第三章 Shannon理论

3.1 密码体制的数学模型

3.2 熵及其性质

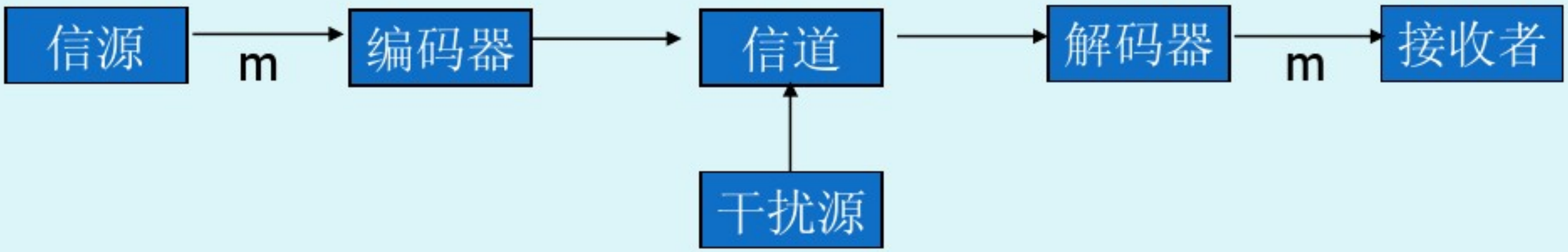
3.3 伪密钥和惟一解距离

3.4 密码体制的完善保密性

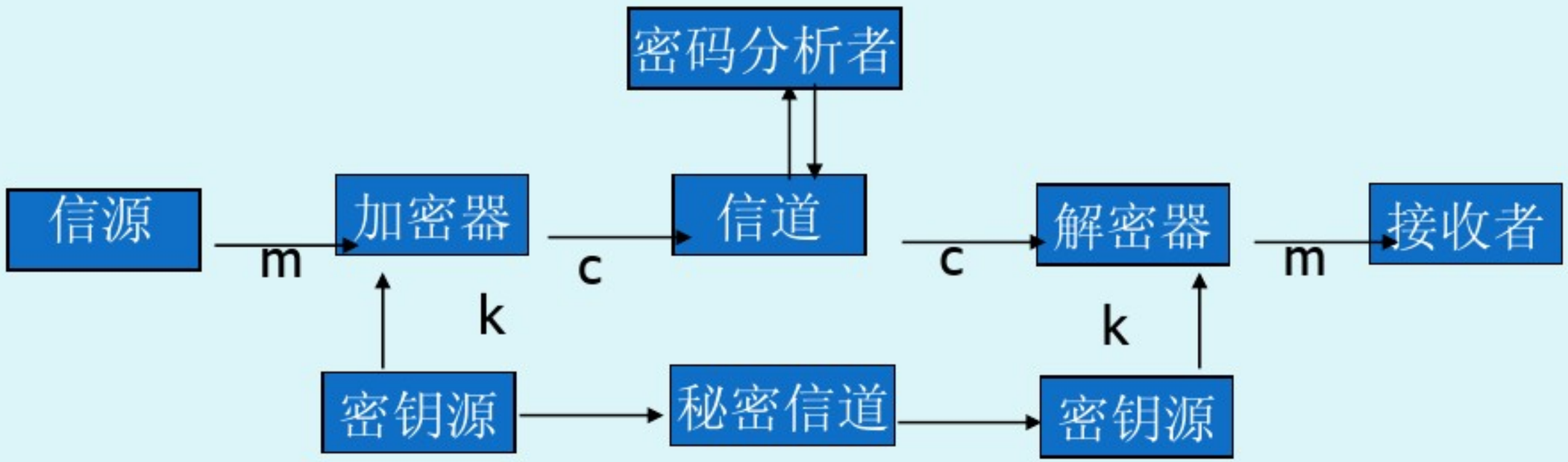
3.5 乘积密码体制

3.1 密码体制的数学模型

密码注定与通信密不可分,不需通信也就不需要密码了。**通信系统：** 设计目的是在信道有干扰的情况下,使接收的信息无差错或差错尽可能地小。



保密系统： 设计目的是使窃听者即使完全准确地接收到了信道上的传输信号也无法恢复原始信息。



信源字母表： $X=\{a_i \mid i=0,1,2,\cdots,q-1\}$ 且 a_i 的概率为

$$\Pr(a_i),0\le \Pr(a_i)\le 1,0\le i\le q-1,\quad \sum_{i=0}^{q-1}\Pr(a_i)=1.$$

明文空间： $M=\{m=(m_1,m_2,\cdots,m_r)\mid m_i\in X,1\le i\le r\}$

1. 如果信源是无记忆的，则

$$\Pr(m)=\Pr(m_1,m_2,\cdots,m_r)=\prod_{i=0}^{r}\Pr(m_i)$$

2.如果信源是有记忆的，则需要考虑明文空间中各元素的概率分布。

密钥源字母表：$B=\{b_i|i=0,1,2,\cdots,p-1\}$，其中$b_i$ 的概率为

$$\Pr(b_i),0\le\Pr(b_i)\le 1,0\le i\le p-1,\quad \sum_{i=0}^{p-1}\Pr(b_i)=1.$$

长为s的密钥：$K=\{k=(k_1,k_2,\cdots,k_s)|k_i\in B,1\le i\le s\}$

加密变换：将明文在密钥的控制下变为密文，即

$$(c_1,c_2,\cdots,c_t)=E_k(m_1,m_2,\cdots,m_r)$$

t和r为何不等呢？

密文空间的统计特性由明文空间和密钥空间的统计特性决定：

由于明文空间与密钥空间是相互独立的，则$\Pr(c)=\sum_{k\in\{k|c\in C_k\}}\Pr(k)\Pr(D_k(c))$

又因为 $\Pr(c|m)=\sum_{k\in\{k|m=D_k(c)\}}\Pr(k)$

故由Bayes公式知

$$\Pr(m|c)=\frac{\Pr(m)\Pr(c|m)}{\Pr(c)}=\frac{\Pr(m)\sum_{k\in\{k|m=D_k(c)\}}\Pr(k)}{\sum_{k\in\{k|c\in C_k\}}\Pr(k)\Pr(D_k(c))}$$

从以上公式可知，知道明文空间和密钥空间的概率分布，就可确定密文空间的概率分布，密文空间关于明文空间的概率分布，以及明文空间关于密文空间的概率分布。

加解密运算可记为$C_k=\{E_k(m)\in C|m\in M\}\quad m=D_k(c)=D_k(E_k(m))$

3.2 熵及其性质

- 定义3.1 $$H(X) \overset{def}{=} -\sum_{i=1}^{n} \Pr(x_i) \log_2 \Pr(x_i)$$

  称为随机变量X的熵。

- 定义3.2 $$H(X,Y) \overset{def}{=} -\sum_{i=1}^{n}\sum_{j=1}^{m} \Pr(x_i, y_j) \log_2 \Pr(x_i, y_j)$$

  称为随机变量X和Y的联合熵。

- 定义3.3 $$H(X \mid y_j) \overset{def}{=} -\sum_{i=1}^{n} \Pr(x_i \mid y_j) \log_2 \Pr(x_i \mid y_j)$$

  称为X在Y取值 y_j 时的条件熵。

- 定义3.4 $$H(X \mid Y) \overset{def}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} \Pr(x_i, y_j) \log_2 \Pr(x_i \mid y_j) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \Pr(y_j) \Pr(x_i \mid y_j) \log_2 \Pr(x_i \mid y_j)$$

  称为X关于Y的条件熵。 $$= \sum_{j} \Pr(y_j) H(X \mid y_j) = \sum_{i} \Pr(x_i) H(Y \mid x_i)$$

- 定义3.5　一个实值函数f 称为在区间I上是凸(严格凸)的，如果对任意

$$x, y \in I \quad \text{,都有} \frac{f(x)+f(y)}{2} \leq (<) f(\frac{x+y}{2}) \quad .$$

- 引理3.1（Jensen不等式）设f是区间I上的一个连续的严格 凸函数

$$, \sum_{i=0}^{n} a_i = 1, \quad a_i > 0, 1 \leq i \leq n. \text{ 则 } \sum_{i=0}^{n} a_i f(x_i) \leq f(\sum_{i=1}^{n} a_i x_i), \quad x_i \in I, 1 \leq i \leq n$$

上式中的等号成立当且仅当 $x_1 = x_2 = \cdots = x_n.$

- 定理3.1 任意信息源X,其熵值满足: $0 \leq H(X) \leq \log_2 n$

$H(X) = 0$　当且仅当存在一个 $\Pr(x_i) = 1,\ 1 \leq i \leq n$ 而对其他 $j \neq i,\ \Pr(x_i) = 0$

$H(X) = \log_2 n,$ 当且仅当对任意 $1 \leq i \leq n$ ， 都有 $\Pr(x_i) = \frac{1}{n}.$

定理3.1证明过程如下：

$$H(X)=-\sum_i \Pr(x_i)\log_2^{\Pr(x_i)}=\sum_i \Pr(x_i)\log_2^{\frac{1}{\Pr(x_i)}}\le \log_2^{\sum_i \Pr(x_i)\frac{1}{\Pr(x_i)}}=\log_2^n$$

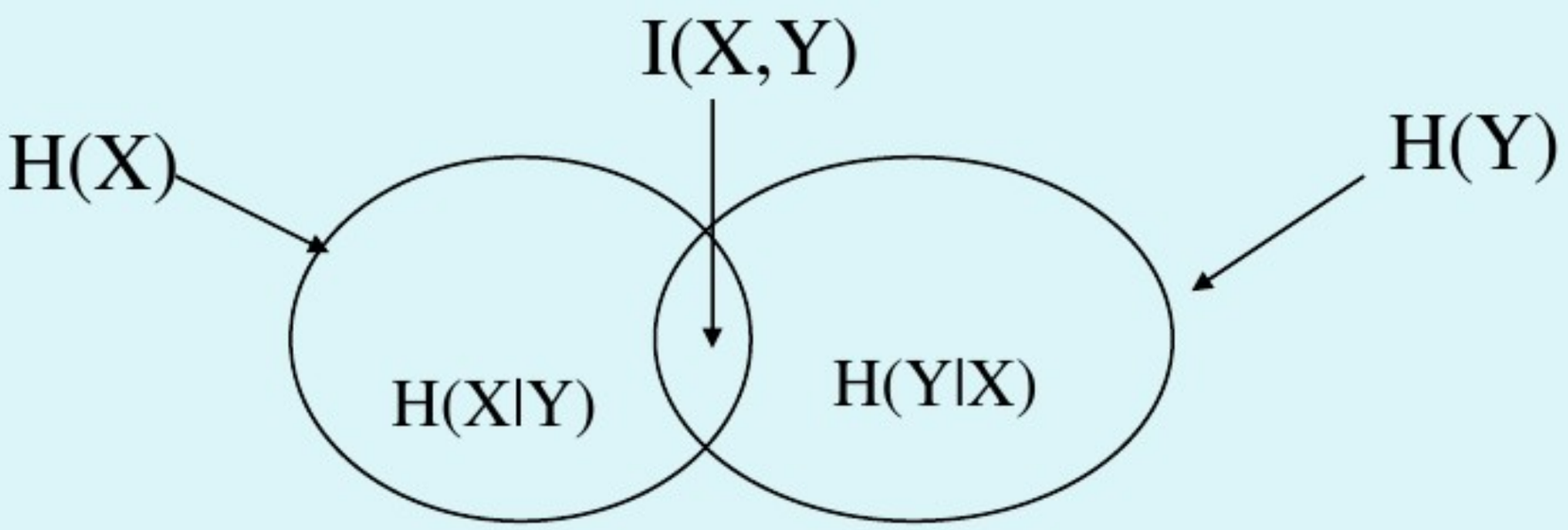
- 定理3.2　　$H(X,Y) \le H(X)+H(Y)$

  等号成立当且仅当X与Y相互独立。

- 定理3.3

$$H(X,Y)=H(Y)+H(X|Y)=H(X)+H(Y|X)$$

- 推论3.1　　$H(X|Y) \le H(X)$　　等号成立当且仅当X与Y相互独立.



定理3.2证明过程如下：

$$H(X)+H(Y)=-\sum_i \Pr(x_i)\log_2^{\Pr(x_i)}-\sum_j \Pr(y_j)\log_2^{\Pr(y_j)}$$

$$=-\sum_i\sum_j \Pr(x_i,y_j)\log_2^{\Pr(x_i)\Pr(y_j)}$$

$$H(X,Y)-H(X)-H(Y)$$

$$=\sum_i\sum_j \Pr(x_i,y_j)\log_2^{\frac{1}{\Pr(x_i,y_j)}}+\sum_i\sum_j \Pr(x_i,y_j)\log_2^{\Pr(x_i)\Pr(y_j)}$$

$$=\sum_i\sum_j \Pr(x_i,y_j)\log_2^{\frac{\Pr(x_i)\Pr(y_j)}{\Pr(x_i,y_j)}}$$

$$\le \log_2^{\sum_i\sum_j \Pr(x_i)\Pr(y_j)} = \log 1 = 0$$

- 例3.1 设$\Phi=(M,C,K,\varepsilon,D)$是一个密码体制，明文空间$M=\{a,b\}$密文空间 $C=\{1,2,3,4\}$，密钥空间 $K=\{k_1,k_2,k_3\}$

  加密变换为

$$E_{k_1}(a)=1 \quad E_{k_2}(a)=2 \quad E_{k_3}(a)=3$$
$$E_{k_1}(b)=2 \quad E_{k_2}(b)=3 \quad E_{k_3}(b)=4$$

设明文的概率分布 $\Pr(a)=\frac{1}{4}$ $\Pr(b)=\frac{3}{4}$

密钥的概率分布为 $\Pr(k_1)=\frac{1}{2}$ $\Pr(k_2)=\frac{1}{4}$ $\Pr(k_3)=\frac{1}{4}$

设m是明文空间 M上的随机变量，c是密文空间 C 上的随机变量，k是密钥空间 K上的随机变量。下面来计算熵

$$H(M), H(K), H(C), H(M|C), H(K|C)$$

根据熵的定义，我们有

$$H(M) = -\Pr(a)\log_2 \Pr(a) - \Pr(b)\log_2 \Pr(b)$$

$$= -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4} \approx 0.81$$

$$H(K) = -\Pr(k_1)\log_2 \Pr(k_1) - \Pr(k_2)\log_2 \Pr(k_2) - \Pr(k_3)\log_2 \Pr(k_3)$$

$$= -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{4}\log_2 \frac{1}{4} - \frac{1}{4}\log_2 \frac{1}{4} = \frac{3}{2}$$

为了计算$H(C)$，我们需要先计算密文的概率分布。假设明文和密钥相互独立是合理的。根据

$$\Pr(c) = \sum_{k \in \{k | c \in C_k\}} \Pr(k)\Pr(D_k(c))$$

有

$$\Pr(2)=\Pr(a)\Pr(k_2)+\Pr(b)\Pr(k_1)=\frac{7}{16}$$

$$\Pr(3)=\Pr(a)\Pr(k_3)+\Pr(b)\Pr(k_2)=\frac{1}{4}$$

$$\Pr(4)=\Pr(b)\Pr(k_3)=\frac{3}{16}$$

于是

$$H(C)=-\frac{1}{8}\log_2\frac{1}{8}-\frac{7}{16}\log_2\frac{7}{16}-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{16}\log_2\frac{3}{16}=1.85$$

为计算 $H(M|C)$，需要先计算在已知密文的情况下明文的概率分布，则

$$\Pr(1|a)=\Pr(k_1)=\frac{1}{2}\qquad \Pr(1|b)=0\qquad \Pr(2|a)=\Pr(k_2)=\frac{1}{4}\qquad \Pr(2|b)=\Pr(k_1)=\frac{1}{2}$$

$$\Pr(3|a)=\Pr(k_3)=\frac{1}{4}\qquad \Pr(3|b)=\Pr(k_2)=\frac{1}{4}\qquad \Pr(4|a)=0\qquad \Pr(4|b)=\Pr(k_3)=\frac{1}{4}$$

再由Bayes公式 $\Pr(x|y)=\dfrac{\Pr(x)\Pr(y|x)}{\Pr(y)}$

可计算得

$$\Pr(a|1)=1 \qquad \Pr(b|1)=0$$

$$\Pr(a|2)=\frac{1}{7} \qquad \Pr(b|2)=\frac{6}{7}$$

$$\Pr(a|3)=\frac{1}{4} \qquad \Pr(b|3)=\frac{3}{4}$$

$$\Pr(a|4)=0 \qquad \Pr(b|4)=1$$

于是 $H(M|C)=\Pr(1)H(M|1)+\Pr(2)H(M|2)+\Pr(3)H(M|3)+\Pr(4)H(M|4)$

≈ 0.46

最后计算 $H(K|C)$. 首先计算在已知密文的情况下密钥的概率分布。

$$\Pr(1|k_1)=\Pr(a)=\frac{1}{4} \qquad \Pr(1|k_2)=0 \qquad \Pr(1|k_3)=0$$

$$\Pr(2|k_1)=\Pr(b)=\frac{3}{4} \qquad \Pr(2|k_2)=\Pr(a)=\frac{1}{4} \qquad \Pr(2|k_3)=0$$

$$\Pr(3|k_1)=0 \qquad \Pr(3|k_2)=\Pr(b)=\frac{3}{4} \qquad \Pr(3|k_3)=\Pr(a)=\frac{1}{4}$$

$$\Pr(4|k_1)=0 \qquad \Pr(4|k_2)=0 \qquad \Pr(4|k_3)=\Pr(b)=\frac{3}{4}$$

由Bayes公式知：

$$\Pr(k_1|1)=1 \qquad \Pr(k_2|1)=0 \qquad \Pr(k_3|1)=0$$

$$\Pr(k_1|2)=\frac{6}{7} \qquad \Pr(k_2|2)=\frac{1}{7} \qquad \Pr(k_3|2)=0$$

$$\Pr(k_1|3)=0 \qquad \Pr(k_2|3)=\frac{3}{4} \qquad \Pr(k_3|3)=\frac{1}{4}$$

$$\Pr(k_1|4)=0 \qquad \Pr(k_2|4)=0 \qquad \Pr(k_3|4)=1$$

于是，可计算出

$$H(K|C)=\Pr(1)H(K|1)+\Pr(2)H(K|2)+\Pr(3)H(K|3)+\Pr(4)H(K|4) \quad =0.46$$

3.3 伪密钥和惟一解距离

- 定理3.4 设 $\Phi=(M,C,K,\varepsilon,D)$ 是一个密码体制，

则 $H(K|C)=H(K)+H(M)-H(C)$

这里M,K,C分别是 M，K，C 上的随机变量。

证明：$H(K,M,C)=H(C|K,M)+H(K,M)=H(M|K,C)+H(K,C)$

$H(C|K,M)=H(M|K,C)=0$

$\therefore H(K,M)=H(K,C)$

$H(K,C)=H(C)+H(K|C)$

$H(K,M)=H(M)+H(K|M)$,而$H(K|M)=H(K)$,K与M独立

$\therefore H(K|C)=H(M)+H(K)-H(C)$

上述定理说明了知道密文确定密钥的不确定性相当大。密码分析者可能得到很多可能的密钥，但其中只有一个是正确的，其他的密钥我们都称为是**伪密钥**。对于任意密文用不同的密钥解密，如果得到有意义明文越多，则表明伪密钥越多，从而增加判断真正唯一密钥的难度。从密码分析师的角度来看当然希望伪密钥个数为0,即$H(K|C)=H(K)+H(M)-H(C)=0$.

定义3.5　设L是一种自然语言，$H_L \overset{def}{=} \lim_{n\to\infty} \frac{H(X^{(n)})}{n}$　称为语言L的熵。

$R_L \overset{def}{=} 1 - \frac{H_L}{\log_2^{|X|}}$　称为语言L的冗余度。

其中X表示语言L的字母表，X(n)表示Xn上的随机变量。

H_L表示自然语言L中每个字母所携带的平均信息量的度量。

如果该自然语言的关联度不大的话，语言中前后字母的搭配没有相关性，相当于每个字母都是独立无关的，这时

$$H_L = \lim_{n\to\infty} \frac{H(X^n)}{n} = \lim_{n\to\infty} \frac{-\sum_{x\in X^n} \Pr(x)\log_2 \Pr(x)}{n} = \frac{-\sum_{x\in X^n} \frac{1}{|X|^n}\log_2 \frac{1}{|X|^n}}{n} = \frac{n\log_2^{|X|}}{n} = \log_2^{|X|}$$

而实际上，任何自然语言前后字母均有一定的关联度，前面字母可能反应下一个字母，这时冗余度导致H_L实际值远小于log|X|。H_L实际值越接近log|X|，表明该语言的冗余度越小，反之越大。

- 定理3.5 设 $\Phi=(M,C,K,\varepsilon,D)$ 是一个密码体制，设X是明文字母表，Y是密文字母表，并且 $|X|=|Y|$ ，设 R_L 是明文语言的冗余度。假设密钥的选取满足均匀分布。则对于任意一个长度为n的密文字母串，当n充分大时，伪密钥的期望数目 $\overline{S}_n$ 满足 $\overline{s}_n \geq \frac{|K|}{|X|^{nR_L}}-1$

该定理主要讨论了伪密钥的平均数目的下界，伪密钥的个数主要与密钥空间的大小基本成正比，而与冗余度成反比。冗余度越大，表明语言的关联度大，语言关联度能帮助我们识别真正密钥。

该定理有两个注意的地方，第一，|X|=|Y|,这个条件一般自然语言均满足，第二是当密文长度n本大的情况下，该定理对伪密钥平均数目的猜测是不准确的。

定义3.6 一个密码体制的唯一解距离unicity distance定义为使得伪密钥的期望数目s_n等于零的密文长度n_0.

$$令\overline{s_n}=0,\ 则\frac{|K|}{|X|^{nR_L}}-1=0,可计算出 n_0\approx\frac{\log_2^{|K|}}{R_L\log_2^{|X|}}.$$

一个密码体制的唯一解距离就是密码分析者在足够的计算时间的情况下，能够惟一的计算处正确 密钥所需的密文的平均长度。

分析上面n_0的表达式，语言的冗余度越大，唯一解距离越小，密文分析者在进行唯密文攻击时越容易得到正确密钥。因此从编码者的角度来讲，要提高密码体制的安全性，应尽量减少明文语言的冗余度，即在对明文加密之前，最后用huffman编码对明文进行一次压缩。

也就是用最简短的语言表达精要的内容!!(发电报)

3.4 密码体制的完善保密性

- 定义3.7 设 $\Phi=(M,C,K,\varepsilon,D)$ 是一个密码体制。如果对任意 $x\in M$ 和任意 $y\in C$，都有 $\Pr(x|y)=\Pr(x)$

  则 Φ 称具有完善的保密性(perfect secrecy)。

- 显然，一个密码体制具有完善的保密性当且仅当明文与密文是相互独立的．也就是说，在唯密文攻击的情况下，从密文得不到关于明文的任何信息。

- 定理3.6 设$\Phi=(M,C,K,\varepsilon,D)$是一个密码体制, 并且 $|M|=|C|\leqslant|K|$ 则密码体制 Φ 具有完善的保密性当且仅当密钥的选取满足均匀分布, 并且对任意 $x\in M$ 和任意$y\in C$, 都存在惟一的密钥$k\in K$使得

$$E_k(x)=y.$$

- “一次一密”密码体制

在 Vernam 密码体制中，如果对不同的明文用不同的密钥进行加密，则这时的 Vernam 体制就是所谓的“一次一密”密码体制。对于“一次一密”密码体制，密码分析人员无法仅从密文获得关于明文或密钥的任何信息， 即使密码分析人员获得了一些密文所对应的明文，他也只是能得到这些密文所对应的密钥，而不能获得其他密文所对应的明文或密钥。因此，“一次一密”密码体制在理论上被认为是不可破译的。但在实际应用中 ，“一次一密”密码体制要求每传送一个明文，都必须产生一个新的密钥并通过一个安全的信道传送给接收方， 这给密钥管理带来了一定的困难。因此，“一次一密 ” 密码体制并不很实用，具有很大的局限性。

3.5 乘积密码体制

- 定义3.8 设 $\Phi_1=(M,M,K_1,\varepsilon_1,D_1)$和 $\Phi_2=(M,M,K_2,\varepsilon_2,D_2)$ 是两个密码体制。它们的明文空间和密文空间全相同。Φ_1 和 Φ_2的乘积定义为密码体制 $(M,M,K_1\times K_2,\varepsilon,D)$ 记为 $\Phi_1\times\Phi_2$ 。
- 加密变换：对任意的明文 $x\in M$ 和密钥 $k=(k_1\times k_2)\in K_1\times K_2$ 加密变换为$E_k(x)=E_{k_2}(E_{k_1}(x))$.
- 解密变换：对任意的密文 $y\in M$ 和密钥 $k=(k_1\times k_2)\in K_1\times K_2$ 解密变换为 $D_k(y)=D_{k_1}(D_{k_2}(y))$.
- 对于乘积体制中密钥空间的概率分布，假设 K_1 中密钥的选取和 K_2 中密钥的选取是相互独立的。因此，对任意的 $k=(k_1\times k_2)\in K_1\times K_2$, 都有 $\Pr(k_1,k_2)=\Pr(k_1)\Pr(k_2)$.

- 定义3.9 如果 $\Phi_1 \times \Phi_2 = \Phi_2 \times \Phi_1$ 也就是说乘积密码体制 $\Phi_1 \times \Phi_2$ 和 $\Phi_2 \times \Phi_1$ 是两个相同的密码体制, 则我们称密码体制 Φ_1 和 Φ_2 是可交换的。
- 显然,密码体制的乘积运算满足结合律,即对任意的 Φ_1, Φ_2, Φ_3 ,都有 $(\Phi_1 \times \Phi_2) \times \Phi_3 = \Phi_1 \times (\Phi_2 \times \Phi_3)$,这里 Φ_1, Φ_2 和 Φ_3 是明文空间和密文空间全相同的密码体制
- 定义3.10 设 Φ 是一个明文空间和密文空间相同的密码体制，则 $\Phi^n \overset{def}{=} \underbrace{\Phi \times \Phi \times \cdots \times \Phi}_{n}$ 称为迭代密码体制。如果 $\Phi^2 = \Phi$ ，则称其是幂等的密码体制。

The End ! Thanks a lot !