Lecture 7
October 11, 2023
Substitution Ciphers

• Change characters in plaintext to produce ciphertext

• Example (Caesar cipher)
  • Plaintext is HELLO WORLD
  • Change each letter to the third letter following it (X goes to A, Y to B, Z to C)
    • Key is 3, usually written as letter ‘D’
  • Ciphertext is KHOOR ZRUOG
Attacking the Cipher

• Exhaustive search
  • If the key space is small enough, try all possible keys until you find the right one
  • Caesar cipher has 26 possible keys

• Statistical analysis
  • Compare to 1-gram model of English
Statistical Attack

• Compute frequency of each letter in ciphertext:
  
<table>
<thead>
<tr>
<th>Letter</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.1</td>
</tr>
<tr>
<td>H</td>
<td>0.1</td>
</tr>
<tr>
<td>K</td>
<td>0.1</td>
</tr>
<tr>
<td>O</td>
<td>0.3</td>
</tr>
<tr>
<td>R</td>
<td>0.2</td>
</tr>
<tr>
<td>U</td>
<td>0.1</td>
</tr>
<tr>
<td>Z</td>
<td>0.1</td>
</tr>
</tbody>
</table>

• Apply 1-gram model of English
  • Frequency of characters (1-grams) in English is on next slide
# Character Frequencies

<table>
<thead>
<tr>
<th>Character</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.07984</td>
</tr>
<tr>
<td>b</td>
<td>0.01511</td>
</tr>
<tr>
<td>c</td>
<td>0.02504</td>
</tr>
<tr>
<td>d</td>
<td>0.04260</td>
</tr>
<tr>
<td>e</td>
<td>0.12452</td>
</tr>
<tr>
<td>f</td>
<td>0.02262</td>
</tr>
<tr>
<td>g</td>
<td>0.02013</td>
</tr>
<tr>
<td>h</td>
<td>0.06384</td>
</tr>
<tr>
<td>i</td>
<td>0.07000</td>
</tr>
<tr>
<td>j</td>
<td>0.00131</td>
</tr>
<tr>
<td>k</td>
<td>0.00741</td>
</tr>
<tr>
<td>l</td>
<td>0.03961</td>
</tr>
<tr>
<td>m</td>
<td>0.02629</td>
</tr>
<tr>
<td>n</td>
<td>0.06876</td>
</tr>
<tr>
<td>o</td>
<td>0.07691</td>
</tr>
<tr>
<td>p</td>
<td>0.01741</td>
</tr>
<tr>
<td>q</td>
<td>0.00107</td>
</tr>
<tr>
<td>r</td>
<td>0.05912</td>
</tr>
<tr>
<td>s</td>
<td>0.06333</td>
</tr>
<tr>
<td>t</td>
<td>0.09058</td>
</tr>
<tr>
<td>u</td>
<td>0.02844</td>
</tr>
<tr>
<td>v</td>
<td>0.01056</td>
</tr>
<tr>
<td>w</td>
<td>0.02304</td>
</tr>
<tr>
<td>x</td>
<td>0.00159</td>
</tr>
<tr>
<td>y</td>
<td>0.02028</td>
</tr>
<tr>
<td>z</td>
<td>0.00057</td>
</tr>
</tbody>
</table>
Statistical Analysis

- $f(c)$ frequency of character $c$ in ciphertext
- $\varphi(i)$ correlation of frequency of letters in ciphertext with corresponding letters in English, assuming key is $i$
  - $\varphi(i) = \sum_{0 \leq c \leq 25} f(c)p(c - i)$ so here,
    - $\varphi(i) = 0.1 \ p(6 - i) + 0.1 \ p(7 - i) + 0.1 \ p(10 - i) + 0.3 \ p(14 - i) + 0.2 \ p(17 - i) + 0.1 \ p(20 - i) + 0.1 \ p(25 - i)$
  - $p(x)$ is frequency of character $x$ in English
Correlation: $\varphi(i)$ for $0 \leq i \leq 25$

<table>
<thead>
<tr>
<th></th>
<th>$\varphi(i)$</th>
<th></th>
<th>$\varphi(i)$</th>
<th></th>
<th>$\varphi(i)$</th>
<th></th>
<th>$\varphi(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0469</td>
<td>7</td>
<td>0.0461</td>
<td>13</td>
<td>0.0505</td>
<td>19</td>
<td>0.0312</td>
</tr>
<tr>
<td>1</td>
<td>0.0393</td>
<td>8</td>
<td>0.0194</td>
<td>14</td>
<td>0.0561</td>
<td>20</td>
<td>0.0287</td>
</tr>
<tr>
<td>2</td>
<td>0.0396</td>
<td>9</td>
<td>0.0286</td>
<td>15</td>
<td>0.0215</td>
<td>21</td>
<td>0.0526</td>
</tr>
<tr>
<td>3</td>
<td>0.0586</td>
<td>10</td>
<td>0.0631</td>
<td>16</td>
<td>0.0306</td>
<td>22</td>
<td>0.0398</td>
</tr>
<tr>
<td>4</td>
<td>0.0259</td>
<td>11</td>
<td>0.0280</td>
<td>17</td>
<td>0.0386</td>
<td>23</td>
<td>0.0338</td>
</tr>
<tr>
<td>5</td>
<td>0.0165</td>
<td>12</td>
<td>0.0318</td>
<td>18</td>
<td>0.0317</td>
<td>24</td>
<td>0.0320</td>
</tr>
<tr>
<td>6</td>
<td>0.0676</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>0.0443</td>
</tr>
</tbody>
</table>
The Result

• Most probable keys, based on $\varphi$:
  
  • $i = 6$, $\varphi(i) = 0.0676$
    
    • plaintext EBIIL TLOLA
  
  • $i = 10$, $\varphi(i) = 0.0631$
    
    • plaintext AXEEH PHKEW
  
  • $i = 14$, $\varphi(i) = 0.0561$
    
    • plaintext WTAAD LDGAS
  
  • $i = 3$, $\varphi(i) = 0.0586$
    
    • plaintext HELLO WORLD

• Only English phrase is for $i = 3$
  
  • That’s the key (3 or ‘D’)
Caesar’s Problem

• Key is too short
  • Can be found by exhaustive search
  • Statistical frequencies not concealed well
    • They look too much like regular English letters

• So make it longer
  • Multiple letters in key
  • Idea is to smooth the statistical frequencies to make cryptanalysis harder
Vigenère Cipher

• Like Caesar cipher, but use a phrase
  • So it’s effectively multiple Caesar ciphers

• Example
  • Message A LIMERICK PACKS LAUGHS ANATOMICAL
  • Key BENCH
  • Encipher using Caesar cipher for each letter:
    
    | key       | BENCHBENCHBENCHBENCHBENCHBENCHBENCH |
    | plain     | ALIMERICKPACKSLAUGHSANATOMICAL       |
    | cipher    | BPVOLSMPMWBGXUSBYTJZBRNVVNMPCS       |
The Vigenère Tableau

Vigenère: rows are keys, columns are plaintext
message: HELLOWORLD
key: ECSECSECSE
cipher: LGDPQOSTDH

Beaufort: left letters are plaintext; trace inwards until you find the key letter; the column is the plaintext
message: HELLOWORLD
key: ECSECSECSE
cipher: XYHTOWQLHB

Variant Beaufort: left letters are keys; trace inwards until you find the plaintext letter; the column is the ciphertext
message: HELLOWORLD
key: ECSECSECSE
cipher: DCTHMEKPTZ
Relevant Parts of Tableau

- Tableau shown has relevant rows, columns only
  - Columns correspond to letters from the message
  - Rows correspond to letters from the message
- Example encipherments:
  - key R, letter B: ciphered letter is where row R and column B meet, giving “S”
  - Key L, letter H: ciphered letter is where row L and column H meet, giving “S”
Useful Terms

• *period*: length of key
  • In earlier example, period is 3

• *tableau*: table used to encipher and decipher
  • Vigenère cipher has key letters on top, plaintext letters on the left

• *polyalphabetic*: the key has several different letters
  • Caesar cipher is monoalphabetic
Attacking the Cipher

• Approach
  • Establish period; call it $n$
  • Break message into $n$ parts, each part being enciphered using the same key letter
  • Solve each part; you can leverage one part from another

• We will show each step
The Target Cipher

• We want to break this cipher:
  ADQYS MIUSB OXKKT MIBHK IZOOO EQOOG IFBAG KAUMF
  VVTAA CIDTW MOCIO EQOOG BMBFV ZGGWP CIEKQ HSNEW
  VECNE DLAAV RWKXS VNSVP HCEUT QOIOF MEGJS WTPCH
  AJMOC HIUIX
Establish Period

- Kaskski: *repetitions in the ciphertext occur when characters of the key appear over the same characters in the plaintext*

- Example:

  key  VIGVIGVIGVIGVIGV
  plain THEBOYHASTHEBALL
  cipher OPKW WECIY OPKWIRG

  Note the key and plaintext line up over the repetitions (underlined). As distance between repetitions is 9 (dotted line), the period is a factor of 9 (that is, 1, 3, or 9)
# Repetitions in Example

<table>
<thead>
<tr>
<th>Letters</th>
<th>Start</th>
<th>End</th>
<th>Gap Length</th>
<th>Gap Length Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>OEQOOG</td>
<td>24</td>
<td>54</td>
<td>30</td>
<td>2, 3, 5</td>
</tr>
<tr>
<td>MOC</td>
<td>50</td>
<td>122</td>
<td>72</td>
<td>2, 2, 2, 3, 3</td>
</tr>
</tbody>
</table>
Estimate of Period

• OEQOOG is probably not a coincidence
  • It’s too long for that
  • Period may be 1, 2, 3, 5, 6, 10, 15, or 30

• MOC is also probably not a coincidence
  • Period may be 1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, or 72

• Period of 2 or 3 is probably too short (but maybe not)

• Begin with period of 6
Check on Period

• Index of coincidence is probability that two randomly chosen letters from ciphertext will be the same

• Tabulated for different periods:

<table>
<thead>
<tr>
<th></th>
<th>0.0660</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0520</td>
</tr>
<tr>
<td>2</td>
<td>0.0473</td>
</tr>
<tr>
<td>3</td>
<td>0.0427</td>
</tr>
</tbody>
</table>
Compute IC for an Alphabet

\[ \text{IC} = \left[ \frac{n(n-1)}{n} \right] \sum_{0 \leq i \leq 25} [F_i (F_i - 1)] \]

- where \( n \) is length of ciphertext and \( F_i \) the number of times character \( i \) occurs in ciphertext

- For the given ciphertext, IC = 0.0433
  - Indicates a key of length 5 or 6
  - A statistical measure, so it can be in error, but it agrees with the previous estimate (which was 6)
Splitting Into Alphabets

alphabet 1: AIKHOIATTOBGEERNEOSAI
alphabet 2: DUKKEFUAWEMGKWDSUFWJU
alphabet 3: QSTIQBMAMQBWQVLKVTMTMI
alphabet 4: YBMZOAFCOOFPHEAXPQEPOX
alphabet 5: SOIOOGVICOVCSVASHOGCC
alphabet 6: MXBOGKVDIGZINNVVCIJJHH

• ICs (#1, 0.0692; #2, 0.0779; #3, 0.0779; #4, 0.0562; #5, 0.1238; #6, 0.0429) indicate all alphabets have period 1, except #4 (between 1 and 2) and #6 (between 5 and 6); assume statistical variance
## Frequency Examination

|   | 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 |
| 1 | 3 | 1 | 0 | 0 | 4 | 0 | 1 | 1 | 3 | 0 | 1 | 0 | 0 | 1 | 3 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 2 | 2 | 2 | 1 | 0 | 0 | 1 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 4 | 0 | 0 | 0 |
| 3 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 4 | 0 | 0 | 0 | 4 | 0 | 1 | 3 | 0 | 2 | 1 | 0 | 0 | 0 |
| 4 | 2 | 1 | 1 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 |
| 5 | 1 | 0 | 5 | 0 | 0 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 3 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 1 | 1 | 0 | 0 | 2 | 2 | 3 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 1 |

The last row has general letter frequencies (H high, M medium, L low)
Begin Decryption

• First matches characteristics of unshifted alphabet
• Third matches if I shifted to A
• Sixth matches if V shifted to A
• Substitute into ciphertext (bold are substitutions)

ADIYS RIUKB OCKKL MIGHK AZOTO EIOOL IFTAG
PAUEF VATAS CIITW EOCNO EIOOL BMTFV EGGOP
CNEKI HSSEW NECSE DDAAA RWCXS ANSNP HHEUL
QONOF EEGOS WLPQM AJEOC MIUAX
Look For Clues

• **AJE** in last line suggests “are”, meaning second alphabet maps A into S:

  ALIYS  RICKB  OCKSL  MIGHS  AZOTO  MIOOL  INTAG
  PACEF  VATIS  CIITE  EOCNO  MIOOL  BUTFV  EGOOP
  CNESI  HSEE  NECSE  LDAAA  RECXS  ANANP  HHECL
  QONON  EEGOS  ELPCM  AREOC  MICAX
Next Alphabet

• **MICAX** in last line suggests “mical” (a common ending for an adjective), meaning fourth alphabet maps $O$ into $A$:

  ALIMS RICKP OCKSL AIGHS ANOTO MICOL INTOG
  PACET VATIS QITE ECCKNO MICOL BUTTV EGOOD
  CNESI VSSEE NSCSE LDOAA RECLS ANAND HHECL
  EONON ESGOS ELDCM ARECC MICAL
Got It!

• QI means that U maps into I, as Q is always followed by U:

ALIME RICKP ACKSL AUGHS ANATO MICAL INTOS
PACET HATIS QUITE ECONOMIC MICAL BUTTH EGOOD
ONESI VESEE NSOSE LDOMA RECLE ANAND THECL
EANON ESSOS ELDOM ARECO MICAL
One-Time Pad

• A Vigenère cipher with a random key at least as long as the message
  • Provably unbreakable
  • Why? Look at ciphertext DXQR. Equally likely to correspond to plaintext DOIT (key AJIY) and to plaintext DONT (key AJDY) and any other 4 letters

• Warning: keys must be random, or you can attack the cipher by trying to regenerate the key
  • Approximations, such as using pseudorandom number generators to generate keys, are not random
Overview of the DES

• A block cipher:
  • encrypts blocks of 64 bits using a 64 bit key
  • outputs 64 bits of ciphertext

• A product cipher
  • basic unit is the bit
  • performs both substitution and transposition (permutation) on the bits

• Cipher consists of 16 rounds (iterations) each with a 48 bit round key generated from the user-supplied key
Structure of the DES

• Input is first permuted, then split into left half (L) and right half (R), each 32 bits

• Round begins; R and round key run through function $f$, then xor’ed with L
  • $f$ expands R to 48 bits, xors with round key, and then each 6 bits of this are run through S-boxes (substitution boxes), each of which gives 4 bits of output
  • Those 32 bits are permuted and this is the output of $f$

• R and L swapped, ending the round
  • Swapping does not occur in the last round

• After last round, L and R combined, permuted, forming DES output
Controversy

• Considered too weak
  • Diffie, Hellman said in a few years technology would allow DES to be broken in days
    • Design using 1999 technology published

• Design decisions not public
  • S-boxes may have backdoors
Undesirable Properties

• 4 weak keys
  • They are their own inverses
• 12 semi-weak keys
  • Each has another semi-weak key as inverse
• Complementation property, where $x'$ is the bitwise complement of $x$
  • is the bitwise complement of $x$
    • $\text{DES}_k(m) = c \Rightarrow \text{DES}_k(m') = c'$
• S-boxes exhibit irregular properties
  • Distribution of odd, even numbers non-random
  • Outputs of fourth box depends on input to third box
Differential Cryptanalysis

- A chosen ciphertext attack
  - Requires $2^{47}$ plaintext, ciphertext pairs

- Revealed several properties
  - Small changes in S-boxes reduced the number of pairs needed
  - Making every bit of the round keys independent did not impede attack

- Linear cryptanalysis improves result
  - Requires $2^{43}$ plaintext, ciphertext pairs