BG ChII Bài 2. Thám mã đối với các hệ mật mã cổ điển.

1. Một vài nhận xét chung.

Như đã trình bày trong chương 1, mục đích của việc thám mã là dựa vào thông tin về bản mật mã có thể thu thập được trên đường truyền tin mà phát hiện lại được bản rõ của thông báo. Vì sơ đồ của hệ mật mã được sử dụng thường khó mà giữ được bí mật, nên ta thường giả thiết thông tin xuất phát của bài toán thám mã là sơ đồ hệ mật mã được sử dụng và bản mật mã của thông báo, nhiệm vụ của thám mã là tìm bản rõ của thông báo đó. Ngoài các thông tin xuất phát đó, tuỳ trường hợp cụ thể, còn có thể có thêm các thông tin bổ sung khác, vì vậy bài toán thám mã được phân thành các loại bài toán khác nhau như: thám mã chỉ dựa vào bản mã, thám

mã khi biết cả bản rõ, thám mã khi có bản rõ được chọn, thám mã khi có bản mã được chọn (xem mục 1.5, chương 1).

Trong tiết này ta sẽ trình bày một vài phương pháp thám mã đối với các hệ mật mã cổ điển mô tả trong tiết trước. Và ta cũng giả thiết các bản rõ cũng như bản mã đều được xây dựng trên bảng ký tự tiếng Anh, và hơn nữa các thông báo là các văn bản tiếng Anh. Như vậy, ta luôn có $P = C = \boldsymbol{Z}_{26}$ hay Z_{26}^{m}, và có thêm thông tin là các bản rõ tuân theo các qui tắc từ pháp và cú pháp của ngôn ngữ tiếng Anh. Đây là một căn cứ quan trọng của các phương pháp thám mã đối với các hệ mật mã cổ điển. Tiếc là việc dùng mật mã để truyền đưa thông tin tiếng Việt không để lại cho ta nhiều tư liệu để nghiên cứu, và những nghiên cứu về từ pháp và cú pháp cũng

chưa cho ta những qui tắc thống kê xác suất đủ tin cậy, nên trong tài liệu này ta chưa trình bày được trên các thí dụ mật mã bằng ngôn ngữ Việt, ta đành tạm mượn các thí dụ bằng văn bản tiếng Anh để minh hoạ, mong được bạn đọc bổ sung sau. Các kết quả chủ yếu được sử dụng nhiều nhất trong thám mã là các qui tắc thống kê tần suất xuất hiện các ký tự hay các bộ đôi, bộ ba,...ký tự liên tiếp trong các văn bản tiếng Anh. Trên cơ sở phân tích các số liệu thống kê từ một lượng rất lớn các văn bản thư từ, sách vở, báo chí, v.v... người ta đã thu được những kết quả mà các tác giả Beker và Piper đã tổng hợp lại như sau:

Phân bố xác suất xuất hiện của các ký tự được sắp xếp theo thứ tự: 1. Ký tự *e* có xác suất xuất hiện cao nhất là 0. 127, 2. Các ký tự *t*, *a*, *o*, *i*, *n*, *s*, *h*, *r* có xác

suất từ 0. 060 đến 0. 090, 3. Các ký tự *d* , *l* có xác suất khoảng 0. 04, 4. Các ký tự *c*, *u*, *m*,*w*, *f*, *g*, *y*, *p*, *b* có xác suất từ 0. 015 đến 0.028, 5. Các ký tự *v*, *k*, *j*, *x*, *q*, *z* có xác suất dưới 0. 01.
Ba mươi bộ đôi ký tự có xác suất xuất hiện cao nhất là (kể từ cao xuống): *th*, *he*, *in*, *er*, *an*, *re*, *ed*, *on*, *es*, *st*, *en*, *at*, *to*, *nt*, *ha*, *nd*, *ou*, *ea*, *ng*, *as*, *or*, *ti*, *is*, *et*, *it*, *ar*, *te*, *se*, *hi*, *of*.
Mười hai bộ ba ký tự có xác suất xuất hiện cao nhất là: *the*, *ing*, *and*, *her*, *ere*, *ent*, *tha*, *nth*, *was*, *eth*, *for*, *dth*.

Sau đây là bảng phân bố xác suất của tất cả các ký tự:

A (0) 0.082 B (1) 0.015 C (2) 0.028
D (3) 0.043 E (4) 0.127
F (5) 0. 022 G (6) 0.020 H (7)

0. 061I (8) 0.070 J (9) 0.002
K (10) 0.008 L (11) 0.040
M (12) 0.024 N (13) 0.067 O
(14) 0.075 P (15) 0.019
Q (16) 0.001 R (17) 0.060 S
(18) 0.063 T (19) 0.091 U
(20) 0.028 V (21) 0.010 W
(22) 0.023 X (23) 0.001 Y
(24) 0.020 Z (25) 0.001.

2. Thám mã đối với mã apphin.

Khoá mã apphin có dạng $K = (a,b)$ với $a, b \in \boldsymbol{Z_{26}}$ và gcd(a,26)=1. Ký tự mã y và ký tự bản rõ x tương ứng có quan hệ

$$y = a.x + b \mod 26.$$

Như vậy, nếu ta biết hai cặp (x, y) khác nhau là ta có được hai phương trình tuyến tính để từ đó tìm ra giá trị hai ẩn số a,b, tức là tìm ra K.

Thí dụ: Ta có bản mật mã:
fmxvedkaphferbndkrxrsrefmorudsdkdvsh vufedkaprkdlyevlrhhrh . Hãy tìm khoá mật mã và bản rõ tương ứng.

Ta thấy trong bản mật mã nói trên, *r* xuất hiện 8 lần, *d* 7 lần, *e*, *k*, *h* mỗi ký tự 5 lần, *f*, *s*, *v* mỗi ký tự 4 lần, v.v...; vậy có thể phán đoán *r* là mã của *e* , *d* là mã của *t*, khi đó có

$$4a + b = 17 \mod 26,$$
$$19a + b = 3 \mod 26,$$

giải ra được $a = 6$, $b = 19$. Vì $\gcd(a, 26) = 2 \neq 1$, nên (a, b) không thể là khoá được, phán đoán trên không đúng. Ta lại thử chon một phán đoán khác: *r* là mã của *e*, *h* là mã của *t* . Khi đó có:

$$4a + b = 17 \mod 26,$$
$$19a + b = 7 \mod 26,$$

ta giải ra được $a = 3$, $b = 5$. Vì $(a, 26) = 1$ nên $K = (3,5)$ có thể là khóa cần tìm. Khi đó phép lập mật mã là $e_K(x) = 3x + 5 \bmod 26$, và phép giải mã tương ứng là $d_K(y) = 9) = 9y - 19 \bmod 26$. Dùng phép giải mã đó cho bản mã ta sẽ được (dưới dạng ký tự) bản rõ là:

algorithmsarequitegeneraldefinitionsofarithmeticprocesses .

Ta có thể kết luận khoá đúng là $K = (3, 5)$ và dòng trên là bản rõ cần tìm.

3. Thám mã đối với mã Vigenère.

Mã Vigenère có thể coi là mã chuyển dịch đối với từng bộ m ký tự. Khoá mã là một bộ $K = (k_1,..., k_m)$ gồm m số nguyên mod 26. Việc thám mã gồm hai bước:

bước thứ nhất xác định độ dài m, bước thứ hai xác định các số k_1,..., k_m.

Có hai phương pháp để xác định độ dài m : phép thử Kasiski và phương pháp dùng chỉ số trùng hợp.

Phép thử Kasiski (đề xuất từ 1863). Phép thử dựa vào nhận xét rằng hai đoạn trùng nhau của bản rõ sẽ được mã hoá thành hai đoạn trùng nhau của bản mã, nếu khoảng cách của chúng trong văn bản rõ (kể từ ký tự đầu của đoạn này đến ký tự đầu của đoạn kia) là bội số của m. Mặt khác, nếu trong bản mã, có hai đoạn trùng nhau và có độ dài khá lớn (≥ 3 chẳng hạn) thì rất có khả năng chúng là mã của hai đoạn trùng nhau trong bản rõ. Vì vậy, ta thử tìm một đoạn mã (có ba ký tự trở lên) xuất hiện nhiều lần trong bản mã, tính

khoảng cách của các lần xuất hiện đó, chẳng hạn được $d_1, d_2..., d_t$; khi đó ta có thể phán đoán $m = d = \gcd(d_1, d_2,..., d_t)$ - ước số chung lớn nhất của $d_1, d_2..., d_t$; hoặc m là ước số của d.

Phương pháp dùng chỉ số trùng hợp: (định nghĩa chỉ số trùng hợp do W.Friedman đưa ra năm 1920).

Định nghĩa 3.1. Cho $x = x_1, x_2... x_n$ là một dãy gồm n ký tự. Xác suất của việc hai phần tử của x trùng nhau được gọi là *chỉ số trùng hợp* của x , ký hiệu là $I_C(x)$.

Ký hiệu $f_0, f_1,..., f_{25}$ lần lượt là tần suất xuất hiện của $a, b, ... ,z$ trong x, ta có:

$$I_C(x) = \frac{\sum_{i=0}^{25}\binom{f_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=0}^{25} f_i(f_i+1)}{n(n+1)}.$$

Giả sử x là một dãy ký tự (tiếng Anh). Ta có thể hy vọng rằng:

$$I_C(x) \approx \sum_{i=0}^{25} p_i^2 = 0{,}065 ,$$

trong đó p_i là xác suất của ký tự ứng với số hiệu i cho bởi bảng phân bố xác suất các ký tự (trang 61)

Nếu x là một dãy ký tự hoàn toàn ngẫu nhiên thì ta có:

$$I_C \approx 26.\ (1/26)^2 = 1/26 = 0{,}038 .$$

Dựa vào các điều nói trên, ta có phương pháp đoán độ dài m của mã Vigenère như sau: Cho bản mã $y = y_1y_2..., y_n$. Ta viết lại y theo bảng có m ($m \geq 1$) hàng như sau:

$$\begin{array}{c} y = \quad y_1 \ y_{m+1} \ldots \ y_{tm+1} \\ y_2 \ y_{m+2} \ldots \ y_{tm+2} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y_m \ y_{em} \ldots \ y_{(tm+1)m} \end{array}$$

nghĩa là viết lần lượt theo các cột m ký tự cho đến hết. Ta ký hiệu $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_m$ là các xâu ký tự theo m hàng trong bảng đó. Chú ý rằng các ký tự ở mỗi hàng $\boldsymbol{y}_i$ đều thu được từ các ký tự ở văn bản gốc bằng cùng một phép dịch chuyển nếu m đúng là độ dài của khoá, do đó nếu m là độ dài của khoá thì ta có thể hy vọng rằng với mọi i, $1 \le i \le m$:

$$I_C(\boldsymbol{y}_i) \approx 0{,}065 .$$

Để đoán độ dài m, ta lần lượt chia y theo cách trên thành $m = 1, 2, 3...$ hàng, và tính các $I_C(y_i)$ $(1 \le i \le m)$, cho đến khi nào được một số m mà với mọi i, $1 \le i \le m$, đều có $I_C(y_i) \approx 0{,}065$ thì ta có thể chắc m là độ dài của khoá.

Thí dụ: Cho bản mã

*chr*eevoahmaeratbiaxxwtnxbeeophbsb
qmqeqerbwrvxuoakxaosxxweahbwgj
mmqmnkgrfvgxwtrzxwiaklxfpskaute
mndemgtsxmxbtuiadngmgpsrelxnjelx
vrvprtulhdnqwtwdtygbphxtfaljhasvbf
xngll*chr*zbwelekmsjiknbhwrignmgjsg
lxfeyphagnbieqjtmrvlcrremndglxrrim
gnsnrw*chr*qhaeyevtaqebbipeewevkak
oewadremxmtbhh*chr*tkdnvrz*chr*clqoh
pwqaiiwxnrmgwoiifkee.

Dùng phép thử Kasiski, ta nhận thấy rằng *chr* xuất hiện 5 lần, khoảng cách của các lần xuất hiện liên tiếp là 165, 70, 50, 10. Ước số chung của các số đó là 5. Vậy ta có thể phán đoán độ dài khoá mã là 5.

Dùng phương pháp chỉ số trùng hợp, với $m = 1$ ta có một chỉ số trùng hợp là 0,045; với $m = 2$ có hai chỉ số là 0,046 và 0,041; với $m = 3$ có ba chỉ số là 0,043; 0,050 và 0,047 ; với $m = 4$ có bốn chỉ số là 0,042; 0,039; 0,046 và 0,043; với $m = 5$, ta thu được năm chỉ số là 0,063; 0,068; 0,069; 0,061 và 0,072, đều khá gần với 0,065. Vậy có thể phán đoán độ dài khoá là 5. Cả hai phương pháp cho kết quả như nhau.

Bây giờ đến bước thứ hai là xác định các giá trị k_1, k_2,...k_m. Ta cần một khái niệm mới là *chỉ số trùng hợp tương hỗ*, được định nghĩa như sau:

Định nghĩa 3.2. Giả sử $\boldsymbol{x} = x_1x_2... x_n$ và $\boldsymbol{y} = y_1y_2... y_n$ là hai dãy ký tự có độ dài n và n'.

Chỉ số trùng hợp tương hỗ của ***x*** và ***y***, ký hiệu $MI_C(\boldsymbol{x},\boldsymbol{y})$, được định nghĩa là xác suất của việc một phần tử của ***x*** trùng với một phần tử của ***y***.

Ký hiệu $f_0, f_1,..., f_{25}$ và $f_0', f_1',..., f_{25}'$ là tần suất xuất hiện của a, b,...,z trong ***x*** và ***y*** tương ứng. Khi đó, ta có:

$$MI_C(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=0}^{25} f_i \cdot f_i'}{n.n'}.$$

Bây giờ với *m* đã xác định, ta viết bản mã ***y*** lần lượt theo từng cột để được *m* hàng $y_1,...y_m$ như ở phần trên. Ta tìm khoá mã $K = (k_1,k_2,...k_m)$.

Giả sử ***x*** là bản rõ và $x_1,...,x_m$ là các phần bản rõ tương ứng với $y_1,...,y_m$. Ta có thể

xem phân bố xác suất của các ký tự trên ***x***, và cũng trên các x_1,..., x_m là xấp xỉ với phân bố xác suất của các ký tự trên văn bản tiếng Anh nói chung. Do đó, xác suất của việc một ký tự ngẫu nhiên của y_i bằng a là p_{-k_i}, bằng b là p_{1-k_i}, v.v... Và ta có thể đánh giá

$$MI_C(y_i, y_j) \approx \sum_{h=0}^{25} p_{h-k_i} \cdot p_{h-k_j} = \sum_{h=0}^{25} p_h \cdot p_{h+k_i-k_j}.$$

Đại lượng đó chỉ phụ thuộc vào k_i - k_j, ta gọi là *dịch chuyển tương đối* của y_i và y_j. Ta chú ý rằng biểu thức:

$$\sum_{h=0}^{25} p_h \cdot p_{h+l}$$

có giá trị lớn nhất khi $l = 0$ là 0,065, và có giá trị biến thiên giữa 0,031 và 0,045 với mọi $l > 0$.

Nhận xét rằng y_j phải dịch chuyển $l = k_i - k_j$ bước (hay dịch chuyển l ký tự trong bảng chữ cái) để được y_i , nên nếu ký hiệu y_j^g là dịch chuyển g bước của y_j , thì ta có hy vọng khi tính lần lượt các đại lượng $MI_C(y_i, y_j^g)$ với $0 \le g \le 25$, ta sẽ đạt được một giá trị xấp xỉ 0,065 với $g = l$, và các giá trị khác đều ở khoảng giữa 0,031 và 0,045. Điều đó cho ta một phương pháp để ước lượng các dịch chuyển $k_i - k_j$, tức là được một số phương trình dạng $k_i - k_j = l$, từ đó giúp ta tính ra các giá trị $k_1, k_2, ..., k_m$.

Trong thí dụ của bản mã đang xét, ta tính được các giá trị $MI_C(y_i , y_j^g)$ với $1 \le i \le j \le 5$, $0 \le g \le 25$, như trong bảng ở trang sau đây (trong bảng đó, ở bên phải mỗi

cặp (i, j) là một ngăn gồm có 26 giá trị của $MI_C(y_i , y_j^g)$ ứng với các giá trị của g = 0,1,2,..., 25).

Nhìn bảng đó, ta thấy các giá trị $MI_C(y_i , y_j^g)$ xấp xỉ 0.065 (như được in đậm và gạch dưới ở trong bảng) ứng với các bộ giá trị (i, j,g) lần lượt bằng (1,2,9), (1,5,16), (2,3,13), (2,5,7), (3,5,20) và (4,5,11).

i	j	Giá trị của $MI_C(y_i , y_j^g)$
1	2	.028 .027 .028 .034 .039 .037 .026 .025 .052 **<u>.068</u>** .044 .026 .037

		.043	.037	.043	.037	.028
		.041	.041	.034	.037	
		.051	.045	.042	.036	
1	3	.039	.033	.040	.034	.028
		.053	.048	.033	.029	.056
		.050	.045	.039		
		.040	.036	.037	.032	.027
		. 037	.036	.031	.037	
		.055	.029	.024	.037	
1	4	.034	.043	.025	.027	.038
		.049	.040	.032	.029	.034
		.039	.044	.044		

		.034	.039	.045	.044	.037
		.055	.047	.032	.027	.039
		.037	.039	.035		
1	5	.043	.033	.028	.046	.043
		.044	.039	.031	.026	.030
		.036	.040	.041		
		.024	.019	.048	**.070**	.044
		.028	.038	.044	.043	.047
		.033	.026	.046		
2	3	.046	.048	.041	.032	.036
		.035	.036		030	.024
		.039	.034	.029	.040	

		.067	.041	.033	.037	.045
		.033	.033	.027	.033	.045
		.052	.042	.030		
2	4	.046	.034	.043	.044	.034
		.031	.040	.045	040	.048
		.044	.033	.024		
		.028	.042	.039	.026	.034
		.050	.035	,032	.040	.056
		.043	.028	.028		
2	5	.033	.033	.036	.046	.026
		.018	.043	**.080**	.050	.029
		.031	.045	.039		

		.037	.027	.026	.031	.039
		.040	.037	.041	.046	.045
		.043	.035	.030		
3	4	.038	.036	.040	.033	.036
		.060	.035	.041	.029	.058
		.035	.035	.034		
		.053	.030	.032	.035	.036
		.036	.028	.046	.032	.051
		.032	.034	.030		
3	5	.035	.034	.034	.036	.030
		.043	.043	.050	.025	.041
		.051	.050	.035		

		.032	.033	.033	.052	.031
		.027	.030	**.072**	.035	.034
		.032	.043	.027		
4	5	.052	.038	.033	.038	.041
		.043	.037	.048	.028	.028
		.036	**.061**	.033		
		.033	.032	.052	.034	.027
		.039	.043	.033	.027	.030
		.039	.048	.035		

Từ đó ta có các phương trình (theo mod26):

$k_1 - k_2 = 9$ $k_2 - k_5 = 7$

$k_1 - k_5 = 16$ $k_3 - k_5 = 20$

$k_2 - k_3 = 13$ $k_4 - k_5 = 11$.

Hệ phương trình đó chỉ có 4 phương trình độc lập tuyến tính, mà có 5 ẩn số, nên lời giải phụ thuộc một tham số, ta chọn là k_1, và được

$$(k_1, k_2, k_3, k_4, k_5) = (k_1, k_1 + 17, k_1 + 4, k_1 + 21, k_1 + 10) \bmod 26.$$

Thử với các giá trị có thể của k_1 ($0 \le k_1 \le 26$), cuối cùng ta có thể tìm được bản rõ như sau đây với khoá là JANET ($k_1 = 9$):

the *almond tree was in tentative blossom the days were longer often ending with magnificent evenings of corrugated pink skies the hunting season was over with hounds and guns put away for six months* the *vineyards were busy again as* the *well*

organized farmers treated their vines and the *more lackadaisical neighbors hurried to do the pruning* they *should have done in november.*

4. Thám mã đối với mã Hill.

Mật mã Hill khó bị khám phá bởi việc thám mã *chỉ dựa vào bản mã*, nhưng lại là dễ bị khám phá nếu có thể sử dụng phép thám mã kiểu *biết cả bản rõ*. Trước hết ta giả thiết là đã biết giá trị *m*. Mục đích của thám mã là phát hiện được khoá mật mã *K*, trong trường hợp mã Hill là một ma trận cấp *m* có các thành phần trong $\boldsymbol{Z_{26}}$.

Ta chọn một bản rõ có chứa ít nhất *m* bộ *m* khác nhau các ký tự:

$$x_1 = (x_{11}, ..., x_{1m}), ..., x_m = (x_{m1}, ..., x_{mm}),$$

và giả thiết biết mã tương ứng của chúng là:

$$y_1 = (y_{11},..., y_{1m}),...., y_m = (y_{m1},..., y_{mm}).$$

Ta ký hiệu X và Y là hai ma trận cấp m , X=(x_{ij}), Y= (y_{ij}). Theo định nghĩa mã Hill, ta có phương trình Y =X.K. Nếu các x_i được chọn sao cho ma trận X có nghịch đảo X^{-1} thì ta tìm được $K = X^{-1}.Y$, tức là tìm được khoá của hệ mã được sử dụng.

Thí dụ: Giả sử mã Hill được sử dụng có m =2, và ta biết bản rõ *friday* cùng bản mã tương ứng *pqcfku*. Như vậy ta biết

$e_K(5,17)$ =(15,16), $e_K(8,3)$ =(2,5), và $e_K(0,24)$ =(10,20).

Từ hai phương trình đầu ta được

$$\begin{pmatrix} 15 & 16 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 17 \\ 8 & 3 \end{pmatrix}.K,$$

từ đó được $K = \begin{pmatrix} 7 & 19 \\ 8 & 3 \end{pmatrix}$. Với K đó phương trình thứ ba cũng nghiệm đúng.

Trở lại với vấn đề xác định m. Nếu m không qua lớn, ta có thể thử cách trên lần lượt với $m = 2,3,4,...$ cho đến khi tìm được khoá, và khoá K xem là tìm được nếu ngoài m cặp bộ m $(x_1,y_1),...,(x_m , y_m)$ dùng để tìm khoá, K vẫn nghiệm đúng với các cặp bộ m khác mà ta có thể chọn để thử.