Index of Coincidence

CS 5158/6058 Data Security and Privacy

Spring 2018

Instructor: Boyang Wang

Index of Coincidence (IC)

- IC (by William Friedman in 1920):
  - Give a sequence of chars, the <u>probability</u> that two randomly selected chars are <u>identical</u>

- Example: 100 chars in total, 20 `a`
  - Take one `a` from 100 chars, $p_1 = 20/100$
  - Take another `a` from 99 chars, $p_2 = 19/99$
  - Select two `a`: $p_1 * p_2$

Index of Coincidence (IC)

- N is the total number of chars
- n_a is the number of char **`a`**
- The probability that two randomly selected chars are both **`a`** is $$\frac{n_a}{N} \cdot \frac{n_a - 1}{N - 1} \approx \left(\frac{n_a}{N}\right)^2$$

- 26 unique chars in total

$$\mathrm{IC} \approx \left(\frac{n_a}{N}\right)^2 + \cdots + \left(\frac{n_z}{N}\right)^2 = \sum_{i=0}^{25} \left(\frac{n_i}{N}\right)^2$$

Index of Coincidence (IC)

- $p_i = n_i/N$ is known for all the 26 English chars

$$\mathrm{IC}_{English} \approx 0.065$$

- <u>Practice:</u> If 26 chars uniformly distributed, IC?
  - $n_i/N = 1/26$, for all i in [0,25]

$$\mathrm{IC}_{uni} = \sum_{i=0}^{25} \left(\frac{n_i}{N}\right)^2 = ?$$

Index of Coincidence (IC)

- Answer:

$$\mathrm{IC}_{uni} = \sum_{i=0}^{25} \left(\frac{n_i}{N}\right)^2 = 26 \cdot \frac{1}{26^2} = \frac{1}{26} \approx 0.038$$

- IC can <u>automate</u> attacks on shift cipher
  - Original attacks: check “make sense” per key
  - Data is not human-readable all the time
  - Decide which key it is by checking IC

Index of Coincidence (IC)

- Practice: given a sequence of 100 chars, there are 20 **a**, 20 **b**, 10 **c**, 10 **d**, 20 **e**, 20 **f**, what is IC?

$$
\begin{aligned}
\mathrm{IC} &\approx \sum_{i=0}^{25} \left(\frac{n_i}{N}\right)^2 \\
&= (0.2)^2 + (0.2)^2 + (0.1)^2 + (0.1)^2 + (0.2)^2 + (0.2)^2 \\
&= 0.04 \times 4 + 0.01 \times 2 \\
&= 0.18
\end{aligned}
$$

IC in Shift Cipher

- Each char is shifted with (unknown) k positions
  - j = i+k mod 26
  - Same frequency. **`a`** <—> **`E`**, If 10 **`a`**s, then 10 **`E`**s

$$\frac{n_i}{N} = p_i = q_{i+k} = q_j = \frac{N_j}{N}$$

- Does shift cipher change IC?
  - No, order is different, but sum is same

$$\mathrm{IC}_{English} = \sum_{i=0}^{25} \left(\frac{n_i}{N}\right)^2 = \sum_{j=0}^{25} \left(\frac{N_j}{N}\right)^2 = \mathrm{IC}_{Shift}$$

Attack on Shift Cipher

- Compute IC_k for all k in {0, 1, …, 25},

$$\mathrm{IC}_k = \sum_{i=0}^{25} p_i^2 = \sum_{i=0}^{25} p_i \cdot q_{i+k}$$

  - p_i is frequency in plaintext, q_{i+k} is frequency in ciphertext, if k is correct, then $p_i = q_{i+k}$
- If IC_k is very close to $IC_{English}$ (0.065)
  - Then k is the key
- Otherwise, IC_k is close to IC_{uni} (0.038)

Attack on Substitution Cipher

```
abcdefghijklmnopqrstuvwxyz
EXAUNDKBMVORQCSFHYGWZLJITP
```

- Does substitution cipher change IC?
  - No, order is different, but sum is same
- Same IC-based attack as the one on shift cipher
  - Key space is much large, i.e., 26!
  - Compute IC_k for all k in {0, 1, …, 26!-1}

Vigenere Cipher

- Preserve frequency, e.g. `r` could map to `F` or `X`
- Several independent instances of Shift Cipher
- Key is a string, e.g. `gouc`

```
plaintext:  drmarccahy
key:        goucgoucgo
ciphertext: JFGCXQWCNM
```

- Key space $|K|=26^t$, t is the length of a key string

Attack on Vigenere Cipher

- Step 1: Decide t, the length of the key.

- Method 1: Kasishi’s Method (1863)
    - Look for repeated sub-strings with a length of 3 or higher in ciphertext
    - Likely were encrypted with a same sub-key
    - The distance between two sub-strings is a multiple of t.

Kasishi's Method

- An example from K&L textbook
  - "**`the`**" is a common English word
  - If key is "**`beads`**", key length is 5

Plaintext:	the	man	and	the	woman	retrieved	the	letter	from	the	post	office
Key:	bea	dsb	ead	sbe	adsbe	adsbeadsb	ead	sbeads	bead	sbe	adsb	eadsbe
Ciphertext:	ULE	PSO	ENG	LII	WREBR	RHLSMEYWE	XHH	DFXTHJ	GVOP	LII	PRKU	SFIADI

  - The distance is 30 (2, 3, 5, 6, 10, 15, 30)
  - Find another sub-string, if distance is 25
  - Key length: t = gcd(25, 30) = 5

Decide Key Length with IC

- Step 1: Decide t, the length of the key.

  $c_1 c_2 c_3 c_4 c_5 c_6 c_7 c_8 c_9 c_{10} c_{11} c_{12} c_{13} c_{14} c_{15} c_{16}$……

- Method 2: Use IC

  $c_1\ c_{1+j}\ c_{1+2j}\ c_{1+3j},$ ……

  - If j = t, sequence is encrypted with <u>shift cipher</u>
  - If j= t, IC_j is approximately $IC_{Shift} = IC_{English}$ (0.065), otherwise it is around IC_{uni} (0.038)
  - Which j is t? Compute IC with different js.

Decide Key Length with IC

```
QPWKA LVRXC QZIKG RBPFA EOMFL  JMSDZ VDHXC XJYEB IMTRQ WNMEA
IZRVK CVKVL XNEIC FZPZC ZZHKM  LVZVZ IZRRQ WDKEC HOSNY XXLSP
MYKVQ XJTDC IOMEE XDQVS RXLRL  KZHOV
```

- An example from Wikipedia
  - Sequence 1 (j=1): `Q P W K A L V R X C ......`
  - Sequence 2 (j=2): `Q   W   A   V   X   ......`
  - Sequence 3 (j=3): `Q     K     V     C ......`

j	1	2	3	4	5	6	7	8	9	10
IC	0.043	0.046	0.040	0.045	0.070	0.038	0.038	0.040	0.045	0.080

Attack on Vigenere Cipher

- Step 2: Once know key length t, it is easy!

- Divide ciphertext into <u>t sequences</u>
    - Sequence 1 c_1 c_{1+t} c_{1+2t} c_{1+3t}, …… is encrypted with shift cipher, attack it with IC, return a key <u>k_1</u>
    - Sequence 2 c_2 c_{2+t} c_{2+2t} c_{2+3t}, …… is encrypted with shift cipher, attack it with IC, return a key <u>k_2</u>

- Vigenere cipher key is <u>$k_1k_2\ldots k_t$</u>

Attack on Vigenere Cipher

- Practice: Assume $IC_{plain} = 0.095$
- Given a sequence of ciphertexts, for j =1, an attacker computes IC_j, j++

j	1	2	3	4	5	6	7	8	9	10
IC	0.043	0.080	0.070	0.090	0.050	0.065	0.038	0.094	0.068	0.080

- What is the key length? 4, since IC_{plain} is 0.095

Attack on Vigenere Cipher

- What if t is much longer (1000, 10000,1000000)?
  - Harder to decide key length
  - Each sub-sequence is shorter, IC may not be very close to 0.065, harder to automate

- This leads to some key idea in One-Time Pad

What We Learn

- Designing secure cipher/encryption is hard

Ciphers	Shift	Substitution	Vigenere
Secure?	No	No	No

- What we learn from Historical Ciphers?
    - Large key size (hard to brute-force)
    - Preserve frequency (deterministic is bad idea)
    - Necessary but not sufficient

An Encryption Scheme

- $k \leftarrow \mathsf{KeyGen}(1^l)$: a probabilistic algorithm that takes a security parameter l, and outputs a key k
- $c \leftarrow \mathsf{Enc}_k(m)$: a deterministic or probabilistic algorithm that takes a key k and a plaintext m as input, and outputs a ciphertext c
- $m \leftarrow \mathsf{Dec}_k(c)$: a deterministic algorithm that takes a key k and a ciphertext c as input, and outputs a plaintext m

Space & Random Variable

- Key space $\mathcal{K}$
- K be a Random Variable for keys
- Pr[K=k]: the probability of a key is k

- Message space $\mathcal{M}$, ciphertext space $\mathcal{C}$
- M is a RV for messages, C is a RV for ciphertexts
- Pr[M=m]: the probability of a message is m
- Pr[C=c]: the probability of a ciphertext is c

Random Variable

- Example: message space $\mathcal{M} = \{\texttt{a}, \texttt{b}, \texttt{c}\}$,
  - `a` (0.5), `b` (0.4), `c` (0.1)
  - *M* is RV for the message space

$$\Pr[M = a] = 0.5$$
$$\Pr[M = b] = 0.4$$
$$\Pr[M = c] = 0.1$$

$$\sum_{m \in \mathcal{M}} \Pr[M = m] = 1$$

- RV *K* and RV *M* are <u>independent</u>

$$\Pr[(K = k) \cap (M = m)] = \Pr[K = k] \cdot \Pr[M = m]$$

Random Variable

- Example: Shift Cipher
  - Message space $\mathcal{M}$ = {`a`, `z`}, `a` (0.7), `z` (0.3),
  - Key space $\mathcal{K}$ = {0, 1, … 25}, Pr[K=k]=1/26, each
  - Ciphertext space C = {`A`, `B`, …, `Z`}

- What is the probability of ciphertext is `B`?
  - Case 1: M=`a` and K=1, `a` + 1 = `B` (mod 26)
  - Case 2: M=`z` and K=2, `z` + 2 = `B` (mod 26)

Random Variable

- Case 1: *M*=`a` and *K*=1

$$\begin{aligned}\Pr[(K=1)\cap(M=a)] &= \Pr[K=1]\cdot\Pr[M=a] \\ &= \frac{1}{26}\cdot 0.7\end{aligned}$$

- Case 2: *M*=`z` and *K*=2

$$\Pr[(K=2)\cap(M=z)] = \frac{1}{26}\cdot 0.3$$

- Probability of *C*=`B`

$$\Pr[C=B]=\Pr[Case1]+\Pr[Case2]=\frac{1}{26}$$

Random Variable

- Practice: Shift Cipher
  - $\mathcal{M}$={`a`, `b`, `c`}, `a` (0.5), `b` (0.3), `c` (0.2)
  - $\mathcal{K}$={0, 1, …, 25}, Pr[K=k]=1/26, each
  - What is the probability of ciphertext is `F`?
- Three cases: 1) M=`a` and K=5; 2) M=`b` and K=4; 3) M=`c` and K=3

$$\Pr[C = F] = 0.5 \cdot \frac{1}{26} + 0.3 \cdot \frac{1}{26} + 0.2 \cdot \frac{1}{26} = \frac{1}{26}$$

Additional Reading

Chapter 1, *Introduction to Modern Cryptography, Drs. J. Katz and Y. Lindell, 2nd edition*