Fall 2019
Chris Christensen
MAT/CSC 483
Section 15

# William Friedman's Index of Coincidence

William Friedman (1891 – 1969) developed statistical methods for determining whether a cipher is monoalphabetic or polyalphabetic and for determining the length of the keyword if the cipher is polyalphabetic.

Friedman retired from the National Security Agency in 1955 after 35 years of service with U.S. cryptological activities. He transformed the methods and approaches of cryptology from the traditional into the modern by applying statistics to cryptology. His wife Elizebeth was also a cryptologist and served at one point with the Coast Guard, cryptanalyzing messages of the rumrunners.

## Determining Whether We Have a Monoalphabetic Cipher or a Polyalphabetic Cipher

Friedman's method for determining whether a cipher is monoalphabetic or polyalphabetic is based upon the probability of randomly selecting two letters from an alphabet and having them be the same.

Consider the probability of randomly selecting two letters from a ciphertext alphabet and having them be the same. Let us say that there are $n$ letters in our ciphertext and $n_a$ `a`s in our ciphertext. Then the probability of selecting two `a`s would be $\frac{n_a}{n} \times \frac{n_a - 1}{n - 1}$.

The probability of choosing two letters the same (i.e., two `a`s or two `b`s or two `c`s or … or two `z`s) would be

$$\frac{n_a}{n} \times \frac{n_a - 1}{n - 1} + \frac{n_b}{n} \times \frac{n_b - 1}{n - 1} + \frac{n_c}{n} \times \frac{n_c - 1}{n - 1} + \ldots + \frac{n_z}{n} \times \frac{n_z - 1}{n - 1}.$$

This number is denoted *I* and called the **index of coincidence** of the ciphertext.

$$I = \frac{n_a}{n} \times \frac{n_a - 1}{n-1} + \frac{n_b}{n} \times \frac{n_b - 1}{n-1} + \frac{n_c}{n} \times \frac{n_c - 1}{n-1} + \ldots + \frac{n_z}{n} \times \frac{n_z - 1}{n-1}$$

The frequencies of the letters in English are:

| Letter | a | b | c | d | e | f | g | h | i | j | k | l | m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | .082 | .015 | .028 | .043 | .127 | .022 | .020 | .061 | .070 | .002 | .008 | .040 | .024 |
| Letter | n | o | p | q | r | s | t | u | v | w | x | y | z |
| Frequency | .067 | .075 | .019 | .001 | .060 | .063 | .091 | .028 | .010 | .023 | .001 | .020 | .001 |

Beker and Piper, *Cipher Systems: The Protection of Communications*, Wiley.

So, if a text were enciphered using a single alphabet, the probability of "drawing" two letters that are the same is:

| aa | or | bb | or | cc | or | … | or | zz |
|---|---|---|---|---|---|---|---|---|
| $.082 \times .082$ | + | $.015 \times .015$ | + | $.028 \times .028$ | + | … | + | $.001 \times .001$ |

This probability of "drawing" two letters that are the same – the index of coincidence -- is approximately $I \approx 0.0656010$.

If more than one alphabet were used, the frequencies of the letters should be more nearly uniform. If they *were* uniform, the probability of "drawing" two letters that were the same would be:

$$I \approx \underbrace{\left(\frac{1}{26} \times \frac{1}{26}\right) + \left(\frac{1}{26} \times \frac{1}{26}\right) + \left(\frac{1}{26} \times \frac{1}{26}\right) + \ldots + \left(\frac{1}{26} \times \frac{1}{26}\right)}_{26 \text{ terms}} = \frac{1}{26} \approx 0.038.$$

Here is the idea of the test.

If the ciphertext were generated by a monoalphabetic cipher, we should determine *I* to be near 0.065 because a monoalphabetic cipher is just a permutation of the letters of a single alphabet. The frequencies of letters for the ciphertext alphabet should be nearly the same as for English – but in a different order.

If the cipher were generated by a polyalphabetic cipher, the frequencies of the letters would become more nearly uniform – more nearly the same for each letter. We should determine *I* to be near 0.038.

We test the ciphertext by calculating *I* based on the ciphertext frequencies. The closer that *I* is to 0.065, the more likely it is that we have a monoalphabetic cipher. The closer that *I* is to 0.038, the more likely that we have a polyalphabetic cipher.

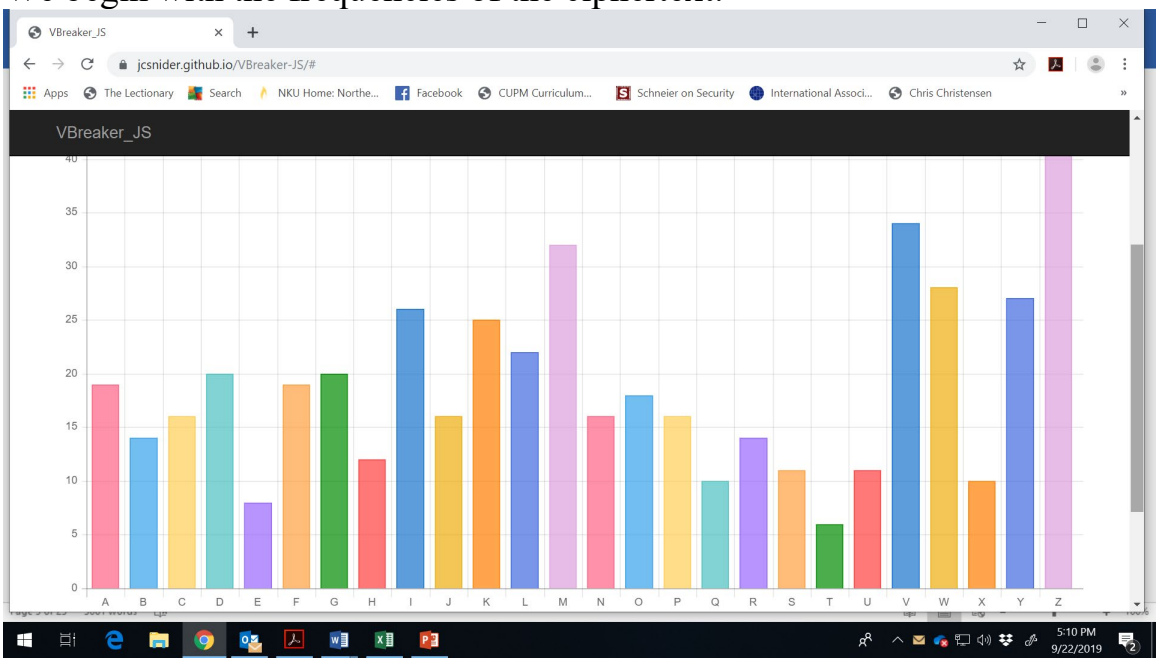
Recall that, using frequency analysis, peaks and valleys of frequencies suggest a monoalphabetic cipher and relatively uniform frequencies suggest a polyalphabetic cipher.

Typically we use both the Friedman test and frequency analysis to determine the kind of cipher we have.

Here is the Vigenère cipher that we cryptanalyzed in a previous section using Kasiski's method:

```
DBZMG   AOIYS   OPVFH   OWKBW   XZPJL   VVRFG   NBKIX
DVUIM   OPFQL   VVPUD   KPRVW   OARLW   DVLMW   AWINZ
DAKBW   MMRLW   QIICG   PAKYU   CVZKM   ZARPS   DTRVD
ZWEYG   ABYYE   YMGYF   YAFHL   CMWLW   LCVHL   MMGYL
DBZIF   JNCYL   OMIAJ   JCGMA   IBVRL   OPVFW   OBVLK
OPVUJ   ZDVLQ   XWDGG   IQEYF   BTZMZ   DVRMM   ANZWA
ZVKFQ   GWEAL   ZFKNZ   ZZVCK   VDVLQ   BWFXU   CIEWW
OPRMU   JZIYK   KWEXA   IOIYH   ZIKYV   GMKNW   MOIIM
KADUQ   WMWIM   ILZHL   CMTCH   CMINW   SBRHV   OPVSO
DTCMG   HMKCE   ZASYD   JKRNW   YIKCF   OMIPS   GAFZK
JUVGM   GBZJD   ZWWNZ   ZVLGT   ZZFZS   GXYUT   ZBJCF
PAVNZ   ZAVWS   IJVZG   PVUVQ   NKRHF   DVXNZ   ZKZJZ
ZZKYP   OIEXX   MWDNZ   ZQIMH   VKZHY   DVKYD   GQXYF
OOLYK   NMJGS   YMRML   JBYYF   PUSYJ   JNRFH   CISYL
N
```

We begin with the frequencies of the ciphertext:




| | |
|---|---|
| A | 19 |
| B | 14 |
| C | 16 |
| D | 20 |
| E | 8 |
| F | 19 |
| G | 20 |
| H | 12 |
| I | 26 |
| J | 16 |
| K | 25 |
| L | 22 |
| M | 32 |
| N | 16 |
| O | 18 |
| P | 16 |
| Q | 10 |
| R | 14 |
| S | 11 |
| T | 6 |
| U | 11 |
| V | 34 |
| W | 28 |
| X | 10 |
| Y | 27 |
| Z | 41 |
| | 491 |

Notice that the relatively uniform frequencies suggest a polyalphabetic cipher. This should be confirmed by our calculation of *I*.

The calculation of *I* is easy.

$$I = \left(\frac{19}{491}\times\frac{18}{490}\right)+\left(\frac{14}{491}\times\frac{13}{490}\right)+\left(\frac{16}{491}\times\frac{15}{490}\right)+\ldots+\left(\frac{41}{491}\times\frac{40}{490}\right)$$
$$= \frac{(19\times18)+(14\times13)+(16\times15)+\ldots+(41\times40)}{491\times490}$$
$$\approx 0.044$$

Because *I* is so near to 0.038 (random alphabet), we can reasonably assume that we have a polyalphabetic cipher. This confirms what we noticed when we saw the relatively uniform frequencies.

## Estimating the Length of the Keyword

Friedman also developed a method for estimating $l$ the length of the keyword.

We will develop an approximation formula for $I$, the index of coincidence; this formula will contain $l$ and $n$, the length of the keyword and the number of letters in the ciphertext. Then, to get an approximation for the length $l$, we will solve for $l$ in terms of $I$ and $n$ (we know $n$, and we can calculate $I$).

First, assume that we know $l$ and that we arrange the ciphertext into $l$ columns. Now each column corresponds to a Caesar cipher. Although the columns might not all have the same length, we will assume that the number of letters in the ciphertext is large enough so that we *can assume* that they each have length $\frac{n}{l}$ (i.e., that the length of the message is large enough so that the error using this number for the length of each column is not large).

If we chose two letters from the ciphertext, what is the probability that they come from the same column and are the same letter? First, we select a letter from the ciphertext. This selection determines a column. The probability that the next letter chosen comes from the same column is $\frac{\frac{n}{l}-1}{n-1}$. Because both letters are selected from the same Caesar cipher alphabet, the probability that both are the same is approximately the same as for standard English 0.065. So, the probability that both letters are selected from the same column and are the same letter is approximately $\frac{\frac{n}{l}-1}{n-1}\times 0.065$.

The other possibility is that we select two letters from the ciphertext that come from different columns but are the same letter. What is that probability? First, we select a letter from the ciphertext. Again, this determines a column. The probability that the next letter comes from a different column is $\frac{n-\frac{n}{l}}{n-1}$. Because the two letters are selected from different Caesar cipher alphabets, the probability that both are the same is

approximately the same as for a random alphabet 0.038. So, the probability that both letters are selected from different columns and are the same letter is approximately $\frac{n-\frac{n}{l}}{n-1}\times 0.038$.

So we have two cases – the two letters are selected from the same column and are the same letter or the two letters are selected from different columns and are the same letter. To get an approximation of the index of coincidence *I*, the probability that the two letters selected are the same, we add these two probabilities:

$$I \approx \frac{\frac{n}{l}-1}{n-1}\times 0.065 + \frac{n-\frac{n}{l}}{n-1}\times 0.038.$$

Doing a bit of algebra to solve for *l*, we obtain:

$$I \approx \frac{\frac{n}{l}-1}{n-1}\times 0.065 + \frac{n-\frac{n}{l}}{n-1}\times 0.038$$

$$(n-1)I \approx \left(\frac{n}{l}-1\right)\times 0.065 + \left(n-\frac{n}{l}\right)\times 0.038$$

$$(n-1)I \approx \frac{n}{l}\times 0.065 - 0.065 + n\times 0.038 - \frac{n}{l}\times 0.038$$

$$(n-1)I + 0.065 - 0.038n \approx \frac{n}{l}\times(0.065-0.038)$$

$$(n-1)I + 0.065 - 0.038n \approx 0.027\frac{n}{l}$$

$$l \approx \frac{0.027n}{(n-1)I + 0.065 - 0.038n}$$

Now let us determine an approximation for the length of the keyword in the ciphertext given above.

We determined that $n = 491$, and we calculated above that $I \approx 0.044$; so,

$$l \approx \frac{0.027 \times 491}{(491-1) \times 0.044 + 0.065 - 0.038 \times 491} \approx 4.468 .$$

Recall that Kasiski method, when applied to this ciphertext suggested that the length of the keyword is 5. In practice, we should consider the results of both tests.

Here are the associations between keyword length and the index of coincidence for a ciphertext message of 200 letters. Notice that it would be difficult to determine the length of the keyword based only on knowing the index of coincidence.

Length of keyword

| Keyword length | Index of Coincidence |
|---|---|
| 1 | 0.06500 |
| 2 | 0.05143 |
| 3 | 0.04691 |
| 4 | 0.04465 |
| 5 | 0.04329 |
| 6 | 0.04239 |
| 7 | 0.04174 |
| 8 | 0.04126 |
| 9 | 0.04088 |
| 10 | 0.04058 |

## The Columns are Monoalphabetic

Before ending, let us go back to the five alphabets of the ciphertext example given above and calculate *I* for alphabet one. *I* should be near 0.065, the monoalphabetic case of *I*.

Alphabet one:

$$I \approx \frac{4 \times 3 + 2 \times 1 + 6 \times 5 + 10 \times 9 + 6 \times 5 + 5 \times 4 + 7 \times 6 + 3 \times 2 + 4 \times 3 + 4 \times 3 + 13 \times 12 + 4 \times 3 + 3 \times 2 + 2 \times 1 + 4 \times 3 + 2 \times 1 + 4 \times 3 + 16 \times 15}{99 \times 98}$$

$I \approx 0.072$.

This confirms what we notice by looking at the frequencies of alphabet number one.

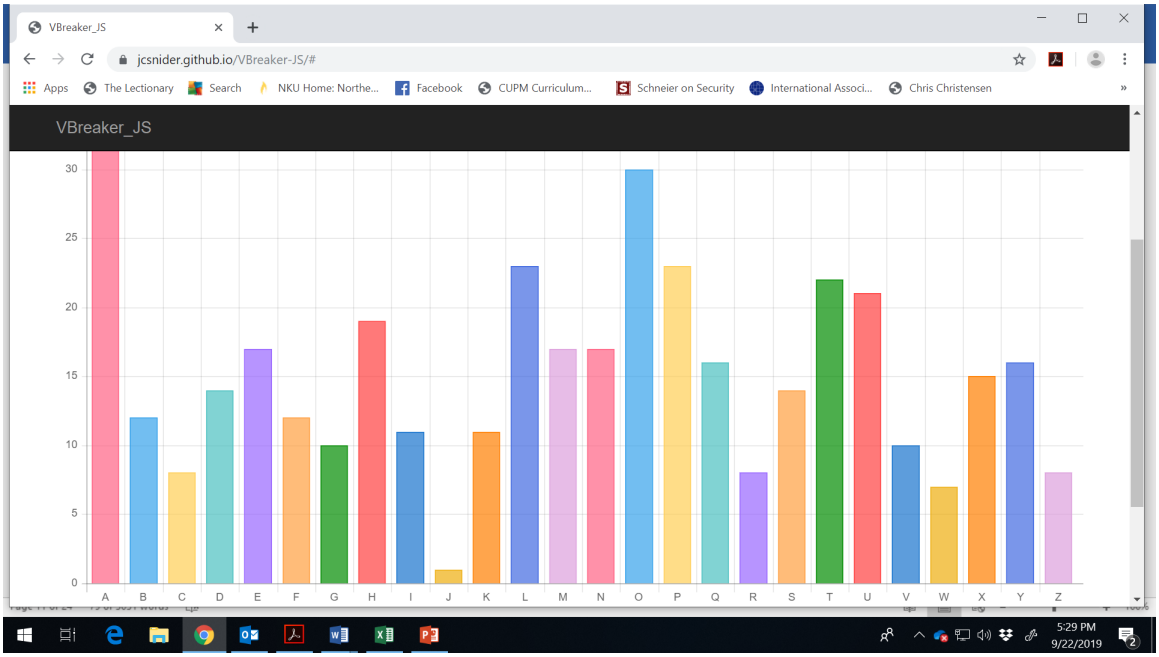
This suggests a brute force way to attack the length of the keyword. Assume that $l = 1, 2, 3, \ldots$; for each of these values of $l$ separate alphabets and calculate *I*; and determine for which value of $l$ all the separated alphabets are monoalphabetic.

Consider the following ciphertext:

```
pukpz gmoqs ihzil ooiop xwtyf hxpfa epmng hhyfh
pelvy enzqo yatev vymxy oitsq nbnya fhohb uqbna
iimop iymqr xuflr durxk domvd stupd bsxyd uaxkl
ooold ewate bufek umlpp digna fmoqs xatel voaes
qdkhu lphke gpsmt omsmo qsttq btzuc laduv agrxh
etaly avoun xbeew iktal uttsu agzno moafm oqsxc
qrlpa nlvrt alqnb nyafh ohbuq wapoh wppnh ataol
blnnn otyps ahpag lztkf pilja npouc aatex sqcmy
uctso ogamc mziek loogu qcmlp thate dlkbh hddbu
fhxvd dxycw xyfzn paalk rgaqw preov uuyl
```

Here are the index of coincidence and the ciphertext frequencies:

$I = .04408$

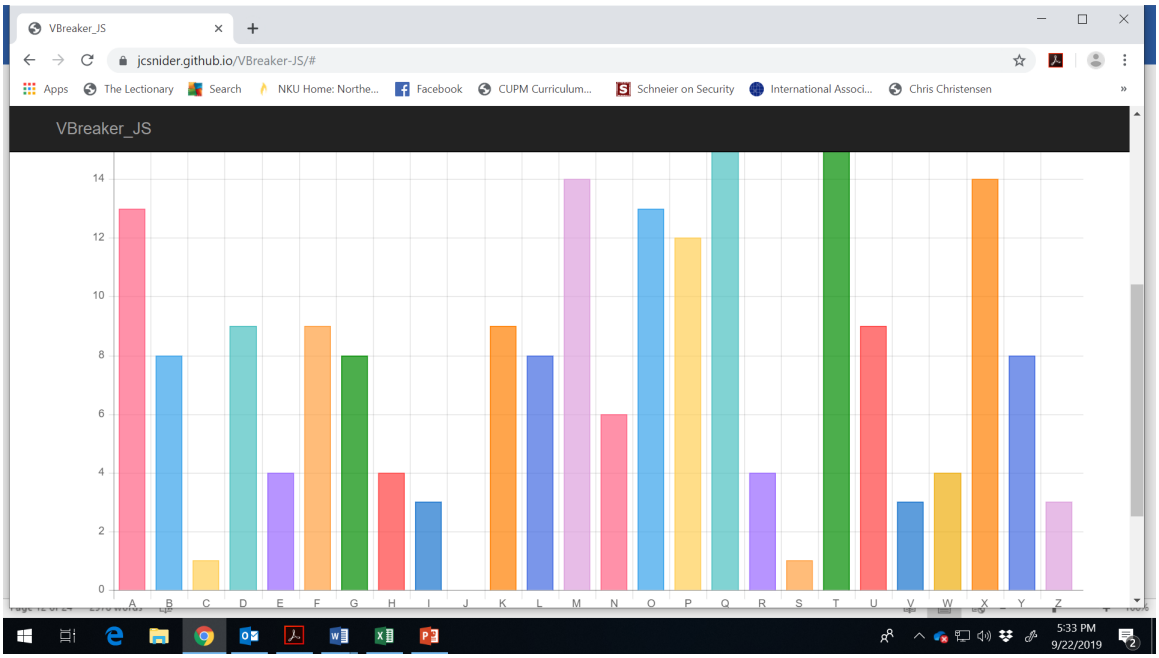



Both the index of coincidence and the ciphertext frequencies suggest that this ciphertext has been encrypted with a polyalphabetic cipher.

Assume that the length of the keyword is 2:

Strip the first alphabet, calculate the index of coincidence, and look at its frequencies:

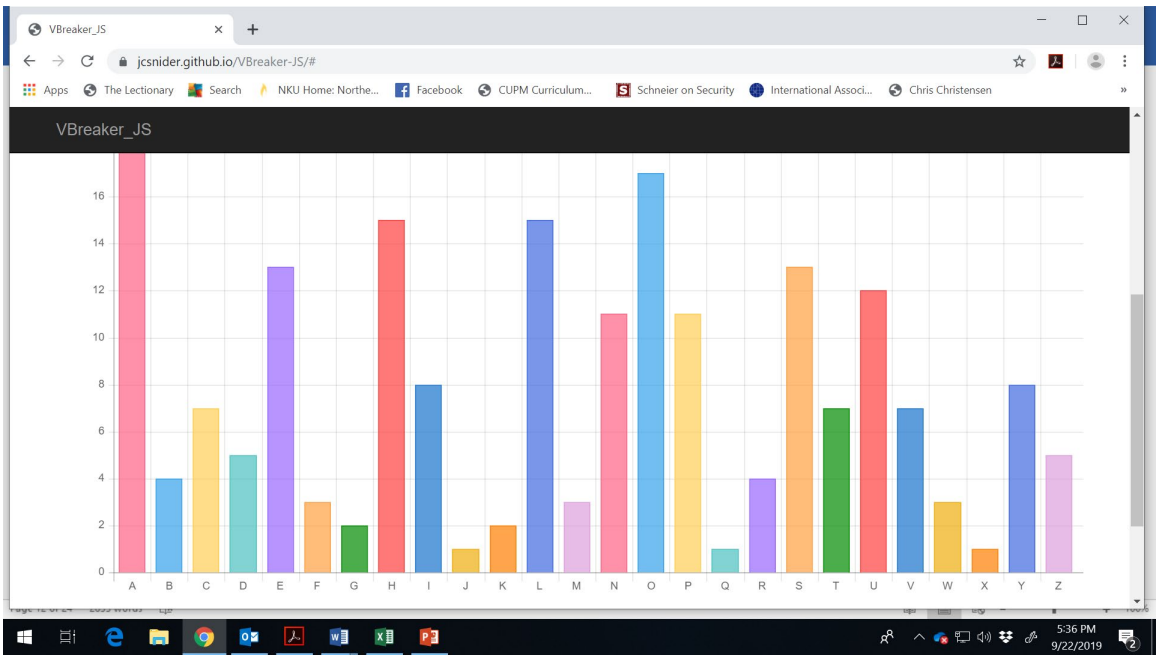
$I = .04731$




This does not appear to be monoalphabetic.

The second alphabet also does not appear to be monoalphabetic

$I = .0521$

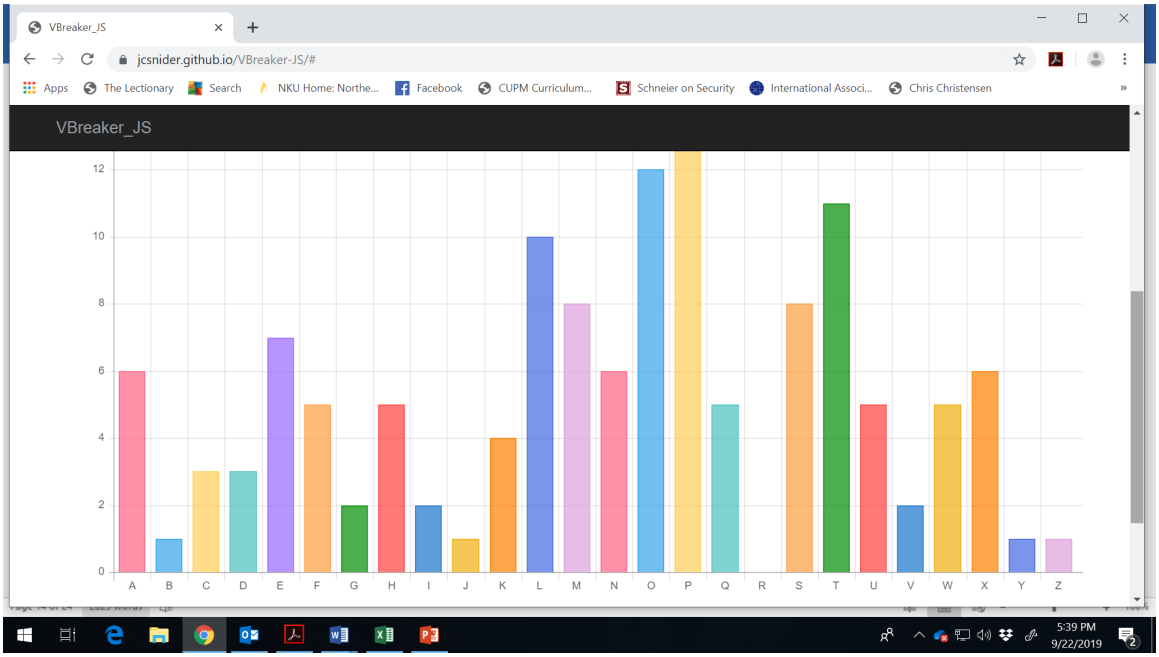



The keyword does not appear to have length 2.

Now assume that the length of the keyword were 3 and strip 3 alphabets:

Alphabet number 1

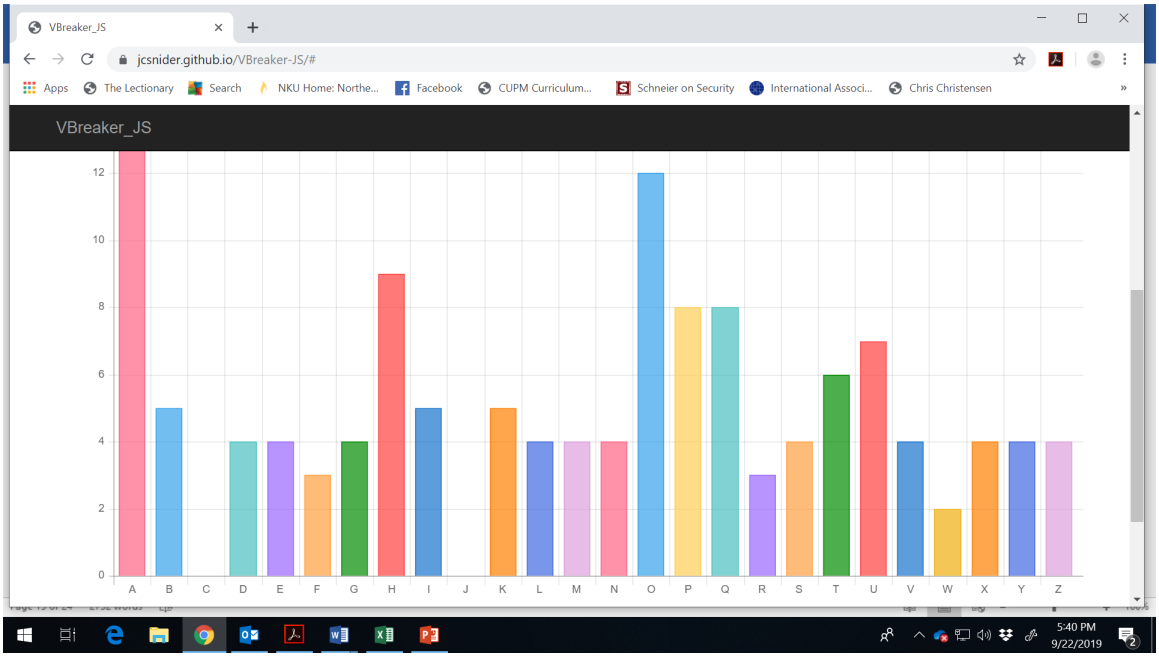
$I = .40947$




This does not appear to be monoalphabetic.

Alphabet number 2

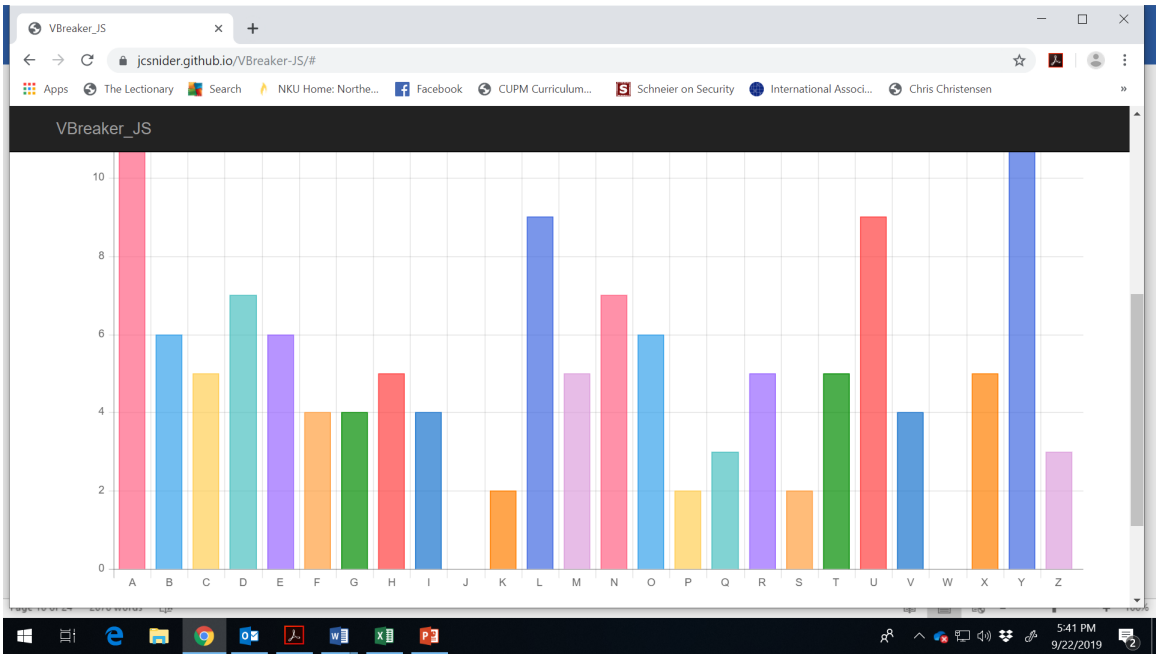
$I = .04522$




This does not appear to be monoalphabetic.

Alphabet number 3

$I = .04347$



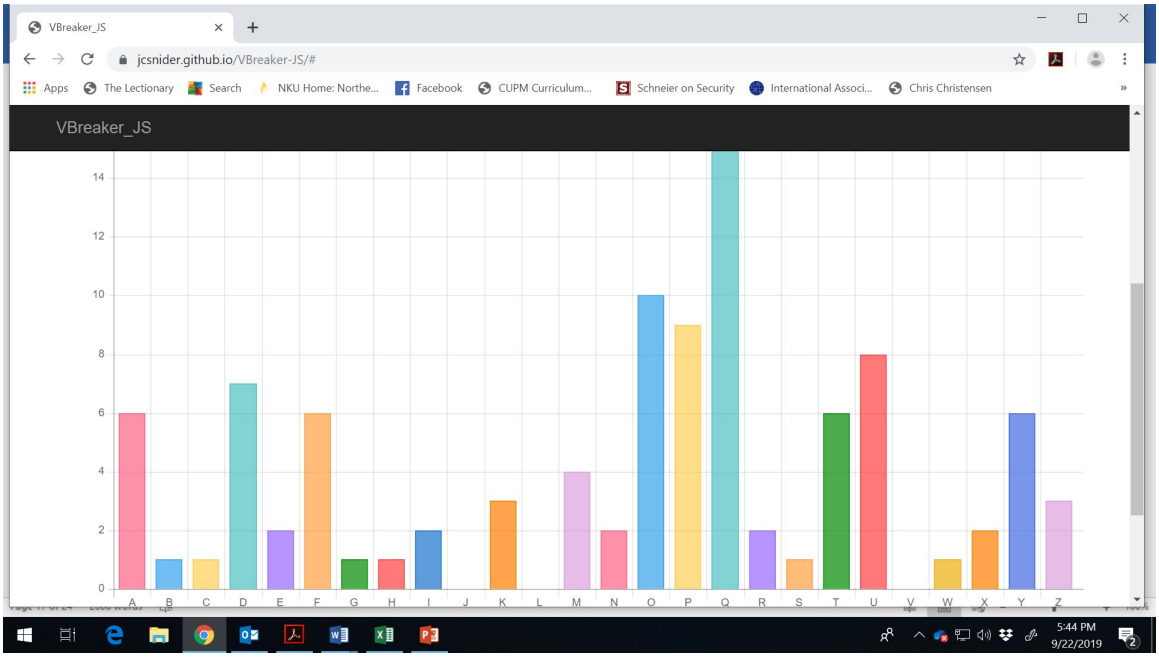

This also does not appear to be monoalphabetic.

The keyword does not appear to have length 3.

Now assume that the length of the keyword were 4 and strip 4 alphabets:

Alphabet number one

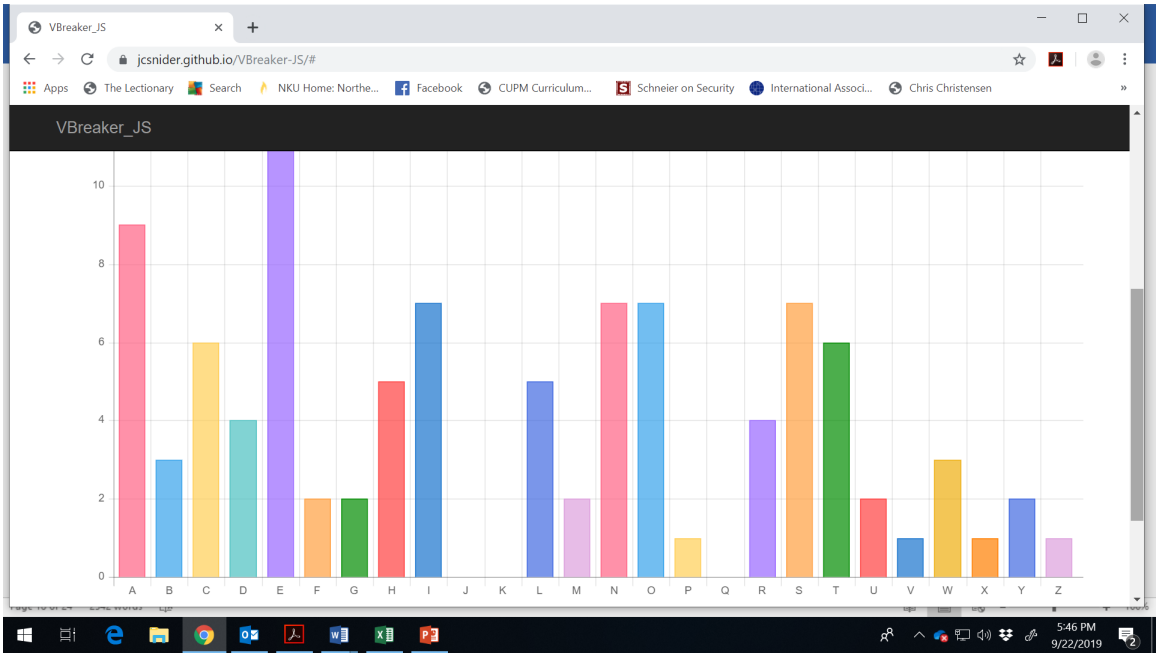
$I = .06367$



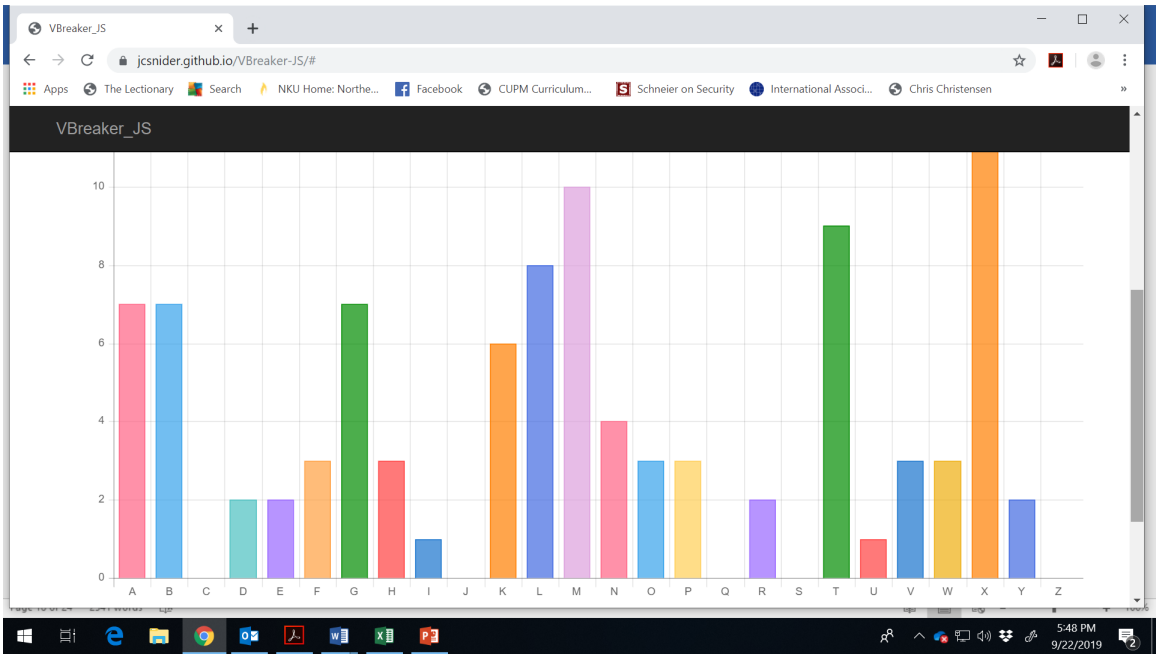

This appears to be monoalphabetic. It seems to correspond to a Caesar cipher. What would be the first letter of the keyword?

Here are the other three alphabets:

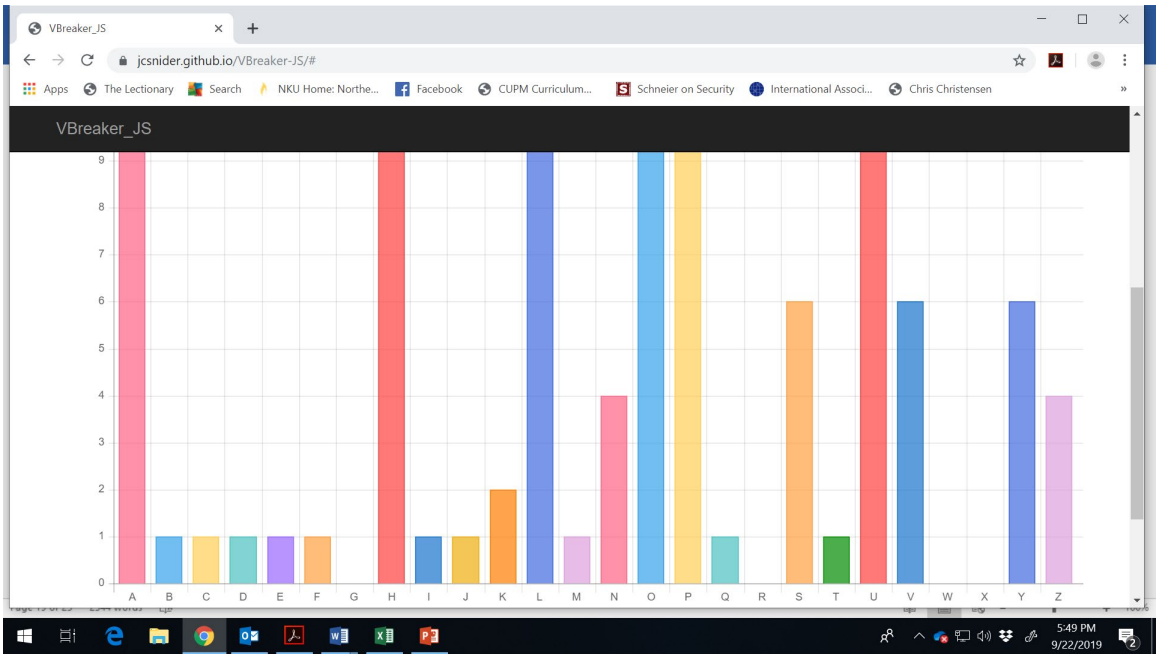
Alphabet number 2:




Alphabet number 3:



Alphabet number 4:




What is the keyword?