导论

哈尔滨工业大学

张宇

2024春

目录

什么是密码学

- 密码学（Cryptography）来自两个希腊单词：kryptos，意为“隐藏，保密”，和graphin，意为“书写”，即秘写。
- 简明牛津字典：书写或破解代码（code）的艺术。
  - Code（代码）是一个预先编排好的符号的系统，特别用于确保消息传输中的秘密。另外中文中“密码”相关的英文单词还包括password（口令）、cipher（加密方案）、key（密钥）。这些术语要注意区分。
- 现在密码学的开端在1980年代。那时在美国以DES、公钥密码学等为代表的成果相继出现，并且在美国密码学也从军用转变为民用。
- 参考教材IMC中定义：用于保护数字信息，系统和分布式计算免于敌对攻击的数学技术的科学研究。本课程主要是学习如何保护信息。

什么是密码学

- 漫画：在美国，密码学曾经被作为武器而被禁止出口；目前各国对密码学产品进出口都有严格限制

- 《中华人民共和国密码法》
  - 第二条：本法所称密码，是指采用特定变换的方法对信息等进行加密保护、安全认证的技术、产品和服务。
  - 第七条：核心密码、普通密码用于保护国家秘密信息，核心密码保护信息的最高密级为绝密级，普通密码保护信息的最高密级为机密级。核心密码、普通密码属于国家秘密。

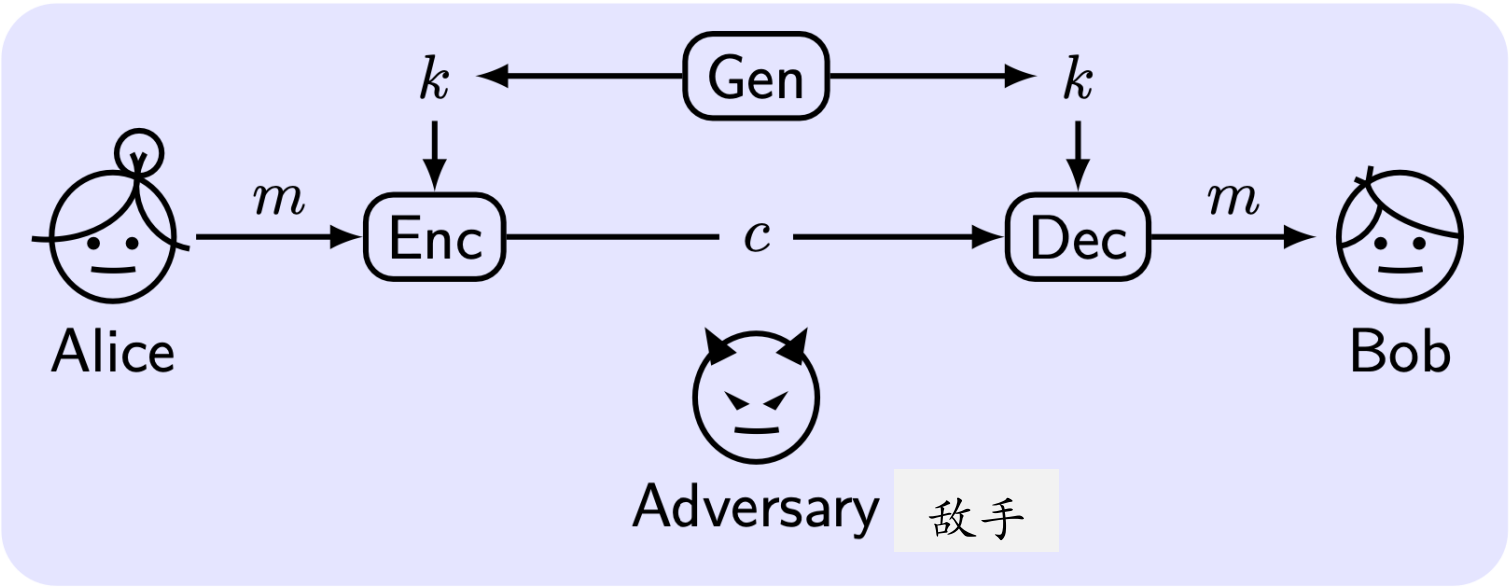
私钥加密（private key encryption）设定

- 目标是构造一个加密方案，用于在预先共享了私钥（对称密钥）的双方之间进行保密通信
- 隐含一个假设：存在某种方法，以保密的方式来分享一个密钥
- 磁盘加密相当于同一个人在不同的时刻间通信

Alice与Bob

- Alice和Bob是密码学领域最出名的人，他们首次出现在1978年的著名的RSA论文，《A Method for Obtaining Digital Signatures and Public-key Cryptosystems》（一种产生数字签名和公开密钥系统的方法）。

加密词法



密钥　　明文　　密文

- key $k \in \mathcal{K}$, plaintext (or message) $m \in \mathcal{M}$, ciphertext $c \in \mathcal{C}$
- **Key-generation** algorithm $k \leftarrow \mathsf{Gen}$　密钥生成算法
- **Encryption** algorithm $c := \mathsf{Enc}_k(m)$　加密算法
- **Decryption** algorithm $m := \mathsf{Dec}_k(c)$　解密算法
- **Encryption scheme**: $\Pi = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$　加密方案
- **Basic correctness requirement**: $\mathsf{Dec}_k(\mathsf{Enc}_k(m)) = m$

基本正确性要求

保护密钥还是隐瞒算法

- 加密方案的“秘密”包括两部分：加密/解密算法和密钥，那么我们应该保密什么？
- 更容易维护一个短密钥的秘密
- 在密钥暴露的情况下，对于诚实方，更换密钥更容易
- 在许多人彼此通信的情况下，更容易采用相同算法，不同密钥
- **Kerchhoffs原则** (**柯克霍夫原则**)：加密方法一定不必是秘密，即便落入敌手也必无不妥
- **香农箴言**：敌人了解系统

为什要“开放密码学设计”？

- 发表的设计经过公开检验会更强健（类似在提倡开源软件时，所给出的一个优点）
- 相对于被攻击者发现，由有道德的黑客来发现安全缺陷会更好
- 即便不公开，代码逆向工程（或被工业间谍泄漏）也构成了严重的安全威胁
- 使标准的建立成为可能
- 即使成为标准也不意味着安全：Dual EC是一个标准化的后门；Dual EC曾经与其他算法一起被NIST, ANSI和ISO标准化来产生随机数；斯诺登披露，以及在关于Bullrun项目和SIGINT使能项目的报告中已经表明，Dual EC是NSA颠覆标准的系统化工作的一部分。路透社报道，NSA在一笔交易中向RSA公司支付了1千万美元，用于将Dual EC设置为BSafe软件中优先或者缺省的数字生成方法。

攻击场景

- 除了窃听密文（称为COA），敌手还有其它手段（敌手能力）
- Ciphertext-only (COA) **唯密文**: 敌手只观察密文
- Known-plaintext (KPA) **已知明文**: 敌手获知同一密钥下的若干明文/密文对
- Chosen-plaintext (CPA) **选择明文**: 敌手有获得所选择明文加密（获得该明文的密文）的能力
- Chosen-ciphertext (CCA) **选择密文**: 敌手有获得所选择的其它密文解密（获得该密文的明文）的能力
- 被动攻击: COA KPA，由于不是所有密文都是机密的
- 主动攻击: CPA CCA，当敌手能够加密/解密任何其所希望的信息

历史上的加密方案及其密码分析（Cryptanalysis）

- 下面学习古典密码，目的是了解加密并没有想象的复杂，但设计安全的加密是很困难的。同时，理解一些密码学设计的基本原则，并思考一个问题：如何确定一个加密方案是安全的？
- 凯撒加密方案（Caesar's Cipher）：凯撒将机密消息加密书写，这是将字母表中字母顺序改变使得没有一个单词可以被理解。若有人要解密，则他必须将字母表中第四个字母，即D，替换成A，并且对其它字母也这么做。

$$\mathsf{Enc}(m) = m + 3 \mod 26$$

- 例子：明文`begintheattacknow`，采用凯撒加密的密文是什么？
- 其弱点是什么？怎么改进？

移位加密（Shift Cipher）

- $\mathsf{Enc}_k(m) = m + k \mod 26$
- $\mathsf{Dec}_k(c) = c - k \mod 26$
- **Weakness**: ?

Example: Decipher the string

`EHJLQWKHDWWDFNQRZ`

- **充足密钥空间原则**：任何安全加密方案必须具有一个经受住穷举搜索的密钥空间
- 问题：如何在穷举过程中自动化地确定密钥？

重合指数法（Index of Coincidence）

Index of Coincidence (IC): the probability that two randomly selected letters (pick-then-return) will be identical.
Let p_i denote the probability of ith letter in English text.

$$I \stackrel{\text{def}}{=} \sum_{i=0}^{25} p_i^2$$

Example

What's the IC of 'apple'?

For a long English text, the IC is ≈ 0.065. For $j = 0, 1, \ldots, 25$, q_j is the probability of jth letter in the ciphertext.

$$I_s \stackrel{\text{def}}{=} \sum_{i=0}^{25} p_i \cdot q_{i+s}$$

Q: For shift cipher, if $s = k$, then $I_s \approx$?

重合指数法

- 对于 $j = 0, 1, \ldots, 25$, 设q_j为密文中第 j 个字母的概率，定义一个带参数s的重合指数

  $I_s \overset{\text{def}}{=} \sum_{i=0}^{25} p_i \cdot q_{i+s}$

  - q_{i+s} 就是明文第 i 个字母被移位 s 个后所得到字母在密文中的概率
- 问题：当得到$I_s = 0.065$时就找到了密钥k?
  - 当$s = k$为密钥时，重合指数最大，因为平方和大于乘积的和，$a^2 + b^2 > 2ab$
  - 例如，将凯撒密码当成$k = 3$的移位密码，a被替换成D，D在密文中的概率与a在明文中的概率是一样的。当$k = 3$时，p_0就是明文中a的概率，q_{0+3}就是密文中D的概率，此时，$p_0 = q_3$。
- **这不正是大数据（足够的英文）+人工智能（是否是英文）吗!?**

单表替换加密（Mono-Alphabetic Substitution）

- 思想：将每个字母以任意方式映射到一个不同字母
- 优点：密钥空间足够大$2^{\{88\}}$。如何计算的？
- 缺点：是什么？怎么改进？

Example

abcdefghijklmnopqrstuvwxyz
XEUADNBKVMROCQFSYHWGLZIJPT
Plaintext: tellhimaboutme
Ciphertext: ????????????????

利用统计模式来攻击

- 1. 将密文中字母的频率制表，得到每个密文字母的频率
- 2. 与英文文本中字符频率比较（英文文本中字母频率）
- 3. 猜测频率最高的字母对应e，如此猜测其他字母
- 4. 挑选“合理的”明文，但并不简单

Table: Average letter frequencies for English-language text

e	12.7%	t	9.1%	a	8.2%	o	7.5%	i	7.0%
n	6.7%	_	6.4%	s	6.3%	h	6.1%	r	6.0%
d	4.3%	l	4.0%	c	2.8%	u	2.8%	m	2.4%
w	2.4%	f	2.2%	g	2.0%	y	2.0%	p	1.9%
b	1.5%	v	1.0%	k	0.8%	j	0.2%	x	0.2%
q	0.1%	z	0.1%						

频率分析例子

```
LIVITCSWPIYVEWHEVSRIQMXLEYVEOIEWHRXEXIPFEMVEWHKVS
TYLXZIXLIKIIXPIJVSZEYPERRGERIMWQLMGLMXQERIWGPSRIH
MXQEREKIETXMJTPRGEVEKEITREWHEXXLEXXMZITWAWSQWXSWE
XTVEPMRXRSJGSTVRIEYVIEXCVMUIMWERGMIWXMJMGCSMWXSJO
MIQXLIVIQIVIXQSVSTWHKPEGARCSXRWIEVSWIIBXVIZMXFSJX
LIKEGAEWHEPSWYSWIWIEVXLISXLIVXLIRGEPIRQIVIIBGIIHM
WYPFLEVHEWHYPSRRFQMXLEPPXLIECCIEVEWGISJKTVWMRLIHY
SPHXLIQIMYLXSJXLIMWRIGXQEROIVFVIZEVAEKPIEWHXEAMWY
EPPXLMWYRMWXSGSWRMHIVEXMSWMGSTPHLEVHPFKPEZINTCMXI
VJSVLMRSCMWMSWVIRCIGXMWYMX
```

Table: Analysis Steps

Ciphertext	Plaintext
I	e
XLI	the
E	a
Rtate	state
atthattMZe	atthattime
heVe	here
remarA	remark

频率分析例子

Hereupon Legrand arose, with a grave and stately air, and brought me the beetle from a glass case in which it was enclosed. It was a beautiful scarabaeus, and, at that time, unknown to naturalists – of course a great prize in a scientific point of view. There were two round black spots near one extremity of the back, and a long one near the other. The scales were exceedingly hard and glossy, with all the appearance of burnished gold. The weight of the insect was very remarkable, and, taking all things into considera- tion, I could hardly blame Jupiter for his opinion respecting it.

– Edgar Allan Poe's "The Gold-Bug"

维吉尼亚（多表移位）加密
Vigenere (poly-alphabetic shift) Cipher

- 思想：通过将明文中相同字母的不同出现映射到密文中不同字母，以此抹平密文中统计分布。

- **Encryption**: $c_i = m_i + k_{[i \bmod t]}$, t is the length (period) of k
- **Cryptanalysis**: Need find t; if t is known, need know whether the decryption "makes sense", but brute force (26^t) is infeasible for $t > 15$

Example (Key is 'cafe')

Plaintext	`tellhimaboutme`
Key	`cafecafecafeca`
Ciphertext	`????????????????`

Kasiski的方法（寻找周期 t）

- 多表移位加密在2百多年内未被有效破解，直到...
- 识别出长度2或3的重复模式，猜想这些重复应该是由于相同的明文片段被相同密钥片段加密的结果；
- 那么，假设密钥中没有重复模式，则这些重复出现之间的距离应该是密钥长度 t 的倍数；
- 那么，假设明文中重复模式是随机的，则密钥长度 t 是所有重复出现间距离的最大公约数；
- 但重复出现也可能是巧合，有没有更有效的方法？

Example (Key is 'beads')

```
themanandthewomanretrievedtheletterfromthepostoffice
beadsbeadsbeadsbeadsbeadsbeansdeadsbeadsbeadsbeadbea
VMFQTPFOHMJJXSFCSSIMTNFZXFYISEIYUIKHWPQMJJQSLVTGJKGF
```

重合指数法（找到周期t）

For $\tau = 1, 2, \ldots$, q_i is the probability of ith letter in $c_1, c_{1+\tau}, c_{1+2\tau}, \ldots$, IC is

$$I_\tau \stackrel{\text{def}}{=} \sum_{i=0}^{25} q_i^2$$

If $\tau = t$, then $I_\tau \approx ?$; otherwise $q_i \approx \frac{1}{26}$ and

- 对于 $\tau = 1, 2, \ldots$ 作为猜测的周期，$c_1, c_{1+\tau}, c_{1+2\tau}, \ldots$，是密文中以$\tau$为固定间隔的字符集合；
- q_i 是该字符集合中第 i 个字母出现的概率，计算该以τ为固定间隔字符集合的重合指数等于 $I_\tau \stackrel{\text{def}}{=} \sum_{i=0}^{25} q_i^2$
- 若 $\tau = t$, 那么 $I_\tau \approx ?$
  - 通过遍历 τ 来寻找 t, 若 $\tau = t$, 则使用了同一个密钥的移位加密所得到的密文 $c_1, c_{1+t}, c_{1+2t}, \ldots$，其中相同的明文字母被映射为相同的密文字母，因此，重合指数I_τ与明文的相同。

重合指数法（找到周期t)

If $\tau = t$, then $I_\tau \approx ?$; otherwise $q_i \approx \frac{1}{26}$ and

$$I_\tau \approx \sum_{i=0}^{25} \left(\frac{1}{26}\right)^2 \approx 0.038$$

Then reuse IC method to find k_i.

- 否则， 认为固定间隔字符集中字符的概率都是相同的，$q_i \approx \frac{1}{26}$ 并且 $I_\tau \approx \sum_{i=0}^{25} \left(\frac{1}{26}\right)^2 \approx 0.038$
  - 此时，假设所挑选出的密文是由明文字母被某个密钥中随机的字母映射所得到的，还假设密钥足够长且其中字母足够多样，则密文中每个字母出现是充分随机的，即出现概率为1/26，（在完美保密部分会进一步学习）。
- 确定周期后，再次使用重合指数法寻找密钥中每个位置 i 的字符 k_i.

小结

- 古典密码学最终都被破解说明一个道理：
  - 任意敌手原则 (Arbitrary Adversary Principle)：对于一类具有指定能力的敌手们，对于其中任意一个敌手，安全必须被确保。换句话说，安全与否只考虑敌手能力，不受敌手具体策略左右。
- 通过古典密码学到的教训：
  - 充分密钥空间原则
  - 设计加密方案是一项艰巨的任务
  - 复杂性不意味着安全
  - 任意敌手原则

现代密码学的三条原则

- 1. 安全和威胁模型的严格定义的形式化
- 2. 当一个加密方案的安全依赖于无法证明的假设时，这个假设必须被精确地描述并且尽可能地小
- 3. 加密方案应该带有一个基于以上定义和假设对安全性的严格证明

1. The formulation of a rigorous **definition** of security / threat model
2. When the security of a cipher relies on an unproven **assumption**, this assumption must be precisely stated and be as minimal as possible
3. Cipher should be accompanied by a rigorous **proof** of security with the above definition and the above assumption

原则1：对精确定义的形式化

. 原则1，对精确定义的形式化

- 如何形式化私钥加密的安全?
- 已知密文，没有敌手能够找到密钥， $\mathsf{Enc}_k(m) = m$
- 没有敌手能够找到与密文所对应的明文，$\mathsf{Enc}_k(m) = m_0 \| \mathsf{AES}_k(m)$，其中$m_0$表示第一个比特
- 没有敌手能够确定与密文所对应的明文中任意字符，$m = 1000$, 但有人能知道 $800 < m < 1200$
- 没有敌手能够从密文中获得关于明文的任何有意义的信息，但如何定义“有意义”?

原则1：如何定义

- 根据图灵对计算的定义，需要直觉，证明定义等价，用该定义来解决例子，以及时间考验。

How To Define Security – Lesson From Alan Turing

- What's computation?[3]
  1. A direct appeal to **intuition**
  2. A **proof of the equivalence** of two definitions (The new one has a greater intuitive appeal)
  3. Giving **examples** solved using a definition
- Additional method for security: **Test of time**

原则2：依赖于精确的假设

- 大多数密码学构造不能够被无条件的证明安全。
- 换句话说，当假设为真，则构造安全。
- 必要性：假设需要验证，方案需要比较，证明需要假设。
- 如何假设？
  - 年久的，经历过时间检验的
  - 简单，低级的假设容易被研究，拒绝和修正。

原则3：安全性的严格证明

- 安全性需要否定式证明，如何证明没有人能破解？如何证明“地球上没有龙”？
- 规约法（Reduction）：给定假设问题X很难，则根据定义构造Y是安全的（破解Y也很难）。
- 证明：将难题X的问题规约到破解Y的问题。
  - 这里X中每个问题都是难题，假设其无法解决。
  - 规约的意思是，可以将每个X中的问题转换为一个破解Y的问题，并且若一个破解Y的问题有答案，则可以由此答案构造一个对X中该问题的解。这里试图证明对Y的破解问题比X中问题要难。假设X中问题无法解决，而破解Y比X中问题还难，则破解Y也不可能。

本节课总结

- 密码学保护信息、事务和计算安全
- Kerckhoffs原则，开放密码学设计
- 凯撒、移位、单表替换、多表替换
- 蛮力，字母频率，Kasiski方法，重合指数（IC）
- 充分密钥空间原则
- 任意敌手原则
- 严格证明安全