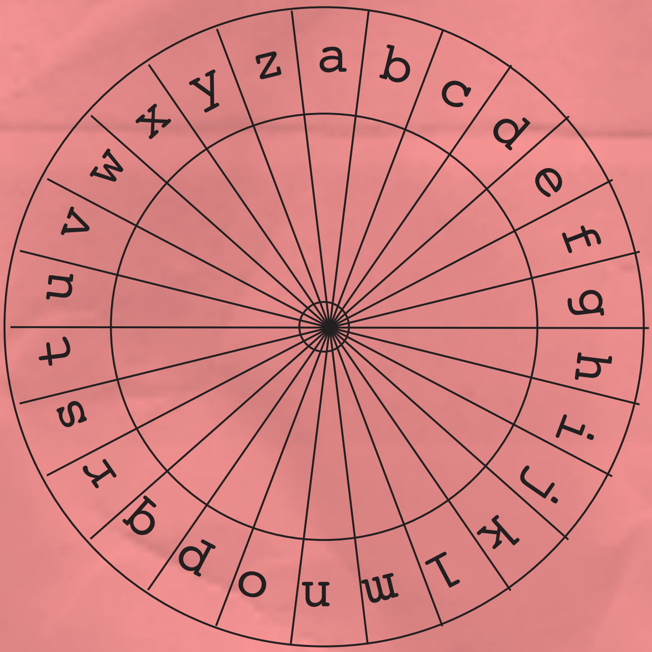
Such a cipher is known as a shift cipher since the letters of the alphabet are shifted by a fixed amount, and as a **Caesar shift** since such ciphers were used by Julius Caesar.

You can build a simple machine, a cipher wheel, to help apply the Caesar cipher consisting of two paper discs, one small and one large, and each carrying the alphabet evenly spaced around their rim as on the next figure. Print them, cut them out and join them at the centre with a paper clip. Next fill in the letters on the outer rim of the large wheel and you are ready to go.

HARRY



BOSS
CIPHER WHEEL



First, choose your shift. There are 25 Caesar shift ciphers, (and another one of them taking a to A which does nothing to the text) which you get by turning the wheel to its 26 different positions. To carry out the Caesar shift cipher above, turn the wheel so the D on the outer rim matches up with a on the inner wheel. Find each letter of the message, on the inner wheel and read its counterpart from the outer rim without moving the wheel. To decrypt a message this way you reverse the reading direction, finding each letter of the ciphertext on the outer wheel and reading its twin on the inner wheel.

BOSS Challenge: Decipher the following message, which has been encrypted using the Caesar shift cipher which takes a to M.

NAEE FAB EQODQF

With only 25 shift ciphers to try, it is not too hard to decipher a Caesar cipher by brute force. This just means we try each of the possible ciphers in turn until we find one that works. This process is a lot simpler using the cipher wheel.

BOSS Challenge: Brute force the following message to see what it says.

HTSLWFYZQFYNTSX, DTZ FWJ STB FS JCUJWY FY GWJFPNSL YMJ HFJXFW HNUMJW, MFWWD

A FIRST EXPLOIT

Just because we can use brute force to solve the cipher doesn't mean we have to. If that was all there was to codebreaking it would be entirely the province of computer scientists and engineers who are very smart at speeding up that sort of computation. At the cutting edge of cryptography, it is the interaction of those disciplines with mathematics which enables governments (and criminal hackers) to read poorly encrypted communications, and we can begin to see where mathematics comes into the picture even when considering a simple cipher like the Caesar shift.

HARRY

Notice that in order to know which shift cipher has been used it is enough to work out where one of the letters has been shifted. That tells us the amount of shift and therefore the entire cipher. This can be done, for example, by discovering which character has replaced the plaintext letter e. The letter e has been chosen here for a reason, it is the single most common letter to be found in English text (curiously, it is largely because the word the is one of the most common words - we will come back to that point in a minute).

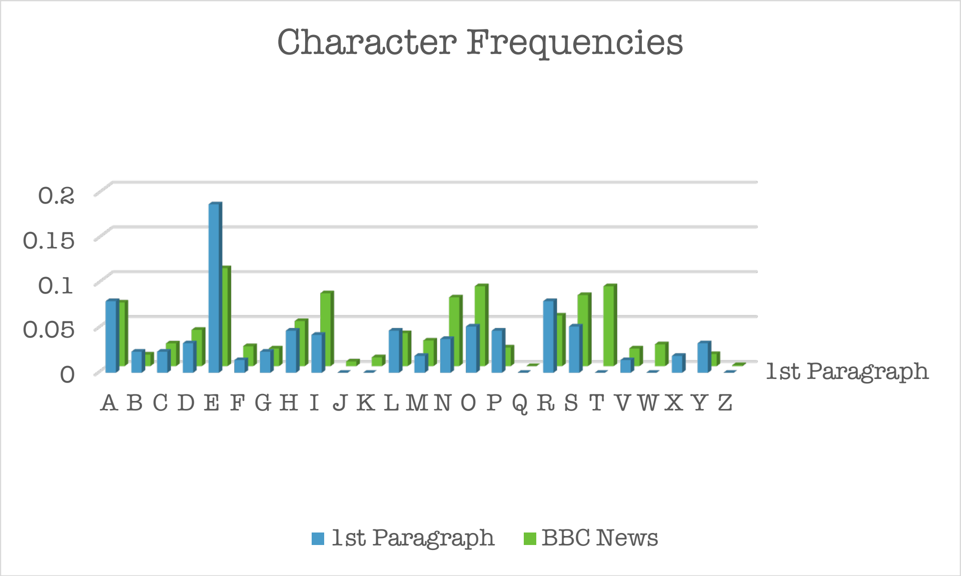
BOSS Challenge: Count the letter frequencies in the following ciphertext to see for yourself which character is the most common.

WKH HDVLHVW PHWKRG RI HQFLSKHULQJ D WHAW PHVVDJH LV WR
UHSODFH HDFK OHWWHU EB DQRWKHU XVLQJ D ILAHG UXOH, VR IRU
HADPSOH HYHUB OHWWHU D PDB EH UHSODFHG EB G, DQG HYHUB
OHWWHU E EB WKH OHWWHU H DQG VR RQ.

You can speed this up using the very useful online text analyser at

http://www.dcode.fr/frequency-analysis

You should see that the letter H appears more than twice as frequently than any other letter, more than 20% of the time. That is because this is the encryption of the first paragraph of this guide using the Caesar shift which takes e to H



If we run the online frequency counter on the first paragraph of the guide we can compare them with the frequencies from other passages of English text. Here we have done that with a lead story from the BBC news site.

The frequencies are not quite the same partly because the plaintext for this cipher was carefully written to ensure it had a lot of the letter e in it to make a point. All the same, you can see the same general pattern, and this is often a good way to identify which letter has been used to encrypt the letter e and for a Caesar shift cipher that is all we need.

This is clearly a weakness in the cipher, and finding such a weakness is called finding an exploit in the trade.

There is another weakness in the message we studied above which is not so much to do with the cipher as the way it has been implemented. Whoever encrypted the message left spaces in the text so that we can see the shape of the words. This gives us another exploit. We can guess that the three-letters starting the sentence form a 3-letter word, and, as remarked above, the most common 3 letter word in English is the. This fits with our frequency count which suggests (correctly) that e has been replaced by H, and a quick check shows that the Caesar shift by 3 does indeed encode the word the as WKH, and it is easy to complete the decryption.

HARRY

BOSS Challenge: Try to break this Caesar cipher using one of the exploits described above and without brute force.

DW WKUHH R’FORFN SUHFLVHOB L ZDV DW EDNHU VWUHHW, EXW KROPHV KDG QRW BHW UHWXUQHG. WKH ODQGODGB LQIRUPHG PH WKDW KH KDG OHIW WKH KRXVH VKRUWOB DIWHU HLJKW R’FORFN LQ WKH PRUQLQJ. L VDW GRZQ EHVLGH WKH ILUH, KRZHYHU, ZLWK WKH LQWHQWLRQ RI DZDLWLQJ KLP, KRZHYHU ORQJ KH PLJKW EH.

MATHEMATICAL CIPHERS

Despite the advantages for an agent in using keyword substitution, most of the ciphers produced and studied by modern cryptographers are automated and rely on mathematics to provide the encryption. On the other side of the cyber battle lie the cryptanalysts, who spend every waking hour applying more mathematics to try to break these ciphers.

A good way to start thinking mathematically is to revisit the Caesar shift cipher which can be viewed as a type of addition.

Each letter is first encoded by its numerical position in the alphabet. For reasons lost in the mists of time the convention is that we take a to lie in position 0, b in position 1 and so on, and the Caesar shift is then given by adding a constant to each of the positions. So if a is shifted to D that corresponds to moving a three places along the alphabet, or in other words to adding 3 to all the positions. You have to think a bit about how to deal with x, y and z here since 23+3=26, 24+3=27 and 25+3=28, and there are no letters in those positions, but since those letters move to A, B, C in positions 0,1,2 we can fix that by changing our addition so that when the answer is bigger than 26 we subtract 26 to put it back in the required range.

You can think of this as putting the numbers 0 – 25 on a clock face and adding them by counting round the face. We do this when telling the time. Three hours after 11 o'clock is 2 o'clock, which we get by adding 3+11=14, then subtracting 12 to get 2.

The military use the 24-hour clock, and 4 hours after what they call "23 Hundred Hours" is 3am, which is given by adding 4+23=27, then subtracting 24.

Mathematicians call this modular arithmetic. The standard 12-hour clock is just arithmetic mod 12, the military clock is arithmetic mod 24, and the Caesar cipher is carried out using arithmetic mod 26.

BOSS Challenge: Carry out the following additions mod 26

$$3+8$$

$$21+6$$

$$13+18$$

$$21+21$$

Of course, other languages have different alphabets. The Norwegian alphabet has 29 letters, so for them the Caesar cipher is carried out using arithmetic mod 29, while modern Greek has 24 letters, so the Caesar shift cipher in Greek looks just like using the 24-hour clock.

Just to recap, returning to our Caesar shift cipher, the shift by 3 sends 6 to $6+3=9$, which corresponds to mapping the plaintext letter g to the ciphertext letter J. At the end of the alphabet we have x mapping to A, y mapping to B and z mapping to C which correspond to the modular arithmetic $23+3=0 \bmod 26$, $24+3=1 \bmod 26$ and $25+3=2 \bmod 26$.

There is a convenient shorthand for the Caesar shift by n, given by

$$x \to x+n.$$

Here we are using x to stand for the position of a letter, and n to stand for the shift amount, i.e., x and n are each one of the values 1-26, rather than letters in the English alphabet. The shift is defined by the integer n, which can take any one of 26 values, and this gives all 26 Caesar shift ciphers. Putting n=0 we don't move the letters at all, which is not much use, so in practice, as we noted in the first section, there are only 25 different useable Caesar shift ciphers.

HARRY

Now we know about modular arithmetic we can introduce a new class of mathematical ciphers known as the affine shift ciphers. These exploit the fact that we can multiply as well as add in modular arithmetic. If you look back at the last BOSS challenge you will see that you were asked to compute 21+21. Clearly, we should think of this as 2x21, and the answer is 16 mod 26 since 2x21 =42 and we subtract 26 to get 16 mod 26. If we want to compute 3x21 then we just add 21+21+21 then subtract 26 until we get a number in the range 1-25.

BOSS Challenge: Compute the following products mod 26:

2x23

3x9

7x15

19x15

It is slightly complicated to set up modular arithmetic, but, once you have the hang of it, it is rather easy to do so we will practice by working through an example.

THE AFFINE SHIFT X→3X+5

We start as before with the position table.

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

This time instead of replacing a position x with the number x+3 we will replace it by the number 3x+5, where this number is interpreted appropriately. So, for example, 2 → (3x2)+5 = 11, means that the plaintext letter c maps to the ciphertext letter L, while 8 → (3x8)+5 = 29 = 26+3, so we interpret this as 3, and the plaintext letter i maps to the ciphertext letter C.

As with the Caesar shift cipher, whenever the result of the computation is larger than 25, we subtract 26 to make

HARRY

it smaller. Unlike the Caesar shift, we might have to do this more than once until it becomes small enough.

For example, consider what happens to the letter u. Its position in the alphabet is 20, so if we apply the affine shift $x \to 3x+5$ it moves position 20 to position $65 = (3x20)+5$. This is too big, so we subtract 26 to get position $65-26 = 39$. This is also too big, so we subtract 26 again to get $39-26=13$.

So the plaintext letter u is replaced by the ciphertext letter N.

We can also think of this as computing

$$(3x20)+5=65$$

and then take the remainder after division by 26.

Applying this to the whole alphabet gives us the following encryption/decryption table.

The top row gives the plaintext characters; the second row their positions; the third row gives their positions after applying the affine shift; the final row shows the corresponding ciphertext character.

The affine shift table corresponding to $x \longrightarrow 3x+5$

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
5	8	11	14	17	20	23	0	3	6	9	12	15	18	21	24	1	4	7	10	13	16	19	22	25	2
F	I	L	O	R	U	X	A	D	G	J	M	P	S	V	Y	B	E	H	K	N	Q	T	W	Z	C

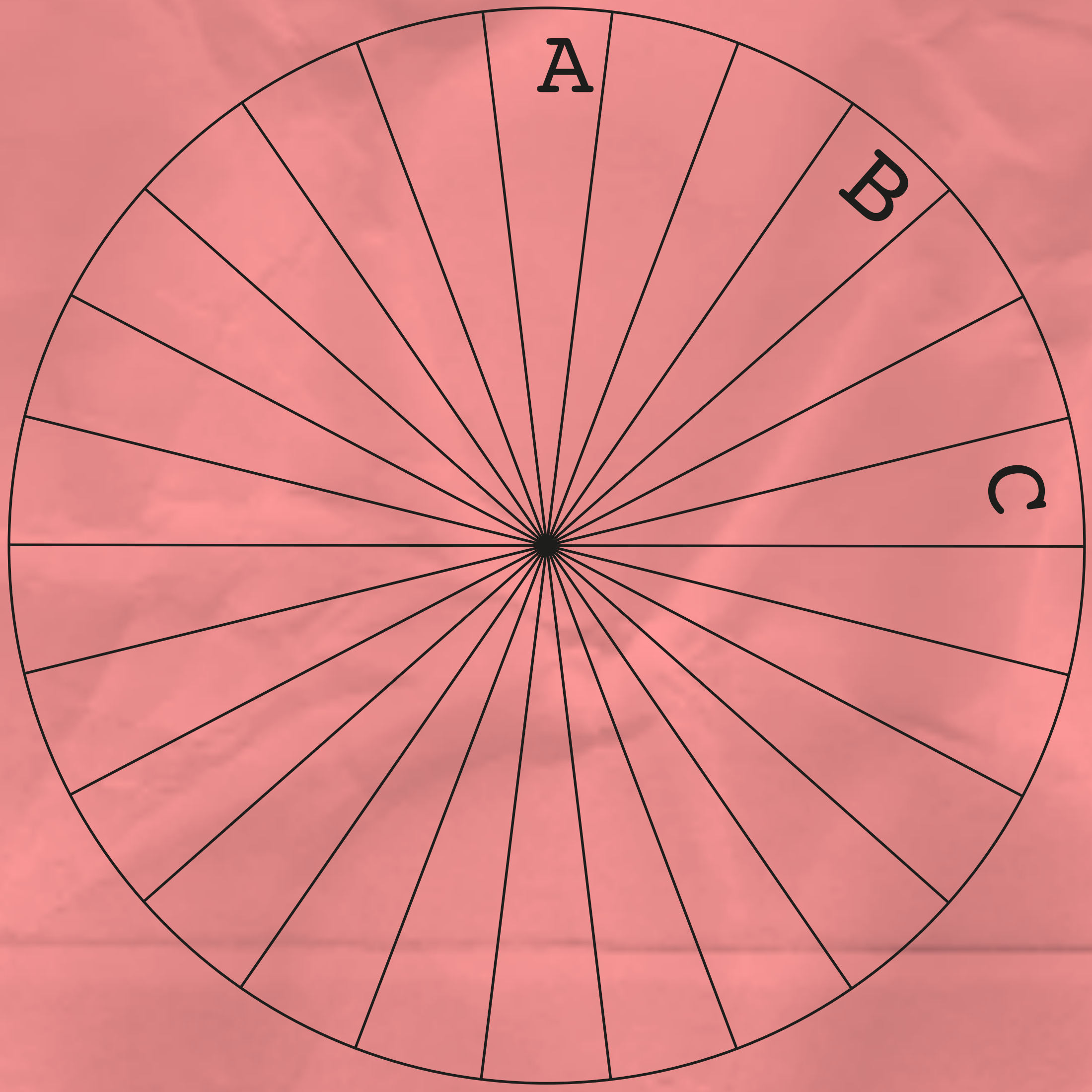
BOSS Challenge: Encrypt the following message from the Sherlock Holmes novel Moriarty by Anthony Horowtiz, using the affine shift cipher $x \to 3x+5$.

Professor Moriarty. Meet me at the Cafe Royal, London. One o'clock, May the twelfth. Wear a red tulip.

You can make a new cipher wheel to help you generate this affine shift cipher as follows. Fill out the outer ring

HARRY

with the letters A-Z starting at the top and skipping three steps at a time so it looks like this to start.



BOSS Challenge: Complete the wheel

Now you can replace the Caesar shift outer wheel with this one. Lining up the a on the inner rim with A on the new outer rim will give you a wheel that implements the affine shift cipher x->3x. Moving the wheel round so that a lines up with B implements x->3x+1, and lining it up with F implements the cipher x->3x+5. Once you have built your cipher wheel you can check that this works by comparing the result with the table above.

BOSS Challenge: Use your cipher wheel to encrypt Harry's name using the affine shift cipher x->3x+5. Check your

HARRY

answer by using the encryption table above. These tables are sometimes called lookup tables.

There are a whole host of other affine shift ciphers given by replacing the multiplier 3 and the shift 5, giving us a rich source of different ciphers. The Caesar shift ciphers are special cases where we take the multiplier to be 1.

Using the blank outer wheel you can make your own cipher wheels for any of the affine shift ciphers. Choose a pair of numbers (a multiplier a and a shift b) we get an affine shift cipher which we write in shorthand form as

$$x \to ax+b.$$

You can build the cipher wheel for this affine shift cipher by filling in the letter A at the top of the outer rim and then skip “a” places around the wheel to put B, a further “a” places to put C and so on. (Here the letter “a” stands for the multiplier you chose for the affine shift cipher, NOT the letter a!)

Having introduced the affine shift ciphers we can see that it becomes much harder to break a message. Compared to the 25 usable Caesar ciphers there are 311 usable affine shift ciphers making it much harder to guess the correct decryption. (We will explain where the number 311 comes from later on.) To see the confusion in practice consider the Caesar shift table corresponding to $x \to x+13$

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
13	14	15	16	17	18	19	20	21	22	23	24	25	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	A	B	C	D	E	F	G	H	I	J	K	L	M

As with the affine shift $x \to 3x+5$ the letter e is enciphered as R so if we intercepted a message enciphered with an affine shift in which the most common letter was R we wouldn’t be sure which of these two ciphers had been

HARRY

used. Our exploit has been weakened, but all is not lost. Consider the second most common letter in English, the letter t. In the Caesar shift table it is encrypted as the ciphertext letter G, while in the affine shift table it is encrypted as K, so these are definitely different ciphers, and frequency checking will tell us which one has been used.

BOSS Challenge: Decide which of the following affine shift ciphers, $x \to x+13$ or $x \to 3x+5$ has been used to encrypt this message:

"KAR YVMDLR HFDO AR TFHS'K TRFEDSX ADH HRFK IRMK." FMRW KNESRO KV MVVJ FK GFLJ.

Thinking of our Caesar shift cipher $x \to x+13$ as an affine shift cipher with multiplier 1, we can see that two different affine shift ciphers can encrypt the letter e in the same way, so it is no longer sufficient to discover the letter substituting for e in order to crack the message.

This happened because there are two things we need to work out to determine which affine shift cipher has been used, the multiplier a, and the shift b. Mathematicians say that there are “two degrees of freedom” in the choice of cipher.

To nail down these two numbers we might hope that deciphering two letters is sufficient. Luckily if we know two values of the expression $ax+b$ we can often solve the two corresponding equations to find the numbers a and b.

We may be more familiar with doing this to solve pairs of equations using the more traditional “real” numbers, where the solution involves subtraction and division. The same method works for modular arithmetic, with one important warning.

We cannot always divide in modular arithmetic.

To see why this might be a problem, suppose we try to use the rule x→2x, doubling all the positions in the alphabet.

The affine shift table corresponding to x→2x

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	2	4	6	8	10	12	14	16	18	20	22	24	0	2	4	6	8	10	12	14	16	18	20	22	24
A	C	E	G	I	K	M	O	Q	S	U	W	Y	A	C	E	G	I	K	M	O	Q	S	U	W	Y

We see that the first thirteen letters of the alphabet get encrypted exactly the same way as the last thirteen, so we can't expect to be able to decrypt the message. If we did get a message encrypted using this table and tried to decipher the ciphertext letter I we wouldn't be sure whether we should read that as the plaintext letter e or the alternative, r.

In terms of the affine shift function x→2x both 4 and 17 are taken to the same answer, 8 mod 26.

This shows that we have to be careful with our choice of multiplier a when defining an affine shift cipher

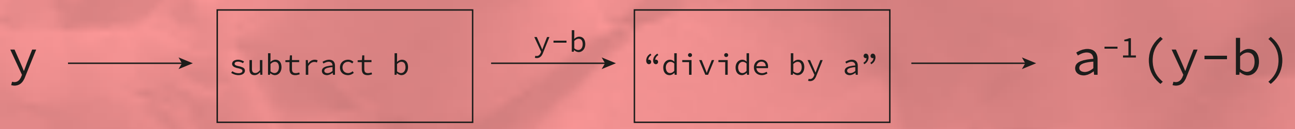
$$x \to ax + b$$

Every encryption needs to be reversible so it can be decrypted by the person we are sending the message to, which means we have to be able to reverse the steps in it. We can regard affine shift encryption as a two-stage process illustrated by the following diagram

$$x \longrightarrow \boxed{\text{multiply by } a} \xrightarrow{ax} \boxed{\text{add } b} \longrightarrow ax+b$$

Decryption is then the reverse of this:

HARRY



It doesn’t matter which value we choose for the addition term b as we can always subtract it, but if we choose the multiplication factor a carelessly we might not be able to divide by it. If we look back at our example on the last page, where we chose a = 2, we saw that

$$2x4 = 2x17 \bmod 26.$$

If we try to cancel the multiplier 2 we get that

$$4=17 \bmod 26,$$

which is clearly wrong.

What is going on here? It is just that in modular arithmetic you cannot always divide! The same thing happens in usual arithmetic if you try to divide by 0, it is just that in modular arithmetic there are other cases where division goes wrong too.

Luckily the number theorists worked all this out a long time ago, and there is a simple rule:

In mod 26 arithmetic you can divide by any odd number which is not a multiple of 13.

It is easiest to think of division (when it is allowed) as given by multiplying by an inverse. In more the more familiar world of standard arithmetic, the inverse of 2 is ½, so dividing by 2 is the same as multiplying by ½. When carrying out arithmetic mod 26 the inverses are given below:

a	1	3	5	7	9	11	15	17	19	21	23	25
a^{-1}	1	9	21	15	3	19	7	23	11	5	17	25

HARRY

Division by $3 \bmod 26$ is the same as multiplication by 9 because

$$3 \times 9 = 1 \bmod 26.$$

so 9 is the equivalent of 1/3 in arithmetic $\bmod 26$.

Similarly, according to this table, division by 15 is given by multiplying by 7, and this is because

$$7 \times 15 = 105 = (4 \times 26) + 1 = 1 \bmod 26.$$

So 7 is the equivalent of 1/15 in arithmetic $\bmod 26$. Notice that also, 3 is the equivalent of 1/9 and 15 is the equivalent of 1/7, so these "modular inverses" behave a lot like the fractions we are used to.

BOSS Challenge: Solve the equation $19y = 1 \bmod 26$ to find a value of y that works. (Hint you need to "divide both sides by 19", but remember that this means multiplying by the inverse of $19 \bmod 26$.

All this is to say that the enciphering rule defines a function from the alphabet to itself, and if the message is to be decipherable then this encryption function needs to be reversible. In the world of modular arithmetic, the multiplication function $x \to ax \bmod n$ can be reversed if and only if the only number that divides into both the multiplier a and the modulus n is 1. There is a fancy term for this. We say that

a is coprime to n.

For our affine shift ciphers we can use any multiplier which coprime to 26, and there are 12 of them: all the odd numbers not divisible by 13 as shown in the table above.

This means that we have 12 possible choices for the multiplier a, and 26 choices for the shift b, yielding 312 affine shift ciphers. As with the Caesar cipher, one of these, the function

$$x \to 1x + 0$$

is unusable as it does not disguise the letters, so we have 311 usable affine shift ciphers as reported above.

The numbers a and b used to define an affine shift cipher x→ax+b are called the **key** for the cipher. We usually write the key as the pair (a,b). As we just saw, there are a total of 311 possible keys for an affine shift cipher, compared with the 25 possible keys for a Caesar shift cipher.

This all makes a brute force attack (without frequency analysis) less practical than the much simpler situation for the Caesar shift ciphers, where there are only 26 possibilities, but certainly a small team could carry out such an attack quickly, and frequency analysis can help us to speed things up as we will now see.

A MATHEMATICAL SOLUTION TO THE AFFINE SHIFT CIPHER

In order to solve a Caesar shift cipher without using brute force, we needed an exploit, and we used the fact that the letter e is the most common letter in English. Once we have found the most common letter in the ciphertext we can assume that it represents e, and use that to tell us which Caesar shift has been used.

In order to pin down an affine shift cipher x→ax+b we need to identify two of the plaintext letters. It is easiest to spot common letters and the two most common letters in an English text are usually e and t, with e the most common.

Because of this we assume that the two most common ciphertext letters stand in for the plaintext letters e and t and try to solve the corresponding equations to find the encryption key pair (a,b).

BOSS Challenge: Find the two most common letters in the following ciphertext:

CER'G UTTCKQ SEG E DUUO DMRRCRW JZQ TMLL LQRWJZ UT JZQ ZUMGQ, ZCWZ MX UR JZQ JUX. CJ SEG JZQ URLY DUUO JZEJ SEG ELSEYG LUKIQN—ELQV ZEN URLY HQQR CR JZQDQ JZDQQ UD TUMD

HARRY

JCOQG, ERN RQPQD UR ZCG USR. SZQR ZQ SEG YUMRWQD, ZQ ZEN TERJEGCBQN JZEJ JZQDQ OCWZJ HQ GUOQJZCRW GJDERWQ MX JZQDQ . . . E JCOQ OEKZCRQ UD E MTU. HMJ CJ SEG URLY ER UTTCKQ SCJZ E NQGI, E KUMXLQ UT TCLCRW KEHCRQJG, GZQLPQG TMLL UT XEXQDG ERN HUUIG. HERI GJMTT—JZEJ'G SZEJ CER GECN. QPQR GU, ELQV SERJQN JU WU MX JZQDQ RUS.

"JZQ XULCKQ GECN ZQ SEGR'J SQEDCRW ZCG GQEJ HQLJ." ELQV JMDRQN JU LUUI EJ FEKI.

GZQ RUNNQN. "YQEZ. JZEJ'G SZEJ JZQY GECN."

"NUQGR'J JZEJ GQQO GJDERWQ JU YUM? YUM IRUS ZUS KEDQTML ZQ SEG. ZQ ELSEYG SUDQ ZCG GQEJ HQLJ. ZQ SUMLNR'J QPQR NDCPQ OQ EDUMRN JZQ KUDRQD SCJZUMJ OEICRW OQ XMJ OCRQ UR."

FEKI JZUMWZJ TUD E OUOQRJ, JZQR GZDMWWQN. "YQEZ, CJ CG GJDERWQ," GZQ GECN. "HMJ JZEJ OMGJ ZEPQ HQQR JZQ SEY CJ SEG. SZY SUMLN JZQ XULCKQ ZEPQ LCQN?"

The two most common letters in the cipher text are Q and J appearing respectively, roughly, 13% and 10% of the time. We guess that this means e is encrypted as Q and t as J.

Now we roll up our sleeves and start to do some mathematics. Remember, e is in position 4 and Q is in position 16 in the alphabet, so in terms of the modular arithmetic this tells us that $4a + b = 16 \bmod 26$.

Similarly, t is in position 19 and J is in position 9 in the alphabet, so we also know that $19a + b = 9 \bmod 26$.

If we take the difference of the terms on the left-hand side of these equations we get $19a - 4a = 15a$, while the difference on the right hand side is $9 - 16 = -7$, so we get

$$15a = -7 \bmod 26.$$

How do we interpret $-7 \bmod 26$? Just as we had to subtract 26 to reduce a “too big” number, we have to add it to make a “too small” number big enough, and since $-7+26=19$ we get the equation

$$15a = 19 \bmod 26\,.$$

Now we want to “divide both sides by 15”, which, as we noted above, is the same thing as multiplying by $7 \bmod 26$.

Multiplying the left-hand side of our equation by 7 gives

$$7x(15a) = (7x15)a = 105a = (104+1)a = ((8x26)+1)a = a \bmod 26\,.$$

so, the left hand side is just a, which is what we are looking for.

Now multiply the right-hand side by 7 as well

$$7x19=133=130+3=(5x26)+3=3 \bmod 26,$$

so we see that a=3. Substituting that value into the first equation

$$5a+b=19 \bmod 26$$

gives us 15+b=19, which gives b=4, so our affine shift cipher is $x \to 3x+4$.

We have found the key to our affine shift cipher, the pair (3,4).

The Affine shift table corresponding to $x \to 3x+4$

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
4	7	10	13	16	19	22	25	2	5	8	11	14	17	20	23	0	3	6	9	12	15	18	21	24	1
E	H	K	N	Q	T	W	Z	C	F	I	L	O	R	U	X	A	D	G	J	M	P	S	V	Y	B

Remember that the bottom row corresponds to the ciphertext while the top row gives the corresponding plaintext letters. Now consider the first word of the ciphertext: CER'G.

HARRY

We look up each of these letters in the bottom row of the table to get the decrypt: ian’s

This is looking hopeful. If we carry on we find an extract from the excellent Alex Rider novel Stormbreaker by Anthony Horowtiz. As Anthony said to Harry one day, "Alex Rider loves codes and ciphers!”

Here is the start of the full decrypt:

Ian's office was a room running the full length of the house, high up on the top. It was the only room that was always locked—Alex had only been in there three or four times, and never on his own. When he was younger, he had fantasized that there might be something strange up there . . . a time machine or a UFO. But it was only an office with a desk, a couple of filing cabinets, shelves full of papers and books. Bank stuff – that's what Ian said. Even so, Alex wanted to go up there now.

"The police said he wasn't wearing his seat belt." Alex turned to look at Jack.

She nodded. "Yeah. That's what they said."

"Doesn't that seem strange to you? You know how careful he was. He always wore his seat belt. He wouldn't even drive me around the corner without making me put mine on."

BOSS Challenge: Complete the decrypt by deciphering the last paragraph

FEKI JZUMWZJ TUD E OUOQRJ, JZQR GZDMWWQN. "YQEZ, CJ CG GJDERWQ," GZQ GECN. "HMJ JZEJ OMGJ ZEPQ HQQR JZQ SEY CJ SEG. SZY SUMLN JZQ XULCKQ ZEPQ LCQN?"

ENLARGING THE KEYSPACE

The affine shift cipher was a lot more work to crack than the Caesar shift because there were so many more possibilities. About twelve times as many, corresponding to the twelve possible choices of the multiplier a in the formula

$$x \rightarrow ax + b.$$

The set of all possible keys for a given cipher is known as its keyspace. So the keyspace for the Caesar cipher is the set of usable shifts {1,…,25}, while the keyspace for the affine shift cipher is the set of pairs {(1,1),…,(1,25),(3,0),(3,1),…,(3,25),…,(25,0),…(25,25)}.

It is a general principle of modern cryptography that the bigger the keyspace, the safer the cipher. We already saw that the affine shift cipher was harder to break than the Caesar cipher, and this is why. This idea was first proposed by Dutch cryptographer Auguste Kerchoffs around 1883, and is known as Kerchoffs' principle.

> "A cryptographic system should be secure even if everything about the system, except the key, is public knowledge."

If the keyspace is small then the system can't be secure, because we can use computers to automate the brute force attack, so modern ciphers have huge keyspaces as part of their design. A modern Advanced Encryption Standard cipher (AES) has over 10^{77} possible keys. Even with the fastest current computers chained together it would take over 10^{51} years to crack such a cipher by brute force, which is the reason it is one of the standards for high security encryption.

One disadvantage of AES is that you need to know how to do arithmetic in several different number systems as well as to carry out vector and matrix calculations, all of which make it an advanced topic at University level. It is certainly not the sort of thing an agent could use on the run. Luckily there are other ciphers that provide some level of security while still being usable in the field.

HARRY

KEYWORD SUBSTITUTION CIPHERS

These were introduced by security services as a highly secure, reliable and easy to use field cipher for agents. Of course, security depends on the ability of the enemy to crack the cipher and they would be hopelessly inadequate against computer attacks, but they are still many times harder to break than the Caesar shift cipher as we will later see.

To build a keyword substitution cipher we design an encryption table by choosing a keyword or phrase which is used to jumble the alphabet as follows.

First write down the phrase, with no spaces between the letters and omitting any repeated characters, then continue round the alphabet in order until every letter appears exactly once, and write the list under the standard alphabet:

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

Here we have chosen the key word SIMPSONS, which appears in the shaded boxes above. The second and third letter S are omitted and we continue the alphabet from the first unused letter after the last used letter, N, which is Q.

Of course, if the key phrase is carefully chosen (for example “The quick brown fox jumps over the lazy dog”) there might not be any letters left to use up but such a choice is not necessary. If instead of using a genuine word or phrase we allow ourselves to use any ordering of the letters in our ciphertext alphabet then the number of such ciphers is

$$26! = 26 \times 25 \times 24 \times 23 \times \ldots \times 2 \times 1$$

Which mathematicians calls “26 factorial”. It is surprisingly big – approximately 10^{26}. As Tom Briggs from Bletchley pointed out, 26! is about a quadrillion times

HARRY

the number of neurons in the human brain, or about 25 times the total amount of data storage in the world (40 Zettabytes). With so many possibilities for the cipher it is impossible to try them all by hand, so brute force cannot be used to attack the problem.

On the other hand a fully random encryption table would be very difficult for an agent to reliably memorize (they work under conditions of extreme stress after all) so a genuine word or phrase is likely to have been used, and this reduces the number of ciphers considerably. According to the Oxford English Dictionary authorities “there are, at the very least, a quarter of a million distinct English words”, which would still make a brute force attack impossible without the aid of a computer, but frequency analysis works, especially if we can see the word shapes.

Consider the text

VEP HYXHLVHTP MO AWFJYFLT H RFNEPS HJNEHAPV FL VEFU ZHC FU VEHV FV FU PHUC VM KPKMSFUP VEP IPCZMSY MS IPCNESHUP, HLY EPLRP VEP RFNEPS HJNEHAPV. VEFU FU FKNMSVHLV, APRHWUP FO VEP UPLYPS EHU VM IPPN VEP RFNEPS HJNEHAPV ML H NFPRP MO NHNPS, VEP PLPKC RHL RHNVWSP VEP NHNPS, YFURMXPS VEP IPC, HLY SPHY HLC RMKKWLFRHVFMLU VEHV EHXP APPL PLRSCNVPY ZFVE FV. EMZPXPS FO VEP IPC RHL AP RMKKFVVPY VM KPKMSC FV FU JPUU JFIPJC VM OHJJ FLVM PLPKC EHLYU.

As before notice that the first word has three letters and, since it occurs several times, VEP may well be the word the. This gives a strong hint that the letter e is enciphered as the letter P in the cipher.

Of course, other three letter words are possible, e.g., “and” or “but”. Nonetheless a quick check shows us that the letter P is the most common letter in the ciphertext, so it is reasonable to assume that the correct decryption translates P to e.

This also suggests that V stands for t and E for h, allowing us to begin to decipher the message.

HARRY

We will continue to use the convention that UPPERCASE LETTERS denote enciphered letters and lowercase letters denote plaintext characters:

```
VEP HYXHLVHTP MO AWFJYFLT H RFNEPS HJNEHAPV FL VEFU
the      t  e                  he     h  et    th

ZHC FU VEHV FV FU PHUC VM KPKMSFUP VEP IPCZMSY MS
       th t  t    e    t   e     e the  e

IPCNESHUP, HLY EPLRP VEP RFNEPS HJNEHAPV. VEFU FU
 e  h   e,     he  e the    he     h  et. th

FKNMSVHLV, APRHWUP FO VEP UPLYPS EHU VM IPPN VEP
     t  t,  e    e    the  e  e  h   t   ee  the

RFNEPS HJNEHAPV ML H NFPRP MO NHNPS, VEP PLPKC RHL
   he     h  et        e e       e , the e e

RHNVWSP VEP NHNPS, YFURMXPS VEP IPC, HLY SPHY HLC
   t  e the    e ,       e  the  e ,      e

RMKKWLFRHVFMLU VEHV EHXP APPL PLRSCNVPY ZFVE FV.
         t     th t h  e  ee  e     te    th  t.

EMZPXPS FO VEP IPC RHL AP RMKKFVVPY VM KPKMSC FV FU
h  e e     the  e       e      tte  t   e      t

JPUU JFIPJC VM OHJJ FLVM PLPKC EHLYU.
 e      e   t         t  e e   h    .
```

Reading carefully we see the single letter word H, and the four letter word thHt circled above, and guess that H is almost certainly the letter a. Making that replacement throughout the ciphertext we get the following, where we are using our upper/lowercase convention to save space but leaving the uppercase encrypted letters black so we can see them more easily:

HARRY

the aYXaLtaTe MO AWFJYFLT a RFNheS aJNhaAet FL thFU ZaC
FU that Ft FU eaUC tM KeKMSFUe the IeCZMSY MS IeCNhSaUe,
aLY heLRe the RFNheS aJNhaAet. thFU FU FKNMStaLt, AeRaWUe
FO the UeLYeS haU tM IeeN the RFNheS aJNhaAet ML a NFeRe
MO NaNeS, the eLeKC RaL RaNtWSe the NaNeS, YFURMXeS the
IeC, aLY SeaY aLC RMKKWLFRatFMLU that haXe AeeL eLRSCNteY
ZFth Ft. hMZeXeS FO the IeC RaL Ae RMKKFtteY tM KeKMSC Ft
FU JeUU JFIeJC tM OaJJ FLtM eLeKC haLYU.

Now in English the two 2-letter words ending with "t" are "at" and "it", and we already think that the plaintext letter a is encrypted as H, so the letter F probably stands for the letter i and the word Ft circled above decrypts as it. This is followed by another 2-letter word, FU beginning with the same letter so together these are likely to decrypt to it is, meaning that U is the cipher for s.

Making these substitutions throughout the text we get:

the aYXaLtaTe MO AWiJYiLT a RiNheS aJNhaAet iL this ZaC
is that it is easC tM KeKMSise the IeCZMSY MS IeCNhSase,
aLY heLRe the RiNheS aJNhaAet. this is iKNMStaLt, AeRaWse
iO the seLYeS has tM IeeN the RiNheS aJNhaAet ML a NieRe
MO NaNeS, the eLeKC RaL RaNtWSe the NaNeS, YisRMXeS the
IeC, aLY SeaY aLC RMKKWLiRatFMLs that haXe AeeL eLRSCNteY
Zith it. hMZeXeS iO the IeC RaL Ae RMKKitteY tM KeKMSC it
is Jess JiIeJC tM OaJJ iLtM eLeKC haLYs.

Before reading on it is worth looking at this to see if you can spot any other likely substitutions of your own.

Appropriate guesses would be:

tM = to, so M = o

haXe = have, so X = v

easC = easy, so C = y

As we identify more letters it gets easier to guess even more and we can decipher the text to get the following

HARRY

extract from Simon Singh's excellent history of codes and ciphers, The Code Book:

"The advantage of building a cipher alphabet in this way is that it is easy to memorize the keyword or key-phrase, and hence the cipher alphabet. This is important, because if the sender has to keep the cipher alphabet on a piece of paper, the enemy can capture the paper, discover the key, and read any communications that have been encrypted with it. However, if the key can be committed to memory it is less likely to fall into enemy hands."

BOSS CHALLENGE: Can you identify the keyword for the cipher we have just broken?

SHARPENING THE ATTACK

FREQUENCY ANALYSIS

We have already seen how frequency analysis can help us to identify common letters and common words. We can go further with this analysis, comparing the number of occurrences of each character in the cipher text with an expected frequency for the standard English alphabet. In the plain text above a character count gives us the following table of occurrences.

a	b	c	d	e	f	g	h	i		k	l	m	n	o	p		r	s	t	u	v	w		y	
32	7	14	11	55	5	2	26	27		6	9	11	20	18	16		17	17	35	4	4	4		12	

The consonants h,s,t are relatively common in plaintext as are the vowels a,e,i and o.

The vowel u is much less common and any occurrence of q is almost guaranteed to be followed by a u. It is also possible to analyze common letter pairs and triples. as we have seen the triple “the” is the most common in English.

Cryptographers refer to triples of letters as trigrams and pairs of letters as digraphs or bigrams, and you can look up standard, bigram and trigram frequency tables on the web, for example at:

http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/

HARDENING THE CIPHER

DISGUISING THE WORD STRUCTURE

A chink in the armour of our ciphers so far has been the preservation of word structure. This allows us to spot common words. In order to avoid such weakness cryptographers usually remove punctuation and block the characters together in groups of four or five, so our previous cipher text looks like

```
VEPHY XHLVH TPMOA WFJYF LTHRF NEPSH JNEHA PVFLV EFUZH
CFUVE HVFVF UPHUC VMKPK MSFUP VEPIP CZMSY MSIPC NESHU
PHLYE PLRPV EPRFN EPSHJ NEHAP VVEFU FUFKN MSVHL VAPRH
WUPFO VEPUP LYPSE HUVMI PPNVE PRFNE PSHJN EHAPV MLHNF
PRPMO NHNPS VEPPL PKCRH LRHNV WSPVE PNHNP SYFUR MXPSV
EPIPC HLYSP HYHLC RMKKW LFRHV FMLUV EHVEH XPAPP LPLRS
CNVPY ZFVEF VEMZP XPSFO VEPIP CRHLA PRMKK FVVPY VMKPK
MSCFV FUJPU UJFIP JCVMO HJJFL VMPLP KCEHL YU
```

Usually the length of the text groups does not matter, however, in analyzing some ciphers (like the infamous Vigenère cipher which we will study later) a carelessly chosen block length may make the length of the key more apparent, since it can reveal important pattern repeats more easily.

To attack a message that that has been grouped in this way we have to work with letters not words. To do so we use the frequency analysis described above, together with a little judgement (or luck!). The process can be hard, but wars have been won or lost on the back of it, and so have fortunes. As remarked by Jericho, the lead character in Robert Harris's novel "Enigma",

"It was hard going, but Jericho didn't mind. He was taking action, that was the point. It was the same as code-breaking. However hopeless the situation, the rule was always to do something. No cryptogram, Alan Turing used to say, was ever solved by simply staring at it."

HARRY

For a cryptographer, judgement comes in two forms discovering exploits like frequency analysis, but also in finding cribs. A crib is a word or phrase that we expect to find in the plaintext somewhere. For example if we intercept an encrypted report from a weather station then we might expect to find the word windspeed in the message. If we are planning to intercept a lot of messages from that source and we have some idea of what type of cipher to expect we can encrypt the crib using lots of possible keys, and keep a list of the results to hand so we can check them against the message. This is quite feasible on a modern computer once you have learned to programme. You can write a small programme that encrypts the crib using all the keyword substitution ciphers arising from words in a standard dictionary, and get the programme to look for the results in the ciphertext to narrow down which keyword might have been used.

Even without the power of modern computers, this was partly how Dilly Knox, Alan Turing and other codebreakers at Bletchley managed to break into the Enigma cipher and other Nazi codes. It was while trying to automate their attack on the Lorentz cipher that Turing, together with Max Newman and Tommy Flowers developed one of the first modern computers, Colossus, which you can see in action at the National Computer Museum at Bletchley.

BOSS Challenge: Use a crib to determine which Caesar shift cipher has been used to encrypt the following extract from an enemy weather ship, and decrypt it.

DPUKZWLLK MVYAF RUVAZ

MULTILINGUAL CODEBREAKING

The task facing Turing and the others at Bletchley was made even harder by the fact that they were trying to break a cipher in a foreign language. Luckily, many of the staff there had some knowledge of German which they could use to find their own exploits. As we noted above, different languages have different alphabets and different idiosyncracies which can help to weaken ciphers used to protect communications written in them.

Sometimes a professional cryptanalyst doesn't even know what language the message was written in. A radio message intercepted in Africa during the Second World War might have been written in Italian, French, German or a local African dialect. Luckily frequency analysis can help to identify the language.

One of the most unusual is the Khmer language, the official dialect of Cambodia, which has the world's largest alphabet. With 74 characters. You can read all about this fascinating script at

https://www.worldatlas.com/articles/which-language-has-the-largest-alphabet.html

POLYALPHABETIC CIPHERS

The main weakness allowing us to tackle a substitution cipher is the irregularity in the distribution of letters in English text

In order to remove this weakness from a cipher it is necessary to disguise the frequencies of letters in the plaintext and the easiest way to do this is by using a polyalphabetic cipher. In such a cipher each plaintext letter may be encoded in more than one way so that, for example, the letter e may be enciphered as both X and G within the ciphertext. One problem with this approach is that if X and G both encode for e we don't have enough letters left to encode the other 25 letters. One elegant solution to this problem is a famous French diplomatic cipher.

THE VIGENERE CIPHER

The Vigenère cipher uses multiple Caesar shift ciphers in a repeating pattern to encrypt the text, with the pattern often described by a keyword or key phrase. So for example, if the keyword is HARRY then the first letter is encrypted by the shift a→H, the second by the shift a→a (so it is not changed at all), the third and fourth by the shifts given by a→R and the fifth by the shift a→Y. Then the pattern repeats so that the sixth character of the plaintext is encrypted by the shift a→H, the seventh by a→A, the eighth by a→R and so on. As long as the key word is kept secret this cipher is very hard to break.

To implement this effectively the cipher clerk in the embassy would have used a lookup table called the Tabula Recta. Each row of the table corresponds to one of the Caesar shift ciphers making it very easy to look them up quickly and accurately.

HARRY

	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

Given the task of enciphering the message attack at dawn using the keyword NOW, the cipher clerk would produce the following table

N	O	W	N	O	W	N	O	W	N	O	W
a	t	t	a	c	k	a	t	d	a	w	n
N	H	P									

Looking up the letter in the row labelled N and the column labelled a gives the ciphertext letter N, looking up row O column t gives H, row W column t gives P and so on.

BOSS Challenge: Complete the encryption of this message.

HARRY

Notice the key feature of the Vigenère cipher, that the same letter of the plaintext can be enciphered in different ways! The letter t is encrypted as H then P in the same message which messes up the frequency counts. This makes it much tougher to crack, and for centuries it was thought to be unbreakable. The same example above shows us something else though, the first t in attack and the t in at are both encrypted as H because they are both underneath the letter O in the repeated keyword. This repeat appears 6 letters apart. Also the letter a appears three times under the keyword letter N, in the 1st, 4th, 7th and 10th positions and these repeats appear 3 letters apart. This is happening because the keyword has length 3. This allows us to guess that the keyword has length 3. We have found an exploit! Of course, not every recurrence of a ciphertext character corresponds to the same plaintext letter, that is the whole point of the cipher, but by looking at the gaps between repeats you can often make a good guess at the repeat length, since it will be a factor of many of the repeat distances.

BOSS Challenge: By computing the gaps between repeated occurrences of the letter a in the following Vigenère cipher, make a guess at the length of the keyword.

EHZVW UVQAF ROLFZ QYWGO ACWPK RVUJQ IBLHF JNOGK UHYDP
HYHXW XJAFK KAYVB LGGDY SKKAG RIAFK KAQZS DSI

Even knowing the length of the keyword there is still a lot of work to do, but it turns out that this is largely a matter of carrying out a version of the frequency analysis attack we used so successfully on substitution ciphers like the shift and keyword ciphers.

This exploit, based on the analysis of repeats in the text was discovered independently by the British mathematician Charles Babbage and the Polish cryptographer Friedrich Kasiski. While you can’t always just use repeating patterns of single letters, an analysis of repeated strings of letters can often be used to determine the length of the keyword, and once this is done a standard frequency analysis is applied to each

HARRY

part of the ciphertext encoded by a single cipher. A very good account of Babbage-Kasiski deciphering can be read in Chapter 2 of Simon Singh's The Code Book.

We will now use this method together to try to break the message in the previous BOSS Challenge, making use of the fact that our agents have broken into the embassy and recovered a note by the cipher clerk in which he has left the word structure intact.

Eh zv w uvqafro lfzqywgoa cw pkrvuj qiblhfjnogku hydp hyh xwxjaf kka yvblggdys, kka griaf kka qzsdsi.

In the previous BOSS challenge, you should have noticed that the gaps between the consecutive occurrences of the letter A are

12 32 4 12 4 4

These are all multiples of 4, suggesting that the keyword has length 4, so we expect that the first, fifth, ninth, thirteenth letters and so on should all have been encrypted by the same Caesar shift cipher. These letters appear as

E...W...A...L...Y...A...K...J...L...N...U...P...X...A...A ...L...Y...A...A...A...D...S...I...

Looking at the letter W in position 5 of the ciphertext we notice that it appears as a single letter word, so W probably represents a or i, meaning that the Caesar shift used here either maps i or a to W. The first of these would take e to M, which doesn't appear in this sequence, and even though we are only looking at a small collection of letters there are enough that we would expect to see at least one representing e. (Recall that the letter e appears around 12% of the time in English text.) So we consider the second option which takes a to W, so takes e to A. This deciphers that part of the message as

I...A...E...P...C...E...O...N...P...R...Y...T...B...E...E ...P...C...E...E...E...H...W...M...

Converting the ciphertext to uppercase and making the W to a decrypt for the first, fifth, ninth letters and so on, writing them in lower case we see the following:

iH ZV a UVQeFRO pFZQcWGOe CW PoRVUn QIBpHFJrOGKy HYDt HYH bWXJeF KKe YVBpGGDcS, KKe GRIeF KKe QZShSI.

Notice the three letter word KKe. The most common three letter word in English ending with the letter e is “the”, which suggests that the first K corresponds to the Caesar shift taking t to K, while the second corresponds to the shift taking h to K. Trying these we get a possible partial message decrypt as follows.

It Zs a gVnerRl prZnciGle oW modErn cIyptFgraGhy tYat tYe biXger Khe kVypsGace, Khe sRfer Khe cZpheI.

You should recognize this as part of Kerchoffs’ principle discussed before, and we can complete the decrypt by inspection.

BOSS Challenge: Complete the decryption of this text, and work out which keyword was used for this Vigenère cipher.

THE INDEX OF COINCIDENCE

There is another very clever method to tackle the Vigenère cipher using something called the index of coincidence (ioc). Like frequency analysis, this is based on the idea that there are hidden patterns in English which make certain letter combinations more likely than others. In this case we study the likelihood that if we pick two letters at random in the ciphertext then they will be the same. This idea was first suggested by William Friedman in 1922.

BOSS Challenge: What do you think the index of coincidence should be for the following sentence?

Hi Harry!

There are 7 letters in this sentence, and so there are 42 ways of picking a pair of letters at random. Of these only H and r repeat and they only repeat once each. The odds of picking the letter H twice are 2/42 (we can pick the H in Hi first then the H in Harry, or the other way round) and the odds of picking the two Rs is the same, so the odds of picking two letters the same is 2/42+2/42=2/21, or about 10%. We usually give the index of coincidence as a decimal. Here it is 0.09524, while for a standard English text, the index of coincidence is usually about 0.0686 so this is a little high.

The index of coincidence can be computed for any text. For each letter of the alphabet you work out the probability that two letters picked at random are that letter. To do this you must first count the number of letters in the text, which we will denote by n, then count the number of copies of the letter you are interested in, which we will denote by N. Now the odds of picking our given letter as the first one of the pair is N/n, and the odds that the second one is also the same letter is (N-1)/(n-1), so the odds of picking it twice at random is the product

$$\frac{N}{n}\frac{(N-1)}{(n-1)}.$$

We can work this out for each letter of the alphabet and add those together to get the odds of picking two letters the same.

As we said above, if we consider a standard long English text then the index of coincidence will be around 0.0686. On the other hand, if we compute it for a genuinely random sequence then each letter will appear roughly 1 in every 26 characters so the index of coincidence for a random string of letters will be about 0.038466.

A typical ciphertext will have an ioc of something between these values. If the value is close to 0.0686 then it is likely that we are considering a substitution cipher like the shift or keyword ciphers we considered above. Another possibility is that we are looking at a jumbled text like one of the transposition ciphers we will study later in this handbook.

HARRY

BOSS Challenge: Compute the index of coincidence in the following phrase: This is no coincidence!

If the ioc is lower than you expect and the text you are studying is not just random, then something subtle is going on and we may well be looking at a polyalphabetic cipher. The weakness in the Vigenère cipher is that for some number k every kth letter in the ciphertext has been encrypted with the same Caesar shift, so if we knew k we could carry out a standard frequency analysis on those letters to figure out the shift. The index of coincidence can help us find k in the following way.

For each k from 1 to around 9, we split the text into blocks of size k and stack the blocks. Next, we read down the columns to extract the sequences of letters k apart in the text. This is essentially what we did when we attacked the Vigenère example above, though we didn’t bother typing it in that form.

Now we compute the index of coincidence for each column. If our k is not the repeat length used by the Vigenère cipher then the letters in the column will have been encrypted by a number of different shifts, so they will be more random and we will see a low index of coincidence. When we arrange the text into the correct number of columns, corresponding to the key length, then the ioc should be much closer to 0.0686 for each of the columns.

The process is a little laborious, but also somewhat miraculous, and once you learn to code or to use a spreadsheet, it can be automated, making it easy to find the length of a Vigenère cipher key. Once that is done, you can carry out frequency analysis on the columns to work out which cipher was used for each one.

This is a very powerful exploit. The Vigenère cipher was the main diplomatic cipher across Europe for a long time, and was widely considered intractable. The index of coincidence attack makes it almost routine to decipher messages using it.

BOSS Challenge: By computing the index of coincidence of the following ciphertext decide whether it is likely to have been encrypted using a keyword substitution cipher or a Vigenère cipher. You do not need to decipher it (unless you want to!).

Ckaxop h okcy el vou kpk el vou ccy, jng zuvcyqzg uqbcs qtf tyrkaqxa pdzgsboildig vhmcuyycayupz xgf tuxilt zq iuiqtu zjl Wuxlhtoldz Evtk cut Iawxkt Zsnqvb, MEOG’y qyymkuqr phck. Kai uxlhz hbdivpet yhi zq whuvlsz Dyyzkzx Mqcuxptutv jesobdoehjoqui, ckax g ulsxga couzyup ae jgjhera ckuzqmgz ikpa re hvhkknd iqbdztpuy. Wutkt axk nlqjgyinkw el Csqyvhyx Fldtkzjup, pjy hphyv ougf, axk qywgppigvpet yhi mkcut uwqig pd Ccauxihjk Jvkyg hdj qmvoepqrnf sgol ytvv rkkuw up axk hphyv vv Tqcusdlh tkuuzgld tkuuzgld.

THE ENIGMA MACHINE

The most famous polyalphabetic cipher in the world must be the Enigma cipher, implemented by the engineer Arthur Scherbius as the Enigma machine. The machine was first designed around 1918 and was sold in large numbers to banks and commercial enterprises, before becoming the new standard for secure communications in the German military.

Polish cryptographers, no doubt worried about the expansion of the powerful military next door, studied the machine in depth throughout the 1930s and as war broke out they risked their lives to share what they knew with their colleagues at the Government Code and Cypher School(GCCS) in the UK. This was the fore-runner of GCHQ, the modern Government Communications Headquarters.

The Enigma machine is an electro- mechanical device which implements a polyalphabetic cipher which in practice encrypts each individual letter of a message with its own custom cipher. Each of these is a cipher which switches the letters of the alphabet in pairs, so if a particular letter a is encrypted as G then if g had been at that position in the message it would have been encrypted as A. To know how a letter would be encrypted you need to know not just what letter it was, but whereabouts it is in the message and how the machine was configured when you started. If you set up two Enigma machines in exactly the same way and ask them to encrypt a message they will both encrypt it the same. Furthermore, if one of them is given the encrypted message from the first one and reset to the original settings, then rather than encrypt it further it will decrypt it instead. That is because letters had been switched in pairs as we remarked above.

HARRY

The “key” for an Enigma cipher describes the way the machine is to be set up for a particular communication, and there are an incredibly large number of ways this can be done. The military Enigma had over 103,325,660,891,587,134,000,000 different settings so brute force was never going to be enough to crack this code even with huge teams of codebreakers. (Compare this to the 250,000 keyword ciphers we discussed above).

Even getting hold of a military Enigma machine was impossible, but luckily the Polish cryptographer Marian Rejewski had worked on reconstructing it from the little information the Polish cipher bureau could glean from intelligence reports and cipher analysis.

Just nine days before the outbreak of war he and his colleagues handed everything they knew about the machine, its workings and its weaknesses to Commander Denniston, head of the British Government Codes and Cypher School and to Dilly Knox, the British chief cryptographer. Together with the large team of brilliant experts at Bletchley they completed the work that Rejewski had started and cracked the Enigma machine. The story has been told many times now, though for decades it remained the most closely guarded secret, not least because rotor machines like the Enigma continued to be used by governments around the world until the end of the 1970s.

A genuine Enigma machine would cost a lot of money now, but you can download several emulators for your computer, tablet or phone. Just search for Enigma in your app store. You can even find a fully featured emulator on the web at

http://summersidemakerspace.ca/projects/enigma-machine/

These programmes are beautiful, and it can be great fun playing with the settings on them, but they don’t necessarily help to understand the workings of the machine. Back in 2005, as part of the story for the National Cipher Challenge we invented the Pringle Can Enigma, which we think illustrates much better how it

HARRY

worked. Our design was based on the paper slips used by Turing and others in their original work on the cipher, but I am sure that if Pringles had existed they would have used the can!

You can make one yourself for the cost of a can of Pringles (and who doesn't like them?). Since we introduced the Pringle Can Enigma to tackle the Fialka cipher in the National Cipher Challenge in 2005 a number of other people have produced their own variations on the theme and you will find them across the web. You can find out more on the resources page at

www.cipherchallenge.org/resources/

Even 70 years after the war this obsolete cipher machine continues to fascinate and enchant fans of spy-craft, codes and ciphers, and steampunk engineering.

TRANSPOSITION CIPHERS

One way to minimize the impact of frequency analysis on the strength of a cipher is to ensure that it doesn't give much away. Transposition ciphers work by jumbling letters rather than replacing them, so a frequency analysis is likely to show that the letter frequencies match standard English, even though the message is unreadable. There are several variations on this theme and we start with the simplest. It is a very clever cipher that is reliable in the field, and has the same size key space as the keyword substitution cipher.

B	A	D
t	h	e
q	u	i
c	k	b
r	o	w
n	f	o
x	j	u
m	p	s
o	v	e
r	t	h
e	l	a
z	y	d
o	g	x

We start by writing our keyword at the head of a table, removing duplicate letters as with the keyword substitution. Here we are using the key BAD, giving three columns. We then enter the plaintext in the boxes below. The last, empty, boxes, if any are padded with an X (usually - there is no fixed rule for which character is used) so that all the boxes are full.

Next, we rearrange the columns so that the letters in the keyword are in alphabetic order, ABD, and read off the rows grouping the letters in blocks of 5 for easy and accurate transmission:

HTEUQ IKCBO RWFNO JXUPM SVOET RHLEA YZDGO X

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

ATTACKING A TRANSPOSITION CIPHER

Clearly the length of the keyword is quite crucial. You should be able to guess this from the length of the ciphertext, which will be a multiple of it. So, in our example the ciphertext has length 36 which has factors 2,3,4,6,9, 12,18 and 36. So we could try laying out the text in grids of these widths and examining the rows.

The best hope for a quick solution is to find a crib. If there is a word you think ought to appear in the cipher text then you could try looking for anagrams of that word. This is made difficult by the fact that the table splits the text into blocks (blocks of three in the example), and if your crib word does not take up an entire block then even the characters from the crib that do appear will be jumbled with other nearby characters, so you need a reasonably long crib. On the other hand, if it is too long only part of the word will appear in that block so you are looking for anagrams of parts of the crib.

In our example if we knew, for some reason, that the text was likely to contain the word “jumps” we could look for anagrams of JUM, UMP or MPS.

BOSS Challenge: Find an anagram of one of these strings in the table above.

HARRY

Having found it a cryptanalyst is in a position to guess that the first and second columns have been transposed while the third has remained fixed. Checking this we find have cracked the cipher.

Things are harder with longer keywords but the principle remains the same. Things get even tougher if the plaintext is not in our own language, since it is harder to say what makes sense. Of course, even in this case it may be that part of the message is in your language and the rest in another. In this case you might hope to crack the ciphertext corresponding to your native language, and apply the knowledge that gives you about the cipher to write down a decrypt of the entire message, even when the text is unfamiliar.

Other (subtle) cribs: In English the letters q and u occur together so if they are separated either you are not looking at English text or they should be brought back together by undoing the anagram.

Numbers often represent dates, so for example the letters/numbers 2, 1, S, T in proximity might represent 21st, while 2,1,T,H might represent 12th since we do not write 21th in English.

HARDENING THE TRANSPOSITION CIPHER

The transposition cipher described above can be made much more secure by reading the ciphertext off by columns rather than rows. So, our message will read

HUKOF JPVTL YGTQC RNXMO REZOE IBWOU SEHAD X

Now the three letters P, M, S are nowhere near one another so you might think that anagramming won't help, but it can help us to work out the length of the keyword. This is the key to solving the cipher as it allows us to lay out the text in the appropriate grid. To see this in action look again for the letters P, M and S. They appear in the 7th, 19th and 31st position in the ciphertext, so they are 12 apart. Thinking about how the cipher works suggests that the encryption table could have 12 rows which is enough for us to get started. Even without a

HARRY

crib like the word JUMPS we could use the numerous cribs provided by the English language. The word “the” for example, or the fact mentioned above that “q” is almost always accompanied by the letter “u”.

Here is an example to try

SIEID ATTPW ADIVL SOLWO IYMRD AOSTT TDUHM AGTTT HSEOO
TAEST EOGNU AEDLN HNRDH KIWOA MENEE INEAS NPAIT SLIAI
AOJDN TCAET SOKEE EIULD HRAUE WSYSA IRBCT WNNSN TARHH
SUHAS MNOAG SVEPI AGINE IOAIS EBGRS TTWYO GTLNO EVMRT
WGTOI SAHHI ECAWP HTRAO TCRTS YRBYG

The ciphertext has 210 characters and 210=2x3x5x7 so possible key lengths are 2,3,5,6,7,10,14,15,21,30,35,42,70,105,210.

We will first try the simple crib “the” finding the positions of the letters T and H in the text and then the distances between them.

These are tabulated below. The first row of table 1 gives the positions of the letter H in the text, the first column gives the position of T and the entries in the table are the absolute values of the difference of the positions, telling us how far apart each T is from each H.

We are looking for patterns thrown up by the fact that T and H would often have been adjacent in a row, and so after permuting the columns the distance between them will be a multiple of the column height. Also, the column height will be a factor of 210, so we are looking for common occurrences of a multiple of a factor of 210. To do that, in table 2, we take the entries in the body of table and compute the highest common factor of each one with 210. (Again, the first column denotes the position of an occurrence of T, while the first row denotes the position of an occurrence of H.)

There are a number of different entries in the body of table 2 corresponding to different possible column heights, and the number of times each appears is given in table 3.

Table 1: the distances between occurrences of T and H in the text

Position of the letter H in the text

Position of the letter T in the text

	34	41	61	65	111	134	135	138	188	189	196
7	27	34	54	58	104	127	128	131	181	182	189
8	26	33	53	57	103	126	127	130	180	181	188
29	5	12	32	36	82	105	106	109	159	160	167
30	4	11	31	35	81	104	105	108	158	159	166
31	3	10	30	34	80	103	104	107	157	158	165
38	4	3	23	27	73	96	97	100	150	151	158
39	5	2	22	26	72	95	96	99	149	150	157
40	6	1	21	25	71	94	95	98	148	149	156
46	12	5	15	19	65	88	89	92	142	143	150
50	16	9	11	15	61	84	85	88	138	139	146
85	51	44	24	20	26	49	50	53	103	104	111
96	62	55	35	31	15	38	39	42	92	93	100
100	66	59	39	35	11	34	35	38	88	89	96
125	91	84	64	60	14	9	10	13	63	64	71
131	97	90	70	66	20	3	4	7	57	58	65
166	132	125	105	101	55	32	31	28	22	23	30
167	133	126	106	102	56	33	32	29	21	22	29
172	138	131	111	107	61	38	37	34	16	17	24
180	146	139	119	115	69	46	45	42	8	9	16
183	149	142	122	118	72	49	48	45	5	6	13
197	163	156	136	132	86	63	62	59	9	8	1
201	167	160	140	136	90	67	66	63	13	12	5
204	170	163	143	139	93	70	69	66	16	15	8

Table 2: The highest common factors of the distances in table 1 with 210

Position of the letter H in the text

Position of the letter T in the text

	34	41	61	65	111	134	135	138	188	189	196
7	3	2	6	2	2	1	2	1	1	14	21
8	2	3	1	3	1	42	1	10	30	1	2
29	5	6	2	6	2	105	2	1	3	10	1
30	2	1	1	35	3	2	105	6	2	3	2
30	3	10	30	2	10	1	2	1	1	2	15
38	2	3	1	3	1	6	1	10	30	1	2
39	5	2	2	2	6	5	6	3	1	30	1
40	6	1	21	5	1	2	5	14	2	1	6
46	6	5	15	1	5	2	1	2	2	1	30
50	2	3	1	15	1	42	5	2	6	1	2
85	3	2	6	10	2	7	10	1	1	2	3
96	2	5	35	1	15	2	3	42	2	3	10
100	6	1	3	35	1	2	35	2	2	1	6
125	7	42	2	30	14	3	10	1	21	2	1
131	1	30	70	6	10	3	2	7	3	2	5
166	6	5	105	1	5	2	1	14	2	1	30
167	7	42	2	6	14	3	2	1	21	2	1
172	6	1	3	1	1	2	1	2	2	1	6
180	2	1	7	5	3	2	15	42	2	3	2
183	1	2	2	2	6	7	6	15	5	6	1
197	1	6	2	6	2	21	2	1	3	2	1
201	1	10	70	2	30	1	6	21	1	6	5
204	10	1	1	1	3	70	3	6	2	15	2

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

Table 3: the frequency of entries in the body of table 2

Height	1	2	3	5	6	7	10	14	15	21	30	35	42	70	105
Count	60	64	26	15	27	6	12	5	7	6	9	4	6	3	3

Column heights of 1,2,3,5,6,7,10 all seem unlikely given that the keyword or phrase would then have to have at least 21 letters in it. On the other hand, a column height of 30 would correspond to a keyword of length 7, which is quite feasible, and gives rise to a good number (9) of TH adjacencies, as marked in green in the corresponding 30x7 grid.

HARRY

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

Notice that in three cases, rows 8,9 and 16 the T and H appear in columns 2 and 7 respectively. This suggests that whatever order the columns should be in we should end up with column 2 next to (and to the left of) column 7. In two of the rows, 9 and 16, there is an E in the fourth entry so we are led to try putting these three columns together in the order 2,7,4.

Assuming this is not an Olde English text, ruling out “Twas” as a word, these three columns are not likely to be the first three, so we need something to the left and the possibilities for that put S,H, I or A to the left of the string TWA in the first row. Trying each in turn we get STWA, HTWA, ITWA or ATWA and the first two seem unlikely. Playing the odds and considering the possibilities for arranging the remaining three letters on the top row we are led to consider ITWAS.

One possibility for the remaining columns reads AHITWAS but then the next row reads GNRDGOI which is clearly wrong. There is one other way to rearrange the columns to get this first row, but that is also unlikely as it gives the second row ONRDGGI. On the other hand, these same letters might suggest the word GOING in row 2 and a rearrangement and further experimentation gives the final arrangement, which you might recognize from earlier:

HARRY

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

"It was hard going, but Jericho didn't mind. He was taking action, that was the point. It was the same as code-breaking. However hopeless the situation, the rule was always to do something. No cryptogram, Alan Turing used to say, was ever solved by simply staring at it."

We stared pretty hard at this, but there was nothing simple about breaking it. I think Jericho, and maybe even Alan Turing, would approve.

HARRY

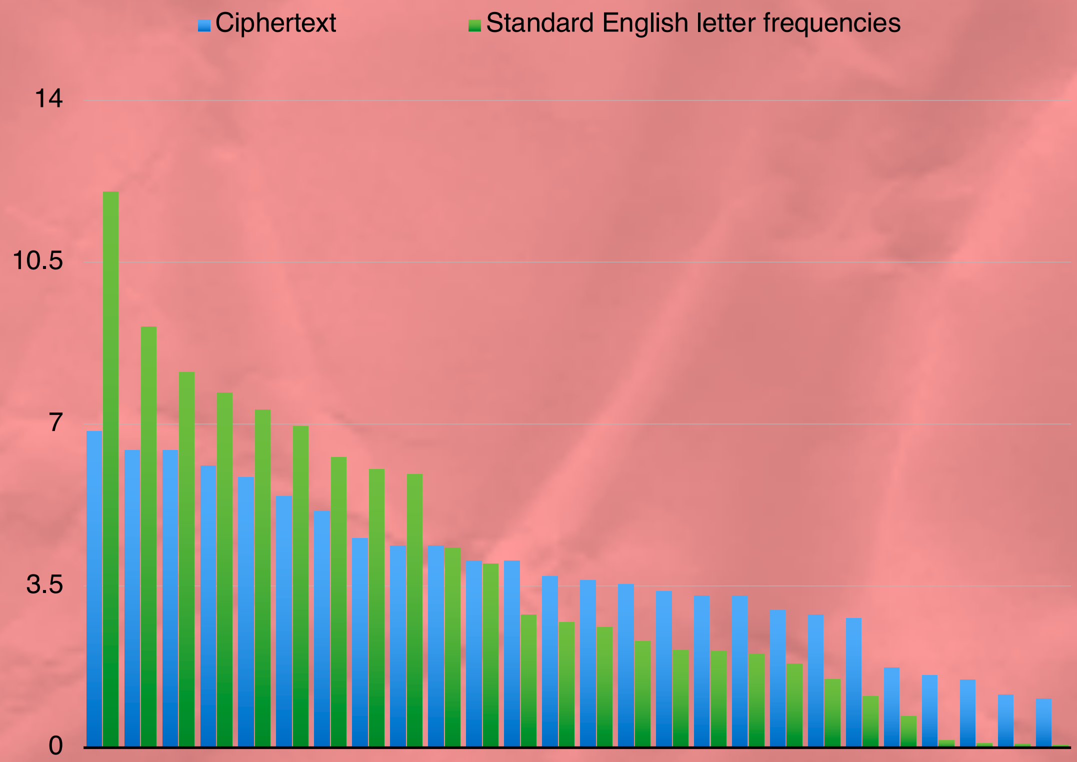
A CHALLENGE

Now let's try to put together everything we have learned to crack the following ciphertext which is from a BBC news article about encryption in the modern world.

```
XSFJD JMNRF RUDJV LMYFT GWWHP TUDIA HWRMS XXAHJ DNBRH
QTOFF NWFGH GLDJJ ATQWH UEQEM DMHRH LMCGL ZAYBT HUWIC
MHDJI CGFVZ TJHWR FYBXB HTTLX AHFLY MHDKM ZKTPS SUMRH
FHLRU WATHU JVLTQ LZSGS NAFWL WUGXD UYCHS WZJWH SIAIY
GYLSQ CMDDF IMXHX JNNRY REFEX NWHTM LNEDJ CYDRM HIGXL
VJLXQ HUYLH SLUYL TSVSH NBTQK FHWTQ DNHXU DQRYG YVSQF
MMRKJ QHZOV SIMGH HTMLN EDJQB YKGZN XSFJD JMNRF XUBIG
JRUKP PSSOE NVSXY GNRJQ YVYXJ JLBSF JDJMT JJFJA DDLYB
XZQAA YKXLL DIYXX JWYRF WAYML NPHQY LYHFH LRUWA THBXD
DQUUT XLYLT SVXTL FNQYN HMJOD NABGO WSOFG HJXIK YHPYM
HZQVX UGILE FAXXL FYITX WJJUF TIFTH LJQKJ NAJUW FLXRD
FDGTS BOFSL YRHJL YTUEY BTYWJ FHLKR JRUMN RFXIF JVLWU
BLKLK IKBDJ IUGIV GRYOJ UQHIF UOWCG HXWAS PHQYW XQTUS
ASAEJ WLJLL KRJSO FGHJX UGIXK JGTYK KYIWT WZJNK FQKKI
KRDLN IGMRO JPXWQ GRUMY HJBBB HKEJN ATGAX OLJGL MYKJV
MQNBS JKHLT REDJX WFWSX NKJDE XBHZO VLCOJ QGMCG YVSGI
NYKGB CMBDK JHVWB HYYWI XJNHZ BRJQX PFUAN NAJDD QCXXV
UTLXI VGRYG TWSGF XALUY IKNHK FATNQ KYNAJ JWWGT SVTJW
TZVWY BXNUW SWKDS LNIGX BKYYF XGAIH HYVMK ZBHLW SNEDV
UWUFG OWRYL XDYJM KNJGW INXPS YBXRD LNWTQ DFFFR XLKGS
TQOAJ XVTGW HLTHN WWMEF LVGUK JSSYN XWQKM CWIHF BCMML
FYBXR HKXUZ JVSSX NXHVY BXRWG WYVWH SYYMM HEFWA NQWZM
XIWGJ HVWBH YNAJP LMILJ FGIYL WHNTF OJGSW INSGL MYNXH
GKMXH UWYEX DVLMU MBHJJ MAFUW IUFTQ YYBHX HOMIG JHVJX
MTFGR GNSLU FNXXH UZLXQ BLMYL JDJJE GTZFF MLDPE JNKNF
WSWKD SLNIG XBKYY FXDFI BTAHS BYTPQ WXMBS WZFNX AHJDI
GJLFA IEAHV MULYR HTMLJ VKYBX XDEJM XYRXX YVWHL PYRX
```

First, we carry out frequency analysis of the whole text and compare it to the standard distribution in English.

The following chart of the frequencies, ranked most to least common in each case, allows us to compare the frequency distributions. We see that the ciphertext letter distribution, while not uniform is much flatter and lacks the distinctive spike at the left, suggesting that the frequency distribution of letters is not a good match to the standard English language. From this we conclude that the text is not encrypted with a transposition or a mono alphabetic substitution cipher like one of the shifts, or the keyword substitution we studied above.



So, we guess that the text has been encrypted with a polyalphabetic cipher, and since we only know about the Vigenère cipher we will assume that is what we have here.

The first step is to try to find the likely keyword length, which we will denote k, which is at least 2 since we are not considering a mono-alphabetic substitution. To do this we will compute the index of coincidence for sequences of letters spaced k apart in the ciphertext. Start by taking k=2. We consider the sequence of every other letter X.F.D.M.R.R.D.V.M.F.W … H.P.R, starting at

HARRY

the first. This sequence has index of coincidence 0.04695494261123 which is not close to the ioc of standard English text.

Next, we try k=3 and examine the sequence XJMF… of every third letter. This has ioc 0.054357657988021 which is closer to the standard but still not good. Taking k=4,5,6,7 in turn we get the following table of index of coincidence.

k	2	3	4	5	6	7	8	9
ioc	0.04695	0.05435	0.04616	0.047614	0.069209	0.046907	0.047228	0.04809

Notice that for k=6 we obtain an ioc of 0.069209, which is very close to the expected value of 0.0668 for English text, whereas the other values of k give a much lower value, which suggests that key length is k=6.

The next step is to split the text into blocks of 6 and to carry out frequency analysis on each of the 6 lists of ciphertext characters this gives us. Here is the first of the six lists that gives us:

XMDFPHXBFHAEHGTMGHXXMKMRULAGHHGMXRXLYGXHTBWXGMHMLBXMBKEGV
BMAXKYRLLRBUTFMBFKHGXTFLAXTLYTLMFBKGOFHHTELFGTWKKGXMBALKB
TWKHOGNMVWZPAXXGXKTATTXKGFHBEFLKNXTRTVTEKXWMXZNXVMAXVALLO
NNXXMAFHGMNXXLGLKKGFTTBXGELLXXVR

Here, X is the first ciphertext character, M is the seventh, D is the thirteenth and so on.

The most common letter by far is X, so we deduce that e has been encrypted by X in this sequence, and since the Vigenère cipher uses Caesar shift ciphers this gives a decrypt of

HARRY

etkmwoeimohlonatnoeetrtybshnoonteyesfneoaidentotsietirlnc
itherfyssyibamtimroneamsheasfastmirnvmooalsmnadrrnetihsri
adrovnutcdgwheenerahaaernmoilmsrueayacalredteguecthechssv
uueethmontueesnsrrnmaaienlsseecy

If this looks like nonsense, don’t worry, it is. This just gives us the decrypt of the first, seventh, thirteenth, twentieth letters of the plaintext and so on. We have to decipher the other five columns and intersplice them to get the decrypt.

The next sequence, which begins SNJT is a little tougher. The frequency analysis shows that J is the most common letter at 11.33% and then X at 10.84%, so either of the shifts e to J or e to X is possible. We will leave it there for the moment and move on to the third list beginning FRV … In this list H is clearly the most common letter so we assume that the shift cipher mapping e to H has been used and see if we can apply our knowledge that “the” is a very common triple to settle the ambiguity over list 2.

To do so we can write each list as a column and look for the pattern t_e across the first three columns after decrypting columns 1 and 3. We find this pattern in rows 23, 48,109,164 and 176 where the encryption string is MRH, MGH, MYH, MMH and MBH, so if any of these are an encryption of the then h must be encrypted as R, G, Y, M or B. These correspond to the affine shifts mapping e to O, D, V, J or Y. We have already seen from our frequency analysis that the most likely encryption of e is either to map it to J or to X, and putting this together with the list we just produced that makes the mapping to J more likely so we assume that our second column is enciphered using a shift mapping e to J and make that substitution.

HARRY

```
 enc_ _ _tio_ _ _kes_ _ _mod_ _ _wor_ _ _oro_ _ _eve_ _
_ime_ _ _mak_ _ _obi_ _ _hon_ _ _llb_ _ _ome_ _ _ngw_ _
_acr_ _ _tca_ _ _nas_ _ _oro_ _ _ewe_ _ _eve_ _ _tca_ _
_rom_ _ _tme_ _ _ypt_ _ _bes_ _ _sup_ _ _hat_ _ _nsa_ _
_ont_ _ _onf_ _ _nti_ _ _tya_ _ _ecu_ _ _yto_ _ _eit_ _
_sib_ _ _fyo_ _ _nsi_ _ _ele_ _ _oni_ _ _ans_ _ _ion_ _
_don_ _ _epa_ _ _nts_ _ _tho_ _ _oul_ _ _tbe_ _ _sib_ _
_ith_ _ _enc_ _ _tio_ _ _idd_ _ _rkm_ _ _lis_ _ _nio_ _
_ctu_ _ _inc_ _ _tog_ _ _hya_ _ _eun_ _ _rsi_ _ _fsu_ _
_yat_ _ _sim_ _ _ste_ _ _ypt_ _ _isa_ _ _bou_ _ _ans_ _
_min_ _ _tel_ _ _ibl_ _ _mbe_ _ _rte_ _ _oun_ _ _ndi_ _
_esi_ _ _ast_ _ _mof_ _ _sen_ _ _her_ _ _ema_ _ _any_ _
_sto_ _ _for_ _ _att_ _ _sfo_ _ _tio_ _ _mes_ _ _igh_ _
_rwa_ _ _nds_ _ _ver_ _ _mpl_ _ _ost_ _ _olv_ _ _app_ _
_let_ _ _sfo_ _ _mbe_ _ _ndu_ _ _ath_ _ _dot_ _ _ran_ _
_rma_ _ _nho_ _ _ern_ _ _tte_ _ _ich_ _ _hod_ _ _sed_ _
_res_ _ _ing_ _ _amb_ _ _dat_ _ _rea_ _ _oul_ _ _ven_ _
_nts_ _ _uth_ _ _twa_ _ _cry_ _ _ddu_ _ _gwo_ _ _war_ _
_hea_ _ _ess_ _ _eds_ _ _not_ _ _evi_ _ _rie_ _ _ain_ _
_heg_ _ _ans_ _ _aus_ _ _eir_ _ _ryp_ _ _nsy_ _ _msd_ _
_ots_ _ _ici_ _ _lys_ _ _mbl_ _ _ssa_ _ _rig_ _ _usm_ _
_ema_ _ _ala_ _ _ysi_ _ _all_ _ _cod_ _ _ack_ _ _lai_ _
_rep_ _ _ern_ _ _dde_ _ _thi_ _ _eme_ _ _ges_ _ _use_ _
_emt_ _ _cre_ _ _the_ _ _hin_ _ _edt_ _ _cry_ _ _hem_ _
_sec_ _ _sre_ _ _ved_ _ _und_ _ _use_ _ _ecr_ _ _eys_ _
_twe_ _ _har_ _ _mon_ _ _ose_ _ _nee_ _ _toc_ _ _uni_ _
_ese_ _ _ely_ _ _sea_ _ _now_ _ _sym_ _ _ric_ _ _ryp_ _
_nsy_ _ _msa_ _ _ave_ _ _akn_ _ _int_ _ _eve_ _ _nei_ _
_lve_ _ _sto_ _ _ses_ _ _esa_ _ _eto_ _ _cre_ _ _ys
```

Now we are getting somewhere. We know this is an article about encryption, and right at the start of the text we see the pattern enc_ _ _tio_. This corresponds to the ciphertext XSFJD JMNRF suggesting that ryp in positions 4,5,6 have been enciphered as JDJ in turn using shifts

HARRY

mapping r to J, y to D and p to J. Trying the J to r shift as the decrypt on the fourth column, the D to y to the fifth and the J to p on the sixth gives us the following

```
encryptionmakesthemodernworldgoroundeverytimeyoumakeamobi
lephonecallbuysomethingwithacreditcardinashoporonthewebor
evengetcashfromanatmencryptionbestowsuponthattransactiont
heconfidentialityandsecuritytomakeitpossibleifyouconsider
electronictransactionsandonlinepaymentsallthosewouldnotbe
possiblewithoutencryptionsaiddrmarkmanulisaseniorlecturer
incryptographyattheuniversityofsurreyatitssimplestencrypt
ionisallabouttransformingintelligiblenumbersortextsoundsa
ndimagesintoastreamofnonsensetherearemanymanywaystoperfor
mthattransformationsomestraightforwardandsomeverycomplexm
ostinvolveswappinglettersfornumbersandusemathstodothetran
sformationhowevernomatterwhichmethodisusedtheresultingscr
ambleddatastreamshouldgivenohintsabouthowitwasencrypteddu
ringworldwariithealliesscoredsomenotablevictoriesagainstt
hegermansbecausetheirencryptionsystemsdidnotsufficientlys
cramblemessagesrigorousmathematicalanalysisbyalliedcodecr
ackerslaidbarepatternshiddenwithinthemessagesandusedthemt
orecreatethemachineusedtoencryptthemthosecodesrevolvedaro
undtheuseofsecretkeysthatweresharedamongthosewhoneededtoc
ommunicatesecurelytheseareknownassymmetricencryptionsyste
msandhaveaweaknessinthateveryoneinvolvedhastopossessthesa
mesetofsecretkeys
```

The shift ciphers used have therefore been shifts by 19, 5, 3, 18, 5, 20 respectively. How would the spies have remembered this sequence? It might have been chosen as the lottery numbers one week, but actually it spells out the word SECRET, with the unusual convention a=1, b=2, c=3 and so on, perhaps to confused the enemy.

It may feel like we cheated a bit using the crib, but that is how real-world cipher cracking works. Modern ciphers are highly sophisticated algorithms designed, as far as possible, to conceal the patterns and rhythms of language so that simple frequency analysis is at best unreliable, and on its own hopelessly inadequate. Sometimes a crib is what you need, and since there is always a context to any communication a crib is often available.

In the war much was made of the fact that the naval enigma was used to transmit weather reports. Comparing these with reports from Allied vessels in the same area was sometimes all it took to crack open the key to that day's transmissions. Careless use of call signs or mission codewords can also fatally weaken the security of a cipher.

This was a far from easy exercise, and it used everything we know about letter frequencies, common patterns, cribs and the index of coincidence. Combining them has allowed us to decipher a message that would have defeated all but the best cryptographers in the past.

Deciphering a secure message is a combination of hard work, luck, knowledge and skill. But above all it takes perseverance. When one tool lets you down you need to try another, and another. And another.

As Turing said,

“No cryptogram was ever solved by simply staring at it”

NOTES

INDEX