Cryptanalysis of the Vigenere Cipher

For a recap of how the Vigenere Cipher works, see here

The Vigenere cipher was though to be completely unbreakable for hundreds of years, and indeed, if very long keys are used the vigenere cipher can be unbreakable. But if short keys are used, or if we have a lot of ciphertext compared to the key length, the vigenere cipher is quite solvable.

Cryptanalysis of the Vigenere cipher has 2 main steps: identify the period of the cipher (the length of the key), then find the specific key. To identify the period we use a test based on the Index of Coincidence, to find the specific key we use the Chi-squared statistic. For the purposes of this explanation, we will try to break the following message:

```
vptnvffuntshtarptymjwzirappljmhhqvsubwlzzygvtyitarptyiougxiuydtgzhhvvmum
shwkzgstfmekvmpkswdgbilvjljmglmjfqwioiivknulvvfemioiemojtywdsajtwmtcgluy
sdsumfbieugmvalvxkjduetukatymvkqzhvqvgvptytjwwldyeevquhlulwpkt
```

The first thing to note is that there is no guarantee that the period of key that we find is the actual key used. If the message is very long we can be almost certain of being correct, but the methods provided here are approximate.

Finding the Period

The Vigenere cipher applies different Caesar ciphers to consecutive letters. If the key is 'PUB', the first letter is enciphered with a Caesar cipher with key 16 (P is the 16th letter of the alphabet), the second letter with another, and the third letter with another. When we get to the 4th letter, it is enciphered using the same cipher as letter 1. As a result, if we gather letters 1,4,7,10,... we should get a sequence of characters, all of which were enciphered using the same Caesar cipher. The sequence of characters 2,5,8,11,... and 3,6,9,12,... will also be enciphered with their own Caesar cipher. The exact sequence will of course depend on the period of the cipher i.e. the key length.

The Index of Coincidence (I.C.) is a statistical technique that gives an indication of how English-like a piece of text is. One of the useful properties of the technique is that the result of the I.C. does not change if you apply a substitution cipher to the text. This is because the I.C. is based on letter frequencies, and simple substitution ciphers do not modify the individual letter frequencies. If text is similar to english it will have an I.C. of around 0.06, if the characters are uniformly distributed the I.C. is closer to 0.03-0.04.

To determine the period of a Vigenere cipher we first assume the key length is 2. We extract the two sequences 1,3,5,7,... and 2,4,6,8,... from the ciphertext. For the example we are working with we get the following result (note that the I.C. is calculated using the whole sequences, not just the part shown)

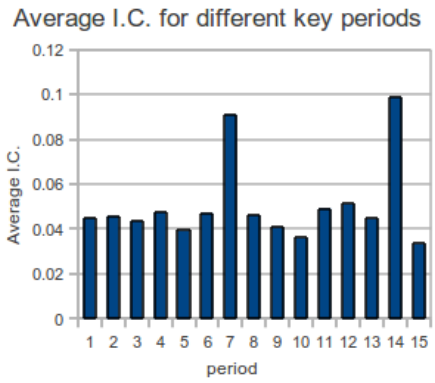
```
                                                  I.C.
original:    vptnvffuntshtarptymjwzirappljmhhqvsubw...     0.049
if key were length 2:
sequence 1:  v t v f n s t r t m w i a p j h q s b ...     0.049
sequence 2:   p n f u t h a p y j z r p l m h v u w...     0.046
                                           average:  0.048
if key were length 3:
sequence 1:  v  n  f  t  t  p  m  z  a  l  h  v  b ...     0.049
sequence 2:   p  v  u  s  a  t  j  i  p  j  h  s  w...     0.046
sequence 3:    t  f  n  h  r  y  w  r  p  m  q  u  ...     0.046
                                           average:  0.047
```

This procedure of breaking up the ciphertext and calculating the I.C. for each subsequence is repeated for all the key lengths we wish to test. What we are most interested in is the average I.C. for a particular period, for the case of period = 2, the average I.C. is around 0.048. If you were to continue this procedure up to a period of 15 we get the following average I.C. values:

```
period         avg I.C.
-----------------------
1 :     0.0449443523561
2 :     0.0457833618884
3 :     0.0435885364312
4 :     0.0474962292609
```

Contents

```
5 :     0.0393612078978
6 :     0.0471437059672
7 :     0.0909922589726
8 :     0.0461858974359
9 :     0.0407804755631
10:     0.0361152882206
11:     0.0491603339901
12:     0.0512663398693
13:     0.0446886446886
14:     0.0988487702773
15:     0.0334554334554
```



We have 2 rows that have very high values of average I.C. This indicates the key is probably of length 7, but could also be of length 14. Both of these probabilities should be tested.

Finding the Key

Since we now know the period is 7, we only have 7 Caesar ciphers to break, which is fairly easy. For this task we will use the Chi-squared statistic, which will compare the frequency distribution of our subsequences to the expected English frequency distribution.

Using every seventh letter starting with the first, our first sequence is 'VURZJUGRGGUGVGJQKEOAGUGKKQVWQP'. This is text all enciphered with the same Caesar cipher, we want to know what the key is. The procedure for this is described fully in the page on the Chi-squared statistic. In essence, we try deciphering this sequence with each of the 25 possible Caesar ciphers, and compare the frequency distribution of the deciphered text with the frequency distribution of English for each key. If we perform this, we get 26 values for the Chi-squared statistic. The correct key will correspond to the deciphered text with the lowest Chi-squared statistic (we hope, due to the statistical nature of the problem it may be the second or third lowest value). We show the results of this procedure here:

```
key         deciphered sequence           chi-sq
-------------------------------------------------
0       VURZJUGRGGUGVGJQKEOAGUGKKQVWQP     595.42
1       UTQYITFQFFTFUFIPJDNZFTFJJPUVPO     466.86
2       TSPXHSEPEESETEHOICMYESEIIOTUON      41.22
3       SROWGRDODDRDSDGNHBLXDRDHHNSTNM      67.73
4       RQNVFQCNCCQCRCFMGAKWCQCGGMRSML     642.37
5       QPMUEPBMBBPBQBELFZJVBPBFFLQRLK     451.49
6       POLTDOALAAOAPADKEYIUAOAEEKPQKJ     121.97
7       ONKSCNZKZZNZOZCJDXHTZNZDDJOPJI    2441.20
8       NMJRBMYJYYMYNYBICWGSYMYCCINOIH     190.46
9       MLIQALXIXXLXMXAHBVFRXLXBBHMNHG    1142.90
10      LKHPZKWHWWKWLWZGAUEQWKWAAGLMGF     358.87
11      KJGOYJVGVVJVKVYFZTDPVJVZZFKLFE     962.13
12      JIFNXIUFUUIUJUXEYSCOUIUYYEJKED     354.18
13      IHEMWHTETTHTITWDXRBNTHTXXDIJDC     241.97
14      HGDLVGSDSSGSHSVCWQAMSGSWWCHICB     107.36
15      GFCKUFRCRRFRGRUBVPZLRFRVVBGHBA     136.40
16      FEBJTEQBQQEQFQTAUOYKQEQUUAFGAZ    1801.65
17      EDAISDPAPPDPEPSZTNXJPDPTTZEFZY     531.22
18      DCZHRCOZOOCODORYSMWIOCOSSYDEYX     247.66
19      CBYGQBNYNNBNCNQXRLVHNBNRRXCDXW     377.60
20      BAXFPAMXMMAMBMPWQKUGMAMQQWBCWV     489.12
21      AZWEOZLWLLZLALOVPJTFLZLPPVABVU     815.45
22      ZYVDNYKVKKYKZKNUOISEKYKOOUZAUT     648.33
23      YXUCMXJUJJXJYJMTNHRDJXJNNTYZTS    1476.11
24      XWTBLWITIIWIXILSMGQCIWIMMSXYSR     279.93
25      WVSAKVHSHHVHWHKRLFPBHVHLLRWXRQ     158.53
```

This means our first Vigenere key letter is 'C' (A=0,B=1,C=2,...). We have to repeat this procedure for each of the 7 key letters. If we continue this procedure of finding the keys corresponding to the Chi-squared minima, we get the sequence 2,8,0,7,4,17,18. This spells out 'CIAHERS', which is wrong. This goes to show you can't rely on the technique fully unless very long ciphertexts are available. The correct key was 'CIPHERS', and indeed the Chi-square test had two very low values for that subsequence. Unfortunately the incorrect one was slightly lower.

Wrapping up

As shown above, statistical techniques can give you wrong answers. To get around this you may have to try decrypting the ciphertext with each of several likely candidates to find the true key.

For a more reliable approach, and one which is conceptually a bit simpler, see Cryptanalysis of the Vigenere Cipher, Part 2.

9 Comments **Practical Cryptography** Login

Recommend 3 Share Sort by Best

Join the discussion…

LOG IN WITH OR SIGN UP WITH DISQUS

Name

JS • 3 years ago

I am a newbie to cryptography, Can you please explain how you got other

deciphered sequence.

2 • Reply • Share ›

CKD • 3 years ago

Once you've found the period using the Index of Coincidence method, try finding the actual key with Simulated Annealing. It takes about 1.3 seconds to get a 30-byte random key using Free Pascal, (and approximately 30 seconds for a 50-byte one) on a Pentium 4 PC dating from 2005! Character perfect every time too. Ridiculously good.

1 • Reply • Share ›

Alex • 18 days ago

I don't understand how deciphered sequences for k > 0 are obtained.

• Reply • Share ›

Alexandre • a month ago

Hi. I am looking for a source code that could help me to determine the period and especially to generate the different sequences according to the length of the key (I'm writting a C program). Thx :)
=> http://www.practicalcryptog...
=> https://www.dcode.fr/index-...

• Reply • Share ›

soad • 2 years ago

A dirty implementation done in python.

* This method will work only with cipher texts with enough length.

```
#!/usr/bin/python3
def get_ic(s):
    n=len(s)
    ic=0
    if n-1:ic=(1/(float(n)*(n-1)))*(sum([s.count(a)*(s.count(a)-1) for a in set(s)]))
    return ic

def get_possible_key_ls(avg_ic_arr):
    cpy=avg_ic_arr.copy()
    avg_ic_arr.sort(reverse=True)
    key_ls=[cpy.index(avg_ic_arr[0])+2,cpy.index(avg_ic_arr[1])+2]
    return key_ls

def get_key_len(c,max_guess):
    avg_ic_arr=[]
    for n in range(2,max_guess+1):
        ic_sum=0.0
        avg_ic=0.0
        for i in range(n):
            s=""
            for j in range(0,len(c[i:]),n):
                s+=(c[i:][j])
            ic=get_ic(s)
            ic_sum+=ic
        avg_ic=ic_sum/n
```

see more

• Reply • Share ›

Anthony Dane • 3 years ago

Is there a way to use the vigenere cipher using a two digit number for one letter and if so how would it be set up if the highest number was 55 or so

• Reply • Share ›

santiagop593 • 3 years ago

Hello guys, this webpage is awesome! It helped in the creation of my first Android App. Thanks practicalcryptography.com !
My app encrypts, decrypts and breaks the encryptions of Caesar and Vigenere cypher. You can save those texts, share them with friends and view Bar Charts with character counts, Index of Coincidence, get some info relevant to the encryption... and more to come soon

Give it a try, I'll be adding new schemas, and other cool features realated with cryptography

https://play.google.com/sto...

Thanks in advance

▲ | ▼ • Reply • Share ›

Elias • 4 years ago

Thanks very much. i really like this post i was stuck here but now i understand how kasisky and index of coincidence finds out the key length of a given ciphertext key to decrypt it...;)

▲ | ▼ • Reply • Share ›

brother • 4 years ago

It would be great if you could extend the floating point IC's in the section where you show the IC's for the indiviual sequences; its only 2 digits after the point so far and a little more could help people like me who like to code stuff up.

▲ | ▼ • Reply • Share ›

Subscribe Add Disqus to your siteAdd DisqusAdd Privacy

y ngp'i zpgo avce ge lgm avce vj oscc vj y jagmcn cyzs; vpn y cyzs csjj iavp avce ge lgm avce vj oscc vj lgm nsjsuds – q.u.u. igczysp. (ias esccgojayk ge ias uyph)

Copyright & Usage

Copyright James Lyons © 2009–2012
No reproduction without permission.

Questions/Feedback

Notice a problem? We'd like to fix it!
Leave a comment on the page and we'll take a look.