1 Introduction to Cryptography

1.1 Language

Cryptology is the study of sending (hopefully) secure messages over non-secure channels. We sometimes use the term *cryptography* for the study of designing such systems; *cryptanalysis* is the study of breaking the security of such systems.

There are three basic types of cryptography.

1.1.1 Steganography

Steganography is the concealing of your message so that evesdroppers can't tell you're sending a message. When this works it is incredibly effective; if no one knows you're sending a message at all, then no one is even trying to read it.

On the other hand, steganography doesn't work very well at all if people know what you're doing. Once they know where to look, your steganography has failed, so it is very fragile.

Famously, the ancient Greek tyrant Histiaeus wanted to send secret message plotting a revolt against the Persians to his nephew Aristagoras. He shaved the head of one of his slaves, tattooed a message on his head, then waited for the hair to grow back. Aristagoras was instructed to shave the slave's head and read the message.

This worked very effectively, but would not have succeeded at all if slave-head-tattooing had been a common method of sending secret messages. It also didn't have very good performance: it took months to send a short message.

Modern security practice rarely involves the tattooing of slaves, but we do have more sophisticated steganographic techniques. A common one involves hiding a message in an image file by adjusting the least significant digit of the color intensity of each pixel. This will change the image imperceptibly but allows the encoding of about one character for every three pixels of image. (If you've read the Orson Scott Card novel *Shadow of the Hegemon*, this sort of technique is an important plot point there).

More prominently, a Russian hacking group called “Turla” recently used the comments in Britney Spears's instagram account to secretly send instructions and updates to infected computers. Messages like `#2hot make loved to her, uupss #Hot #X` were secretly encoded addresses for new control servers for the malware.

We won't spend a lot of time in this course discussing steganography, because the actaul steganographic part of the message tends to involve abstractly clever hiding places more than the use of interesting mathematics.

1.1.2 Codes

Codes are prearranged set of signals and representations for specific meanings.

Codes make substitutions at the level of words or phrases; the stereotypical spy-talk “The eagle flies at midnight” is a code phrase because it sends a specific and complex prearranged thought.

A code is effectively a parallel language, and in fact uncommon and poorly-known languages have been used as codes. In World War II, the US Army often used Native Americans speaking to each other in their native languages (“code talkers”) to transmit information that their opponents could not decipher through any sort of mathematical analysis.

We will again mostly not discuss codes, because they primarily aren’t mathematically interesting; they are abstract enough that in effect, you either know the correspondences or you don’t. (The primary disadvantage of codes is that you have to communicate and protect very extensive codebooks for them to be practicable). However, we may discuss the idea of *encoding* and *coding theory*, which are related but distinct.

1.1.3 Ciphers

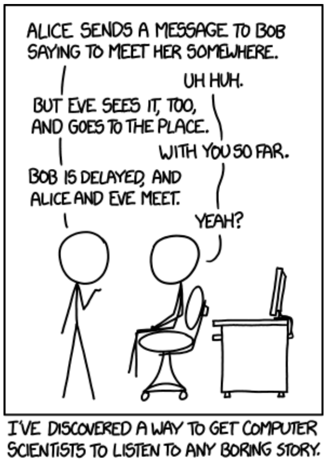
Ciphers are the primary subject of this course.

In contrast to codes, which make substitutions at the level of meanings and words and phrases, ciphers replace individual characters (or bits) of data.

Ciphers are very flexible and can encrypt arbitrary messages, since they replace character-by-character.

The basic setup we wish to study features two parties (usually named *Alice* and *Bob*) who wish to communicate with each other, but wish to prevent a third party “evesdropper” or “assailant” (usually named *Eve*) from understanding their messages.

The message Alice wishes to send to Bob is called the *plaintext*. She will use some prearranged encryption method to convert it into a *ciphertext*, which she sends to Bob. Bob uses his knowledge of the encryption method to convert the ciphertext back into the plaintext; even if Eve acquires the ciphertext, she



`https://xkcd.com/1323`

shouldn't be able to perform the same conversion to acquire (or, in more malicious scenarios, replace) the plaintext.

1.2 Caesar Ciphers

One of the earliest known ciphers is what's known as the *Caesar cipher* or *Shift cipher*. (Julius Caesar was hardly the first to use them; there are examples dating back to at least Sparta, and of variants used by the Hebrews before 500 BCE).

The Caesar cipher works by picking a number n between 1 and 25, called the *shift*, and replacing each letter with the letter n places to the left. (Note: some sources shift to the right instead; the direction doesn't really matter as long as you're consistent). The traditional story has Caesar using a shift of three to the left:

Plaintext	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
Ciphertext	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

> 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.
>
> Suetonius, Life of Julius Caesar 56

Example 1.1. Suppose we have the ciphertext "QEFP JBPPXDB EXP YBBK BKZFMEBOBA YV X ZXBPXO ZFMEBO TFQE X PEFCQ LC QEOBB", and we know it has been enciphered by a Caesar cipher with a shift of three. We can decipher it by shifting each letter forward by three places. we get "THIS MESSAGE HAS BEEN ENCIPHERED BY A CAESAR CIPHER WITH A SHIFT OF THREE."

Remark 1.2. The most commonly used Caesar cipher today is ROT13, for "rotation 13", which is a shift by thirteen. Notice that in this case the algorithms for encryption and decryption are actually identical.

ROT13 was in common use on the early internet for spoilers: it's easy to decrypt so it doesn't prevent anyone who wants to read it from reading it, but it keeps you from accidentally reading the message without intending to. Thus ROT13 is built into a number of text editors and other computing tools.

People sometimes joke about the encryption algorithm double-ROT13, which is "obviously" twice as secure.

There are 26 possible shift ciphers. So if you know you're dealing with a shift cipher it's actually pretty easy to decode.

Example 1.3. Suppose we have the message IWXH RXEWTG XH HWXUITS QN TATKTC, and we know it has been encoded by some Caesar cipher. Since there are only twenty-six possibilities we can just try each one. In fact, we can apply each shift to just the first word and see which shifts make sense:

0	IWXH	7	PDEO	14	WKLV	21	DRSC
1	JXYI	8	QEFP	15	XLMW	22	ESTD
2	KYZJ	9	RFGQ	16	YMNX	23	FTUE
3	LZAK	10	SGHR	17	ZNOY	24	GUVF
4	MABL	11	THIS	18	AOPZ	25	HVWG
5	NBCM	12	UIJT	19	BPQA		
6	OCDN	13	VJKU	20	CQRB		

Looking down this table, we see the only shift that produces a recognizable word is a shift of eleven, giving "THIS" as output. If we shift the entire message by eleven, we get the message "THIS CIPHER IS SHIFTED BY ELEVEN".

In order to make this process slightly harder, most enciphered text messages will be sent without spaces between words. (This was also easier on early transmission technology). Consider the following example:

Example 1.4. CWUHLYUXNBCMWUHSIO

We could try every shift on the entire message. In this case, the message is short enough that that's reasonable, but for longer messages we might not want to convert every letter with every possible shift.

Instead, we can test the first few letters with each shift, and see which ones look like they might contain or be part of words. Taking the first three letters "CWU" we get the following table:

0	CWU	7	JDB	14	QKI	21	XRP
1	DXV	8	KEC	15	RLJ	22	YSQ
2	EYW	9	LFD	16	SMK	23	ZTR
3	FZX	10	MGE	17	TNL	24	AUS
4	GAY	11	NHF	18	UOM	25	BVT
5	HBZ	12	OIG	19	VPN		
6	ICA	13	PJH	20	WQO		

Based on just these letters, it's hard to tell which options are good; 4 gives us "gay" which is a word, and 6, 12, and 24 all give us collections of letters that at least look like they could be English. But we can be pretty confident at throwing out strings like "FZX" and "JDB" if we expect the message to be words.

Using a few more letters gives us better data. If we used the first six instead of the first three, we would get the table

0	CWUHLY	7	JDBOSF	14	QKIVZM	21	XRPCGT
1	DXVIMZ	8	KECPTG	15	RLJWAN	22	YSQDHU
2	EYWJNA	9	LFDQUH	16	SMKXBO	23	ZTREIV
3	FZXKOB	10	MGERVI	17	TNLYCP	24	AUSFJW
4	GAYLPC	11	NHFSWJ	18	UOMZDQ	25	BVTGKX
5	HBZMQD	12	OIGTXK	19	VPNAER		
6	ICANRE	13	PJHUYL	20	WQOBFS		

and the only terribly useful-looking shift is 6, which gives "I can re"as plausible English. So using a shift of 6, we get "ICANREADTHISCANYOU". After squinting, we see the message: "I can read this. Can you?"

1.2.1 Modular arithmetic

Recall from Discrete Math that we say two integers a, b are equivalent modulo m, and write $a \equiv b \mod m$, if $m|b - a$. This is an equivalence relation, and is preserved by addition, subtraction, and multiplication.

The Caesar cipher is easy to describe in terms of modular arithmetic. We can identify each letter with a number modulo 26. There are different conventions here, but we'll follow Hoffstein, Pipher, and Silverman:

Definition 1.5. There is a bijection between the letters of the alphabet and numbers modulo 26, sending A to 0, B to 1, and so on, up to sending Z to 25.

Letter	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
Number	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

Then we can see a Caesar cipher as choosing some fixed number n, and subtracting that number from each letter in your message. For the rest of the class we'll probably *add* the number instead, because it's easier to think about.

From this perspective, a key is just a number from 0 to 25 inclusive.

Example 1.6. If the message is "MEET ME AT MIDNIGHT", we can render this as the list of numbers 12 4 4 19 12 4 0 19 12 8 3 13 8 6 7 19.

If we want to encrypt this with a shift of plus 15, we get 1 19 19 8 1 19 15 8 1 23 18 7 23 21 22 8, which translates back to "BTTIBTPIBXSHXVWI" which is our ciphertext.

To decrypt, we simply subtract 15 from each letter and get back "MEETMEATMIDNIGHT".

1.3 Monoalphabetic substitution

The shift ciphers above are limited by the fact that you have to keep the letters in order—you only change which one comes first. But there's no reason we can't reorder them. A *monoalphabetic substitution cipher* does away with that constrant, and simply pairs each plaintext letter up with a distinct ciphertext letter.

Example 1.7. An example monoalphabetic substitution cipher would be

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

To encipher a message, we find each letter of the plaintext in the first row, and replace it with the corresponding letter in the second row. So the message "THIS IS A SIMPLE SUBSTITUTION CIPHER" becomes "OBHM HM G MHSTNE MVIMOHOVOHFD LHTBEU". We can decrypt the ciphertext by looking up each letter of the ciphertext in the second row and finding the corresponding character in the first row. It can be helpful to have a table sorted by the ciphertext characters instead of the plaintext characters:

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

Thus a choice of monoalphabetic substitution cipher is equivalent to a choice of a permutation of the alphabet (or to an element of the group S_{26}). Therefore there are $26! \approx 4 \times 10^{26}$ possible monoalphabetic substitution ciphers, so *unlike* in the case of a Caesar cipher, you can't realistically try all the possibilities until you find one that works. We need to be a bit cleverer.determine the substitution.

1.4 Frequency Analysis of Monoalphabetic Ciphers

Fortunately, language has a structure to it: not all strings of letters are equally common. (If I tell you my message is either "brown" or "yoltk", you can make a pretty good guess as to which it is).

The first and simplest tool we have is the relative frequency of letters in English text. This approach is usually credited to the Arab philosopher Abu Yusuf Ya'qub ibn Ishaq Al-Kindi in the ninth century CE. The basic idea is that not all letters occur equally often; if your ciphertext has one letter appearing ten times in fifty total letters, it's probably not a "q" or "z".

There are also more sophisticated approaches we can take to frequency analysis, because English does not have its letters distributed in a random order. (That is, our model of the plaintext has more information than just the frequencies of individual letters). That is, if we see the same pair of letters appearing in the same order many times, we might guess that this pair is "th" or "he" or "an". If we see the same trio of letters appearing many times, we might guess that it is "the" or "and".

Below are tables showing the frequency with which each letter appears in English texts (Figure 1.1), and the frequencies of the most common English bigrams (Figure 1.2), drawn from Hoffstein, Pipher, and Silverman. (Note that different sources will have slightly different numbers due to using different corpuses).

E	13.11%	M	2.54%	A	8.15%	N	7.10%
T	10.47%	U	2.46%	B	1.44%	O	8.00%
A	8.15%	G	1.99%	C	2.76%	P	1.98%
O	8.00%	Y	1.98%	D	3.79%	Q	0.12%
N	7.10%	P	1.98%	E	13.11%	R	6.83%
R	6.83%	W	1.54%	F	2.92%	S	6.10%
I	6.35%	B	1.44%	G	1.99%	T	10.47%
S	6.10%	V	0.92%	H	5.26%	U	2.46%
H	5.26%	K	0.42%	I	6.35%	V	0.92%
D	3.79%	X	0.17%	J	0.13%	W	1.54%
L	3.39%	J	0.13%	K	0.42%	X	0.17%
F	2.92%	Q	0.12%	L	3.39%	Y	1.98%
C	2.76%	Z	0.08%	M	2.54%	Z	0.08%

Figure 1.1: English Letter Frequencies

th	he	an	re	er	in	on	at	nd	st	es	en	of	te	ed
168	132	92	91	88	86	71	68	62	53	52	51	49	46	46

Figure 1.2: Most common English bigrams (frequency per 1000 words)

Example 1.8. Let's try to decrypt the ciphertext:

```
JNRZR BNIGI BJRGZ IZLQR OTDNJ GRIHT USDKR ZZWLG OIBTM NRGJN IJTZJ LZISJ NRSBL
QVRSI ORIQT QDEKJ JNRQW GLOFN IJTZX QLFQL WBIMJ ITQXT HHTBL KUHQL JZKMM LZRNT
OBIMI EURLW BLQZJ GKBJT QDIQS LWJNR OLGRI EZJGK ZRBGS MJLDG IMNZT OIHRK MOSOT
QHIJL QBRJN IJJNT ZFIZL WIZTO MURZM RBTRZ ZKBNN LFRVR GIZFL KUHIM MRIGJ LJNRB
GKHRT QJRUU RBJLW JNRZI TULGI EZLUK JRUST QZLUK EURFT JNLKJ JNRXR S
```

We begin by counting the frequency of each letter, and put our results in figure 1.3

Letter	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
Frequency	0	15	0	5	5	6	15	8	27	30	13	25	12	19	10	0	16	33	9	20	12	2	7	3	0	24

Letter	R	J	I	L	Z	T	N	Q	B	G	K	U	M	O	S	H	W	F	E	D	X	V
Frequency	33	30	27	35	24	20	19	16	15	15	13	12	12	10	9	8	7	6	5	5	3	2

Bigram	JN	NR	TQ	LW	RB	RZ	JL
Frequency	11	8	6	5	5	5	5

Figure 1.3: Frequency count for example 1.8

The most common letter is "R" so we'll guess that "R" is encrypting "e". We notice that the most common bigrams in the ciphertext are "JN" and "NR", and the most common bigrams in English are "th" and "he"; this leads us to guess that "JNR" is "the". (This is also reassuring since the second most common English letter is "t" and the second-most-common letter in the ciphertext is "J", to which we've just assigned the letter "t").

```
theZe BhIGI BteGZ IZLQe OTDht GeIHT USDKe ZZWLG OIBTM heGth ItTZt LZISt heSBL
QVeSI OeIQT QDEKt theQW GLOFh ItTZX QLFQL WBIMt ITQXT HHTBL KUHQL tZKMM LZehT
OBIMI EUeLW BLQZt GKBtT QDIQS LWthe OLGeI EZtGK ZeBGS MtLDG IMhZT OIHeK MOSOT
QHItL QBeth ItthT ZFIZL WIZTO MUeZM eBTeZ ZKBhh LFeVe GIZFL KUHIM MeIGt LtheB
GKHeT QteUU eBtLW theZI TULGI EZLUK teUST QZLUK EUeFT thLKt theXe S
```

There are a few things we could do now. We can look at our other list of common bigrams. We see that "JL" is common in the ciphertext, which means we need a common English bigram whose first letter is "t"; "te" is the most common, but is ruled out since we know that "e" is "R". No others appear on our list.

We also have "RB" and "RZ" as common bigrams, and we know that "R" is "e". Looking at our list, it looks like these bigrams are probably "er" and "es". It's not entirely clear which should be which; one would make the first word of our text "there" and the second would give "these", which are both perfectly reasonable. It's not clear what to do here.

We can also just go ahead and guess that our next-most-common ciphertext letters "I" and "L" are our next-most-common plaintext letters "a" and "o". (This also makes "JL" into “to”, which seems quite plausible!) That would give us

```
theZe BhaGa BteGZ aZoQe OTDht GeaHT USDKe ZZWoG OaBTM heGth atTZt oZaSt heSBo
QVeSa OeaQT QDEKt theQW GoOFh atTZX QoFQo WBaMt aTQXT HHTBo KUHQo tZKMM oZehT
OBaMa EUeoW BoQZt GKBtT QDaQS oWthe OoGea EZtGK ZeBGS MtoDG aMhZT OaHeK MOSOT
QHato QBeth atthT ZFaZo WaZTO MUeZM eBTeZ ZKBhh oFeVe GaZFo KUHaM MeaGt otheB
GKHeT QteUU eBtoW theZa TUoGa EZoUK teUST QZoUK EUeFT thoKt theXe S
```

At this point we might want to go looking through the text for guessable words. We see “at” several times, and might start noticing some patterns. What sticks out to me is the string “eth atth”, which looks like it ends in “something-eth at th-”. Almost certainly, the next letter should be a vowel; since we have “e,a,o” already spoken for, it should be “i” or “u”. The ciphertext letter “T” is quite common; since “i” is a common English letter and “u” is not, we’ll guess that “T” becomes “i”.

```
theZe BhaGa BteGZ aZoQe OiDht GeaHi USDKe ZZWoG OaBiM heGth atiZt oZaSt heSBo
QVeSa OeaQi QDEKt theQW GoOFh atiZX QoFQo WBaMt aiQXi HHiBo KUHQo tZKMM oZehi
OBaMa EUeoW BoQZt GKBti QDaQS oWthe OoGea EZtGK ZeBGS MtoDG aMhZi OaHeK MOSOi
QHato QBeth atthi ZFaZo WaZiO MUeZM eBieZ ZKBhh oFeVe GaZFo KUHaM MeaGt otheB
GKHei QteUU eBtoW theZa iUoGa EZoUK teUSi QZoUK EUeFi thoKt theXe S
```

Looking back at our list of bigrams, we see that “TQ” is common. An English bigram whose first letter is “i” is probably “in”, so we might guess that “Z” is “n”; but this gives some unlikely strings in our message like “thene BhaGa” or “that in tonaSt” or “-eth at thinFano” or “toW thenaiUoGa”. While any one of these is possible, they don’t seem likely. Unfortunately we don’t have any other common i-initial bigrams to look at.

So let’s go back to our idea that maybe “Z” is “s” or “r”. “r” is the more common English letter, so we might try that first. “there BhaGa” seems plausible; but “that irtora” is improbable, as is “-eth at thirFaro” or “toW theraiUoGa”.

So we try “s”, and we get “these BhaGa”; “that is to sa”, “-eth at this Faso” and “toW thesaiUoGa”. The first three seem extremely likely, and the fouth possible, so we guess that “Z” is “s”.

We can now look to sort out words, or just go back to our frequency charts. The next most common bigram is “TQ” which is “iQ”, and the most common English bigram beginning with “i” is “in. Also, the next most common letter is “Q”, and the next most common English letter is “n”, so we might guess from both of these things that “Q” is “n”.

```
these BhaGa BteGs asone OiDht GeaHi USDKe ssWoG OaBiM heGth atist osaSt heSBo
nVeSa Oeani nDEKt thenW GoOFh atisX noFno WBaMt ainXi HHiBo KUHno tsKMM osehi
```

```
OBaMa EUeoW Bonst GKBti nDanS oWthe OoGea EstGK seBGS MtoDG aMhsi OaHeK MOSOi
nHato nBeth atthi sFaso WasiO MUesM eBies sKBhh oFeVe GasFo KUHaM MeaGt otheB
GKHei nteUU eBtoW thesa iUoGa EsoUK teUSi nsoUK EUeFi thoKt theXe S
```

The next most common letter after this is "B". Our first thought is that we can take the next most common English letter of "r", but then we get "rhaGarte" which really doesn't look likely. On the other hand, if we replace our "G" with "r" we get "BharaBters" which suggests "B" for "c".

```
these chara cters asone OiDht reaHi USDKe ssWor OaciM herth atist osaSt heSco
nVeSa Oeani nDEKt thenW roOFh atisX noFno WcaMt ainXi HHico KUHno tsKMM osehi
OcaMa EUeoW const rKcti nDanS oWthe Oorea EstrK secrS MtoDr aMhsi OaHeK MOSOi
nHato nceth atthi sFaso WasiO MUesM ecies sKchh oFeVe rasFo KUHaM Meart othec
rKHei nteUU ectoW thesa iUora EsoUK teUSi nsoUK EUeFi thoKt theXe S
```

At this point we really should be looking to recognize words in teh text. The phrase "that is to saS theS" strongly suggests that "S" is "y". We see the string "constrKct" which tells us "K" has to be a vowel; the only one left is "u", and "construct" is a reasonable word.

```
these chara cters asone OiDht reaHi UyDue ssWor OaciM herth atist osayt heyco
nVeya Oeani nDEut thenW roOFh atisX noFno WcaMt ainXi HHico uUHno tsuMM osehi
OcaMa EUeoW const ructi nDany oWthe Oorea Estru secry MtoDr aMhsi OaHeu MOyOi
nHato nceth atthi sFaso WasiO MUesM ecies suchh oFeVe rasFo uUHaM Meart othec
ruHei nteUU ectoW thesa iUora EsoUu teUyi nsoUu EUeFi thout theXe y
```

"sMecies" tells us that "M" is probably "p". "Fithout" suggests that "F" is "w", and then "without the Xey" seriously limits what "X" can stand for; we guess "X" is "k". "the saiUor" is probably "the sailor", so "U" is probably "l".

```
these chara cters asone OiDht reaHi lyDue ssWor Oacip herth atist osayt heyco
nVeya Oeani nDEut thenW roOwh atisk nowno Wcapt ainki HHico ulHno tsupp osehi
Ocapa EleoW const ructi nDany oWthe Oorea Estru secry ptoDr aphsi OaHeu pOyOi
nHato nceth atthi swaso WasiO plesp ecies suchh oweVe raswo ulHap peart othec
ruHei ntell ectoW thesa ilora Esolu telyi nsolu Elewi thout theke y
```

"aEsolutely" implies that "E" is "b". The most common leftover letter is "O"; we see it in "these characters, as one OiDht reaHily..." and in "not suppose hiO capable" and "the Oore abstruse cryptoDraph". From these, we guess "O" is "m" (especially to fill out "him") and then guess that "D" is "g", giving us "might" and "cryptograph".

```
these chara cters asone might reaHi lygue ssWor macip herth atist osayt heyco
nVeya meani ngbut thenW romwh atisk nowno Wcapt ainki HHico ulHno tsupp osehi
```

```
mcapa bleoW const ructi ngany oWthe morea bstru secry ptogr aphsi maHeu pmymi
nHato nceth atthi swaso Wasim plesp ecies suchh oweVe raswo ulHap peart othec
ruHei ntell ectoW thesa ilora bsolu telyi nsolu blewi thout theke y
```

Finally, “reaHily guess” implies that “H” is “d”. We’re running out of letters now; we look at the “V” and see it twice, in “that is to say they conVey a meaning” and “a simple species such howeVer as would”, and it looks like “V” is actually “v”! We just need to translate the “W”, and “but then Wrom” tells us that “W” is “f”.

```
these chara cters asone might readi lygue ssfor macip herth atist osayt heyco
nveya meani ngbut thenf romwh atisk nowno fcapt ainki ddico uldno tsupp osehi
mcapa bleof const ructi ngany ofthe morea bstru secry ptogr aphsi madeu pmymi
ndato nceth atthi swaso fasim plesp ecies suchh oweve raswo uldap peart othec
rudei ntell ectof thesa ilora bsolu telyi nsolu blewi thout theke y
```

Going through and respacing, we get:

These characters, as any one might readily guess, form a cipher—that is to say, they convey a meaning; but then from what is known of Kidd, I could not suppose him capable of constructing any of the more abstruse cryptographs. I made up my mind, at once, that this was of a simple species—such, however, as would appear to the crude intellect of the sailor, absolutely insoluble without the key.

And the key to the cipher is

Ciphertext	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
Plaintext	-	c	-	g	b	w	r	d	a	t	u	o	p	h	m	-	n	e	y	i	l	v	f	k	-	s

Plaintext	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
Ciphertext	I	E	B	H	R	W	D	N	T	-	X	U	O	Q	L	M	-	G	Z	J	K	V	F	-	S	-

Ciphertext	R	J	I	L	Z	T	N	Q	B	G	K	U	M	O	S	H	W	F	E	D	X	V
Plaintext	e	t	a	o	s	i	h	n	c	r	u	l	p	m	y	d	f	w	b	g	k	v

A few things to notice: by chance, “V” becomes “v”, and “F” and “W” are interchanged. Nothing requires that sort of thing to happen, but nothing prohibits it either.

Also notice that there are some ciphertext and plaintext letters that we don’t have correspondences for. The plaintext simply never used an “x” or a “z”, so we don’t know what rule it would have used for them, had it needed one. But if we got a future message in the same cipher, it would be quite easy to determine the meanings of the “A” and “C” in the message.

Note that this process requires experimentation and can take a number of wrong turns; I personally spent quite a while convinced that “L” was “i” in the preceeding cipher. If something isn't work, revisit one of your earlier guesses and try something different.

This sort of approach can fail in a few ways:

1. The message could be too short. If the message is ten letters long we can't possibly do any useful statistical analysis on it. (In fact a ten-letter message is generally impossible to decipher even in principle; an English message typically must be at least 27.6 letters to be amenable to frequency analysis. We'll discuss this general topic later, in 4).

2. The text could be atypical. The pangram "A quick brown fox jumped over the lazy dog" is often used as a test sentence in many applications. But this message has four "o"s and only two "e"s, so statistical approaches will be somewhat misleading.

   The likelihood of this happening by accident, and the difficulty of it happening on purpose, decrease as the messages get longer. But it's completely possible that even a long message is highly atypical in this way; Ernest Vincent Wright wrote a full novel, called "Gadsby", without using the letter "e".

3. The text might not even be English. For instance, in Portuguese the most common letter is "a", not "e", and the letter "t" barely cracks the top ten. If you do statistical analysis assuming the encrypted message is English, but it's actually Portuguese, you may never even make enough progress to realize your error.

4. The message might not be enciphered with a monoalphabetic substitution cipher at all.

The first problem is a problem of not enough data; the third and fourth problems are problems of having a flawed model. (The second problem is a mix of the two). Both of these are important problems that come up any time we do statistical analysis. If our modelling assumptions are wrong, all the statistics in the world won't help us.

However, this sort of statistical analysis is powerful enough and well-enough understood that monoalphabetic ciphers are considered pretty thoroughly insecure; despite the number of possible keys, anyone who knows what they're doing can break these ciphers easily, and this has been true for over a millennium.

1.5 Polyalphabetic Ciphers

The next major advance in cryptography began the 1500s with the invention of the polyalphabetic cipher.

The basic idea here is simple. A monoalphabetic substitution cipher is vulnerable to attacks based on letter frequency, exploiting the fact that one letter in the ciphertext always represents the same letter in the plaintext. We can (partially) defeat this attack by not always using the same ciphertext letter to represent the same plaintext letter.

The first (known) implementation of this idea was straightforward. Leon Battista Alberti in 1467 would write in a Caesar cipher, but from time to time he would change to a different shift, and signal this shift with the use of a capital letter.

This method of encryption by itself isn't much better than a single Caesar cipher, since you can easily tell when the cipher is shifting, and simply break each portion on its own.

Example 1.9.

However, this idea of using more than one substitution to encrypt a message is very powerful. Basically all strong encryption through the end of World War II is based on this principle.

1.5.1 The Vigenère Cipher

The most common and simplest example of polyalphabetic cipher is usually attributed to Blaise de Vigenère, though it's probably more accurately attributed to Giovan Battista Bellasco in 1553; Vigenère improved on it in 1586 and then got the credit.

To use the Vigenère Cipher we first need a *key word*, which can be any word (or any string of letters). We treat each letter as determining a specific Caesar cipher. There are a number of ways to determine this correspondence, but we will use the numerical correspondence we established earlier, adding the letter of our keystream to the letter of our ciphertext.

Algorithm 1.1 (Vigenère)**.** Start with a key word. Encrypt the first letter of your plaintext using the Caesar cipher corresponding to the first letter of your key word. Encrypt the second letter of your plaintext according to the second letter of your key word, and so on, repeating your key word when you reach the end.

To decrypt the cipher, we do the same thing. We use hte same keyword, but we subtract the keystream instead of adding it.

Example 1.10. Suppose we want to encrypt the plaintext "I LOVE CRYPTOLOGY" using the key word "MATH". We start by writing the message out, with the key word repeated under it (this repeated keyword is called the *keystream*):

Plaintext	I	L	O	V	E	C	R	Y	P	T	O	L	O	G	Y
Keystream	M	A	T	H	M	A	T	H	M	A	T	H	M	A	T

It's probably helpful at this point to replace the letters with their corresponding numbers, which gives

Plaintext	8	11	14	21	4	2	17	24	15	19	14	11	14	6	24
Keystream	12	0	19	7	12	0	19	7	12	0	19	7	12	0	19
Ciphertext	20	11	7	2	16	2	10	3	1	19	7	18	0	6	17
Ciphertext	U	L	H	C	Q	C	K	D	B	T	H	S	A	G	R

To decrypt we do the same thing in reverse, subtracting the keystream from the ciphertext to get the plaintext.

This cipher is highly resistant to the sort of frequency analysis-based attacks we will study in section 1.4. Common letters like "e" and "a" will be be substituted by multiple different letters, and each letter in the ciphertext representes more than one plaintext letter, so knowing which letter appears most frequently in the ciphertext doesn't convey much information. The Vigenère Cipher effectively smooths out the frequencies so that frequency analysis does not work.

In fact, for over two hundred years people considered the Vigenère cipher effectively unbreakable. However in 1854 Charles Babbage successfully completed a public challenge to decrypt a Vigenère ciphertext; however, he never published his techniques. In 1863 Friedrich Kasiski published a method that is effective at breaking the Vigenère cipher, which we will study in section 2.2.

1.5.2 Binary encryption

We should conclude by discussing how this applies to encrypting computer data, which does not consist of Roman letters.

Computers encode data in *binary* strings of ones and zeroes. A *bit* is a single zero-or-one character; a *byte* is a string of eight bits and is the unit computers use to represent a single character of text.

Thus we can treat computers as working with an alphabet of two symbols. In this context monoalphabetic substitution is quite useless: either you change nothing, or you simply switch the ones and zeroes with each other. With only two possible keys, brute forcing is easy.

However, a polyalphabetic cipher like the Vigenère cipher works quite well. We can take as our key some binary string and add it (modulo 2) to our plaintext to get our ciphertext.

Remark 1.11. Adding bits modulo 2 is the same as the XOR operation, which returns 1 if its inputs are different, and 0 if they are the same.

Example 1.12. Suppose our key is 10010011 and we wish to encrypt the plaintext

01010000 01001111 01001011 01000101 00100000 00110101 00111001 00110100 00110101 00111000 00101100 00110110 00110010.

Adding our key to each byte, we get the output

11000011 11011100 11011000 11010110 10110011 10100110 10101010 10100111 10100110 10101011 10111111 10100101 10100001.

To decrypt, we simply subtract (or add, since they're the same thing modulo 2) our key back from our ciphertext to get the plaintext.

1.6 Other Stream Ciphers

Any cipher that works by generating a keystream and then adding it to the plaintext is called a *stream cipher*. There are a number of variants we can discuss.

1.6.1 Autokey ciphers

Blaise de Vigenère actually developed an autokey cipher, which uses the plaintext to generate the keystream. There are a number of variants to this idea.

One simple algorithm is to form a keystream by using a keyword and then following it with the plaintext. So under this algorithm, if the keyword is "MATH" and the message is "I LOVE CRYPTOLOGY", we get the keystream "MATHILOVECRYPTO". Then we encrypt:

Plaintext	I	L	O	V	E	C	R	Y	P	T	O	L	O	G	Y
Keystream	M	A	T	H	I	L	O	V	E	C	R	Y	P	T	O

Plaintext	8	11	14	21	4	2	17	24	15	19	14	11	14	6	24
Keystream	12	0	19	7	8	11	14	21	4	2	17	24	15	19	14
Ciphertext	20	11	7	2	12	13	5	19	19	21	5	9	3	25	12
Ciphertext	U	L	H	C	M	N	F	T	T	V	F	J	D	Z	M

Notice that we got the same first four letters as with the Vigenère cipher (since the key is the same and four letters long), but things change after that.

In order to decrypt a message encrypted in this way, we decrypt in chunks. Knowing the keyword lets us decrypt the first four letters; knowing the first four letters lets us decrypt the second set of four letters; and so on.

1.6.2 Modern Stream Ciphers

There are a number of modern usable stream cipher algorithms.

These usually involve plugging key data into a pseudorandom number generator to generate a keystream.

There are two big weaknesses that limit the use of stream ciphers. In order to maintain security, a stream cipher needs to

- Use a different key for every message; and
- Produce a keystream that has a long period before repeating itself.

Most cryptography in use today uses other principles, which we will discuss later on in the course.

References

[Singh(2015)] Vikram Singh. A practical key exchange for the internet using lattice cryptography. Cryptology ePrint Archive, Report 2015/138, 2015. `http://eprint.iacr.org/2015/138`.