Queen Mary
University of London

MAS 335 Cryptography

Notes 2: Substitution ciphers — Spring 2008

In the simplest (monoalphabetic) type of substitution cipher, we take a permutation of the alphabet in which the plaintext is written, and substitute each symbol by its image under the permutation. The key to the cipher is the permutation used; anyone possessing this can easily apply the inverse permutation to recover the plaintext.

If we take a piece or ordinary English text, ignore spaces and punctuation, and convert all letters to capitals, then the alphabet consists of 26 symbols, and so the number of keys is

$$26! = 403291461126605635584000000.$$

This is a sufficiently large number to discourage anyone making an exhaustive test of all possible keys. (It is approximately equal to the age of the Universe in microseconds!) However, the cipher is usually very easy to break, as we will see.

We can represent a permutation by writing down the letters of the alphabet in the usual order, and writing underneath each letter its image under the permutation. To find the inverse, write the bottom row above the top row, and then sort the columns so that the new top row is in its natural order. For example, the inverse of the permutation

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

is

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

The identity permutation is the very simple permutation which leaves each symbol where it is: not much use for enciphering!

```
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 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
```

Finally, the composition $g \circ h$ of two permutations is obtained by applying first g and then h to the alphabet.

Definition A set G of permutations forms a *group* if

(a) for all $g, h \in G$, $g \circ h \in G$;

(b) the identity permutation e belongs to G;

(c) for every $g \in G$, the inverse permutation g' belongs to G.

The *order* of the group G is the number of permutations it contains.

For example, the set of all permutations of an n-element set is a group, called the *symmetric group* of degree n and denoted by S_n. Its order is $n!$. The symmetric group S_n is the set of keys for substitution ciphers with an n-letter alphabet.

Caesar cipher

The simplest possible substitution cipher is the *Caesar cipher*, reportedly used by Julius Caesar during the Gallic Wars. Each letter is shifted a fixed number of places to the right. (Caesar normally used a shift of three places). We regard the alphabet as a cycle, so that the letter following Z is A. Thus, for example, the table below shows a right shift of 5 places.

```
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
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
```

The message "Send a hundred slaves as tribute to Rome" would be enciphered as `Xjsi f mzsiwji xqfajx fx ywngzyj yt Wtrj`. The key is simply the number of places that the letters are shifted, and the cipher is decrypted by applying the shift in the opposite direction (five places back).

Some practical details make the cipher harder to read. In particular, it would be sensible to ignore the distinction between capital and lower case letters, and also to ignore the spaces between words, breaking the text up into blocks of standard size, for example

```
XJSIF MZSIW JIXQF AJXFX YWNGZ YJYTW TRJXX
```

(We have filled up the last block with padding.)

The Caesar cipher is not difficult to break. There are only 26 possible keys, and we can try them all. In this case we would have

```
XJSIF MZSIW JIXQF AJXFX YWNGZ YJYTW TRJXX
YKTJG NATJX KJYRG BKYGY ZXOHA ZKZUX USKYY
ZLUKH OBUKY LKZSH CLZHZ AYPIB ALAVY VTLZZ
...
```

```
SENDA HUNDR EDSLA VESAS TRIBU TETOR OMESS
...
```

Almost certainly only one of the twenty-six lines will make sense, and it is easy to break it into words and discard the padding.

There are other tricks that can be used, which will be important later. As we will see in the next section, in English text, the commonest letter is usually E. Also, the consecutive letters R, S, T, U are common, and are followed by a block V, W, X, Y, Z of relatively uncommon letters. If we can spot these patterns, then we can make a guess at the correct shift. Our example is too short to show much statistical regularity; but (if we assume that the last two Xs are padding) the commonest letter is J, and the letters W, X, Y, Z are common while A, B, C, D, E are rare, so we would guess that the shift is 5 (which happens to be correct). We will look at this again in the next section.

We will in future use the convention that the plaintext is in lower case and the ciphertext in capitals.

A famous modern instance of a Caesar shift was HAL, the rogue computer in the science-fiction story *2001: A Space Odyssey*. The computer's name is a shift of IBM. (The author, Arthur C. Clarke, denied that he had deliberately done this.)

The Caesar shifts form a group. If the alphabet is $A = \{a_0, a_1, \ldots, a_{q-1}\}$, then the shift by i places can be written as $f_i : a_j \mapsto a_{j+i \bmod q}$, and we have

$$\begin{aligned} f_{i_1} \circ f_{i_2} &= f_{i_1+i_2 \bmod q}, \\ f_0 &= e, \\ f_i' &= f_{-i \bmod q}. \end{aligned}$$

The order of this group is q.

Letter frequencies

In any human language (and in most artificial languages as well), words are not random combinations of symbols, and so they will show various statistical regularities. For example, in English, the commonest letter is E; in a typical (not too short) piece of English, about 12% of all the letters will be E.

As an example, in the text of *Alice's Adventures in Wonderland*, by Lewis Carroll (AAIW for short), the frequencies of the letters (ignoring spaces and punctuation) are given in Table 1 (the figure given is the average number of occurrences among 100 letters), in the column labelled "AAIW". (The figures in the table are the average numbers of occurrences among 100 letters of text.) The columns labelled "Meaker" and "Garrett" are from the books *Cryptanalysis* by Helen Fouché Gaines, and *Making,* *Breaking Codes* by Paul Garrett. Gaines (whose book was published in 1939) took the numbers from a table by O. P. Meaker; Garrett, on the other hand, simply analysed a megabyte of old email. The French and Spanish statistics are also quoted by Gaines, from tables by M. E. Ohaver, *Cryptogram Solving*. The last column will be explained later.

Note that even for English text the figures vary, though not too much: in AAIW the most frequent letters, in order, are E, T, A, O, I, H, N, S, R, D, L, U; in Gaines' table, the order is E, T, A, O, N, I, S, R, H, L, D, U. However, in other languages the order is quite different. For example, in German, the order is typically E, N, I, R, S, A, D, T, U, G, H, O.

Figure 1 shows a histogram of the expected frequencies, together with the actual letter frequencies in the message encrypted by Caesar's cipher. It is clear by eye that the best fit is obtained if the actual message is shifted five places left.

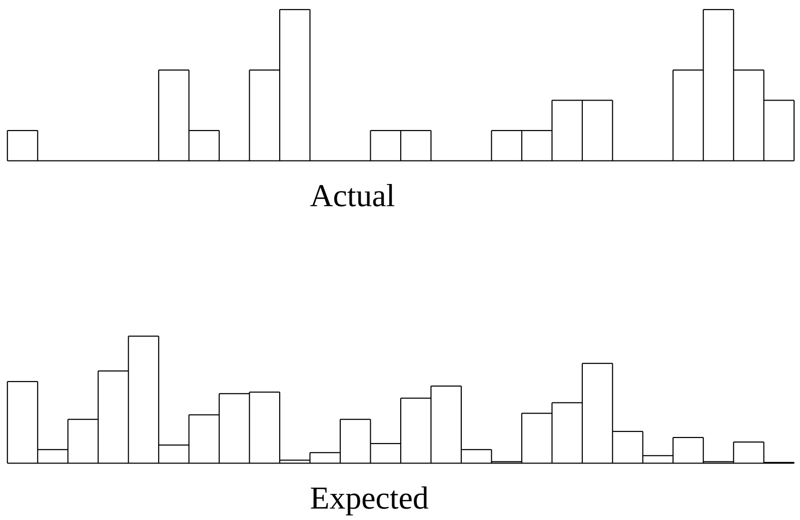

Figure 1: Expected and actual letter frequencies in Caesar cipher

Pairs of letters (referred to as *digrams*) also have their characteristic frequencies. Some of the most common in English are given in Table 2. Meaker's tables, and those of Pratt and Fraprie, are taken from Gaines.

One can also analyse trigrams, or longer sequences. Among the most commmon trigrams in English are THE, ING, THA, AND, ION.

Letter	AAIW	Meaker	Garrett	French	Spanish	Gadsby
A	8.15	8.05	7.73	9.42	12.69	10.96
B	1.37	1.62	1.58	1.02	1.41	2.14
C	2.21	3.20	3.06	2.64	3.93	2.66
D	4.58	3.65	3.24	3.38	5.58	4.12
E	12.61	12.31	11.67	15.87	13.15	0.00
F	1.86	2.28	2.14	0.95	0.46	2.15
G	2.36	1.61	2.00	1.04	1.12	3.61
H	6.85	5.14	4.52	0.77	1.24	4.91
I	6.97	7.18	7.81	8.41	6.25	8.81
J	0.14	0.10	0.23	0.89	0.56	0.23
K	1.07	0.52	0.79	0.00	0.00	1.18
L	4.37	4.03	4.30	5.34	5.94	5.32
M	1.96	2.25	2.80	3.24	2.65	2.07
N	6.52	7.19	6.71	7.15	6.95	8.61
O	7.58	7.94	8.22	5.14	9.49	10.42
P	1.40	2.29	2.34	2.86	2.43	1.91
Q	0.19	0.20	0.12	1.06	1.16	0.05
R	5.02	6.03	5.97	6.46	6.25	4.77
S	6.05	6.59	6.55	7.90	7.60	6.97
T	9.93	9.59	9.53	7.26	3.91	8.50
U	3.22	3.10	3.21	6.24	4.63	4.16
V	0.78	0.93	1.03	2.15	1.07	0.31
W	2.49	2.03	1.69	0.00	0.00	2.80
X	0.13	0.20	0.30	0.30	0.13	0.04
Y	2.11	1.88	2.22	0.24	1.06	3.18
Z	0.07	0.09	0.09	0.32	0.35	0.11

Table 1: Letter frequencies

Digraph	AAIW	Meaker	P & F	Garrett
TH	3.23	3.51	3.16	3.18
HE	3.23	2.51	1.08	2.17
AN	1.48	1.72	1.08	1.59
IN	1.89	1.69	1.57	2.59
ER	1.68	1.54	1.33	1.95
RE	1.07	1.48	1.25	1.85

Table 2: Frequencies of common digrams

As an indication of how these frequencies reflect the language, here are three "random" pieces of text. In each case, in order to split the text into words, a 27-letter alphabet (consisting of the 26 letters and the space character) has been used; any punctuation characters in the original text are regarded as spaces, and a string of spaces is reduced to a single space. In the first piece of text, the computer has generated random text using the same letter and space frequencies as in AAIW. In the second, the digram frequencies have been used; and in the third, trigram frequencies are used. Notice how the random texts resemble the original more closely as longer sequences are used.

Letter frequencies

garyrndtdbleayir hedryeeabeslt tyt watat vnot sooannaheoynoc hhh ndn e n mom scie cehealiiea yneuries u imn h utootpn eomvtet ia ecadehatyba eub e lsrv utl ecnrhmer etwtata nstp thttwttl ht tth dg uyatnpbs toinhpitehttesttthotrehushilwlhtaehyto rovt aget eaeaflrwu gnat asrl eeri luikghreborelephre hhvde egnso nodieiha dcoeothgoa tsabns s cneo ndnhfbtsont ne cpnoed m t old fzl rohuiinirtosthe arrn genialendtr hhntn tsmtr osnol ngohne aiauumnie p hhb te t gtt o araswc tak omlhidtaoi er rlumh ceca tlo acnimal tto sosi ah htoe c sty laaahsouseshi oae oh afasth wnsihnaeoawoi aesnhi yb vresptn gas elplteot or annner en s e dfhat tso nmlr te euhdre ltsnsr f reesd s cchtehavns uhtiwalo tahot lrrnnt

Digram frequencies

tre wherrltau ar a inor hee ly goove aye abinglothased as an nontte fin whike it im yon coveng a per weker ligo d ated ay s red ase ous andldrthi i anory acke owhalist the w an thi tuth abinwaly lyton bofforyilenour t n ns art asod h athostugir telidademifure bing hee hedertliryouricell araks edshe capl asove a asino thaf ar at heldryirry id and aghanorsith anesance age angh oum st athed w waronoubit ir bellea a d a at alle t quceendld hello ag t we mar ncerin avesabout ag thedoed sherkishe ano ai t ithe alkeyorated abomor p rs he ag a itainokittina acerr s abupped iranchendl whecthede awhe athai asus oo i and a s shermfu bar and a thre mer s aig it at a an y b alerd a taryouga shed f aithon iseal anghetheme as put m s n d

Trigram frequencies

ithe pits as but she i hat she peasessid to this begit a said to yout ands i loome four shone shemalice cou at sion to one se al the sped ithe gand nerse shereaverybottly embecon unnoth there pen the droqueelf land gloorger an tol the came in go the could ner so des on a wit ite bee ot the spearep onfor hown aft she is ander han ithe quive cut of ano mut andly wit it wrilice dookinam ther heseen everse ter and owles a saing alice way le jusishe s to its torrock ing teopersed show as dif to happen theirs itte heam whis way vered ant his a sairs handeauteree way murse begs a as sid s yout of ence wo cho and th ord des ned be that speopead the timessizaris ank th all guittelf to holl his and execin hand th t

Breaking a substitution cipher

Breaking a cipher is an art; it cannot be done by applying a formula. But there are some rules to follow when doing this job. Here is a partly worked example of breaking a substitution cipher; you should complete the working.

The ciphertext is:

```
RZOLB QJOWW QBWIR DQFQE VICOB OKOLR UVIDW QFMRO IVTOH
OVZMA UFUIR UVEWM DWOBH UOVYO RQRZO UBWRM TOVRW RZOSZ
ITRQW COIBQ DOTUO VYORQ RZOWR MTOVR BOYRQ BWIVT RQRZO
WRMTO VRAIT OWRIR MROWC ZUYZD QBOHO BSZIB TFSML QVRZO
ARZOL BQJOW WQBCI WJUVO TUJZO DOEIV ZUWRO IYZUV EIAUV
MROFI ROQBY QVRUV MOTIA UVMRO FQVEO BRZIV RZOJU XOTRU
AOIVT WZQMF TRZUW ZILLO VRZOW RMTOV RWCZQ JIUFO TRQFO
IHORZ OFOYR MBOBQ QAUAA OTUIR OFSCO BORZO AWOFH OWJUV
OTUVI TTURU QVRZO LBQJO WWQBC IWJUV OTUJZ OWZUB KOTOX
LFIUV UVEIT UJJUY MFRLI WWIEO QBUJZ OJIUF OTRQE ORRZB
QMEZR ZOWSF FIDMW ZOCIW JUVOT UJZOF OJRRZ OYURS JQBIT
ISCUR ZQMRR UOBOY RQBWL OBAUW WUQVI VTUJZ OAIBB UOTCI
WIFFQ COTQV FSQVO TISQJ JJQBR ZOLMB LQWOR ZOYUR SJQBU
```

```
RWLIB RRQQK IZIVT UVYQV RBQFF UVERZ OLBQJ OWWQB WIVTR
ZOSCO BOJQB YOTRQ RIKOI VQIRZ VQRRQ FOIHO DQFQE VIUVW
OIBYZ QJAQB OFMYB IRUHO QBFOW WQVOB QMWLQ WRWXX
```

We first count the frequencies of the letters. The commonest of the 715 letters, with their frequencies, are given in the table.

O	R	Q	I	U	W	V	B	Z
99	72	59	50	49	48	45	43	43

We also notice that RZ is a very common digram, with 23 occurrences. So we might guess the following identifications: `O = e, R = t, Z = h`. This gives

```
theLB QJeWW QBWIt DQFQE VICeB eKeLt UVIDW QFMte IVTeH
eVhMA UFUIt UVEWM DWeBH UeVYe tQthe UBWtM TeVtW theSh
ITtQW CeIBQ DeTUe VYetQ theWt MTeVt BeYtQ BWIVT tQthe
WtMTe VtAIT eWtIt MteWC hUYhD QBeHe BShIB TFSML QVthe
AtheL BQJeW WQBCI WJUVe TUJhe DeEIV hUWte IYhUV EIAUV
MteFI teQBY QVtUV MeTIA UVMte FQVEe BthIV theJU XeTtU
AeIVT WhQMF TthUW hILLe VtheW tMTeV tWChQ JIUFe TtQFe
IHeth eFeYt MBeBQ QAUAA eTUIt eFSCe Bethe AWeFH eWJUV
eTUVI TTUtU QVthe LBQJe WWQBC IWJUV eTUJh eWhUB KeTeX
LFIUV UVEIT UJJUY MFtLI WWIEe QBUJh eJIUF eTtQE etthB
QMEht heWSF FIDMW heCIW JUVeT UJheF eJtth eYUtS JQBIT
ISCUt hQMtt UeBeY tQBWL eBAUW WUQVI VTUJh eAIBB UeTCI
WIFFQ CeTQV FSQVe TISQJ JJQBt heLMB LQWet heYUt SJQBU
tWLIB ttQQK IhIVT UVYQV tBQFF UVEth eLBQJ eWWQB WIVTt
heSCe BeJQB YeTtQ tIKeI VQIth VQttQ FeIHe DQFQE VIUVW
eIBYh QJAQB eFMYB ItUHe QBFeW WQVeB QMWLQ WtWXX
```

The other common letters probably include `a`, `i`, `o` and `n`. Various clues help us to make the right identification. For example, consider the string `tQthe`, which occurs several times. Here, `the` is probably either a word or the beginning of a word like `then`. If this is right, `tQ` ends a word, and the most likely possibility is that `Q = o`.

Another clue is that `WW` occurs four times in the text. Double letters are not very common in English; `ee`, `ll` and `ss` are the most common, so probably `W = s`.

After a certain amount of guesswork of this sort, we begin to recognise more complicated words, and we find eventually that the substitution is

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

and the message is

> The professors at Bologna were kept in absolute and even humiliating subservience to their students. They had to swear obedience to the

> student rectors and to the student-made statutes, which bore very hardly upon them. The professor was fined if he began his teaching a minute late or continued a minute longer than the fixed time, and should this happen the students who failed to leave the lecture-room immediately were themselves fined. In addition, the professor was fined if he shirked explaining a difficult passage, or if he failed to get through the syllabus; he was fined if he left the city for a day without the rector's permission, and if he married, was allowed only one day off for the purpose. The city, for its part, took a hand in controlling the professors, and they were forced to take an oath not to leave Bologna in search of more lucrative or less onerous posts.

This description of employment conditions for academics in the Middle Ages is taken from J. D. Knowles, *The Evolution of Mediaeval Thought*.

Two fictional accounts of substitution ciphers are the stories "The Gold Bug", by Edgar Allen Poe, and "The Adventure of the Dancing Men", a Sherlock Holmes story by Sir Arthur Conan Doyle.

Worked example Solve the following substitution cipher.

```
)}&@^ {;‘?@ (‘@,( ^{?}# $‘{+^ ‘;#:^ ,(‘@? }#‘:^
;[^‘= ){*‘! }#@‘{ %^.[: ^;;)@ ){{#+ !^:;? }#={}
,;}+^ @(){* ‘!}#@ )@!#@ ,(^{? }#$‘{ {}@+^ ‘;#:^
)@,(^ {?}#$ ‘{{}@ ^.[:^ ;;)@) {{#+! ^:;?} #:={}
,_^%* ^);}& ‘+^‘* :^‘{% #{;‘@ );&‘$ @}:?= ){%..
```

Solution: This cipher is surprisingly difficult, as you will find if you try it for yourself! A hint makes it much easier. The conclusion of the message, `..`, is padding; you are told that the letter used for padding is `x`. This gives a lot of information, since `.` occurs twice in the rest of the message, and `x` is usually preceded by `e` in English; so we guess that `^` is `e`. Now we have the sequence `ex[:e;;` which is probably going to be `express`, giving us three more letters. Now finish the rest yourself!

The moral of this is that a seemingly innocent trait of the cryptographer, such as always using `x` as a filler, may give away crucial information.

Affine substitutions

The sharp-eyed will have noticed something special about the substitution used here. It maps `a` to `I`, `b` to `D`, `c` to `Y`, and so on; advancing the plain letter one place moves the cipher letter back five places (or forward 21 places). In otherwords, if the letters of the alphabet are numbered from 0 to 25, so that a is represented by 0, b by 1, ..., z by 25, then the substitution takes the form

$$x \mapsto 8 + 21x \pmod{26}.$$

Such a substitution, or the cipher it generates, is called *affine*.

The Caesar shift is a special case of an affine cipher, having the form

$$x \mapsto x + b \pmod{26}$$

for some fixed b. The general form of an affine cipher is

$$x \mapsto ax + b \pmod{26}$$

for some fixed a and b. The advantage is that the key is simple; instead of needing a general permutation of the letters, we only need the numbers a and b mod 26.

What affine ciphers are possible, and how can they be inverted?

First we must decide when an affine substitution is a permutation. Consider the substitution $\theta : x \mapsto ax + b \pmod{n}$. It will fail to be a permutation if there exist two distinct x_1, x_2 with

$$ax_1 + b \equiv ax_2 + b \pmod{n},$$

that is, $ay \equiv 0 \pmod{n}$, where $y = x_2 - x_1$. So θ is a permutation if and only if the congruence $ay \equiv 0 \pmod{n}$ has a solution $y \not\equiv 0 \pmod{n}$.

Let d be the greatest common divisor of a and n. Then, say, $a = a_1 d$ and $n = n_1 d$ for integers a_1, n_1. Suppose that $d > 1$, so that $n_1 < n$. Putting $y = n_1$, we have

$$ay = a_1 d n_1 = a_1 n \equiv 0 \pmod{n},$$

so θ fails to be a permutation.

Conversely, suppose that $d = \gcd(a, n) = 1$. By Euclid's Algorithm (see the end of this chapter), there exist integers u, v such that $au + nv = 1$. Now, if $ay \equiv 0 \pmod{n}$, then

$$y = (au + nv)y = u(ay) + n(vy) \equiv 0 \pmod{n},$$

so θ is a permutation.

We conclude:

Theorem 1 *The affine map $x \mapsto ax + b$ is a permutation if and only if $\gcd(a, n) = 1$.*

What happens if we compose two such maps? Write $\theta_{a,b}$ for the map $x \mapsto ax + b \pmod{n}$, where $\gcd(a, n) = 1$. We have

$$\theta_{a,b} \circ \theta_{a',b'} : x \mapsto ax + b \mapsto a'(ax + b) + b',$$

so $\theta_{a,b} \circ \theta_{a',b'} = \theta_{aa',ba'+b'}$.

The identity permutation $x \mapsto x$ is the map $\theta_{1,0}$. So to find the inverse of $\theta_{a,b}$ in the form $\theta_{a',b'}$, we have to solve the congruences

$$\begin{aligned} aa' &\equiv 1 \pmod{n}, \\ ba' + b' &\equiv 0 \pmod{n}. \end{aligned}$$

The first congruence has a unique solution mod n, which can be found by Euclid's Algorithm as before. Then the second congruence also has a unique solution, namely $b' \equiv -ba' \pmod{n}$.

In particular, with $n = 26$, we want to invert the map $\theta_{21,8}$. By trial and error (or by Euclid's Algorithm), $21 \cdot 5 \equiv 1 \pmod{26}$; and then $-5 \cdot 8 \equiv 12 \pmod{26}$. So the inverse of $\theta_{21,8}$ is $\theta_{5,12}$.

Definition *Euler's totient function* ϕ is the function on the natural numbers given by

$$\phi(n) = \begin{cases} \text{number of congruence classes } a \text{ mod } n \\ \text{such that } \gcd(a,n) = 1. \end{cases}$$

We give a formula for it, which will be proved later.

Theorem 2 *Let $n = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$, where $p_1, p_2, \ldots, p_r$ are distinct primes and $a_1, a_2, \ldots, a_r > 0$. Then*

$$\phi(n) = p_1^{a_1-1}(p_1 - 1) p_2^{a_2-1}(p_2 - 1) \cdots p_r^{a_r-1}(p_r - 1).$$

For example, $26 = 2 \cdot 13$, so $\phi(26) = 1 \cdot 12 = 12$. The congruence classes coprime to 26 are represented by the odd numbers from 1 to 25 excluding 13.

Theorem 3 *The set of affine permutations mod n is a group of order $n \cdot \phi(n)$.*

We have verified the group properties in the earlier argument. For the order, note that there are $\phi(n)$ choices for a and n choices for b.

There are thus $26 \cdot 12 = 312$ affine permutations. If we know or suspect that a substitution cipher is affine, we could try all 312 keys, though this is not trivial by hand. The method of looking for patterns of consecutive letters (as used to crack the Caesar cipher) does not apply. Like any substitution cipher, an affine cipher is vulnerable to frequency analysis. Its advantage is the small size of the key (two numbers rather than a complete permutation.)

Worked example Decrypt the following affine substitution cipher:

```
JZQOU DQGKZ UULYU MKUOX LQJQJ ZQZCW ZQDYU MDXUJ
QRJCE LQEDR CRWGL UUIEJ JZQEP QDEWQ QEDRC RWGCR
JZCGK ZEDJJ ZQYJQ LLJZQ GJUDY
```

You are given that the frequency distribution in the ciphertext is as follows:

C	D	E	G	I	J	K	L	M	O	P	Q	R	U	W	X	Y	Z
6	8	7	5	1	13	3	6	2	2	1	15	6	10	4	2	4	10

Solution The commonest letter `Q` in the given cipher is likely to be `e`. We also see that the trigram `JZQ` occurs five times and so is likely to be `the`. This gives `J`=`t` and `Z`=`h`.

The letters `Q` and `Z` are x_{16} and x_{25} (where $q = 26$ here), while `e` and `h` and x_4 and x_7. Thus the parameters c and d satisfy

$$\begin{aligned} 4c+d &\equiv 16 \pmod{26}, \\ 7c+d &\equiv 25 \pmod{26}, \end{aligned}$$

from which we find $c = 3$ and $d = 4$. Now we can compute the inverse of this affine transformation, which will be the decryption map: if the inverse is $i \mapsto c'i + d'$, then we have (using the formula we worked out earlier)

$$\begin{aligned} 3c' = 1, &\quad \text{so} \quad c' = 9; \\ 4c' + d' = 0, &\quad \text{so} \quad d' = 16. \end{aligned}$$

From this the entire substitution can be worked out, and we find the plaintext to be

```
themo resch oolyo ucomp letet hehig heryo urpot
entia learn ingsl ookat theav erage earni ngsin
thisc hartt heyte llthe story
```

or, correctly spaced and with punctuation,

> The more school you complete, the higher your potential earnings. Look at the average earnings in this chart; they tell the story!

Making a substitution cipher safer

A substitution cipher can be solved by frequency analysis, and so is insecure for all but the shortest messages. However, there are some improvements that can be made. The first two rely on using a different alphabet for the ciphertext, with more characters than the plaintext alphabet. For example we could use an alphabet of 100 characters, represented by symbols $00, 01, \ldots, 99$.

Nulls: These are additional symbols in the cipher alphabet which do not have any meaning but are inserted in random positions to confuse the frequency analysis.

Homophones: We can translate the same letter in plaintext by several different letters in ciphertext. For example, if we use a 100-character cipher alphabet, we can associate about as many characters with each plaintext letter as its percentage frequency in normal text (say, 12 characters for `e`, 9 for `t`, and so on). Then we randomly decide which character to substitute for each occurrence of a letter. In the ciphertext, each character will occur approximately the same number of times. However, the ciphertext is still not random, and patterns of digraphs and trigraphs can be recognised.

Use of language: We can further confuse the analysis by using words from other languages, or by careful choice of words. As an example of what can be done, at least two English novels have been written containing no occurrence of the letter `e`, the commonest letter in English. One of these is *Gadsby*, by Ernest Vincent Wright. The author tied down the `E` key of his typewriter to write the book. The first paragraph reads as follows:

> If youth, throughout all history, had had a champion to stand up for it; to show a doubting world that a child can think; and, possibly, do it practically; you wouldn't constantly run across folks today who claim that "a child don't know anything." A child's brain starts functioning at birth; and has, amongst its many infant convolutions, thousands of dormant atoms, into which God has put a mystic possibility for noticing an adult's act, and figuring out its purport.

To a casual glance, there is nothing odd about this; but it would pose an obvious problem for a cryptanalyst if encrypted with a substitution cipher. A frequency analysis of *Gadsby* is included in Table 1.

The novel *A Void* is even more remarkable, having been translated by Gilbert Adair from the French novel *La Disparition* by Georges Perec, which also lacked the letter `e`.

Another trick is to write words "phonetically", or to use text-messaging abbreviations.

Features of text messaging language such as phonetic spelling (such as "nite" for "night"), the common omission of vowels ("txt" for "text"), use of abbreviations (such as AFAIK for "as far as I know"), use of numerals `2`, `4` and `8` for `to`, `for` and `ate`, and use of "emoticons" such as `;-)` as an essential part of the text, would give frequency analysis quite different from standard English. I don't know whether such analysis of a body of text messages has been done.

Transposition: The substitution can be combined with *transposition*, that is, permuting the order of the characters in the ciphertext in a specified way. This will help to destroy the patterns of digram and trigram frequencies.

With these improvements, even a substitution cipher can be effective for a short message which will not receive very sophisticated analysis.

Number theory

In this section we give more details of some of the number theory which underlies our discussion of affine ciphers.

Euclid's algorithm

Euclid's algorithm is a procedure to find the greatest common divisor of two integers. In the form of a one-line recursive program it can be written as follows:

`if` $b = 0$ `then` $\gcd(a,b) := a$ `else` $\gcd(a,b) := \gcd(b, a \bmod b)$ `fi`

where $a \bmod b$ means the remainder on dividing a by b.

For example,

$$\gcd(30,8) = \gcd(8,6) = \gcd(6,2) = \gcd(2,0) = 2.$$

The algorithm can also be used to write $\gcd(a,b)$ in the form $ua + vb$ for some integers u, v. We express each successive remainder in this form until we reach the last non-zero remainder, which is the gcd. In the above example,

$$\begin{aligned} 6 &= 30 - 3 \cdot 8 \\ 2 &= 8 - 1 \cdot 6 \\ &= 8 - (30 - 3 \cdot 8) \\ &= (-1) \cdot 30 + 4 \cdot 8, \end{aligned}$$

so $u = -1$, $v = 4$.

This can be used to find inverses mod n. For example, $\gcd(21,26) = 1$, and Euclid's algorithm shows that $1 = (-4) \cdot 26 + 5 \cdot 21$; so $5 \cdot 21 \equiv 1 \pmod{26}$, and the inverse of 21 mod 26 is 5.

Euler's function

In this section we prove Theorem 2. We begin with the theorem known as the *Chinese Remainder Theorem*.

The following discussion is based on the section on Chinese mathematics in George Gheverghese Joseph, *The Crest of the Peacock: Non-European Roots of Mathematics*, Penguin Books 1992. The fourth-century text *Sun Tsu Suan Ching* (Master Sun's Arithmetic Manual) contains the following problem:

> There is an unknown number of objects. When counted in threes, the remainder is 2; when counted in fives, the remainder is 3; when counted in sevens, the remainder is 2. How many objects are there?

The problem asks for an integer N such that $N \equiv 2 \pmod 3$, $N \equiv 3 \pmod 5$, and $N \equiv 2 \pmod 7$. One solution is given as

$$N = 2 \cdot 70 + 3 \cdot 21 + 2 \cdot 15 = 233;$$

it is clear that adding or subtracting a multiple of 105 from any solution gives another solution; so the smallest solution is

$$N = 233 - 2 \cdot 105 = 23.$$

A folk-song gives the mnemonic:

> Not in every third person is there one aged three score and ten,
> On five plum trees only twenty-one boughs remain,
> The seven learned men meet every fifteen days,
> We get our answer by subtracting one hundred and five over and
> over again.

Why does it work? Observe that 70 is congruent to 1 mod 3, to 0 mod 5, and to 0 mod 7, and similarly for 21 and 15; then $70a + 21b + 15c$ is congruent to a mod 3, to b mod 5, and to c mod 7, as required. But how do we find these numbers 70, 21 and 15? Well, the first number is supposed to be divisible by 5 and 7, so is a multiple of 35; then 35 is congruent to 2 mod 3, so 2.35 is congruent to 2.2, which is congruent to 1 mod 3, as required. (In this last step we multiplied by the *inverse* of 2 modulo 3. In more difficult cases we can use Euclid's algorithm to find the appropriate inverse.)

A similar procedure works in general. The fact that we can always find numbers with the required congruence conditions is not entirely obvious, but follows from Euclid's algorithm using the fact that the moduli are coprime. We give the result just for two moduli: it is easily extended to any number by induction.

Let $\mathbb{Z}/(n)$ denote the set of congruence classes mod n. It is clear that, if $x \equiv x' \pmod{mn}$, then $x \equiv x' \pmod m$; so, for $x \in \mathbb{Z}/(mn)$, there is a well-defined element x mod m of $\mathbb{Z}/(m)$. Similarly with n replacing m.

Theorem 4 (Chinese Remainder Theorem) *If* $\gcd(m,n) = 1$, *then the map* F *from* $\mathbb{Z}/(mn)$ *to* $\mathbb{Z}/(m) \times \mathbb{Z}/(n)$ *defined by*

$$F(x) = (x \bmod m, x \bmod n)$$

is a bijection.

Proof: Suppose that $F(x) = F(x')$. Then $x \bmod m = x' \bmod m$, that is, m divides $x - x'$. Similarly n divides $x - x'$. Since m and n are coprime, it follows that mn divides $x - x'$, so that $x = x'$ (as elements of $\mathbb{Z}/(mn)$). Thus F is one-to-one.

Now $|\mathbb{Z}/(mn)| = mn = |\mathbb{Z}/(m) \times \mathbb{Z}/(n)|$; so F must also be onto, and thus a bijection.

This proof is non-constructive, whereas the original Chinese argument gave an algorithmic way to compute the inverse of F. This can be generalised as follows. Since $\gcd(m,n) = 1$, Euclid's algorithm shows that there exist numbers a and b such that $am + bn = 1$. Now we see that

$$\begin{aligned} am &\equiv 0 \pmod{m}, & am &\equiv 1 \pmod{n}, \\ bn &\equiv 1 \pmod{m}, & bn &\equiv 0 \pmod{n}, \end{aligned}$$

so the solution to the simultaneous congruences

$$x \equiv y \pmod{m}, \qquad x \equiv z \pmod{n}$$

is given by

$$x \equiv bny + amz \pmod{mn}.$$

Remark: In fact F is a *ring isomorphism*: this simply means that $F(x + x') = F(x) + F(x')$ and $F(xx') = F(x)F(x')$.

Now $\gcd(x, mn) = 1$ if and only if $\gcd(x,m) = 1$ and $\gcd(x,n) = 1$. Since Euler's function gives the number of congruence classes coprime to the modulus, the Chinese Remainder Theorem implies that

$$\phi(mn) = \phi(m)\phi(n)$$

if $\gcd(m,n) = 1$.

This extends to products of any number of pairwise coprime factors. Thus

$$\phi(p_1^{a_1} \cdots p_r^{a_r}) = \phi(p_1^{a_1}) \cdots \phi(p_r^{a_r})$$

if $p_1, \ldots, p_r$ are distinct primes.

So, to complete the proof of the theorem, we have to show only that $\phi(p^a) = p^{a-1}(p-1) = p^a - p^{a-1}$ for p prime and $a > 0$. This holds because, of the p^a congruence classes mod p^a, the ones not coprime to p^a are precisely those divisible by p, of which there are p^{a-1}.