Mật mã và độ phức tạp thuật toán
Chủ đề 3: Lý thuyết mã và ứng dụng

TS. Ngô Thị Hiền (hien.ngothi@hust.edu.vn)
PGS.TS. Nguyễn Đình Hân (han.nguyendinh@hust.edu.vn)

Viện Toán ứng dụng và Tin học
Đại học Bách khoa Hà Nội

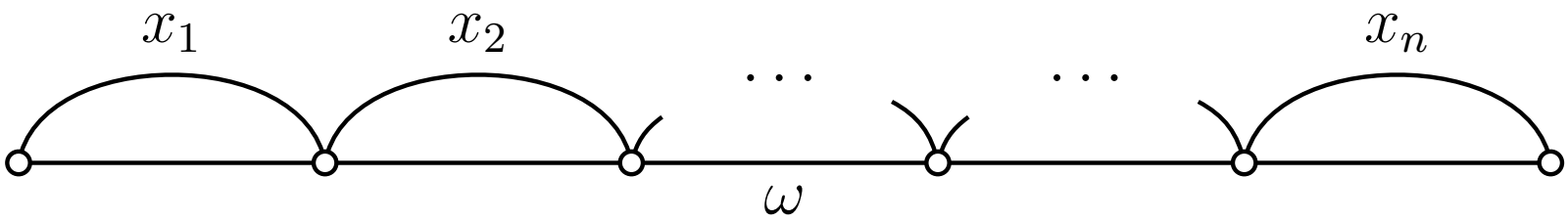
Lý thuyết mã và ứng dụng

Vai trò của lý thuyết mã

- Mã có vai trò thiết yếu trong nhiều lĩnh vực như xử lý thông tin, nén dữ liệu, truyền thông dữ liệu và mật mã.
- Lý thuyết mã, khởi nguồn từ lý thuyết thông tin, là một bộ phận không thể thiếu của khoa học máy tính, công nghệ thông tin và truyền thông.
- Nhu cầu sử dụng mã trong biểu diễn, bảo mật thông tin ngày càng cấp thiết về thực tiễn đòi hỏi phải đào tạo một lực lượng rất lớn cán bộ kỹ thuật chuyên trách, có trình độ cao.
- Các bài toán được quan tâm trong lý thuyết mã và ứng dụng: kiểm định mã, nghiên cứu mở rộng mã, phân bậc ngôn ngữ, xây dựng các hệ mật mới ứng dụng trong mật mã học, v.v.

Ngôn ngữ và mã của các từ hữu hạn

- Cho Σ là bảng hữu hạn các chữ cái. Σ^* là vị nhóm tự do của tất cả các từ hữu hạn sinh bởi Σ. Từ rỗng được ký hiệu là ε và $\Sigma^+ = \Sigma^* - \{\varepsilon\}$.
- Mỗi tập con của Σ^* được gọi là *một ngôn ngữ* trên Σ.
- Một *X-phân tích* của một từ $\omega \in \Sigma^*$ theo X, với $X \subseteq \Sigma^*$, là một dãy $\omega = u_1 u_2 \cdots u_n$ với $u_1, u_2, \ldots, u_n \in X$, $n \geq 1$.



- Một tập $X \subseteq \Sigma^+$ được gọi là *mã* trên Σ nếu mọi từ $\omega \in \Sigma^+$ có nhiều nhất một X-phân tích.

Ngôn ngữ và mã của các từ hữu hạn

Ta có thể hình dung một mã X là một ngôn ngữ của các từ hữu hạn sao cho bất kì một thông điệp nào, được mã hóa thành các từ của X, sẽ được giải mã theo *một cách duy nhất*.

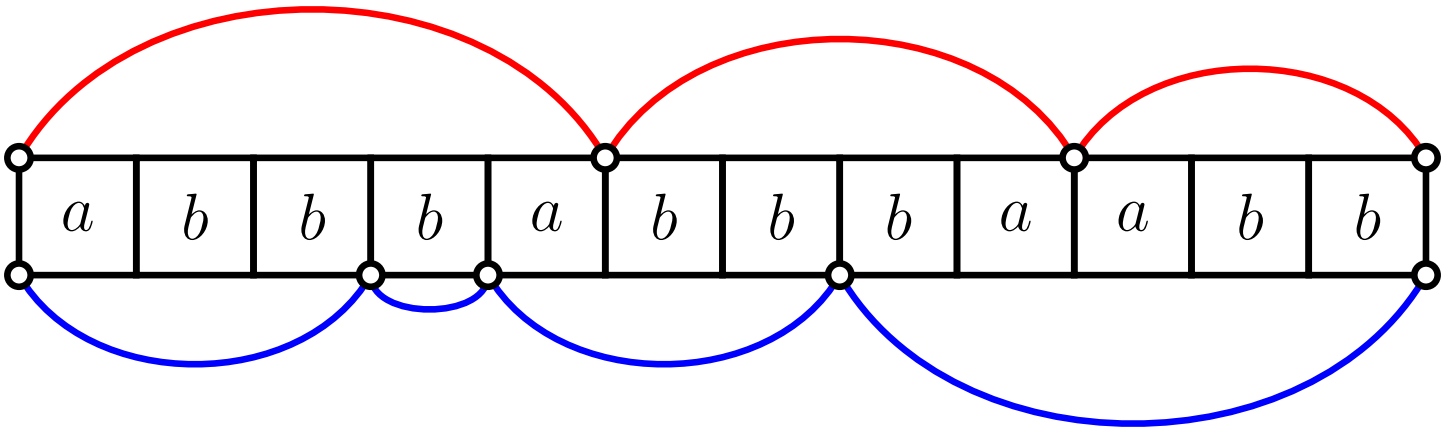
Ví dụ 3.1 Cho Σ là bảng hữu hạn các chữ cái. Khi đó:

- Các ngôn ngữ $X = \emptyset$ và $X = \Sigma$ đều là mã. Tổng quát hơn, nếu $p \geq 1$ là một số nguyên thì $X = \Sigma^p$ là mã uniform độ dài p.
- Giả sử $\Sigma = \{0, 1\}$ thì $X = \{00, 100, 10\}$ là mã trên Σ.

Lưu ý $\rightsquigarrow$ Mã không bao giờ chứa từ rỗng ε. Bất kì tập con nào của mã cũng là mã. Phần tử của mã được gọi là *từ mã*.

Ví dụ 3.2 Cho $\Sigma = \{a, b\}$ và $X = \{b, abb, abbba, bbba, baabb\}$. Tập X không là mã vì tồn tại từ $\omega = abbbabbbaabb$ có hai phân tích khác nhau trong X,

$$\omega = (abbba)(bbba)(abb) = (abb)(b)(abb)(baabb).$$

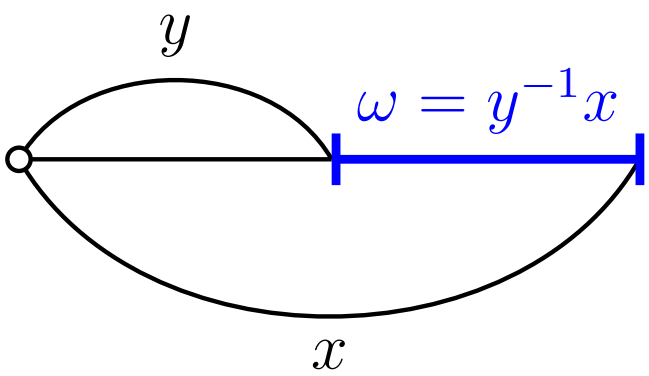


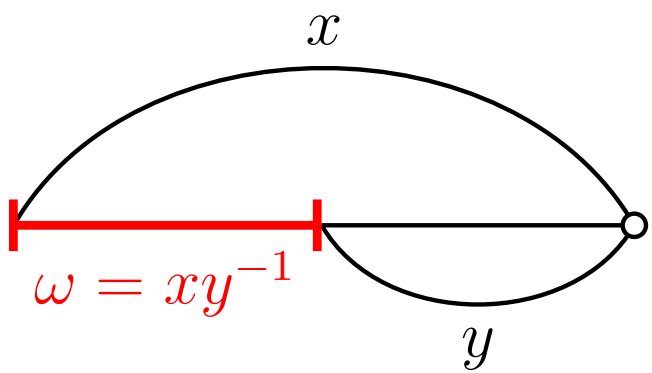
Vị nhóm

Cho M là một vị nhóm.

– Với $x, y \in M$, ta có

$$y^{-1}x = \{\omega \in M \mid y.\omega = x\} \text{ và } xy^{-1} = \{\omega \in M \mid x = \omega.y\}.$$





– Với $S, T \subseteq M$, ta định nghĩa *thương trái* và *thương phải* của S bởi T

$$T^{-1}S = \{u \in M \mid \exists t \in T : t.u \in S\} \text{ và}$$
$$ST^{-1} = \{u \in M \mid \exists t \in T : u.t \in S\}.$$

Mối liên hệ giữa ngôn ngữ và mô hình tính toán

Định lý 3.1 ([4]) *Cho $X \subseteq \Sigma^*$ là một ngôn ngữ. Các mệnh đề sau đây là tương đương*

(i) X *là chính quy.*

(ii) X *đoán nhận được bởi một otomat hữu hạn.*

(iii) *Otomat tối tiểu* $\mathcal{A}(X)$ *là hữu hạn.*

(iv) *Họ các tập* $u^{-1}X$*, với* $u \in \Sigma^*$*, là hữu hạn.*

(v) *Vị nhóm cú pháp* M_X *là hữu hạn.*

(vi) X *thỏa bởi một đồng cấu từ* Σ^* *đến vị nhóm hữu hạn* M*.*

Bài toán kiểm định mã

Kiểm định mã (kiểm định một ngôn ngữ có là mã không) là bài toán *cơ bản, cốt lõi* của lý thuyết mã.

- Sardinas-Patterson (1953): Giới thiệu một thủ tục kiểm định mã gọi là Tiêu chuẩn Sardinas-Patterson.
- Rodeh (1982): Thuật toán kiểm định mã có hữu hạn từ với độ phức tạp thời gian là $\mathcal{O}(n.m)$, n là tổng số từ mã của ngôn ngữ và m là tổng độ dài của chúng.
- Robert (1996): Thuật toán kiểm định mã chính quy có độ phức tạp thời gian là $\mathcal{O}(n^4)$, với n là tổng số trạng thái của otomat đoán nhận ngôn ngữ đầu vào.
- Han, *et al* (2014): Thủ tục kiểm định mã có số bước tính toán cỡ tuyến tính thay cho cỡ hàm mũ của thủ tục Sardinas-Patterson.

Tiêu chuẩn Sardinas-Patterson

Cho một ngôn ngữ $X \subseteq \Sigma^+$, ta xem xét các tập phần dư U_i kết hợp với X được định nghĩa đệ quy như sau

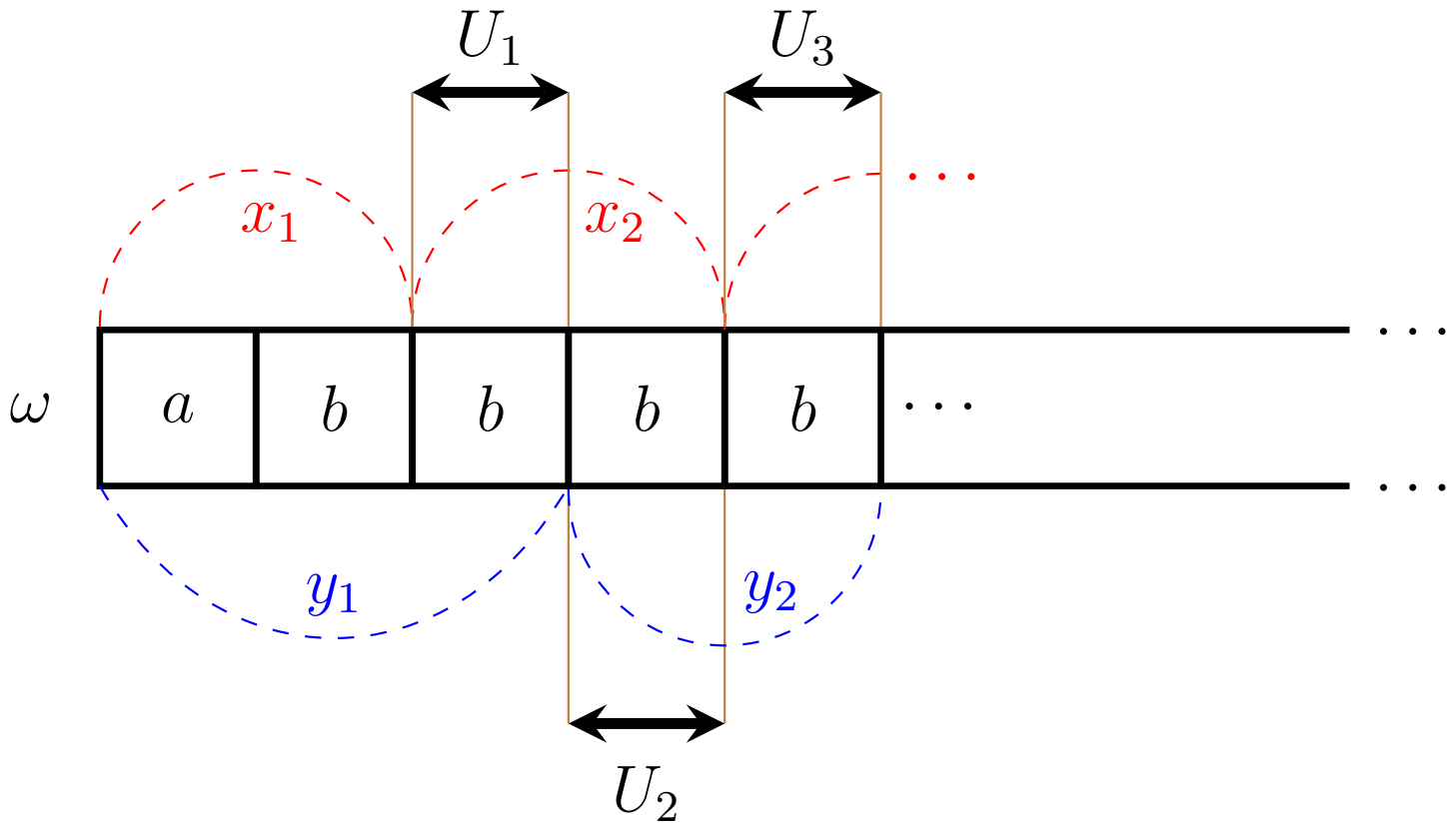
$$\begin{aligned} U_1 &= X^{-1}X - \{\varepsilon\} \\ U_{i+1} &= U_i^{-1}X \cup X^{-1}U_i, \quad i \geq 1. \end{aligned} \tag{3.1}$$

Định lý cơ bản sau đây của Sardinas-Patterson cung cấp một tiêu chuẩn kiểm định mã.

Định lý 3.2 *Tập $X \subseteq \Sigma^+$ là mã khi và chỉ khi các tập U_i $(i \geq 1)$, được định nghĩa theo công thức (3.1) không chứa từ rỗng.*

Tiêu chuẩn Sardinas-Patterson

Ví dụ 3.3 Cho $\Sigma = \{a, b\}$ và $X = \{ab, abb, bb\}$. Ngôn ngữ X là mã vì $U_i = \{b\} \neq \emptyset$ với mọi $i \geq 1$.



Tiêu chuẩn Sardinas-Patterson cải tiến

Cho $X \subseteq \Sigma^+$, ta định nghĩa một dãy các tập phần dư V_i kết hợp với X như sau

$$\begin{aligned} V_1 &= X^{-1}X - \{\varepsilon\} \\ V_{i+1} &= V_i^{-1}X \cup X^{-1}V_i \cup V_i, \quad i \geq 1. \end{aligned} \tag{3.2}$$

Định lý sau đây cung cấp một tiêu chuẩn kiểm định mã, gọi là tiêu chuẩn Sardinas-Patterson cải tiến - A Modification of Sardinas-Patterson (MSP) test.

Định lý 3.3 *Cho $X \subseteq \Sigma^+$ và $V_i\,(i \geq 1)$ được thiết lập theo công thức (3.2). Khi đó, X là mã khi và chỉ khi $\varepsilon \notin V_i$ với mọi $i \geq 1$.*

Thuật toán Sardinas-Patterson cải tiến

Thuật toán 3.1 (MSP) Kiểm định mã cho ngôn ngữ chính quy

Đầu vào: $X \subseteq \Sigma^+$ là một ngôn ngữ chính quy

Đầu ra: "YES" nếu X là mã, "NO" nếu X không là mã

Bước 1: Tính $V_1 = X^{-1}X - \{\varepsilon\}$ và đặt $i = 2$

if $V_1 == \emptyset$ then return "YES"

Bước 2 (lặp): Tính $V_i = V_{i-1}^{-1}X \cup X^{-1}V_{i-1} \cup V_{i-1}$

if $\varepsilon \in V_i$ then return "NO"

if $V_i == V_{i-1}$ then return "YES"

else đặt $i = i + 1$, và quay về Bước 2

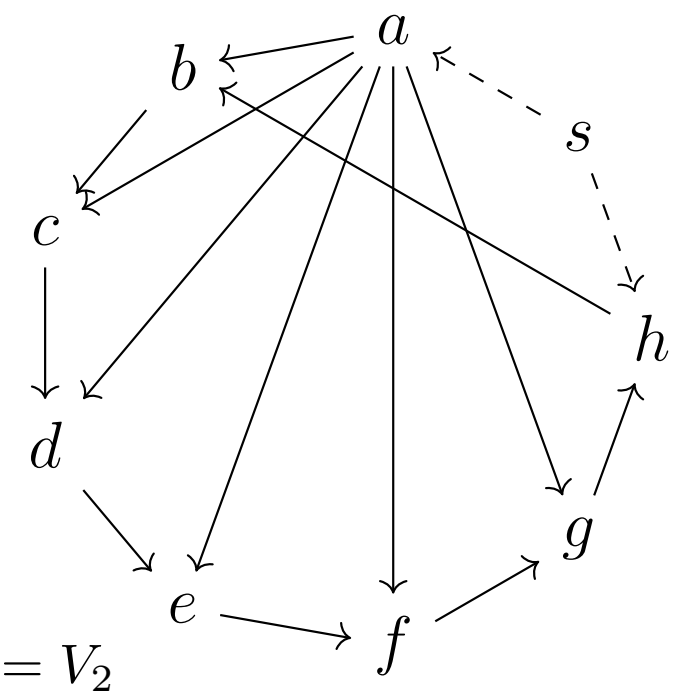
Định lý 3.4 *Cho $X \subseteq \Sigma^+$ là ngôn ngữ chính quy có chỉ số n. Khi đó, thuật toán MSP kiểm định X có là mã không với số bước không quá n.*

Thuật toán Sardinas-Patterson cải tiến

Ví dụ 3.4 Cho $\Sigma = \{s, a, b, c, d, e, f, g, h\}$ và X như sau:

$$X = \{ab, ac, ad, ae, af, ag, bc, cd, de, ef, fg, gh, hb, s, sa, sh\}.$$

Khi đó, để kết luận X là mã theo các dãy U_i cần 9 bước tính toán, còn theo V_i chỉ cần 3 bước. Lưu ý, trong hình vẽ, ta biểu diễn trực quan X bởi điều kiện $x \to y \Leftrightarrow (\exists z : y = x^{-1}z)$ với $x, y, z \in X$.



$$U_1 = \{a, h\},$$
$$U_2 = \{b, c, d, e, f, g\},$$
$$U_3 = \{c, d, e, f, g, h\},$$
$$U_4 = \{b, d, e, f, g, h\},$$
$$U_5 = \{b, c, e, f, g, h\},$$
$$U_6 = \{b, c, d, f, g, h\},$$
$$U_7 = \{b, c, d, e, g, h\},$$
$$U_8 = \{b, c, d, e, f, h\},$$
$$U_9 = \{b, c, d, e, f, g\} = U_2$$

$$V_1 = \{a, h\},$$
$$V_2 = \{a, b, c, d, e, f, g, h\},$$
$$V_3 = \{a, b, c, d, e, f, g, h\} = V_2$$

Độ trễ giải mã

Định nghĩa 3.1 *Tập $X \subseteq \Sigma^+$ được gọi là có độ trễ giải mã hữu hạn nếu có một số nguyên $d \geq 0$ sao cho*

$$\forall x, x' \in X, \forall y \in X^d, \forall u \in \Sigma^*, xyu \in x'X^* \Rightarrow x = x'. \tag{3.3}$$

Nếu X có độ trễ giải mã hữu hạn thì số nguyên nhỏ nhất thoả hệ thức (3.3) được gọi là *độ trễ giải mã* của X, số nguyên bất kỳ thỏa hệ thức (3.3) được gọi là *độ trễ giải mã yếu* của X.

Ví dụ 3.5 Cho $\Sigma = \{a, b\}$. Khi đó, tập $X = \{ab, abb, baab\}$ có độ trễ giải mã $d = 1$ và tập $Y = \{a, ab, b^2\}$ có độ trễ giải mã vô hạn.

Xác định độ trễ giải mã

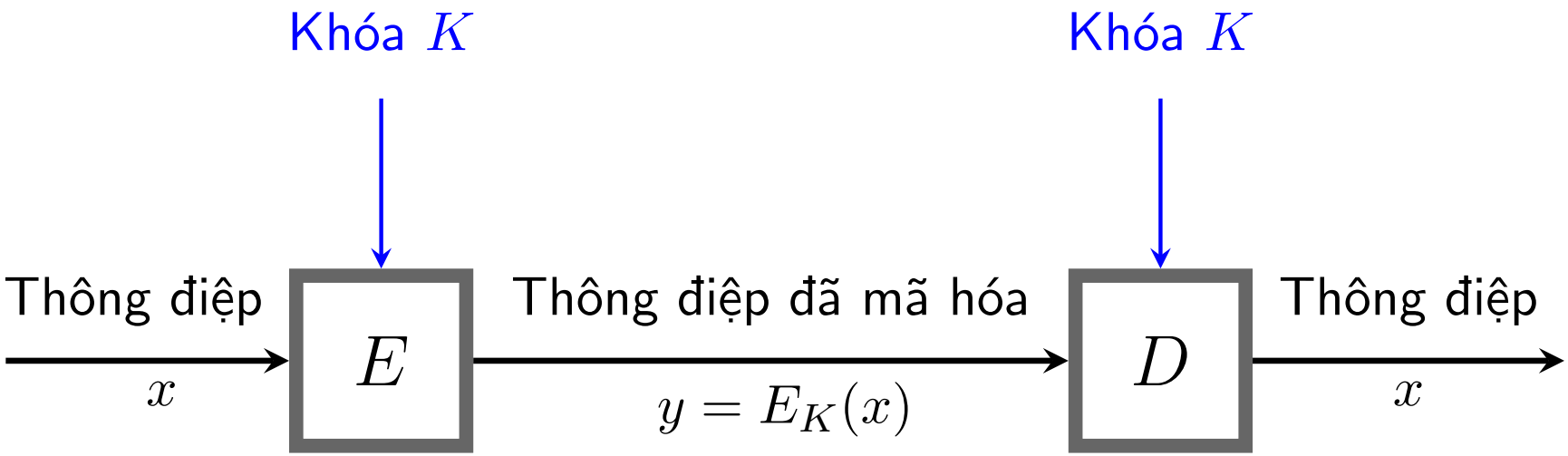
Cho $X \subseteq \Sigma^+$, dựa vào định nghĩa độ trễ giải mã, ta xem xét hai tập phần dư V_i, U_i kết hợp với X được định nghĩa đệ quy như sau

$$\begin{aligned} U_0 &= (X^+)^{-1}X - \{\varepsilon\}, \\ V_1 &= U_0^{-1}X \cup (X^{-1}X - \{\varepsilon\}), \\ U_i &= (V_i X^*)^{-1}X, \\ V_{i+1} &= U_i^{-1}X \cup X^{-1}V_i,\ i \geq 1. \end{aligned} \tag{3.4}$$

Định lý sau đây cung cấp cho ta một tiêu chuẩn xác định độ trễ giải mã của một ngôn ngữ chính quy bất kỳ.

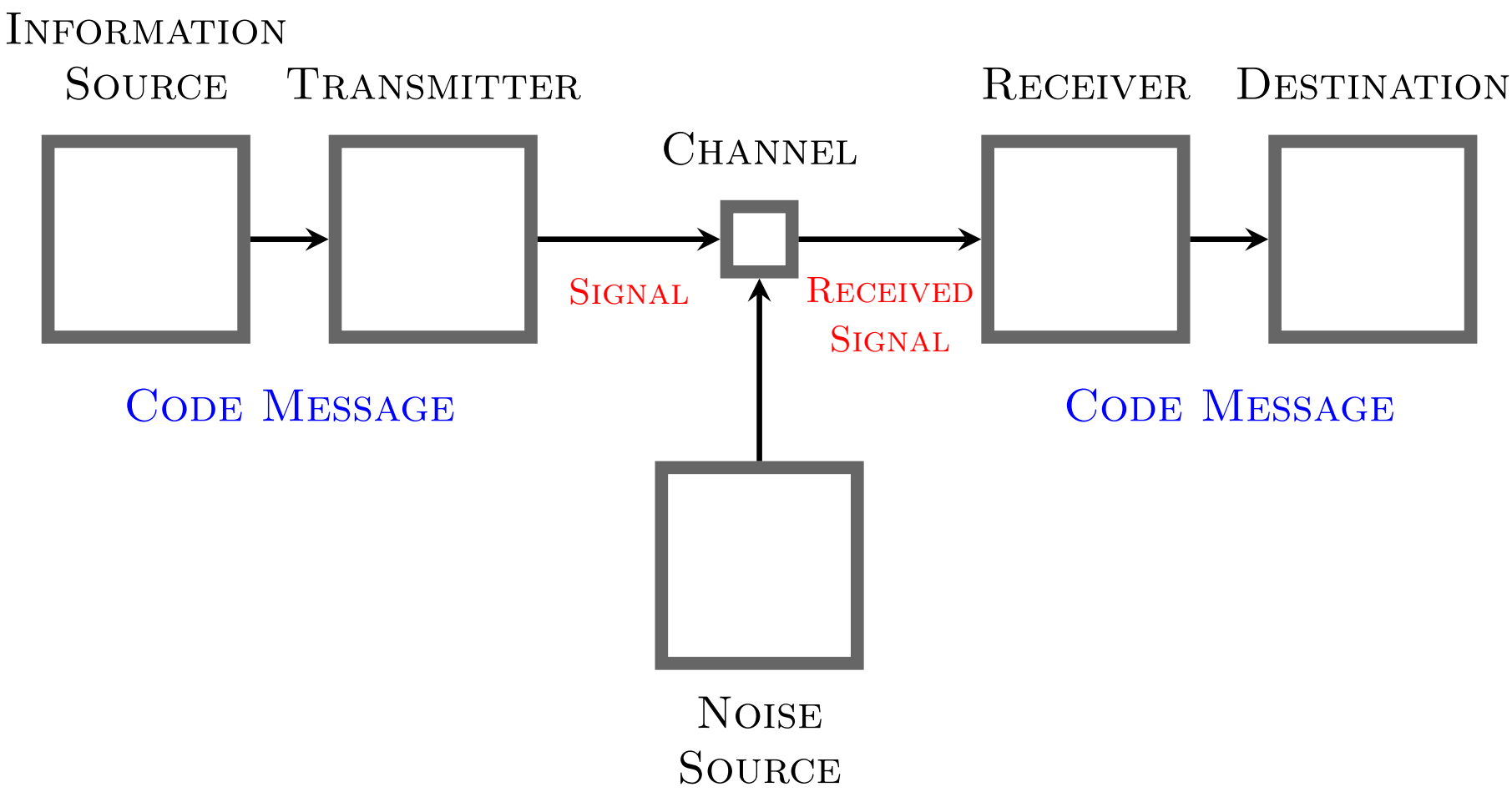
Định lý 3.5 *Cho $X \subseteq \Sigma^+$ và cho $V_i, U_i\ (i \geq 1)$ được định nghĩa theo công thức (3.4). Khi đó, X có độ trễ giải mã hữu hạn $d \geq 0$ khi và chỉ khi $V_{d+1} = \emptyset$ và $V_d \neq \emptyset$.*

Ứng dụng của mã trong mật mã học



Hình 3.1 Mô hình hệ mật mã tổng quát

Ứng dụng của mã trong truyền thông dữ liệu



Hình 3.2 Mô hình truyền thông dữ liệu

Vài nét về lịch sử mật mã

- Mật mã học (Cryptology) gồm Mật mã (Cryptography) và Giấu tin (Steganography).
- 4000 năm trước ở Ai Cập, con người đã biết sử dụng chữ tượng hình.
- Thượng cổ ở Hy Lạp, mã hiệu đã được sử dụng để đánh dấu nô lệ.
- 2000 năm trước, Julius Caesar (La Mã) là người đầu tiên sử dụng hệ thống mật mã trong trao đổi thông tin.
- Phân loại theo yếu tố thời gian, ta có các *hệ mật cổ điển* và *hệ mật hiện đại*.
- Có hai kỹ thuật cơ bản sử dụng trong các hệ mật mã cổ điển, đó là **thay thế** (substitution) và **dịch chuyển**

Khái niệm hệ mật

Định nghĩa 3.2 Một hệ mật (cryptosystem) là một bộ năm $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$, thỏa mãn các điều kiện sau đây:

1. $\mathcal{P}$ là một tập hữu hạn các từ/văn bản gốc (từ hiện/từ rõ)
2. $\mathcal{C}$ là một tập hữu hạn các từ/văn bản mã
3. $\mathcal{K}$ là một tập hữu hạn các khóa
4. Với mỗi $K \in \mathcal{K}$, có một phép mã hóa $e_K \in \mathcal{E}$ và một phép giải mã tương ứng $d_K \in \mathcal{D}$. Mỗi $e_K : \mathcal{P} \to \mathcal{C}$ và $d_K : \mathcal{C} \to \mathcal{P}$ là các ánh xạ sao cho $d_K(e_K(x)) = x$ với mọi $x \in \mathcal{P}$.

Lưu ý $\rightsquigarrow$ Từ nay về sau, khi biểu diễn các hệ mật, ta qui ước sử dụng chữ thường để biểu diễn nội dung các bản rõ và sử dụng chữ hoa để biểu diễn nội dung các bản mã.

Định nghĩa 3.3 Cho a, b là các số nguyên và m là số nguyên dương. Ta viết $a \equiv b \pmod{m}$ nếu $(b - a)$ chia hết cho m. Cụm từ $a \equiv b \pmod{m}$ được đọc là "a đồng dư b modulo m". Số nguyên m được gọi là môđun của đồng dư thức.

Giả sử ta chia a và b cho m để nhận được thương và phần dư nằm trong khoảng 0 và $m - 1$. Tức là, $a = q_1 m + r_1$ và $b = q_2 m + r_2$ với $0 \leq r_1 \leq m - 1$ và $0 \leq r_2 \leq m - 1$. Khi đó, dễ thấy $a \equiv b \pmod{m}$ khi và chỉ khi $r_1 = r_2$. Ta sẽ kí hiệu $a \mod m$ là phần dư của phép chia a cho m. Vậy, $a \equiv b \pmod{m}$ khi và chỉ khi $a \mod m = b \mod m$. Nếu thay a bởi $a \mod m$, ta nói rằng a bị rút gọn theo môđun m.

Lưu ý $\rightsquigarrow$ Để tiện trình bày, ta qui ước kết quả của phép chia $a \mod m$ luôn là số không âm.

Mật mã cổ điển

Bây giờ ta định nghĩa phép toán số học môđun m:

Cho tập hợp $\mathbb{Z}_m = \{0, \ldots, m-1\}$ với hai phép toán cộng $(+)$ và nhân $(*)$. Hai phép toán này tương tự như phép cộng và nhân số nguyên, chỉ khác là tổng và tích tương ứng sẽ được rút gọn theo môđun m. Phép toán cộng và nhân của $\mathbb{Z}_m$ thỏa các tính chất sau đây:

1. Nếu $a, b \in \mathbb{Z}_m, a + b \in \mathbb{Z}_m$
2. Nếu $a, b \in \mathbb{Z}_m, a + b = b + a$
3. Nếu $a, b, c \in \mathbb{Z}_m, (a + b) + c = a + (b + c)$
4. Nếu $a \in \mathbb{Z}_m, a + 0 = 0 + a = a$
5. Nếu $a \in \mathbb{Z}_m, \exists (m - a) \in \mathbb{Z}_m : a + (m - a) = (m - a) + a = 0$
6. Nếu $a, b \in \mathbb{Z}_m, ab \in \mathbb{Z}_m$
7. Nếu $a, b \in \mathbb{Z}_m, ab = ba$
8. Nếu $a, b, c \in \mathbb{Z}_m, (ab)c = a(bc)$

Hệ mật dịch chuyển (Shift Cipher)

Định nghĩa 3.4 Đặt $\mathcal{P} = \mathcal{C} = \mathcal{K} = \mathbb{Z}_{26}$. Với $0 \leq K \leq 25$ định nghĩa:

$$e_K(x) = x + K \mod 26$$

và

$$d_K(y) = y - K \mod 26$$

$(x, y \in \mathbb{Z}_{26})$.

Trường hợp đặc biệt $K = 3$ ứng với hệ mật mã Caesar.

Hệ mật thay thế (Substitution Cipher)

Định nghĩa 3.5 Đặt $\mathcal{P} = \mathcal{C} = \mathbb{Z}_{26}$. Với $\mathcal{K}$ gồm tất cả các hoán vị có thể có của 26 kí hiệu $0, 1, \ldots, 25$. Với mỗi $K \in \mathcal{K}$, định nghĩa:

$$e_K(x) = K(x)$$

và

$$d_K(y) = K^{-1}(y)$$

$(x, y \in \mathbb{Z}_{26})$ và K^{-1} là hoán vị ngược của K.

Hệ mật Affine (Affine Cipher)

Định nghĩa 3.6 Đặt $\mathcal{P} = \mathcal{C} = \mathbb{Z}_{26}$ và đặt

$$\mathcal{K} = \{(a, b) \in \mathbb{Z}_{26} \times \mathbb{Z}_{26} : \gcd(a, 26) = 1\}.$$

Với mỗi $K = (a, b) \in \mathcal{K}$, định nghĩa:

$$e_K(x) = ax + b \mod 26$$

và

$$d_K(y) = a^{-1}(y - b) \mod 26$$

$(x, y \in \mathbb{Z}_{26})$.

Lưu ý $\rightsquigarrow$ Giả sử $a \in \mathbb{Z}_{26}$, số nghịch đảo của a là $a^{-1} \in \mathbb{Z}_{26}$ sao cho

$$aa^{-1} \equiv a^{-1}a \equiv 1 \pmod{26}.$$

Hệ mật Vigenère (Vigenère Cipher)

Định nghĩa 3.7 Cho m là một số nguyên dương cố định. Đặt $\mathcal{P} = \mathcal{C} = \mathcal{K} = (\mathbb{Z}_{26})^m$. Với mỗi $K = (k_1, k_2, \ldots, k_m)$, định nghĩa:

$$e_K(x_1, x_2, \ldots, x_m) = (x_1 + k_1, x_2 + k_2, \ldots, x_m + k_m)$$

và

$$d_K(y_1, y_2, \ldots, y_m) = (y_1 - k_1, y_2 - k_2, \ldots, y_m - k_m)$$

$(x_i, y_i \in \mathbb{Z}_{26}, i = 1, 2, \ldots, m)$.

Lưu ý $\rightsquigarrow$ Ta kết hợp mỗi khóa bí mật K với một xâu chữ, gọi là một *từ khóa* (keyword), có độ dài m.

Thám mã (Cryptanalysis)

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

Ta có các dạng thám mã sau:

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

Lưu ý $\rightsquigarrow$ Nguyên lý Kerckhoffs - người thám mã đã biết rõ hệ mật được sử dụng khi tiến hành phân tích mã.

Thám mã

Tùy thuộc vào hiểu biết của người thám mã đối với hệ mật được sử dụng, mức độ tấn công vào hệ mật có thể được phân loại như sau:

- **Tấn công chỉ biết bản mã** (ciphertext-only): người thám mã chỉ có bản mã.
- **Tấn công biết bản rõ** (known plaintext): người thám mã có bản rõ và bản mã.
- **Tấn công chọn bản 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 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 rõ tương ứng.

Thám mã

Nhiều kỹ thuật thám mã sử dụng đặc điểm thống kê của tiếng Anh để tiến hành phân tích mã, chẳng hạn sử dụng kết quả thống kê của Beker và Piper (1983):

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

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

Thám mã hệ mật Affine

Giả sử Trudy đã lấy được bản mã: FMXVEDKAPHFERBNDKRXRSRE FMORUDSDKDVSHVUFEDKAPRKDLYEVLRHHRH.

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

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

Thám mã hệ mật Affine

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:

???

Thám mã hệ mật thay thế

Giả sử Trudy đã lấy được bản mã:

YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR

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

chữ cái	t.suất	chữ cái	t.suất	chữ cái	t.suất	chữ cái	t.suất
A	0	H	4	O	0	V	5
B	1	I	5	P	1	W	8
C	15	J	11	Q	4	X	6
D	13	K	1	R	10	Y	10
E	7	L	0	S	3	Z	20
F	11	M	16	T	2		
G	1	N	9	U	5		

Thám mã hệ mật thay thế

Thực hiện dự đoán $d_K(Z) = e$ do Z xuất hiện nhiều nhất. Do ZW xuất hiện nhiều lần (4 lần) nên dự đoán tiếp $d_K(W) = d$. Tiếp tục phân tích để dự đoán $d_K(D) \in \{r, s, t\}$ và $d_K(R) = n$.

Đến đây, ta có:

- - - - - - e n d - - - - - - - - - e - - - - n e d - - - e - - - - - - - - - - - -

Y I F Q F M Z R W Q F Y V E C F M D Z P C V M R Z W N M D Z V E J B T X C D D U M J

- - - - - - - - e - - - - e - - - - - - - - - - n - - d - - - e n - - - - e - - - - e

N D I F E F M D Z C D M Q Z K C E Y F C J M Y R N C W J C S Z R E X C H Z U N M X Z

- e - - - n - - - - - - - n - - - - - - - e d - - - e - - - e - - n e - n d - e - e - -

N Z U C D R J X Y Y Y S M R T M E Y I F Z W D Y V Z V Y F Z U M R Z C R W N Z D Z J J

- e d - - - - - - n - - - - - - - - - - - - - e - - - - e d - - - - - - - - d - - - e - - n

X Z W G C H S M R N M D H N C M F Q C H Z J M X J Z W I E J Y U C F W D J N Z D I R

Thám mã hệ mật thay thế

Tiếp tục, thực hiện dự đoán $d_K(N) = h$ do NZ xuất hiện nhiều. Nếu đúng, cụm từ $ne - ndhe$ gợi ý $d_K(C) = a$. Dự đoán tiếp $d_K(M) = i$ hoặc o. Vì ai có vẻ hợp lý hơn ao nên $d_K(M) = i$.

Đến đây, ta có:

- - - - - i e n d - - - - - a - i - e - a - i n e d h i - e - - - - - - a - - - i -

Y I F Q F M Z R W Q F Y V E C F M D Z P C V M R Z W N M D Z V E J B T X C D D U M J

h - - - - - i - e a - i - e - a - - - - - i - n h a d - a - e n - - a - e - h i - e

N D I F E F M D Z C D M Q Z K C E Y F C J M Y R N C W J C S Z R E X C H Z U N M X Z

h e - a - n - - - - - i n - i - - - - e d - - - e - - - e - i n e a n d h e - e - -

N Z U C D R J X Y Y S M R T M E Y I F Z W D Y V Z V Y F Z U M R Z C R W N Z D Z J J

- e d - a - - i n h i - - h a i - - a - e - i - - e d - - - - - a - d - - h e - - n

X Z W G C H S M R N M D H N C M F Q C H Z J M X J Z W I E J Y U C F W D J N Z D I R

Thám mã hệ mật thay thế

Tiếp tục, thực hiện dự đoán $d_K(Y) = o$,
$d_K(D) = s, d_K(F) = r, d_K(H) = c, d_K(J) = t, \ldots$

Cuối cùng, ta nhận được:

Our friend from Paris examined his empty glass with surprise, as if evaporation had taken place while he wasn't looking. I poured some more wine and he settled back in his chair, face tilted up towards the sun.

Thám mã hệ mật Vigenère

- Để thám mã hệ mật Vigenère, trước hết cần xác định độ dài từ khóa (tức 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 ngẫu nhiên.

Lưu ý $\rightsquigarrow$ Phương pháp Kasiski (Friedrich Kasiski, 1863): 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 $d_1, d_2, \ldots$

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

Thám mã hệ mật Vigenère

Ví dụ 3.6 Xét một hệ mật Vigenère với bản rõ, từ khóa và bản mã tương ứng như sau:

- Plaintext: conghoa|danchun|handant|runghoa|sapsuat|hanghoa
- Keyword: abcdefg
- Ciphertext: CPPJLTG DBPFLZT HBPGESZ RVPJLTG SBRVYFZ HBPJLTG

Theo phương pháp Kasiski, ta tìm được xâu lặp độ dài 3 là PJL. Vị trí xuất hiện của dãy PJL lần lượt là: 3, 24, 38. Do vậy, dãy $d_1, d_2, \ldots$ là $21, 35, \ldots$; $\gcd(d_1, d_2, \ldots) = 7$.

Thám mã hệ mật Vigenère

- Phương pháp chỉ số trùng hợp ngẫu nhiên (Wolfe Friedman, 1920): Giả sử $x = x_1x_2\cdots x_n$ là xâu có n ký tự. Chỉ số trùng hợp ngẫu nhiên 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 ta ký hiệu tần suất của $A, B, C, \ldots, Z$ trong x lần lượt là $f_0, f_1, \ldots, f_{25}$. Ta có thể chọn hai phần tử của x theo $\binom{n}{2} = n!/(2!(n-2)!)$ cách. Với mỗi $0 \le i \le 25$, có $\binom{f_i}{2}$ cách chọn các phần tử là i. Vì vậy, 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

$$I_c(x) \approx \sum_{i=0}^{25} p_i^2 = 0.065.$$

Độ mật hoàn thiện

Có hai tiếp cận cơ bản để bàn về bảo mật của một hệ mật:

Bảo mật tính toán

Tiêu chuẩn này quan tâm đến nỗ lực tính toán cần thiết để phá vỡ một hệ mật. Ta có thể định nghĩa một hệ mật là *bảo mật tính toán* nếu thuật toán tốt nhất để phá vỡ nó đòi hỏi ít nhất N phép toán, với N đủ lớn nào đó.

Bảo mật vô điều kiện

Tiêu chuẩn này quan tâm đến bảo mật của các hệ mật khi người thám mã không bị giới hạn về khả năng tính toán. Một hệ mật được xem là *bảo mật vô điều kiện* nếu nó không thể bị phá vỡ với mọi nỗ lực tính toán (kể cả nguồn lực tính toán là vô hạn).

Độ mật hoàn thiện

Định nghĩa 3.8 Một hệ mật $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ có *độ mật hoàn thiện* (perfect secrecy) nếu xác suất hậu nghiệm để bản rõ là x khi đã biết bản mã y đúng bằng xác suất tiền nghiệm để bản rõ là x. Tức là $p_p(x|y) = p_p(x)$ với mọi $x \in \mathcal{P}, y \in \mathcal{C}$.

Định lý 3.6 Giả sử $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ là hệ mật thỏa mãn $|\mathcal{K}| = |\mathcal{C}| = \mathcal{P}|$. Khi đó, hệ mật này có độ mật hoàn thiện khi và chỉ khi tất cả các khóa được sử dụng với xác suất bằng $1/|\mathcal{K}|$ và với mọi $x \in \mathcal{P}$ và $y \in \mathcal{C}$, chỉ có một khóa duy nhất K sao cho $e_K(x) = y$.

Lưu ý $\rightsquigarrow$ Ta sẽ xây dựng lý thuyết hệ mã bảo mật vô điều kiện đối lại hình thức tấn công "chỉ biết bản mã". Rõ ràng công cụ độ phức tạp tính toán không thích hợp để nghiên cứu độ mật vô điều kiện bởi vì ta cho phép thời gian tính toán là vô hạn. Công cụ phù hợp để nghiên cứu là lý thuyết xác suất.

Hệ mật One-time Pad

Định nghĩa 3.9 Cho số nguyên $n \geq 1$ và đặt $\mathcal{P} = \mathcal{C} = \mathcal{K} = (\mathbb{Z}_2)^n$. Với $K \in (\mathbb{Z}_2)^n$, định nghĩa $e_K(x)$ là véctơ tổng môđun 2 của K và x (hoặc tương đương, là kết quả phép XOR của hai xâu bit biểu diễn K và x). Vì vậy, nếu có $x = (x_1, \ldots, x_n)$ và $K = (K_1, \ldots, K_n)$ thì:

$$e_K(x) = (x_1 + K_1, \ldots, x_n + K_n) \mod 2$$

và tương tự nếu $y = (y_1, \ldots, y_n)$ thì

$$d_K(y) = (y_1 + K_1, \ldots, y_n + K_n) \mod 2$$

$(x, y \in (\mathbb{Z}_2)^n)$.

Entropy và mã Huffman

- Ta đã xem xét các hệ mật có độ mật hoàn thiện mà ở đó mỗi bản rõ được mã hóa bằng một khóa duy nhất.

- Đối với các hệ mật mà nhiều bản rõ được mã hóa bằng một khóa chung thì câu hỏi đặt ra là thám mã sẽ được thực hiện như thế nào đối với kiểu tấn công chỉ biết bản mã? Để trả lời câu hỏi này, năm 1948, Shannon đề xuất khái niệm entropy - một đại lượng toán học dùng để đo độ không chắc chắn của thông tin.

- Giả sử biến ngẫu nhiên X nhận giá trị trên một miền hữu hạn theo phân phối xác suất $p(X)$. Khi đó, ta nhận được thông tin gì từ một sự kiện xảy ra theo phân phối xác suất $p(X)$? hoặc tương đương, nếu một sự kiện chưa xảy ra thì độ không chắc chắn là bao nhiêu? Độ đo này là entropy của X.

Entropy và mã Huffman

Định nghĩa 3.10 Giả sử X là một biến ngẫu nhiên nhận giá trị trên một miền hữu hạn, theo phân phối xác suất $p(X)$. Khi đó, entropy của X, kí hiệu là $H(X)$, được xác định như sau:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Nếu các giá trị có thể có của X là x_i, $1 \leq i \leq n$ thì

$$H(X) = -\sum_{i=1}^{n} p(X = x_i) \log_2 p(X = x_i).$$

Lưu ý $\rightsquigarrow$ Khái niệm entropy có liên quan mật thiết tới hiệu quả của phép mã hóa.

Định nghĩa 3.11 Giả sử X là một biến ngẫu nhiên nhận giá trị trên một miền hữu hạn theo phân phối xác suất $p(X)$. Một phép *mã hóa* của X là ánh xạ

$$f : X \to \{0,1\}^*.$$

Cho trước dãy sự kiện $x_1x_2\cdots x_n$, ta có thể mở rộng phép mã hóa f như sau:

$$f(x_1x_2\cdots x_n) = f(x_1) \,||\, \cdots \,||\, f(x_n).$$

Lưu ý $\rightsquigarrow$ Giả sử xâu $x_1x_2\cdots x_n$ phát sinh từ một nguồn thông tin (không nhớ). Khi đó, các giá trị x_i không nhất thiết phải khác nhau. Hơn nữa, ta có

$$p(x_1x_2\cdots x_n) = p(x_1) \times p(x_2) \times \cdots \times p(x_n).$$

Entropy và mã Huffman

Giả sử ta muốn dùng f để mã hóa các thông điệp thì điều cần thiết là f đơn ánh và f hiệu quả.

Hiệu quả của phép mã hóa f được đo lường thông qua độ dài trung bình các từ mã của X. Kí hiệu độ dài này là $l(f)$, ta có

$$l(f) = \sum_{x \in X} p(x)|f(x)|.$$

Vậy, khi $l(f)$ nhỏ nhất thì f đạt hiệu quả tối ưu. Phép mã hóa hay thuật toán Huffman (1952) đáp ứng điều kiện này. Cụ thể, phép mã hóa Huffman f thỏa mãn

$$H(X) \leq l(f) < H(X) + 1.$$

Thuật toán 3.2 (Sinh mã Huffman)

Đầu vào:	Dãy $x_1, x_2, \ldots, x_n, x_i \in X$ được sắp theo thứ tự giảm dần của $p(x_i)$
Đầu ra:	Dãy $\omega_1, \omega_2, \ldots, \omega_n$, ω_i là mã Huffman của $x_i, i = \overline{1,n}$
Bước 1:	Đặt $\omega_i = \emptyset, i = \overline{1,n}$
Bước 2 (lặp):	Hai phần tử có xác suất nhỏ nhất lập thành một phần tử mới có xác suất là tổng xác suất của hai phần tử này. Phần tử có xác suất nhỏ hơn được gán giá trị 0, phần tử có xác suất lớn hơn được gán giá trị 1 và cập nhật các ω_i ($\omega_i = 0 \,\|\, \omega_i$ hoặc $\omega_i = 1 \,\|\, \omega_i$) liên quan. Tiếp tục thực hiện tới khi dãy chỉ còn một phần tử.
Bước 3:	Trả về dãy $\omega_1, \omega_2, \ldots, \omega_n$.

Ví dụ 3.7 Giả sử đầu vào của Thuật toán 3.2 là $X = \{a, b, c, d, e\}$ có phân phối xác suất: $p(a) = 0.60$, $p(b) = 0.13$, $p(c) = 0.12$, $p(d) = 0.10$ và $p(e) = 0.05$. Hoạt động của thuật toán như sau

<table>
<tr><th>a</th><th>b</th><th>c</th><th>d</th><th>e</th><th>ω_i</th></tr>
<tr><td>0.60</td><td>0.13</td><td>0.12</td><td>0.10</td><td>0.05</td><td></td></tr>
<tr><td></td><td></td><td></td><td>1</td><td>0</td><td>$\omega_5 = 0, \omega_4 = 1$</td></tr>
<tr><td>0.60</td><td>0.13</td><td>0.12</td><td colspan="2">0.15</td><td></td></tr>
<tr><td></td><td>1</td><td>0</td><td colspan="2"></td><td>$\omega_3 = 0, \omega_2 = 1$</td></tr>
<tr><td>0.60</td><td colspan="2">0.25</td><td colspan="2">0.15</td><td></td></tr>
<tr><td></td><td colspan="2">1</td><td colspan="2">0</td><td>$\omega_5 = 00, \omega_4 = 01, \omega_3 = 10, \omega_2 = 11$</td></tr>
<tr><td>0.60</td><td colspan="4">0.40</td><td></td></tr>
<tr><td>1</td><td colspan="4">0</td><td>$\omega_5 = 000, \omega_4 = 001, \ldots, \omega_1 = 1$</td></tr>
<tr><td colspan="5">1.0</td><td></td></tr>
</table>

TRÂN TRỌNG CẢM ƠN!