Chapter 2
Cryptology Before 1500: A Bit of Magic

Abstract Cryptology was well established in ancient times, with both Greeks and Romans practicing different forms of cryptography. With the fall of the Roman Empire, cryptology was lost in the West until the Renaissance, but it flourished in the Arabic world. The Arabs invented the first reliable tool for cryptanalysis, frequency analysis. With the end of the Middle Ages and the increase in commerce and diplomacy, cryptology enjoyed a Renaissance of its own in the West. This chapter examines the most common cipher of the period, the monoalphabetic substitution cipher and then looks at the technique of frequency analysis that is used to break the monoalphabetic substitution. An extended example is given to illustrate the use of frequency analysis to break a monoalphabetic.

2.1 Veni, Vidi, Cipher

Julius Caesar, probably the greatest of all Roman generals, was no stranger to cryptology. In his famous *Commentary on the Gallic Wars*, Caesar himself describes using a form of a cipher to hide a message.

> Then with great rewards he induces a certain man of the Gallic horse to convey a letter to Cicero. *This he sends written in Greek characters, lest the letter being intercepted, our measures should be discovered by the enemy.* He directs him, if he should be unable to enter, to throw his spear with the letter fastened to the thong, inside the fortifications of the camp. He writes in the letter, that he having set out with his legions, will quickly be there: he entreats him to maintain his ancient valor. The Gaul apprehending danger, throws his spear as he has been directed. It by chance stuck in a tower, and, not being observed by our men for two days, was seen by a certain soldier on the third day: when taken down, it was carried to Cicero. He, after perusing it, reads it out in an assembly of the soldiers, and fills all with the greatest joy. Then the smoke of the fires was seen in the distance, a circumstance which banished all doubt of the arrival of the legions [1, Chap. 48, italics added].

This, however, is not Caesar's most famous contribution to the history of cryptology. The Roman historian Gaius Suetonius Tranquillus, in his *The Twelve Caesars* describes Julius Caesar's use of a cipher to send messages to his friends

J. F. Dooley, *A Brief History of Cryptology and Cryptographic Algorithms*,
SpringerBriefs in Computer Science, DOI: 10.1007/978-3-319-01628-3_2,

and political allies. This was a cipher that, according to Seutonius, "If he had anything confidential to say, he wrote it in cipher, that is, by so changing the order of the letters of the alphabet, that not a word could be made out. If anyone wishes to decipher these, and get at their meaning, he must substitute the fourth letter of the alphabet, namely D, for A, and so with the others" [3, Chap. 56]. This is the first written description of the modern monoalphabetic substitution cipher using a shifted standard alphabet. Using Caesar's cipher, the cipher alphabet looks like

```
Plain:  abcdefghijklmnopqrstuvwxyz
Cipher: DEFGHIJKLMNOPQRSTUVWXYZABC
```

and Caesar's famous "I came, I saw, I conquered" would be enciphered as `L FDPH, L VDZ, L FRQTXHUHG`.

2.2 Cryptology in the Middle Ages

For 900 years the monoalphabetic substitution cipher was the strongest cipher system in the Western world. The Romans used it regularly to protect their far-flung lines of communication. But after the fall of the Western Roman Empire in 476 C.E. the knowledge of cryptology vanished from the West and wasn't to return until the Italian Renaissance. Indeed, with the decline of literacy and scholarship in Europe during the Dark Ages following the fall of Rome cryptology turned from a useful technique for keeping communications secret into a dark art that bordered on magic.

But interest in cryptology was not dead. In the latter part of the first millennium, there was another place where intellectual curiosity and scholarship flowered and where mathematics and cryptology saw their biggest advances since Caesar—the Arab world. And this was where the next big advance in cryptanalytic techniques would come from.

The period around the 9th century C.E. is considered to be the beginning of the Islamic Golden Age, when philosophy, science, literature, mathematics, and religious studies all flourished in what was then the peace and prosperity of the Abbasid Caliphate. Into this period was born Abu Yūsuf Ya-qūb ibn Isāq as-Sabbāh al-Kindi (801–873 C.E.), a polymath who was the philosopher of the age. Al-Kindi wrote books in many disciplines including astronomy, optics, philosophy, mathematics, medicine, and linguistics, but his book on secret messages for court secretaries, A *Manuscript on Deciphering Cryptographic Messages* is the most important to the history of cryptology. It is in this book that the technique of *frequency analysis* is first described.

2.3 Frequency Analysis, the First Cryptanalytic Tool

In every language, if one is given a text of several hundred or thousand characters and the individual letters in the text are counted, some of the letters will appear more often than others, and some will appear very infrequently. If another text of

Table 2.1 English frequency percentages

Letter	Percentage	Letter	Percentage
A	8.4	N	7.0
B	1.7	O	7.3
C	3.1	P	2.0
D	4.4	Q	0.1
E	12.7	R	6.3
F	2.0	S	6.0
G	2.0	T	9.3
H	5.4	U	2.4
I	7.0	V	1.0
J	0.2	W	2.0
K	0.7	X	0.2
L	4.0	Y	2.2
M	2.5	Z	0.1

similar length is analyzed in the same way, the same letters will pop up as either more frequently occurring or less frequently occurring. Thus, the *frequency of occurrence* of individual letters is a characteristic of the language.

It is also impossible to hide this frequency of occurrence if one substitutes one letter for another in a message. What al-Kindi discovered is that in a message enciphered using a monoalphabetic substitution cipher, the language characteristics are not hidden by the substitution. In particular the letter frequencies will shine through the substitution like a beacon leading the cryptanalyst to the concealed letters of the plaintext.

In English, the most frequently occurring letters are usually given in the order of ETAOINSHRDLU. Table 2.1, which was constructed by counting all 95,512 or so words (450,583 letters) in David Kahn's biography of Herbert O. Yardley, *The Reader of Gentlemen's Mail* illustrates the ordering for modern English usage.

Graphically, this looks like Fig. 2.1.

The technique of frequency analysis is to do the same count of letters for the ciphertext, and then use those counts to guess at the letters of the ciphertext. Thus, the most frequently occurring letter in the ciphertext should represent e. The next most frequently occurring should represent t, then a, etc. al-Kindi laid all this out in a few short paragraphs and with it revolutionized cryptanalysis.

One does not need to be restricted to just single letter frequencies when doing this type of analysis. It turns out that there are also pairs of letters (digraphs) that occur with great frequency and pairs that don't occur at all. For example, in English, the most frequent pairs of letters are *th*, *he*, *in*, *er*, *an*, *re*, and *nd*. And one could continue with the most common three letter words in English, *the*, *and*, *for*, *not*, and *you*.

To illustrate the technique of frequency analysis, lets decrypt an English cryptogram that was created using a monoalphabetic substitution cipher. How should we go about decrypting the following cryptogram?

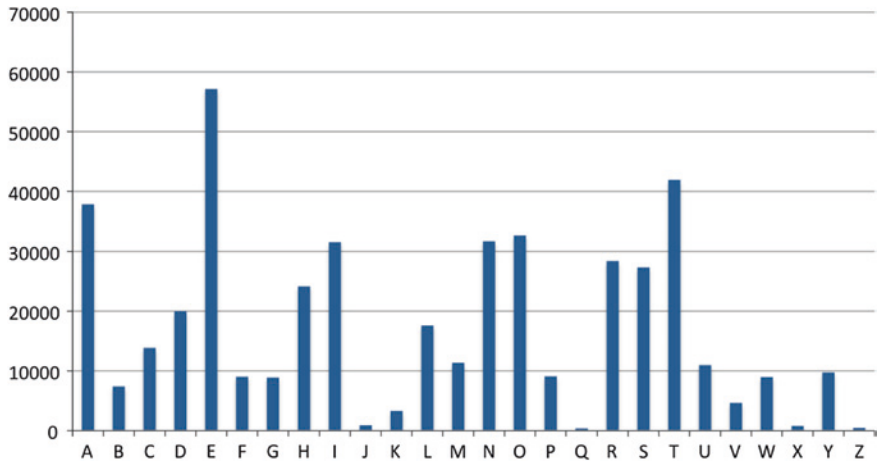

Fig. 2.1 A graph of English letter frequencies

```
SCEAC SKDXA CESDS CKVSO LCDDA GKEMG AMTYK TOVKS OSFNC
FPCEE XMTDA OLTCQ OLGKG ACOKS ADSFN EGFGN KCHLQ HGFOL
TMQRI TYOSF VLSYL SCFCD XMTGF TLQFP KTPCF PMSWO XMTHC
KCOTY SHLTK MRQOS YGFAT MMOLC OOLSM SMTFO SKTDX FTVOG
ETOLT GRITY OGAOL GMTVL GSFUT FOTPO LTMXM OTELC MCHHC
KTFOD XRTTF OGYGF YTCDO LCOOL TMTYL CKCYO TKMYG FUTXC
ETMMC NTCFP OGNSU TOLTS PTCOL COOLT XCKTO LTETK TKCFP
GEMBT OYLTM GAYLS DPKTF CKOLQ KYGFC FPGXD TOLTC PUTFO
QKTGA OLTPC FYSFN ETF
```

We begin by counting all the letters in the cryptogram and producing two things—a frequency table and a frequency chart. The frequency table looks like Table 2.2.

And the frequency chart for the cryptogram looks like Fig. 2.2.

Looking at the many ups and downs in the frequency chart we can easily see that this is a monoalphabetic substitution. With the T being so much higher than any of the other letters, it is our top candidate for *e*. O and C look like candidates to be the next two highest frequency letters *t* and *a*, but which is which we don't know yet. Remember that the frequency count for English is based on a very large number of letters, while the frequency count for a single cryptogram is based on many fewer letters. That fact may skew some of the frequencies and the overall distribution.

Our next step is to try to break down the letters in the cryptogram into at least three different groups—high frequency letters, medium frequency, and low frequency. In standard English, *e, t, a, i, o, n, r, s*, and *h* form the high-frequency letters—defined as those with a frequency percentage of greater than 5 % for our purposes, For the medium frequency group we have *c, d, f, g, l, m, p, u, w*, and *y* and for the low-frequency letters (at less than 2 % of the count each) we have *b, j, k, q, v, x*, and *z*. So if we can identify these groups in the cryptogram we could be

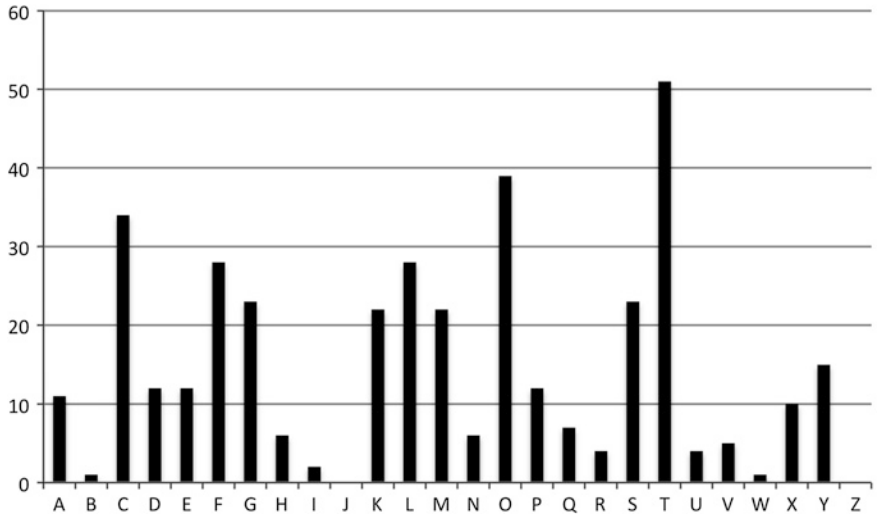

Fig. 2.2 Cryptogram frequency chart

Table 2.2 Cryptogram frequency count

A	B	C	D	E	F	G	H	I	J	K	L	M
11	1	34	12	12	28	23	6	2	0	22	28	22
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
6	39	12	7	4	23	51	4	5	1	10	15	0

Table 2.3 Frequency count in descending order

T	O	C	F	L	G	S	K	M	Y	D	E	P
51	39	34	28	28	23	23	22	22	15	12	12	12
A	X	Q	H	N	V	R	U	I	B	W	J	Z
11	10	7	6	6	5	4	4	2	1	1	0	0

on our way to getting the entire cipher alphabet. If we re-arrange Table 2.2 so that the letters are written in descending order by count, we get Table 2.3.

Ignoring the large dip between the T and the O, the next big dip in frequency is a dip of 7 between the M and the Y, conveniently between the ninth and tenth letters, just where the dip between the high and medium frequency letters is. Now that we have a feel for how the individual letters are arranged, it is time to look at digraphs. Digraphs give us a feel for how the letters arrange themselves next to other letters. We've seen that *th*, *he*, *in*, *er*, *an*, *re*, and *nd* are the most common digraphs, so it should be the case that some pairs of letters in the cryptogram behave similarly.

Looking at the digraphs we see that OL is the most frequently occurring digraph at 18. LT occurs 12 times (and the three-letter group OLT occurs 9 times),

KT eight times, MT, CF, GF, and TF all occur seven times, and TM occurs six times. If we assume that OL is the digraph *th*, and LT is the digraph *he*, we then have good confirmation that O = t, L = h, and T = e.

The next thing is to identify the other high-frequency letters, especially the vowels, *a*, *i*, and *o*. The next three highest frequency ciphertext letters are C, F, and L. We also note that the sequence OLCO occurs three times in the cryptogram. Given what we already know, this sequence decrypts to *th*t*, which could be the word *that*, leaving C = a. This replacement also gives us the popular digraph *ea* five times in the deciphered part of the cryptogram, a good sign.

The next high frequency digraph is *in* which also includes two letters from the high-frequency letter group. Looking carefully through the ciphertext, we see that S occurs 23 times and F occurs 28 times. This might lead us to believe that F = *i* and S = *n*. If we substitute these new pairs, however, we get decrypted sequences like LSCFC = *hnaia* and OLCOOLS = *thatths*, neither of which look promising. If instead we see that the digraph SF occurs 5 times and the trigraph SFN occurs twice we can go further. If SF = *in* then it is possible that SFN = *ing* allowing us to supposed that S = *i*, F = *n*, and N = *g*. This will also give us the trigraph *ent* in 5 different places; another good sign. Putting those guesses into the ciphertext we end up with the partial solution

```
SCEACSKDXACESDSCKVSOLCDDAGKEMGAMTYKTOVKSOSFNCFPCEE
ia  ai    a i ia  itha          e  et  itingan a
XMTDAOLTCQOLGKGACOKSADSFNEGFGNKCHLQHGFOLTMQRITYOSF
  e  thea th    at i  ing  n g a h   nthe    e tin
VLSYLSCFCDXMTGFTLQFPKTPCFPMSWOXMTHCKCOTYSHLTKMRQOS
 hi hiana   e neh n  e an  i t  e a ate i he    ti
YGFATMMOLCOOLSMSMTFOSKTDXFTVOGETOLTGRITYOGAOLGMTVL
  n e  thatthi i enti e  ne t  ethe   e t  th  e h
GSFUTFOTPOLTMXMOTELCMCHHCKTFODXRTTFOGYGFYTCDOLCOOL
 in ente the   te ha a  a ent   eent   n ea thatth
TMTYLCKCYOTKMYGFUTXCETMMCNTCFPOGNSUTOLTSPTCOLCOOLT
e e ha a te    n e a e  agean t gi ethei eathatthe
XCKTOLTETKTKCFPGEMBTOYLTMGAYLSDPKTFCKOLQKYGFCFPGXD
 a ethe e e an     et he    hi   ena th    nan
TOLTCPUTFOQKTGAOLTPCFYSFNETF
ethea  ent  e  the an ing en
```

Of the high frequency letters we still need to assign *o*, *r*, and *s*. We notice that the digraph GF occurs seven times. That represents ?n in plaintext, indicating that the ? is probably a vowel. The only two vowels left are *o* and *u* and the sequence *on* occurs much more frequently in English than *un*, so it is possible that G = *o*. We also see the sequence OLCOOLSMSM, which is currently decrypted as *thatthi?i?* and which might logically decrypt as *that this is* if M = *s*. In addition, there are two double M's in the cryptogram, reinforcing the idea that M = *s*. Finally, for the high-frequency letters we notice that there are 8 KT pairs in the cryptogram. We

already know that T = *e* and we also know that *re* is a high-frequency digraph, so its possible that K = *r*. Adding these to the ciphertext we end up with

```
SCEACSKDXACESDSCKVSOLCDDAGKEMGAMTYKTOVKSOSFNCFPCEE
ia  air   a i iar itha   or so se ret ritingan a
XMTDAOLTCQOLGKGACOKSADSFNEGFGNKCHLQHGFOLTMQRITYOSF
 se  thea thoro atri  ing onogra h  onthes   e tin
VLSYLSCFCDXMTGFTLQFPKTPCFPMSWOXMTHCKCOTYSHLTKMRQOS
 hi hiana  seoneh n re an si t se arate i hers  ti
YGFATMMOLCOOLSMSMTFOSKTDXFTVOGETOLTGRITYOGAOLGMTVL
 on essthatthisisentire  ne to etheo  e to those h
GSFUTFOTPOLTMXMOTELCMCHHCKTFODXRTTFOGYGFYTCDOLCOOL
oin ente thes ste hasa  arent   eento on ea thatth
TMTYLCKCYOTKMYGFUTXCETMMCNTCFPOGNSUTOLTSPTCOLCOOLT
ese hara ters on e a essagean togi ethei eathatthe
XCKTOLTETKTKCFPGEMBTOYLTMGAYLSDPKTFCKOLQKYGFCFPGXD
 arethe ereran o s et heso  hi  renarth r onan o
TOLTCPUTFOQKTGAOLTPCFYSFNETF
ethea  ent reo the an ing en
```

This is the breakthrough we needed. The analysis now depends on guessing possible words that we can see hints of in the partially decoded ciphertext. It is easy to see words like *writing*, *message*, *separate*, *secret*, etc. and we can now uncover the plaintext in short order. The final plaintext is (with punctuation added).

> I am fairly familiar with all forms of secret writing, and am myself the author of a trifling monograph upon the subject, in which I analyse one hundred and sixty separate ciphers, but I confess that this is entirely new to me. The object of those who invented the system has apparently been to conceal that these characters convey a message, and to give the idea that they are the mere random sketches of children. Arthur Conan Doyle, "The Adventure of the Dancing Men" [2].

So what is the process of cryptanalysis here? We begin with two facts, the relative frequency counts in English, and the behavior of digraphs and trigraphs as they appear in words in English. Then we get the actual frequency counts in the cryptogram and use our knowledge to try to identify the high-frequency letters and digraphs in the cryptogram. Once we have a partial reconstruction using the high-frequency letters we can then begin to guess whole words, filling in more letter equivalents as we go.

References

1. Caesar, Julius. 2008. *The Gallic Wars*. Hardcover. Oxford, UK: Oxford University Press.
2. Doyle, Sir Arthur Conan. 1903. The adventure of the dancing men. *The Strand Magazine*.
3. Seutonius. 1957. *The Twelve Caesars*. Paperback. Trans Robert Graves. London, UK: Penguin Classics.