Index of coincidence

$$I_c = \frac{\text{\# of pairs of equal letters in } C}{\text{total \# of pairs of letters in } C}$$

. . . . E . . . . R . . . .

pair of letters
not equal

. . . . E . . . . E . . . . . .

pair of equal letters

$$I_c = \frac{\sum_{\alpha=A}^{Z} N_\alpha (N_\alpha - 1)/2}{N(N-1)/2}$$

where N = length of C
and N_α = # of α in C

Note I_c is the same if you apply a Caesar OR Monalphabetic substitution for Enlish expect $I_c \approx 0.065$

Say that my cyphertext is grouped into p blocks each with the same monoalphabetic substitution: Vigenere

N = total letters in cyphertext

M = total letters in each block

$N = Mp$ $\quad$ p = period

$$I_c = \frac{\sum_{\alpha=A}^{Z} N_\alpha (N_\alpha - 1)}{N(N-1)}$$

$$= \frac{\sum_{\alpha=A}^{Z} \left(\sum_{i=1}^{p} M_\alpha^{(i)}\right)\left(\sum_{j=1}^{p} M_\alpha^{(j)} - 1\right)}{N(N-1)}$$

$N_\alpha = \sum_{i=1}^{p} M_\alpha^{(i)}$ $\quad$ $M_\alpha^{(i)}$ = # of α in the i^{th} block

N_α = # of α in cyphertext

$$\sum_{\alpha=A}^{Z} M_\alpha^{(i)} = M \quad \text{for each } i$$

$M_\alpha^{(i)} = M \cdot p_\alpha^{(i)}$ where $p_\alpha^{(i)}$ is the probability
but it should be approximately
a permutation of English stats

$$M = \frac{N}{P} \quad \text{OR} \quad Mp = N$$

$$\sum_{\alpha=A}^{Z} \left(M_\alpha^{(i)}\right)^2 = M^2 \sum_{\alpha=A}^{Z} \left(p_\alpha^{(i)}\right)^2 \cong M^2 \sum_{\alpha=A}^{Z} \left(p_\alpha^{eng}\right)^2$$

$$\cong .065 \frac{N^2}{p^2}$$

$$i \neq j \quad \sum_{\alpha=A}^{Z} M_\alpha^{(i)} \cdot M_\alpha^{(j)} = M^2 \sum_{\alpha=A}^{Z} p_\alpha^{(i)} \cdot p_\alpha^{(j)} \cong M^2 \frac{1}{26}$$

$$\cong \frac{N^2}{p^2}(.038)$$

$$\frac{\displaystyle\sum_{\alpha=A}^{Z} \left(\sum_{i=1}^{p} M_\alpha^{(i)}\right)\left(\sum_{j=1}^{p} M_\alpha^{(j)} - 1\right)}{N(N-1)}$$

$$= \frac{\displaystyle\sum_{\alpha=A}^{Z} \left(\sum_{i=1}^{p} \left(M_\alpha^{(i)}\right)^2 + 2 \cdot \sum_{1 \le i < j \le p} M_\alpha^{(i)} M_\alpha^{(j)} - \sum_{i=1}^{p} M_\alpha^{(i)}\right)}{N(N-1)}$$

$$= \frac{\sum_{i=1}^{p} \sum_{\alpha=A}^{Z} \left(M_\alpha^{(i)}\right)^2 + 2 \sum_{i<j} \sum_{\alpha=A}^{Z} M_\alpha^{(i)} M_\alpha^{(j)} - \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)}}{N(N-1)}$$

$$\cong \frac{\sum_{i=1}^{p} \frac{N^2}{p^2}(.065) + 2 \sum_{1 \le i < j \le p} \frac{N^2}{p^2}(.038) - N}{N(N-1)}$$

$$= \frac{\frac{N^2}{p}(.065) + 2 \cdot \frac{1}{p^2} \frac{p(p-1)}{2} N^2 (.038) - N}{N(N-1}$$

$$= I_c$$

Day 12

1. Reading week next week

   3-4 more assignments due Feb 28

   Exam 2 Feb 29

2. Index of Coincidence

3. An inequality → break rectangular transposition

4. Break monoalphabetic subsitution

A'LA'IH, DO'NEH'LINI, A'LA'IH, DO'NEH'LINI, A'LA'IH, DO'NEH'LINI, DO'NEH'LINI, DO'NEH'LINI, DO'NEH'LINI, A'LA'IH, DO'NEH'LINI, A'LA'IH, DO'NEH'LINI, A'LAIH, A'LA'IH, DO'NEH'LINI, . . .
FOR ADDED SECURITY, AFTER WE ENCRYPT THE DATA STREAM, WE SEND IT THROUGH OUR NAVAJO CODE TALKER.
...IS HE JUST USING NAVAJO WORDS FOR "ZERO" AND "ONE"?
WHOA, HEY, KEEP YOUR VOICE DOWN!

The index of coincidence is defined as

$$I_c = \frac{\text{number of pairs of equal letters in ciphertext}}{\text{the total number of pairs of letters}}$$

That is if we set

- $N_\alpha =$ the number of occurrences of the letter α in the cyphertext
- $$D_c = \sum_{\alpha=A}^{Z} \binom{N_\alpha}{2}$$

  D_c represents the number of pairs of equal letters in the cyphertext.
- then $I_c = \frac{D_c}{\binom{N}{2}}$
- where $N =$ the number of letters in the cyphertext

The index of coincidence is invariant under monoalphabetic cyphers and we estimate under this condition that $N_\alpha = N * p_{\sigma(\alpha)}$ for some permutation of the alphabet σ and so

$$\begin{aligned} I_c &= \frac{\sum_{\alpha=A}^{Z}(N_\alpha^2 - N_\alpha)}{N(N-1)} \\ &\approx \frac{N^2(\sum_{\alpha=A}^{Z} p_\alpha^2) - N}{N(N-1)} \\ &= \frac{N(.065) - 1}{N-1} \\ &\approx .065 \end{aligned}$$

If the cyphertext was obtained from a polyalphabetic cipher then the index of coincidence can also be used to estimate the period of the cipher.
Let p be the period of the cyphertext and place the letters of the cyphertext into groups of p so that the letters in the i^{th} position of the groups are all encrypted with the same key.

- Let $M_\alpha^{(i)}$ equal the number of occurrences of the letter α that appears in the i^{th} positions in the groups.
- If there are M groups of p, then $\sum_{\alpha=A}^{Z} M_\alpha^{(i)} = M$
- We also have $N = Mp$
- Also we can estimate that $M_\alpha^{(i)} \approx Mp_{\sigma(\alpha)}$ (again for some permutation for the alphabet σ)

Now, we calculate that

$$\begin{aligned}
2D_c &= \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)}(M_\alpha^{(i)} - 1) + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)} M_\alpha^{(j)} \\
&\approx M^2 p(.065) - pM + M^2(.038)p(p-1) \\
&= \frac{N^2}{p}(.027) - N + N^2(.038)
\end{aligned}$$

Note that because $I_c = \frac{D_c}{\binom{N}{2}}$, we have that

$$2D_c = N(N-1)I_c.$$

And we just derived that

$$2D_c \approx \frac{N^2}{p}(.027) - N + N^2(.038)$$

Therefore,

$$N(N-1)I_c \approx \frac{N^2}{p}(.027) - N + N^2(.038)$$

$$(N-1)I_c \approx \frac{N}{p}(.027) - 1 + N(.038)$$

$$(N-1)I_c + 1 \approx \frac{N}{p}(.027) + N(.038)$$

$$(N-1)I_c + 1 - N(.038) \approx \frac{N}{p}(.027)$$

$$p((N-1)I_c + 1 - N(.038)) \approx N(.027)$$

$$p \approx \frac{N(.027)}{(N-1)I_c + 1 - N(.038)}$$

Lets see how accurate this is (it gives an approximation to the period, not the actual period) with text that contains about 21K letters. We use the same text and vigenere cipher with period 3 through 6.

`indcoin < plaintext`
Index of coincidence : 0.063616
Estimate of the period : 1.052158

- `indcoin < cyphertextvig3`
  Index of coincidence : 0.044720
  Estimate of the period : 3.990527
- `indcoin < cyphertextvig4`
  Index of coincidence : 0.042903
  Estimate of the period : 5.455495
- `indcoin < cyphertextvig5`
  Index of coincidence : 0.042236
  Estimate of the period : 6.304608
- `indcoin < cyphertextvig6`
  Index of coincidence : 0.041899
  Estimate of the period : 6.842702

Lets do another experiment with less letters (precisely 3183 letters).

```
indcoin < plaintext
Index of coincidence : 0.069377
Estimate of the period : 0.852563
```

-
  ```
  indcoin < cyphertextvig3
  Index of coincidence : 0.045386
  Estimate of the period : 3.512710
  ```

-
  ```
  indcoin < cyphertextvig4
  Index of coincidence : 0.045457
  Estimate of the period : 3.480884
  ```

-
  ```
  indcoin < cyphertextvig5
  Index of coincidence : 0.045034
  Estimate of the period : 3.681678
  ```

-
  ```
  indcoin < cyphertextvig6
  Index of coincidence : 0.043903
  Estimate of the period : 4.352677
  ```

Lets do another experiment with less letters (precisely 14590 letters).

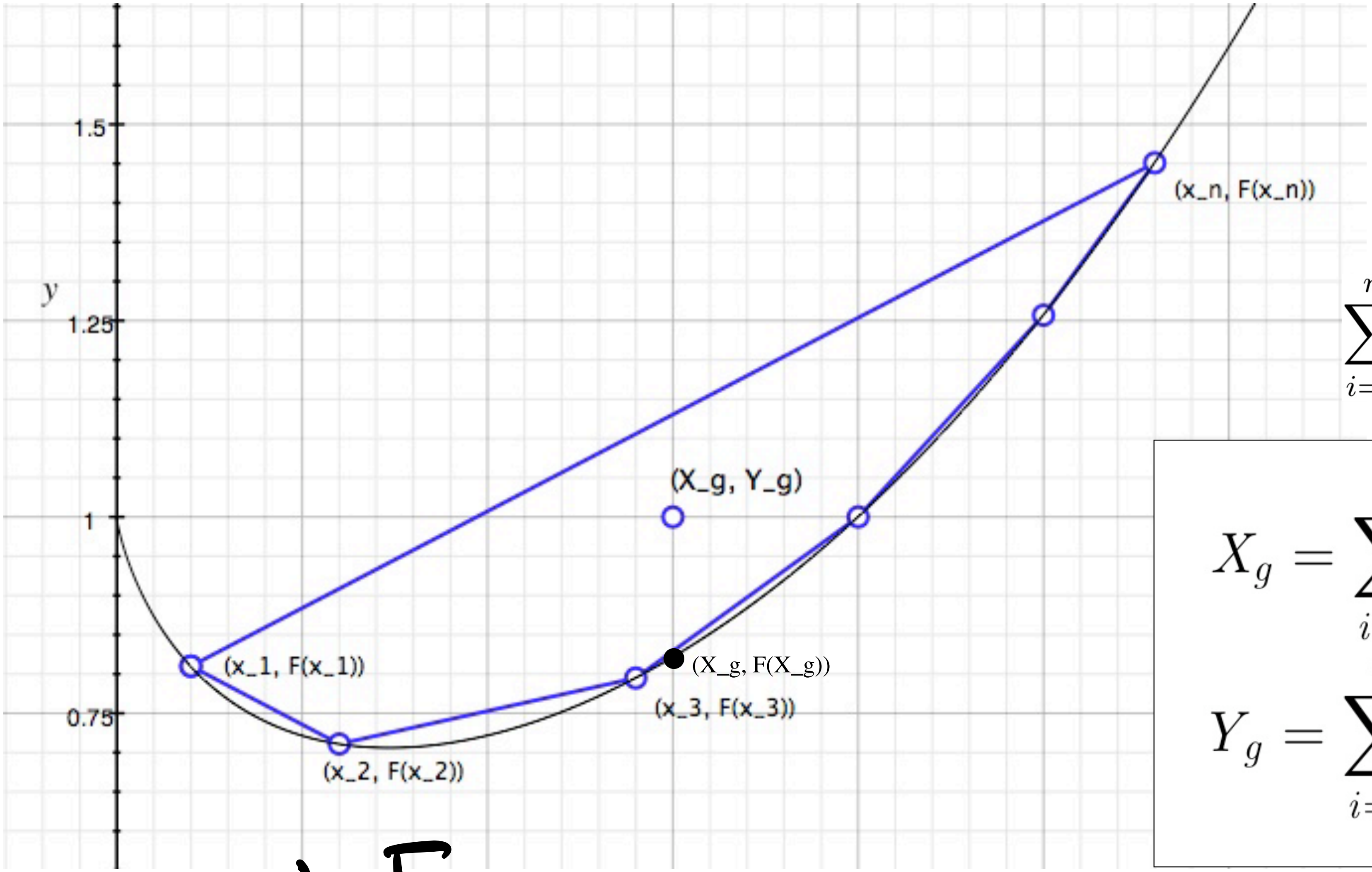
```
indcoin < plaintext
Index of coincidence : 0.064586
Estimate of the period : 1.013137
```

-
  ```
  indcoin < cyphertextvig3
  Index of coincidence : 0.045976
  Estimate of the period : 3.357689
  ```

-
  ```
  indcoin < cyphertextvig4
  Index of coincidence : 0.042790
  Estimate of the period : 5.560689
  ```

-
  ```
  indcoin < cyphertextvig5
  Index of coincidence : 0.041953
  Estimate of the period : 6.718174
  ```

-
  ```
  indcoin < cyphertextvig6
  Index of coincidence : 0.041019
  Estimate of the period : 8.752510
  ```



$$m_i \geq 0$$

$$\sum_{i=1}^{n} m_i = 1$$

$$X_g = \sum_{i=1}^{n} m_i x_i$$

$$Y_g = \sum_{i=1}^{n} m_i F(x_i)$$

1. F is concave up
2. weighted average of points lies in polygon

$$F\left(\sum_{i=1}^{n} m_i x_i\right) = F(X_g) \leq Y_g = \sum_{i=1}^{n} m_i F(x_i)$$

$$\boxed{F(x) = x \log x}$$

$$\boxed{F\left(\sum_{i=1}^{n} m_i x_i\right) \leq \sum_{i=1}^{n} m_i F(x_i)}$$

$$\boxed{x_i = p_i/q_i} \qquad \boxed{m_i = q_i} \qquad 1 \cdot \log(1) = 0$$

$$\sum_{i=1}^{n} \cancel{q_i}\left(p_i/\cancel{q_i}\right) = \sum_{i=1}^{n} p_i = 1 \qquad \begin{array}{c} \| \\ \left(\sum_{i=1}^{n} x_i m_i\right) \cdot \log\left(\sum_{i=1}^{n} x_i m_i\right) \end{array}$$

$$F\left(\sum_{i=1}^{n} m_i \cdot x_i\right) = 0 \leq \sum_{i=1}^{n} q_i\left(p_i/q_i \log\left(p_i/q_i\right)\right)$$

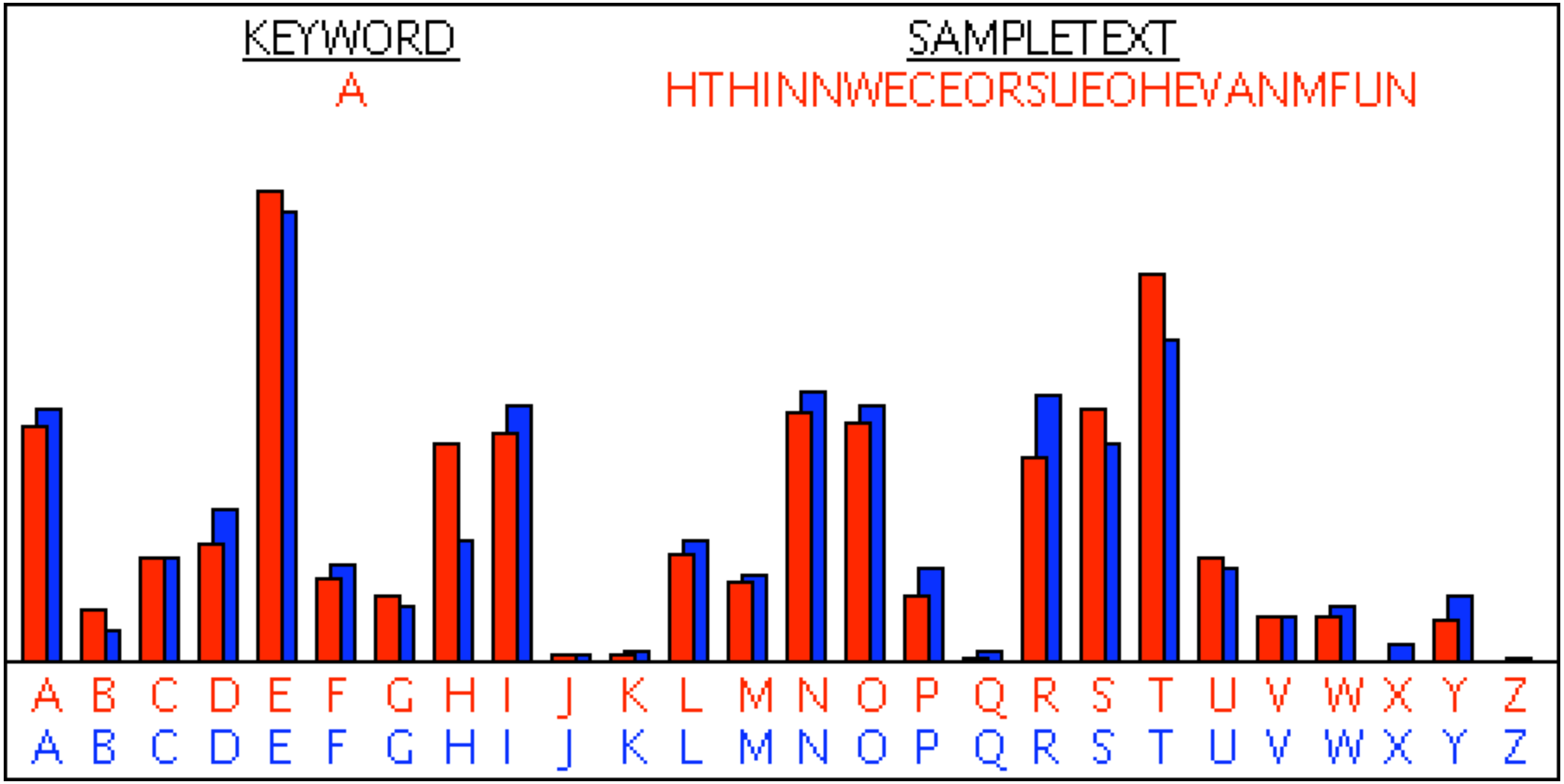
$$0 \leq \sum_{i=1}^{n} p_i \log(p_i/q_i)$$

$$= \sum_{i=1}^{n} p_i(\log p_i - \log(q_i))$$

$$= \sum_{i=1}^{n} p_i \log(p_i) - \sum_{i=1}^{n} p_i \log(q_i)$$

$$\sum_{i=1}^{n} p_i \log(q_i) \leq \sum_{i=1}^{n} p_i \log(p_i)$$

- Consider English text which is transformed by the rectangular transposition cipher.
- If we look at the single letter statistics they continue to look like English and hence they tell us nothing about how to recover the plaintext from the cyphertext.



Guess at the period p of the cyphertext.
Let a and b represent letters and for $1 \leq i, j \leq p$

$N_{a,b}^{(i,j)}$ = the number of pairs of letters equal to a, b where a is in the i^{th} position in the blocks of p and b is in the j^{th} position.
Let N = the number of letters in the cyphertext divided by p
Let $p_{a,b}$ = the probability that ab occurs in English.

- If the letters in the j^{th} position in the cyphertext are supposed to follow the letters in the i^{th} position, then the transition is 'good' and we should expect to see $N_{a,b}^{(i,j)}$ to be roughly equal to $N * p_{a,b}$.
- If the letters in the j^{th} position in the cyphertext are not supoosed to follow the letters in the i^{th} position, then the transition is 'not good' and and we should expect to see $N_{a,b}^{(i,j)} = N * q_{a,b}$ for some other probabilities $q_{a,b}$.

Now calculate

$$\sum_{a,b=A}^{Z} p_{a,b} \log N_{a,b}^{(i,j)}$$

- If the $i \rightarrow j$ transition is 'good' then $N_{a,b}^{(i,j)} \approx N * p_{a,b}$ and

$$\sum_{a,b=A}^{Z} p_{a,b} \log \underset{N\cdot p_{ab}}{N_{a,b}^{(i,j)}} \approx \log N + \sum_{a,b=A}^{Z} p_{a,b} \log p_{a,b}$$

- If the $i \rightarrow j$ transition is 'not good' then $N_{a,b}^{(i,j)} \approx N * q_{a,b}$ and

$$\sum_{a,b=A}^{Z} p_{a,b} \log N_{a,b}^{(i,j)} \approx \log N + \sum_{a,b=A}^{Z} p_{a,b} \log q_{a,b}$$

- Recall that we derived the inequality

$$\sum_{i} p_i \log q_i \leq \sum_{i} p_i \log p_i .$$

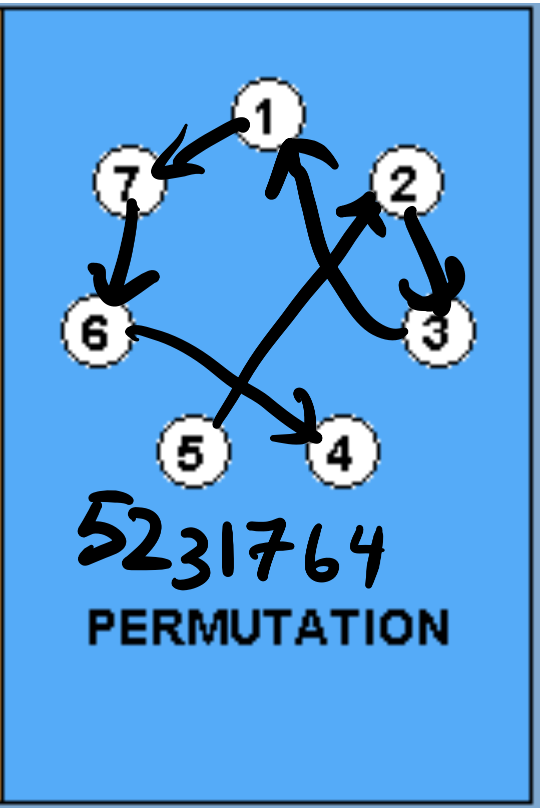
Example of table of $\sum_{a,b=A}^{Z} p_{a,b} \log N_{a,b}^{(i,j)}$ with correct period

english

counts in ith and jth graph

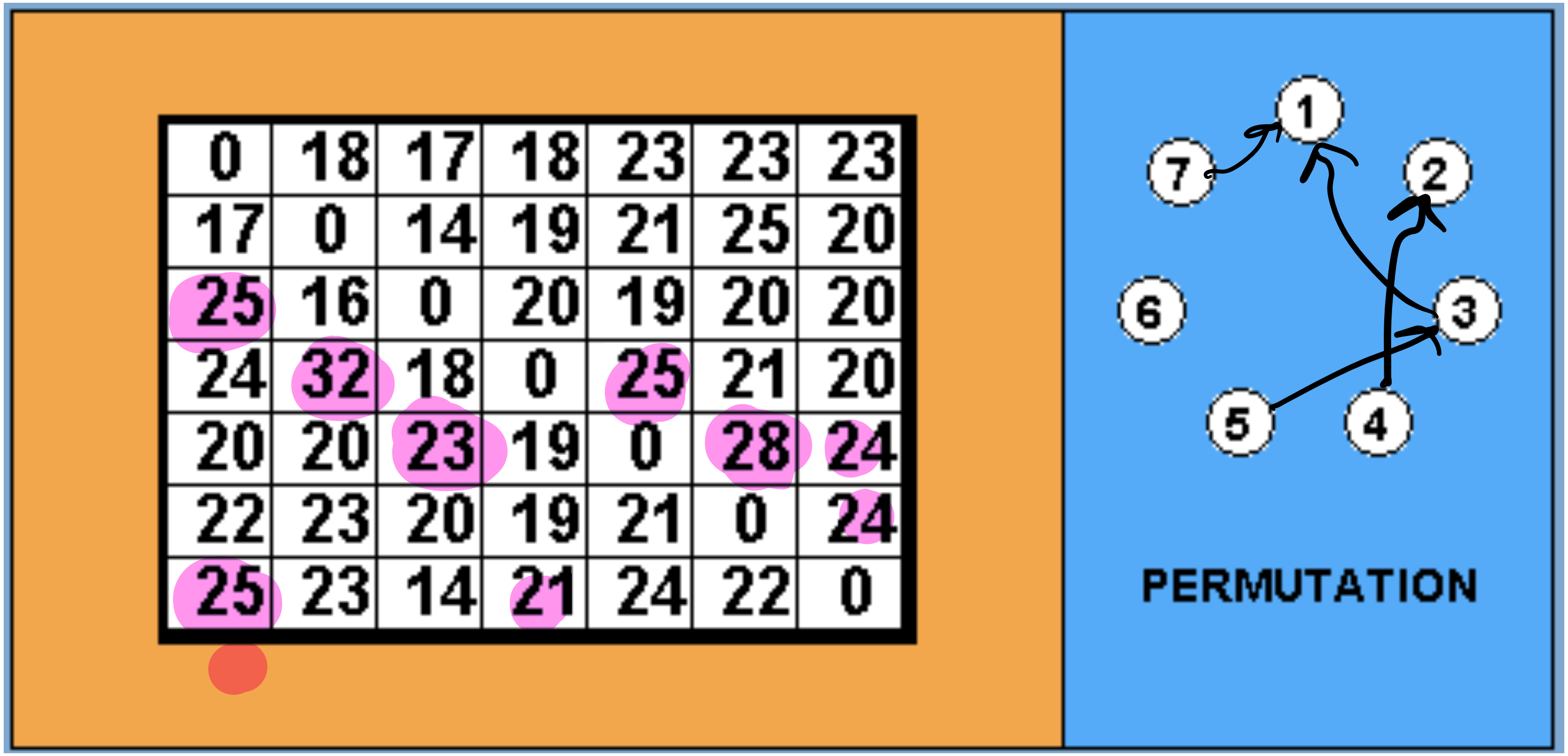
We should see high values in each row and column except one row (the last position of the permutation) and one column (the first position of the permutation).

0	26	31	34	26	20	36
18	0	53	32	24	32	27
39	26	0	26	24	29	18
27	19	33	0	26	28	22
24	39	29	29	0	26	21
21	28	28	44	27	0	23
29	26	28	23	25	43	0



Example of table of $\sum_{a,b=A}^{Z} p_{a,b} \log N_{a,b}^{(i,j)}$ with incorrect period

We should see high and low values evenly distributed in the table.



0	18	17	18	23	23	23
17	0	14	19	21	25	20
25	16	0	20	19	20	20
24	32	18	0	25	21	20
20	20	23	19	0	28	24
22	23	20	19	21	0	24
25	23	14	21	24	22	0

1
2
7
6
3
5
4
PERMUTATION

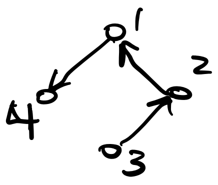
Exercises:

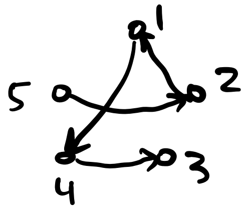
1. Decrypt the following message that was encoded using rectangular transpostion. The matrices provided should be enough to recover the period and key.

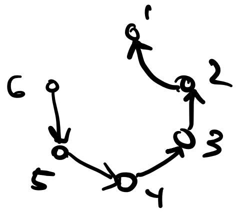
YCWOT NNASA SD|FON YFIEC UHEA|U SALET
OYELH F|OUHU BDHNE TAI|F OTEHH WOWIE

$$\begin{bmatrix} 0 & 66 & 57 & 66 \\ 70 & 0 & 55 & 48 \\ 56 & 80 & 0 & 63 \\ 60 & 57 & 64 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 47 & 44 & 73 & 51 \\ 72 & 0 & 53 & 48 & 54 \\ 43 & 53 & 0 & 52 & 41 \\ 51 & 42 & 72 & 0 & 47 \\ 42 & 66 & 46 & 43 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 47 & 43 & 51 & 41 & 51 \\ 50 & 0 & 43 & 36 & 47 & 41 \\ 38 & 59 & 0 & 47 & 36 & 55 \\ 42 & 38 & 54 & 0 & 40 & 40 \\ 39 & 58 & 41 & 56 & 0 & 50 \\ 37 & 40 & 45 & 41 & 50 & 0 \end{bmatrix}$$

2. What was the permutation used to encrypt the message in the previous question?







3214

OYYT
YFCA
EIWI
LEOT
HCTF

52143

IFYOU
FOCUS
ONWHA
TYOUL

654321