

How to Crack Substitution Ciphers

Eliot Ball

OK, so here is some text that I have encrypted using a substitution cipher:

FRNRWXRNNKFYEJZXIXZGFRLFVFGENVFGXIXFRFKNZEMCRFYFZGDKNCXGEFENKLMUZENNYFRLGMALGMYDGG
ENVXRZXLNGENZVFAANZGYMZZXBANFVMDRGMUUNRINGENNRWXRNNKXZUXKZGENENKLZGENZENNYXRGMFIXK
IANFRLGENRYDGGZGENUNRINFKMDRLGENVLNIAFKXRWFIXKIANCXAADZNGENANFZGUNRINUMKFWXTNRFKNFZ
MGEXZXZGENBNZGZMADGXMRGENYEJZXIXZGXZRNQZGENIKNFGNZFIXKIDAFKUNRINMUXRUXRGNKFLXDZFK
MDRLGENZENNYFRLGENRLKFCZGENUNRINGXWEGFKMDRLGENENKLLNIAFKXRWGXZCXAAXWTNGENZVFAANZG
IXKIDAFKUNRINFKMDRLGENENKLGENVFGENVFGXIXFRXZAFZGFUGNKWXTXRWGENYKMBANVFAXGGANGEMDWE
GENYDGFZVFAAUNRINFKMDRLEXVZNAUFRLGENRLNIAFKNZXLNUXRVJZNAUGMBNMRGENMDGZXLN

I have removed the spaces because it makes it harder! Note that if the characters are in blocks of the same length it actually makes it easier if you remove the spaces, but if the spaces are between words they should be left in and cracking the cipher will be *much* easier. The first step when faced with the ciphertext is to perform frequency analysis. Here is the frequency analysis for this text, showing the percentage of the text taken up by each letter:



If the text shows, like it does above, big differences between the frequencies of each of the letters, and they do not go in roughly this order, from biggest to smallest...

ETAOINSRHLDCUMFPGWYBVKXJQZ

Then it is *likely* to be a simple substitution cipher, with each letter being replaced by a different one consistently throughout the whole message.

Now we can start off a table of which letters we think are replaced for which, using the lowercase version of the letter in the ciphertext until we have an idea of what we are going to replace it with, like this:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
 abcdefghijklmnopqrstuvwxyz

This table just says that we swap each letter for its lowercase version, and therefore simply generates the lowercase version of the ciphertext:

frnrwxrnnkfyejzxixzgrlflvfgenvfgixixfrfknczemcrfyfzgdknxcgfenklmuzennyfrlmgalmgydgg
 envxrzxlngenzvfaanzgymzzxbanfvmrdrgmuunringennrwxrnnkxzuxkzgenenklzgenzennyxrgmfikx
 ianfrlgenrydgzgenunrinfkmdrlgenvlniafkxrwfixkiancxaadzngenanfzgunrinumkfwxtnrnfknfz
 mgexzxzgenbnzgzmadgxmrgeyeyjzxixzgxzrnqgzeniknfgnzfixkidafkunrinmuxruxrxgnkflxdzfk
 mdrllgenzennyfrlgenrlkfczgenunringxwegfkmdrlgenenklniafkxrwgexzcxawxtngenzvfaanzg
 ixkidafkunrinfkmdrlgenenklgenvfgenvfgixixfrxazfzfgugnkwtxrxwgenykmbanvfaxggangemdwe
 genydgzfvfaaunrinfkmdrllexvznaufrlgenrlniafkznxlnuxrnvjznaugmbnmrgenmdgzxln

The next step is to refer back to the frequency analysis. As can be seen on the previous page, the first and second most common letters in normal English are E and T, so it is probably worth substituting the two most common letters in our ciphertext, N and G respectively, for those two letters. This produces the following table and plaintext:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
 abcdefThijklmEopqrstuvwxyz

frErwxrEEkfyejzxixzTfrlflvfTeEvfTxixfrfkEzemcrfyfzTdkEcxEteEklmuzeEEyfrlTmalTmydTT
 eEvxrzxLETeEzvfaaEzTymzzxbaEfvmrdTmuuEriETeEErwxrEEkxzuxkzTeEeEklzTeEzeEEyxrTmfikx
 iaEfrlTErydTzTEuEriEfkmrdlTEvLEiafkxrwfixkiaEcxaadzETEaEfzTuEriEumkfwxtErfkEfz
 mTexxzTeEbEzTzmadTxmrTEyejzxixzTxzrEqTzeEikEfTEzfixkidafkuEriEmuxruxrTEkflxdzfk
 mdrllTeEzeEEyfrlTEerlkfczTEuEriETxwTfkmrdlTEeEklLEiafkxrwTexzcxawxtETEzvfaaEzT
 ixkidafkuEriEfkmrdlTEeEklTEEvfTEvftxixfrxazfzTfuTEkwtxrxwTEykmbeVfaxTTaETemdwe
 TEydtzfzvfauEriEfkmrdllexvzEaufrlTErLEiafkEzxlEuxrEvjzEauTmbEmrTEemdTzxLE

I have made the capitals bold to make it easier to see them. (I wrote a script to do it, I didn't do it by hand!!) Bearing in mind that this *might* not be the correct substitution for those two letters, we can nevertheless see a couple of things that would make us think that this is likely to be correct. There are several uses of TT and EE in the text, and there are not many letters that work as doubles, especially not as commonly as that. Also, there are several occurrences of TeE, and a good guess would be that this stands for THE. So the next substitution that we will try is e for H, giving us the following table and text:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
 abcdHfThijklmEopqrstuvwxyz

frErwxrEEkfyHjzxixzTfrlflvfTHEvfTxixfrfkEzHmcrfyfzTdkEcxTHfHEklmuzHEEYfrlTmalTmydTT
 HEVxrzxLETHEzvfaaEzTymzzxbaEfvmrdTmuuEriETHEErwxrEEkxzuxkzTHEHEklzTHEzHEEYxrTmfikx
 iaEfrlTHErydTzTHEuEriEfkmrdlTHEvLEiafkxrwfixkiaEcxaadzETHEaEfzTuEriEumkfwxtErfkEfz
 mTHxxzTHEbEzTzmadTxmrTHEyHjzxixzTxzrEqTzeHEikEfTEzfixkidafkuEriEmuxruxrTEkflxdzfk
 mdrllTHEzHEEYfrlTHErllkfczTHEuEriETxwHTfkmrdlTHEHEklLEiafkxrwTHxxzcxawxtETHEzvfaaEzT
 ixkidafkuEriEfkmrdlTHEHEklTHEvfTHEvfTxixfrxazfzTfuTEkwtxrxwTHEykmbeVfaxTTaETHmdwH
 THEydtzfzvfauEriEfkmrdlHxvzEaufrlTHErLEiafkEzxlEuxrEvjzEauTmbEmrTHEmdTzxLE

Now we can clearly see the word THE appearing throughout the message, and it is safe to say that we have probably got the first three substitutions correct. There are a couple of things that look a bit strange, like THEHE and ETHE, but remember that we have not put in any spaces yet, and those fragments probably span more than one word. At this point it is hard to see any more words beginning to appear so it is best to return to the frequency analysis and look at the next most frequent letter, which is F in the ciphertext with 8.1% of all the letters, and this could well correspond to the third most frequent letter in English, A. We will try that:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
 abcdHATHijklmEopqrstuvwxyz

ArErwxrEEkAyHjzxixzTArIAvATHEvATxixArAkEzHmcrAyAzTdkEcXTHAHEklmuzHEEYArLTmalTmydTT
 HEVxrzxlETHEzvAaaEzTymzzxbaEAvmdrTmuuEriETHEErwxrEEkxzuxkzTHEHEklzTHEzHEEYxrTmAixk
 iaEARlTHErydTzTHEuEriEAKmdrLTHEvLEiaAkxrwAixkiaEcxaadzETHEaEAzTuEriEumkAwxtErAKEAZ
 mTHxzZzTHEbEzTzmadTxmrTHEyHjzxixzTxzrEqTzHEikEATEzAixkidaAkuEriEmuxruxrxTEkAlxdzAk
 mdrLTHEzHEEYArLTHErLkAczTHEuEriETxwHTAkmdrLTHEHEklLEiaAkxrwTHxzcxaaawxtETHEzvAaaEzT
 ixkidaAkuEriEAKmdrLTHEHEklTHEvATHEvATxixArxzaAzTAuTEkwtxrwTHEykmbaEvAaxTTaETHmdwH
 THEydtZAzvAaaEriEAKmdrLHxvzEauArLTHErLEiaAkEzxlEuxrEvjzEauTmbEmrTHEmdTzxlE

Now we need to study the text carefully and remember that each lowercase letter is substituted for the same uppercase letter throughout the message. Looking carefully, we can spot the section that I have highlighted on the first line. About half of the letters in that section have been revealed, and there are two repeated letters. We can guess that this might actually be the word MATHEMATICIAN. (we could have used a crossword solver but that takes the fun away!) This guess leads us to four more letter substitutions, and performing those should give us a good idea of whether the hunch was correct. Here we go:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
 abcdHATHcJklmEopqNstuMwIyz

ANENwINEEkAyHjzICIZTANIAMATHEMATICIANAkEzHmcNAyAzTdkEcITHAHEklmuzHEEYANLTmalTmydTT
 HEMINzIIEtHEzMAaaEzTymzzIbaEAMmdNTmuuENCETHEENwINEEkIzuIkzTHEHEklzTHEzHEEYINTmACIk
 CaEANlTHENydtzTHEuENCEAkmdNlTHEMLCaAkINwACIkCaEcIaadzETHEaEAzTuENCEumkAwItENAKEAz
 mTHIzIzTHEbEzTzmadTImNTHEyHjzICIZTIzNEqTzHECKEATEzACIkCdaAkuENCEmuINuINITEkAlIdzAk
 mdNlTHEzHEEYANlTHENlKAczTHEuENCETIwHTAkmdNlTHEHEklLECaAkINwTHIzcIaawItETHEzMAaaEzT
 CIkCdaAkuENCEAkmdNlTHEHEklTHEMATHMATICIANIzaAzTAuTEkwItINwTHEykmbaEMAaITTaETHmdwH
 THEydtZAzMAaaEENCEAkmdNlHIMzEauANlTHENlECaAkEzIIEuINEMjzEauTmbEmNTHEmdTzIIE

Now we can take another look at the parts that we have decrypted already and see on the third line, the fragment ENCE. The word mathematician being earlier in the text might make us think that this would be HENCE, but the letter H has already been used, leaving very few possibilities for what this word could actually be. The only words that come to mind are FENCE and PENCE. We can already see THE before the word, so that suggests that it is FENCE, because THE PENCE does not really make sense.

Making the substitution of u for F gives us this table and plaintext:

ABCDEFGHIJKLMNPOQRSTUVWXYZ
abcdHATHcJkLmEopqNstFMwIyz

ANENwINEEKAYHjzICIZTANIAMATHEMATICIANAKeZHmcNAYAZtdkEcITHAHEKlmFzHEEYANlTmalTmydTT
HEMINzIIEtHEzMAaaEZTymzzIbaEAMmdNTmFFENCETHEENwINEEKIZFIKzTHEHEKlzTHEzHEEYINTmACIR
CaEANlTHENydtzTHEFENCEAKmdNlTHEMLlECaAKINwACIKCaEcIaadzETHEaEAzTFENCEFmkAwItENAKeAZ
mTHIZIZTHEbEzTzmadTImNtHEyHjzICIZTIzNEqTzHECKEATEzACIKCdaAKFENCEmFINFINITEkAlIdzAK
mdNlTHEzHEEYANlTHENlKAczTHEFENCETIwHTAKmdNlTHEHEKllECaAKINwTHIZcIaawItETHEzMAaaEZT
CIKcdaAKFENCEAKmdNlTHEHEKlTHEMATHATICIANIZaAZTAFTEkwItINwTHEyKmbaEMaAITTaETHmdWH
THEydtZAZMAaaFENCEAKmdNlHIMzEaFANlTHENlECaAKeZlIEFINEMjzEaFTmbEmNtHEmdTZIIE

This looks convincing because INFINITE has appeared on the fourth line, and FINE has also appeared on the last line. Also, on the first line, we can see a string a letters that couldn't be part of very many different words. It looks like that must stand for ENGINEER. Making those substitutions leaves us with this:

ABCDEFGHIJKLMNPOQRSTUVWXYZ
abcdHATHcJrLmEopqNstFMGIyz

ANENGINEERAYHjzICIZTANIAMATHEMATICIANAREzHmcNAYAZtdREcITHAHERlmFzHEEYANlTmalTmydTT
HEMINzIIEtHEzMAaaEZTymzzIbaEAMmdNTmFFENCETHEENGINEERIZFIRzTHEHERlzTHEzHEEYINTmACIR
CaEANlTHENydtzTHEFENCEARmdNlTHEMLlECaARINGACIRCaEcIaadzETHEaEAzTFENCEFmRAGItENAREAZ
mTHIZIZTHEbEzTzmadTImNtHEyHjzICIZTIzNEqTzHECREATEzACIRCdaARFENCEmFINFINITERAlIdzAR
mdNlTHEzHEEYANlTHENlRACzTHEFENCETIGHTARmdNlTHEHERllECaARINGTHIZcIaaGIItETHEzMAaaEZT
CIRCdaARFENCEARmdNlTHEHERlTHEMATHATICIANIZaAZTAFTERGIItINGTHEyRmbaEMaAITTaETHmdGH
THEydtZAZMAaaFENCEARmdNlHIMzEaFANlTHENlECaAREZlIEFINEMjzEaFTmbEmNtHEmdTZIIE

A couple a sequences of full words have appeared, which looks promising, and shows that we are on the right track. Also, on the penultimate line we can see the sequence CIRCdaARFENCE. It seems very likely that this stands for CIRCULAR FENCE, so that reveals a couple more letter substitutions:

ABCDEFGHIJKLMNPOQRSTUVWXYZ
LbcUHATHcJrLmEopqNstFMGIyz

ANENGINEERAYHjzICIZTANIAMATHEMATICIANAREzHmcNAYAZtUREcITHAHERlmFzHEEYANlTmLlTmyUTT
HEMINzIIEtHEzMALLEzTymzzIbLEAMmUNTmFFENCETHEENGINEERIZFIRzTHEHERlzTHEzHEEYINTmACIR
CLEANlTHENyUTzTHEFENCEARmUNlTHEMLlECLARINGACIRCLEcILLUZETHELEAZTFENCEFmRAGItENAREAZ
mTHIZIZTHEbEzTzmLUTImNtHEyHjzICIZTIzNEqTzHECREATEzACIRCULARFENCEmFINFINITERAlIUzAR
mUNlTHEzHEEYANlTHENlRACzTHEFENCETIGHTARmUNlTHEHERllECLARINGTHIZcILLGIItETHEzMALLEzT
CIRCULARFENCEARmUNlTHEHERlTHEMATHATICIANIZLazTAFTERGIItINGTHEyRmBLEMALITTLELTHmUGH
THEyUTZAZMALLFENCEARmUNlHIMzELFANlTHENlECLAREZlIEFINEMjzELFTmbEmNtHEmUTZlIE

Again, the word LITTLE has appeared near the end of the message, which is reassuring. Right after that is the sequence THmUGHTHE. It seems likely that this stands for THOUGH THE or THOUGHT HE, which should be enough to convince us that m is substituted for O.

Performing that substitution yields this:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
LbcUHATHcJr1OEopqNstFMGIyz

ANENGINEERyHjzICIZTANIAMATHEMATICIANAREzHOcNAyAzTUREcITHAHER1OFzHEEyAN1TOL1ToYUTT
HEMINzI1ETHEzMALLEzTyOzzIbLEAMOUNTOFFENCETHEENGINEERIZFIRzTHEHER1zTHEzHEEyINTOACIR
CLEAN1THENyUTzTHEFENCEAROUN1THEM1ECLARINGACIRCLEcILLUzETHELEAzTFENCEFORAGItENAREAz
OTHIZIZTHEbEzTzOLUTIONTHEyHjzICIZTIzNEqTzHECREATEzACIRCULARFENCEOFINFINITERA1IUzAR
OUN1THEzHEEyAN1THEN1RAczTHEFENCETIGHTAROUN1THEHER11ECLARINGTHIZcILLGIItETHEzMALLEzT
CIRCULARFENCEAROUN1THEHER1THEMATHATICIANIZLAzTAFTERGIItINGTHEyRObLEMALITTLE **THOUGH**
THEyUTzAzMALLFENCEAROUN1HIMzELFAN1THEN1ECLAREzI1EFINEMjzELFTObEONTHEOUTzI1E

The message is beginning to emerge, and we can see many possible sequences of words that could lead us to correct substitutions, but after the part we just looked at is a fairly promising section, so we can work with that. HIMzELF looks almost certain to be HIMSELF, which means that UTz stands for UTS, and that makes it look likely that THOUGHTHEyUTz stands for THOUGHT HE PUTS, considering that the next part reads A SMALL FENCE AROUND HIMSELF if we take l for D. We can now make the substitutions of y for P, z for S and l for D:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
LbcUHATHcJrDOEopqNstFMGIPS

ANENGINEERAPHjSICISTANDAMATHEMATICIANARESHOcNAPASTUREcITHAHERDOFSHEEPANDTOLDTOPUTT
HEMINSIDETHESMALLESTPOSSIBLEAMOUNTOFFENCETHEENGINEERISFIRSTTHEHERDSTHESHEEPINTOACIR
CLEANDTHENPUTSTHEFENCEAROUNDTHEMDECLARINGACIRCLEcILLUSETHELEASTFENCEFORAGItENAREAS
OTHISISTHEbESTSOLUTIONTHEPHjSICISTISNEqTSHECREATESACIRCULARFENCEOFINFINITERADIUSAR
OUNDTHESHEEPANDTHENDRAcSTHEFENCETIGHTAROUNDTHEHERDDECLARINGTHIScILLGIItETHESMALLEST
CIRCULARFENCEAROUNDTHEHERDTHEMATHATICIANISLASTAFTERGIItINGTHEPRObLEMALITTLETHOUGH
THEPUTSASMALLFENCEAROUNDHIMSELFANDTHENECLARESIDEFINEMjSELFTObEONTHEOUTSIDE

Now most of the message has been revealed, and we can insert spaces around most of the words, which will make it very easy to finish off the table of substitutions and reveal the whole message:
AN ENGINEER A PHjSICIST AND A MATHEMATICIAN ARE SHOcN A PASTURE cITH A HERD OF SHEEP AND TOLD TO PUT THEM INSIDE THE SMALLEST POSSIBLE AMOUNT OF FENCE THE ENGINEER IS FIRST HE HERDS THE SHEEP INTO A CIRCLE AND THEN PUTS THE FENCE AROUND THEM DECLARING A CIRCLE cILL USE THE LEAST FENCE FOR A GItEN AREA SO THIS IS THE bEST SOLUTION THE PHjSICIST IS NEqT SHE CREATES A CIRCULAR FENCE OF INFINITE RADIUS AROUND THE SHEEP AND THEN DRAcS THE FENCE TIGHT AROUND THE HERD DECLARING THIS cILL GIItE THE SMALLEST CIRCULAR FENCE AROUND THE HERD THE MATHEMATICIAN IS LAST AFTER GIItING THE PRObLEM A LITTLE THOUGHT HE PUTS A SMALL FENCE AROUND HIMSELF AND THEN DECLARES I DEFINE MjSELF TO bE ON THE OUTSIDE

This gives us the final substitutions of j for Y, c for W, t for V, b for B and q for X with little difficulty.

This leads us to the final table and plaintext:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
LBWUHAT CYRDOE XN VFMGIPS

AN ENGINEER A PHYSICIST AND A MATHEMATICIAN ARE SHOWN A PASTURE WITH A HERD OF SHEEP AND TOLD TO PUT THEM INSIDE THE SMALLEST POSSIBLE AMOUNT OF FENCE THE ENGINEER IS FIRST HE HERDS THE SHEEP INTO A CIRCLE AND THEN PUTS THE FENCE AROUND THEM DECLARING A CIRCLE WILL USE THE LEAST FENCE FOR A GIVEN AREA SO THIS IS THE BEST SOLUTION THE PHYSICIST IS NEXT SHE CREATES A CIRCULAR FENCE OF INFINITE RADIUS AROUND THE SHEEP AND THEN DRAWS THE FENCE TIGHT AROUND THE HERD DECLARING THIS WILL GIVE THE SMALLEST CIRCULAR FENCE AROUND THE HERD THE MATHEMATICIAN IS LAST AFTER GIVING THE PROBLEM A LITTLE THOUGHT HE PUTS A SMALL FENCE AROUND HIMSELF AND THEN DECLARES I DEFINE MYSELF TO BE ON THE OUTSIDE

Note that the letters H, O, P and S did not appear in the ciphertext and therefore we cannot know what they were substituted for, and we do not need to know. The final message has been revealed! Job done!

Notes

The plaintext was a fairly well known Engineer, Physicist, Mathematician joke taken from <http://www.phy.ilstu.edu/~rfm/107F07/EPMjokes.html>:

An engineer, a physicist, and a mathematician are shown a pasture with a herd of sheep, and told to put them inside the smallest possible amount of fence. The engineer is first. He herds the sheep into a circle and then puts the fence around them, declaring, "A circle will use the least fence for a given area, so this is the best solution." The physicist is next. She creates a circular fence of infinite radius around the sheep, and then draws the fence tight around the herd, declaring, "This will give the smallest circular fence around the herd." The mathematician is last. After giving the problem a little thought, he puts a small fence around himself and then declares, "I define myself to be on the outside!"

The Python to make the uppercase letters bold:

```
upper = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
```

```
text = raw_input()
```

```
result = ""
```

```
for i in text:
```

```
    if upper.find(i) != -1:
```

```
        result += "<b>" + i + "</b>"
```

```
    else:
```

```
        result += i
```

```
print "<div style=\"font: 10pt Consolas\">" + result + "</div>"
```

Outputs HTML which can be viewed in a web browser and then the text can be copied into word with the bold characters and correct font.