Một số phương pháp thám mã

Bởi:

Khoa CNTT ĐHSP KT Hưng Yên

Vấn đề thám mã

Khái niệm:

Thám mã là công việc phân tích bản tin mã hóa để nhận được bản tin rõ trong điều kiện không biết trước khóa mã.

Trong thực tế, công việc thám mã gặp nhiều khó khăn hơn khi không biết rõ hệ mật mã nào được sử dụng. Tuy nhiên, để đơn giản hóa, chúng ta giả sử người thám mã đã biết rõ hệ mật mã được sử dụng khi tiến hành phân tích mã (*nguyên lý Kerckhoff*). Mục đích là thiết kế được một hệ mật mã an toàn bảo mật.

Trước hết chúng ta cần phân loại mức độ tấn công vào các hệ mật mã. Mức độ này tùy thuộc vào hiểu biết của người thám mã đối với hệ mật mã được sử dụng. Theo đó, chúng ta có thể chia thành các loại tấn công sau:

- **Tấn công chỉ biết bản mã (ciphertext-only):** người thám mã chỉ có bản tin mã hóa.
- **Tấn công biết bản tin rõ (known plaintext):** người thám mã có bản tin rõ và bản mã.
- **Tấn công chọn bản tin rõ (chosen plaintext):** người thám mã tạm thời có quyền truy xuất tới Bộ mã hóa, do đó anh ta có khả năng chọn bản tin rõ và xây dựng bản mã tương ứng.
- **Tấn công chọn bản mã (chosen ciphertext):** người thám mã tạm thời có quyền truy xuất tới Bộ giải mã, do đó anh ta có khả năng chọn bản mã và xây dựng lại bản tin rõ tương ứng.

Trong mọi trường hợp, mục đích là tìm ra khóa mã được sử dụng. Kiểu tấn công chọn bản mã được thực hiện với hệ mật mã khóa công khai mà chúng ta sẽ xem xét trong chương kế tiếp. Trong phần này chúng ta chỉ thảo luận về kiểu tấn công được xem là "yếu nhất" - Tấn công chỉ biết bản mã.

Nhiều kỹ thuật thám mã sử dụng đặc điểm thống kê của tiếng Anh, trong đó dựa vào tần suất xuất hiện của 26 chữ cái trong văn bản thông thường để tiến hành phân tích mã. Becker và Piper đã chia 26 chữ cái thành năm nhóm và chỉ ra xác suất của mỗi nhóm như sau:

1. E, có xác suất khoảng 0.120
2. T, A, O, I, N, S, H, R, mỗi chữ cái có xác xuất nằm trong khoảng từ 0.06 đến 0.09
3. D, L, mỗi chữ cái có xác xuất xấp xỉ 0.04
4. C, U, M, W, F, G, Y, P, B, mỗi chữ cái có xác xuất nằm trong khoảng từ 0.015 đến 0.023
5. V, K, J, X, Q, Z, mỗi chữ cái có xác xuất nhỏ hơn 0.01

Ngoài ra, tần suất xuất hiện của dãy hai hay ba chữ cái liên tiếp được sắp theo thứ tự giảm dần như sau [11]: TH, HE, IN, ER … THE, ING, AND, HER…

Thám mã tích cực:

Thám mã tích cực là việc thám mã sau đó tìm cách làm sai lạc các dữ liệu truyền, nhận hoặc các dữ liệu lưu trữ phục vụ mục đích của người thám mã.

Thám mã thụ động:

Thám mã thụ động là việc thám mã để có được thông tin về bản tin rõ phục vụ mục đích của người thám mã.

Thám mã Affine

Giả sử Trudy đã lấy được bản mã sau đây:

FMXVEDKAPHFERBNDKRXRSREFMORUDSDKDVSHVUFEDKAPRKDLYEVLRHHRH.

Trudy thống kê tần suất xuất hiện của 26 chữ cái như trong bảng sau:

Chữ cái	Tần suất	Chữ cái	Tần suất
A	2	N	1
B	1	O	1
C	0	P	3
D	6	Q	0
E	5	R	8
F	4	S	3
G	0	T	0
H	5	U	2
I	0	V	4
J	0	W	0
K	5	X	2
L	2	Y	1
M	2	Z	0

Chỉ có 57 chữ cái trong bản mã nhưng phương pháp này tỏ ra hiệu quả để thám mã Affine. Ta thấy tần suất xuất hiện các chữ cái theo thứ tự là: R(8), D(6), E, H, K(5) và F, S, V(4). Vì vậy dự đoán đầu tiên của ta có thể là: R là mã của e, D là mã của t. Theo đó, $e_K(4) = 17$ và $e_K(19) = 3$. Mà $e_K(x) = ax+b$ với a, b là các biến. Để tìm K=(a, b) ta giải hệ phương trình:

4a+b=17

19a+b=3

Suy ra, a = 6, b=19. Đây không phải là khóa vì gcd(a, 26) = 2 > 1. Ta lại tiếp tục phỏng đoán: R là mã của e, E là mã của t. Ta nhận được a = 13, chưa thỏa mãn. Tiếp tục với H, ta có a=8. Cuối cùng, với K ta tìm được K = (3, 5). Sử dụng khóa mã này ta có được bản tin rõ như sau:

algorithmsrequiregeneraldefinitionsofarithmeticprocesses

Thám mã Vigenere

Để thám mã Vigenere, trước hết cần xác định độ dài từ khóa, ký hiệu là m. Sau đó mới xác định từ khóa. Có hai kỹ thuật để xác định độ dài từ khóa đó là phương pháp Kasiski và phương pháp chỉ số trùng hợp (index of coincidence).

Phương pháp Kasiski được đưa ra bởi Friedrich Kasiski năm 1863. Phương pháp này làm việc như sau:

Tìm trên bản mã các cặp xâu kí tự giống nhau có độ dài ít nhất là 3, ghi lại khoảng cách giữa vị trí chữ cái đầu tiên trong các xâu và xâu đầu tiên. Giả sử nhận được d 1 , d 2 ...

Tiếp theo ta phỏng đoán m là số sao cho ước số chung lớn nhất của các d i chia hết cho m.

Ví dụ:

Plaintext: conghoa|danchun|handant|runghoa|sapsuat|hanghoa

Keyword: abcdefg

Ciphertext: CPPJLTG DBPFLZT HBPGESZ RVPJLTG SBRVYFZ HBPJLTG

Vị trí xuất hiện của dãy PJL lần lượt là: *3, 24, 38.* Do vậy, dãy *d1, d2 ... là 21, 35; gcd(d1, d2 ...) = 7*

Phương pháp chỉ số trùng hợp sẽ cho biết các bằng chứng để nhận được giá trị m. Phương pháp này được đưa ra bởi Wolfe Friedman năm 1920 như sau:

Giả sử x=x 1 x 2… x n là xâu có n ký tự. Chỉ số trùng hợp của x, ký hiệu là I c (x), được định nghĩa là xác suất mà hai phần tử ngẫu nhiên của x là giống nhau. Giả sử chúng ta ký hiệu tần suất của A, B, C, ..., Z trong x lần lượt là f 0 , f 1 , ..., f 25 . Chúng ta có thể chọn hai phần tử của x theo $(2^n) = n!/(2!(n-2)!)$ *cách. Với mỗi* $0 \le i \le 25$ *, có* (2^{f_i}) *cách chọn các phần tử là i. Vì vậy, chúng ta có công thức:*

$$I\ c\ (x) = \frac{\sum_{i=0}^{25} f_i(f_i - 1)}{n(n-1)}$$

Bây giờ, giả sử x là xâu văn bản tiếng Anh. Ta có $I_c(x)$ $\sum_{i=0}^{25} p_i^2 = 0.065$

Ví dụ:

Cho bản mã trong hệ mật mã Vigenere

CHREEV	OAHMAE	RATBIA	XXWTNX	BEEOPH	BSBQMQ	EQERBW
RVXUOA	XXAOSXX	…				
LXFPSK						
VRVPRT						…CHR
ZBWELE						
AMRVLO	…		…	WCHRQH	…	
PEEWEV	KAKOE	WADREM	XMTBHHC	HRTKDN	VRZCHR	CLQOHP
WQAIIW	XNRMGW	OIIFKBE				

- Theo phương pháp Kasiski, đầu tiên xâu CHR xuất hiện ở 4 vị trí trong bản mã, lần lượt là: 1, 166, 236 và 286. Khoảng cách giữa các xâu là 165, 235 và 285. Ước số chung lớn nhất của các số này là 5. Vậy ta có m =5.

- Theo phương pháp chỉ số trùng hợp, với m=1 thì chỉ số trùng hợp là $I_c(x)$ = 0.045; m=2, $I_c(x)$=0.046 và 0.041; m=3, $I_c(x)$=0.043, 0.050, 0.047; m=4, $I_c(x)$=0.042, 0.039, 0.046, 0.040; m=5, $I_c(x)$=0.063, 0.068, 0.069, 0.072; Ta dừng và nhận được m = 5.

Để xác định khóa mã, ta sử dụng phương pháp thống kê sau đây:

Giả sử x=x 1 x 2… x n và y=y 1 y 2… y n' là hai xâu có n và n' ký tự . Chỉ số trùng hợp tương quan của x và y, ký hiệu là MI c (x,y), được định nghĩa là xác suất mà một phần tử ngẫu nhiên của x bằng một phần tử ngẫu nhiên của y. Nếu chúng ta ký hiệu tần suất của A, B, C, …, Z trong x và y lần lượt là f 0 , f 1 , …, f 25 . và f' 0 , f' 1 , …, f' 25 . Thì:

$$MI\ c\ (x,y) = \frac{\sum_{i=0}^{25} f_i f'_i}{nn'}$$

Bây giờ, giả sử x,y là xâu văn bản tiếng Anh. Ta có $MI_c(x_i,y_j)$ 0.065

Ví dụ:

Giả sử m=5 như ta đã thực hiện ở trên. Theo phương pháp thống kê [11] ta tìm được khóa mã là: JANET. Vậy bản tin rõ sẽ là: *the almond tree was in ...*