基于信息论的密码学

（Cryptography Based on Information Theory）

1949 年，Shannon发表了题为“保密系统的通信理论”的论文，把信息论引入密码学中，使得信息论成为研究密码学的一个重要理论基础，并将已有数千年历史的密码学推上了科学的轨道，形成了科学的私钥密码学理论，很多学者将此后的时代称为科学密码学时代。

1976 年，Diffie-Hellman发表了“密码学的新方向”，使得科学密码学进入一个新时期：非对称密码体制时期。

保密通信系统设计的目的是使攻击者即使在完全准确地接收到信号的情况下也无法恢复出原始消息。

本章目标：在信息论的框架下，回答一个基本问题：保密通信系统在什么条件下具有无条件安全性（理论安全性）？

安全性

密码系统有两种安全性标准：

一是无条件安全性（理论安全性、完善保密性、完全保密性），指破译者具有无限时间、截获足够多的密文、具有无限计算资源下的抗破译能力；

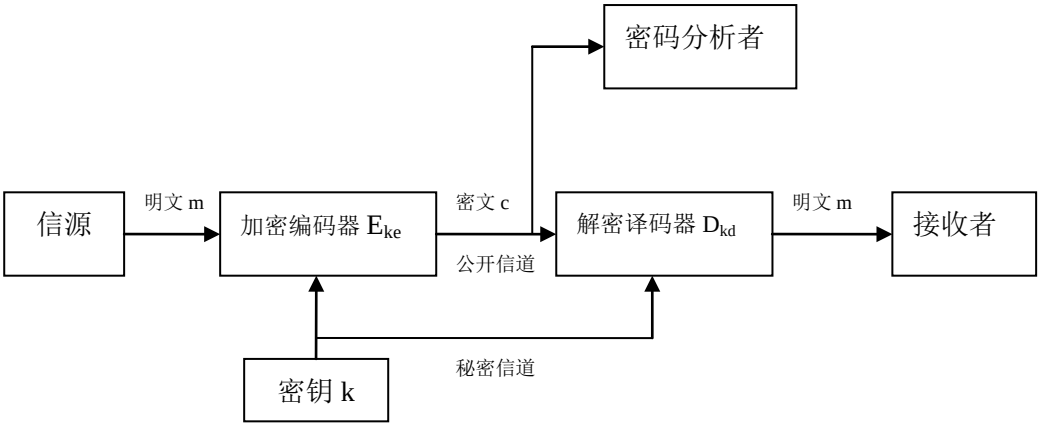
二是实用安全性（实际安全性），指在破译者仅有一定计算资源及其它实际限制下的抗破译能力。可以分为计算安全性（computational secure）和可证明安全性（provable secure）：

计算安全性：如果破译一个系统在原理上是可行的，但是用已知的算法和现有计算工具不可能完成所要求的计算量，就称其为计算上安全的。

可证明安全性：如果能够证明破译某体制的困难性等价于解决某个数学难题，就称其为可证明安全的。

计算安全性和可证明安全性都是从计算量来考虑问题的，不完全相同。

- 计算安全性要算出或估计出破译的计算量下限。
- 可证明安全性要从理论上证明破译的计算量不低于解已知难题的计算量。



$$c = E_{ke}(m)$$
$$m = D_{kd}(c) = D_{kd}(E_{ke}(m))$$
$$D_{kd} \circ E_{ke} = I$$

图：保密通信系统模型

1976 年以前研究的密码体制都具有特征：$k_d = k_e$ 或两者之间存在简单的关系，称这样的密码体制为对称密码体制或单钥密码体制。1976 年以后，出现了一种新的密码体制，其特征是 $k_d \neq k_e$，并且由 k_e 不能简单地计算出 k_d，其计算难度通常都是与一些数学难题的求解有关，称这样的密码体制为非对称密码体制或双钥密码体制。

	加密解密	密钥分发	数字签名	代表体制
对称密码体制	速度快	难	方案复杂	DES,AES,IDEA
非对称密码体制	速度慢	易	方案简单	DH, RSA,ECC
混合密码体制	用非对称密码体制做密钥分发，用对称密码体制做数据的加密和解密			

一个安全的密码系统通常应满足如下条件：

1. 系统即使不是理论上不可破译，至少也应当是实用上不可破译；
2. 系统保密性不是依赖于加密算法与解密算法，而是依赖于密钥的保密性；
3. 加密运算、解密运算简单快速，易于实现；
4. 密钥量适中，密钥的分配、管理方便。

密码分析（cryptanalysis）

密码系统的攻击类型，按照攻击者可获取的信息量来决定.

攻击类型	攻击者掌握的内容			例子
	算法	密文	其它信息	
1. 惟密文攻击（ciphertext only）	加 密算法	截获的部分密文		
2. 已知明文攻击（known plaintext）	加 密算法	截获的部分密文	一个或多个明文-密文对	有固定格式或某些固定内容的文件、文档
3. 选择明文攻击(chosen plaintext)	加 密算法	截获的部分密文	攻击者选择的明文消息以及由密钥产生的相应密文	攻击者有机会使用密码机，在源系统中插入明文，可以蓄意插入揭示密钥结构的消息模式
4. 选择密文攻击(chosen ciphertext)	加 密算法	截获的部分密文	攻击者选择的密文消息以及相应的被解密的明文	攻击者有机会使用密码机，可选择一些密文，并产生明文
5. 选择文本攻击（chosen text）	加 密算法	截获的部分密文	3 和 4 的结合	3 和 4 的结合

按攻击手段分类

攻击类型	英文术语		
分组密码的分析方法			
1. 强力攻击	Brute-force approach	尝试所有可能的密钥	增大密钥空间（key space）
2. 统计测试	Statistical test approach	统计符号模式的频度	扩散、混淆等
3. 差分密码分析	Differential Cryptanalysis	选择明文攻击中选择一些特殊的模式	
4. 线性密码分析	Linear cryptanalysis		
5. 差分--线性密码分析	Differential-Linear Cryptanalysis		
6. 插值攻击	Interpolation Cryptanalysis		
7. 密钥相关攻击			
序列密码的分析方法			
公钥密码的分析方法			

密码学的三个阶段

阶段		假设前提	体制	关键人物
第一阶段 1949 年以前	古典密码学时期	算法不公开 密钥不公开	对称体制	
第二阶段 1949 年以后	科学密码学的开始	算法公开* 密钥不公开	对称体制 （单钥体制）	Shannon
第三阶段 1976 年以后	公钥密码学的开始	算法公开* 加密密钥公开	非对称体制 （双钥体制） 混合体制	Dieffie-Hellman

*密码体制的算法公开，也被称为 Kerckhoff 假设

完全保密系统

下面针对惟密文攻击（ciphertext only）研究对称密码系统的理论安全性。

定义【完全保密】密码系统 (M,B,K,E_k,D_k) 称为完全保密，是指对一切 $m_i \in M, c_j \in B, p(c_j)>0$，有 $p(m_i \mid c_j)=p(m_i)$

即 $I(M;B)=I(B;M)=0$

定理【完全保密的充要条件】密码系统 (M,B,K,E_k,D_k) 为完全保密的充要条件是指对一切 $m_i \in M, c_j \in B, p(c_j)>0$，有 $p(c_j \mid m_i)=p(c_j)$

证明：因为 $p(c_j \mid m_i)p(m_i)=p(m_i \mid c_j)p(c_j)=p(m_i,c_j)$

所以 $\forall m_i \in M, c_j \in B, p(c_j)>0$，$p(m_i \mid c_j)=p(m_i) \Leftrightarrow p(c_j \mid m_i)=p(c_j)$

上述定理表明：对于每个密文 $c_j \in B, p(c_j)>0$，则对任意的 $m_i \in M$，总存在密钥 k，满足 $E_k(m_i)=c_j$，其理由是 $p(c_j \mid m_i)=p(c_j)>0$。

定理【数量关系】在一个完全保密的密码系统 (M,B,K,E_k,D_k) 中，

（1）$|B| \geq |M|$

（2）$|K| \geq |M|$，不同的密钥数目不会少于不同明文的数目。

在证明上述定理之前，首先说明该定理的用途：对于一个分组密码体制而言，若密钥串的长度为 l_K，则明文数据的分组长度 l_M 最好不要超过 l_K，即 $l_M \leq l_K$。

下面证明该定理。

证明：

（1）$\forall k \in K$，由于加密变换为一一变换，

故有 $\forall m_i, m_j \in M, m_i \neq m_j \Rightarrow E_k(m_i) \neq E_k(m_j)$，从而密文数目不会少于明文数目，即 $|B| \geq |M|$。

（2）

$\forall c \in B, p(c) > 0$

$\Rightarrow$

$\forall m_i \in M, \exists k_i \in K, s.t.$

$E_{k_i}(m_i) = c$

并且

若 $i \neq j$，则 $k_i \neq k_j$，

否则 $E_{k_i} = E_{k_j}$ 将两个不同的明文 m_i, m_j 变换成同一个密文 c，

与“编码变换是一一对应的”相矛盾。

从而加密变换数不会少于明文的数目，即密钥数至少同明文数目相等 $|K| \geq |M|$

定理【特殊数量关系】在一个密码系统 (M, B, K, E_k, D_k) 中，若 $|M| = |B| = |K| = n$，则该密码为完全保密的充要条件是：

（1）

$\forall m_i \in M, \forall c_j \in B, \exists | k, s.t.$

$E_k(m_i) = c_j$

（2）

$\forall k \in K, p(k) = \frac{1}{n}$

证明：

充分性证明。

若（1）（2）成立，则对 $\forall m_i \in M, \forall c_j \in B$ 有 $p(c_j \mid m_i) = p(k) = \frac{1}{n}$

$$p(c_j) = \sum_{i=1}^{n} p(c_j \mid m_i) p(m_i) = \frac{1}{n} \sum_{i=1}^{n} p(m_i) = \frac{1}{n} = p(c_j \mid m_i)$$

故由**定理【完全保密的充要条件】知，该密码系统为完全保密系统。**

必要性证明。

（1）若该密码系统是完全保密的，由于 $|M| = |B|$，而每个 E_k 都是一一变换，故 $\forall c_j \in B, p(c_j) > 0$。

进而由**完全保密的充要条件，**对任意的 $m_i \in M$, 至少存在一个密钥 k，满足 $E_k(m_i) = c_j$。

（注解：若用二部图来解释，任意两对顶点之间有边相连）

又若有两个不同的 $k_1, k_2 \in K$，而 $E_{k_1}(m_i) = E_{k_2}(m_i) = c_j$

则由于$\forall m \in M$，均可对应为任一个$c \in B$，此时对于不同的c需要不同的E_k，故至少有$|B|$个不同的E_k，使得m在它们的映射之下互不相同。

而$|K| = |B|$，故不可能有$k_1 \neq k_2$，且$E_{k_1}(m_i) = E_{k_2}(m_i) = c_j$的情况产生，与假设矛盾，从而（1）成立。

（2） 又$\forall j$，有$p(c_j \mid m_1) = p(c_j \mid m_2) = \ldots = p(c_j \mid m_n) = p(c_j)$，

而$p(c_j \mid m_i)$表明将明文m_i加密成密文c_j的概率，它等于将明文m_i加密成密文c_j的所有密钥的概率之和，即：$p(c_j \mid m_i) = \sum_{k:E_k(m_i)=c_j} p(k)$

由（1）可知，满足$E_k(m_i) = c_j$的k只有一个，记为k_{ij}，所以

$$p(k_{ij}) = p(c_j \mid m_i) = p(c_j)$$

当m_i跑遍明文空间时，k_{ij}跑遍密钥空间，所以密钥等概率。

定理【存在性】完全保密系统存在

提示：构造性证明$c = E_k(m) = m \oplus k = (m_1 \oplus k_1, m_2 \oplus k_2, \ldots, m_N \oplus k_N)$

唯一解距离

本节讨论在惟密文攻击下破译一个密码系统时密码分析者必须处理的密文量的理论下界。Shannon从密钥含糊度（疑义度）$H(K \mid C)$出发研究了此问题。$H(K \mid C)$给出了在给定密文下密钥的不确定性。

定义【明文熵、密钥熵、含糊度】

对密码系统(M, B, K, E_k, D_k)，

明文熵定义为$H(M) = -\sum_{m \in M} p(m) \log p(m)$

密钥熵定义为$H(K) = -\sum_{k \in K} p(k) \log p(k)$

明文含糊度定义为$H(X^t \mid Y^n) = -\sum_{m \in X^t, c \in Y^n} p(c, m) \log p(m \mid c)$

密钥含糊度定义为$H(K \mid Y^n) = -\sum_{k \in K, c \in Y^n} p(c, k) \log p(k \mid c)$

定义【惟一解距离】一个密码系统的惟一解距离 N_0 定义为使 $H(K|Y^n)=0$ 的最小正整数 n，即 $N_0=\min\ n\{H(K|Y^n)=0\}$。

直接计算 $H(K|Y^n)$ 或计算 $N_0=\min\ n\{H(K|Y^n)=0\}$ 都是非常困难的。

Shannon 提出利用随机密码模型来估计 $H(K|Y^n)$。随机密码模型满足如下假设：

1) 明文、密文共用同一个字母表 A，长度为 n 的明文集合 A^n 划分成两个集合 B_n 和 $\overline{B_n}=A^n-B_n$，B_n 中的明文是有意义的，而 $\overline{B_n}$ 中的明文是无意义的，且当 $n\to\infty$ 时，$\overline{B_n}$ 中明文出现的概率可忽略不计；
2) 密钥为等概率分布；
3) 对于 $k\in K$，对应的加密变换 E_k 是 $A^n\to A^n$ 的一一映射；
4) B_n 中的明文为等概率分布

定理【密钥含糊度的计算公式】若密码系统 (M,B,K,E_k,D_k) 满足随机密码模型假设、并且将长度为 n 的明文加密成长度为 n 的密文，则

$$H(K|Y^n)=H(K)+H(X^n)-H(Y^n)$$

证明：

对于满足随机密码模型假设的密码系统而言，

（1）$H(Y^n|K,X^n)=0$（已知密钥和明文可以求出密文）

（2）$H(X^n|K,Y^n)=0$（已知密钥和密文可以求出明文）

（3）$H(K,X^n)=H(K)+H(X^n)$（明文和密钥相互独立）

由条件熵与联合熵之间的关系有

（4）$H(Y^n|K,X^n)=H(Y^n,K,X^n)-H(K,X^n)$

（5）$H(X^n|K,Y^n)=H(Y^n,K,X^n)-H(K,Y^n)$

由（1）（2）（4）（5）式得：

（6）$H(K,Y^n)=H(K,X^n)$

所以由（6）和（3）式有

$$H(K|Y^n)=H(K,Y^n)-H(Y^n)$$
$$=H(K,X^n)-H(Y^n)$$
$$=H(K)+H(X^n)-H(Y^n)$$

定义【信息率与剩余度】：信源的绝对信息率定义为$R_0=\log|A|$，信源的近似信息率定义为$R_n=\frac{\log|B_n|}{n}$，信源的近似剩余度定义为$d_n=R_0-R_n$。

由假设条件知，

$$H(X^n)=nR_n=\log|B_n|$$
$$H(Y^n)=nR_0=n\log|A|$$

所以有$H(K|Y^n)=H(K)+nR_n-nR_0$

令$H(K|Y^n)=0$，则有$N_0=n=\frac{H(K)}{R_0-R_n}=\frac{H(K)}{d_n}$

记$\overline{H}=\lim_{n\to\infty}R_n$为信源信息率，n 充分大时，可用$\overline{H}$来代替$R_n$，即有$N_0=n=\frac{H(K)}{R_0-\overline{H}}$

注解：

（1） $N_0=n=\frac{H(K)}{R_0-R_n}=\frac{H(K)}{d_n}$是一个理论值，它说明：在截获的密文长度大于$N_0$时，如果破译者具有无限的计算能力，并且能充分利用信源的全部统计知识，则可唯一确定所使用的密钥，从而可以破译该密码。

（2） 一般而言，破译一个密码系统所需要的密文量均远大于它。

（3） 从理论上讲，惟一解距离与密钥熵成正比，与剩余度成反比。因此，如果希望一个密码系统的破译难度大，就希望惟一解距离大，也就是希望密钥熵大而剩余度小。

（4） 对应随机密码模型假设，如果密钥空间大，则密钥熵大。

$H(K)=-\sum_{k\in K}p(k)\log p(k)=-\sum_{k\in K}\frac{1}{|K|}\log\frac{1}{|K|}=\log|K|$。

（5） 理论安全性毕竟是理想假设下的结论，现实的密码系统则是考虑实用安全性的。实用安全性通常以“计算复杂性”为理论基础。