现代密码学

Modern Cryptography

张方国

中山大学计算机学院

Office: Room 305, IM School Building

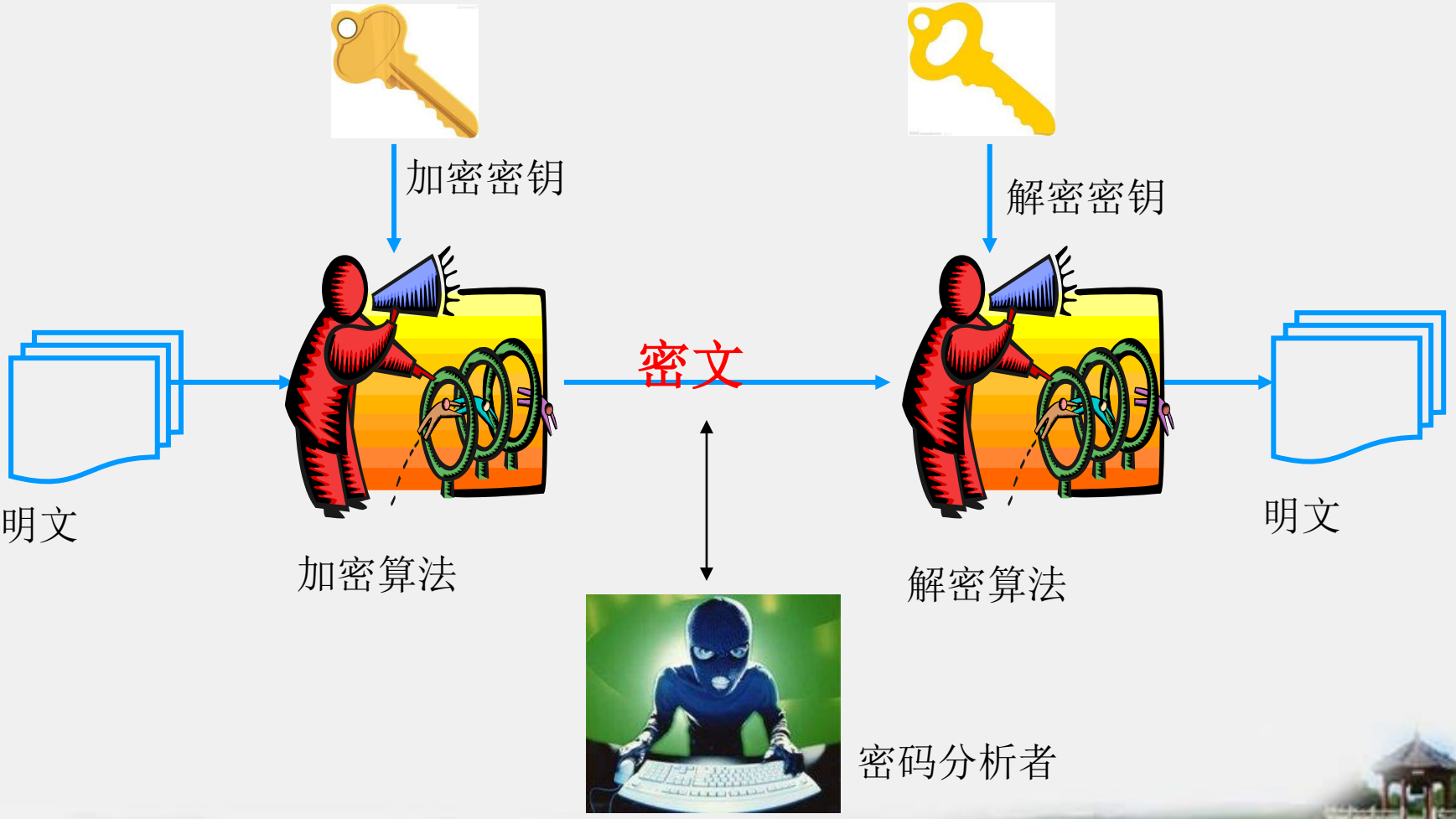
E-mail: isszhfg@mail.sysu.edu.cn

HomePage: https://cse.sysu.edu.cn/content/2460

第二讲 古典密码学1

- 代换密码及其密码分析
  移位密码,
  仿射密码,
  维吉尼亚密码,
  希尔密码

密码学的基本概念



基本概念

一个满足下面条件的五元组(P, C, K, E, D)为一个密码体制：

(1) P是一个非空有限集合，表示所有的明文空间。

(2) C是一个非空有限集合，表示所有的密文空间。

(3) K是一个非空有限集合，表示所有的密钥空间。

基本概念

(4) 对任意的$k \in K$, 都存在一个

加密函数：

$E_k(\in E): P \rightarrow C$

和相应的解密函数：

$D_k(\in D): C \rightarrow P$

对任意的明文$m \in P$均有$D_k(E_k(m))=m$。其中E_k和D_k都必须是单射函数。

基本概念

通信过程：

(1) 通信双方通过协商选择并共享一个密钥 k∈K。

(2) 发送方使用加密函数Ek对明文串进行加密得到密文C。

(3) 当Bob接到密文串C时，他使用解密函数Dk对其进行解密，就可以得到原始明文串m。

密码分析的方法

- Kerckhoff假设：攻击者Oscar知道所使用的密码体制。
- 目标是设计在Kerckhoff假设下安全的密码体质。
- 唯密文攻击
- 已知明文攻击
- 选择明文攻击
- 选择密文攻击
- 自适应选择密文攻击

古典密码（Classical Cipher）

古典密码是密码学的渊源，这些密码大都比较简单，可用手工或机械操作实现加解密，现在已很少采用了。然而，研究这些密码的原理，对于理解、构造和分析现代密码都是十分有益的。

代换密码和置换密码

代换密码（Substitution Cipher)是明文中的每一个字符被替换成密文中的另一个字符。接收者对密文做反向替换就可以恢复出明文。

置换密码(Permutation Cipher)又称换位密码（Transposition Cipher)，加密过程中明文的字母保持相同，但顺序被打乱了。

代换密码(Substitution Cipher)

明文字母表$\mathcal{A}$： $Z_q=\{0, 1, \ldots, q-1\}$

明文消息是长为L个字母串，称为明文组，以m表示，

$$m=(m_0 m_1,\ldots, m_{L-1}) \quad m_i\in Z_q$$

它是定义在Z_q^L上的随机变量，Z_q^L是Z_q上的L维向量空间。明文空间$\mathcal{M}=\{m, m\in Z_q^L\}$。

密文字母集$\mathcal{A}'$：$Z_{q'}=(0, 1,\ldots, q'-1)$表示。密文组

- $$c=(c_0, c_1, \ldots, c_{L'-1}) \quad c\in Z_q$$
- c是定义在L'维向量空间$Z_{q'}^{L'}$上的随机变量。密文空间$\mathcal{C}=\{c, c\in Z_{q'}^{L'}\}$。一般当$\mathcal{A}'=\mathcal{A}$时有$\mathcal{C}=\{c, c\in Z_q^L\}$，即明文和密文由同一字母表构成。

代换密码

加密变换：明文空间到密文空间的映射：

$$f:\ m \to c\ \ m \in \mathcal{M},\quad c \in C$$

在1—1的映射下，存在有逆映射f^{-1}，使

$$f^{-1}(c)=f^{-1}\cdot f(m)=m\quad m \in \mathcal{M},\ c \in C$$

加密变换通常是在密钥控制下变化的，即

$$c=f(m, k)=E_k(m)$$

式中，$k \in \mathcal{K}$，$\mathcal{K}$为密钥空间。一个密码系统就是在f和密钥k作用下，由$Z_q^{\ L} \to Z_{q'}^{\ L'}$的映射，或以$Z_{q'}^{\ L'}$中的元素代换$Z_q^{\ L}$中的元素，在这意义下，称这种密码为**代换密码**(Substitution Cipher)。L=1时，称作**单字母**或**单码代换**(Monogram Substition)，也称为**流密码**(Stream Cipher)。L>1时称作**多字母**或**多码代换**(Polygram Substition)，也称为**分组密码**。

代换密码

一般选择$q=q'$，即明文和密文字母表相同。此时，

- $L=L'$，f可以构造成1—1的映射，密码**没有数据扩展**。
- $L<L'$，则有**数据扩展**，函数f为1→多的映射，明文组可有多个密文组来代换，称为**多名**或**同音**(Homophonic)**代换密码**。
- $L>L'$，则明文数据将被**压缩**(Compression)。函数f不是可逆的，保密通信$L\leq L'$。$L>L'$可用在数据认证系统中。
- **单表代换(Monoalphabetic Substitution)**：在$\mathcal{A}=\mathcal{A}'$、$q=q'$和$L=1$时，对所有明文字母，都用一个固定的代换进行加密。
- **多表代换**(Polyalphabetic Substitution)：在$\mathcal{A}=\mathcal{A}'$、$q=q'$和$L=1$时，用一个以上的代换表进行加密。，

这是古典密码中的两种重要体制，曾得到过广泛的应用。

单表代换密码

单表代换密码：明文字母表到密文字母表的固定映射，

$$f: Z_q \to Z_q$$

令明文$\boldsymbol{m}=m_0m_1\ldots$，则相应密文为

$$\boldsymbol{c}=E_k(\boldsymbol{m})=c_0c_1\ldots=f(m_0)f(m_1)\ldots$$

1．移位代换密码 (Shift Substitution Cipher)

加密变换：$E_k(i)=(i+k)\equiv j \bmod q \quad 0\le i, j<q$

$$K=\{k \mid 0\le k<q\}$$

k=0时为**恒等变换**。

解密变换：$D(j)=E_{q-k}(j)\equiv j+q-k \equiv i+k-k\equiv i \bmod q$

移位代换密码的例子

- 凯撒(Caeser)密码： One of the oldest recorded ciphers, known as Caesar's cipher, is described in ``De Vita Caesarum, Divus Iulius" (``**The Lives of the Caesars, The Deified Julius**"), written in approximately 110 C.E.:

  There are also letters of his to Cicero, as well as to his intimates on private affairs, and in the latter, if he had anything confidential to say, he wrote it in cipher, that is, by so changing the order of the letters of the alphabet, that not a word could be made out. If anyone wishes to decipher these, and get at their meaning, he must substitute the fourth letter of the alphabet, namely D, for A, and so with the others

例 凯撒(Caeser)密码是对英文26个字母进行移位代换的密码，其q=26。例如，选择密钥k=3，则有下述代换表：

$\mathcal{A}$： a b c d e f g h i j k l m n o p q r s t u v w x y z

A‘： DEFGH I J KLMNOP Q RS TUVWXYZ AB C

明文： **m = veni，vidi，vici**

密文： **c=E(m)= YHAL， YLGL，YLFL**

（意思是“我来，我见，我征服”,曾经是恺撒征服本都王法那西斯后向罗马元老院宣告的名言）

解密运算为$D_3=E_{23}$，用密钥k=23的加密表加密就可恢复明文。

又称为加法密码(Additive Cipher)。

- 移位密码是不安全的，显然可用穷搜索方法来攻击；
- 一个密码体制安全的必要条件是能够抵抗穷 搜索攻击，普通的做法是使密钥空间足够大。

密码体制 1.3　仿射密码

令$\mathcal{P}=\mathcal{C}=\mathbb{Z}_{26}$且

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

对任意的 $K=(a,b)\in\mathcal{K}$，$x,y\in\mathbb{Z}_{26}$，定义加密变换为：

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

相应的解密变换为：

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

当$a=1$时就得到移位密码。q=26时可能的密钥数为(26×12)－1＝311个。

任意的单表代换密码

- Each plaintext character is mapped to a different ciphertext character. The mapping is 1–1 in order to enable decryption (i.e.,the mapping is a permutation).

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
X	E	U	A	D	N	B	K	V	M	R	O	C	Q	F	S	Y	H	W	G	L	Z	I	J	P	T

- The key is therefore a permutation of the alphabet (the size of the key space is therefore 26! (about 2^{88}).

密码体制 1.2　代换密码

令$\mathcal{P}=\mathcal{C}=\mathbb{Z}_{26}$。$\mathcal{K}$由 26 个数字 0,1,…,25 的所有可能置换组成。对任意的置换 $\pi\in\mathcal{K}$,定义:

$$e_\pi(x)=\pi(x)$$

再定义:

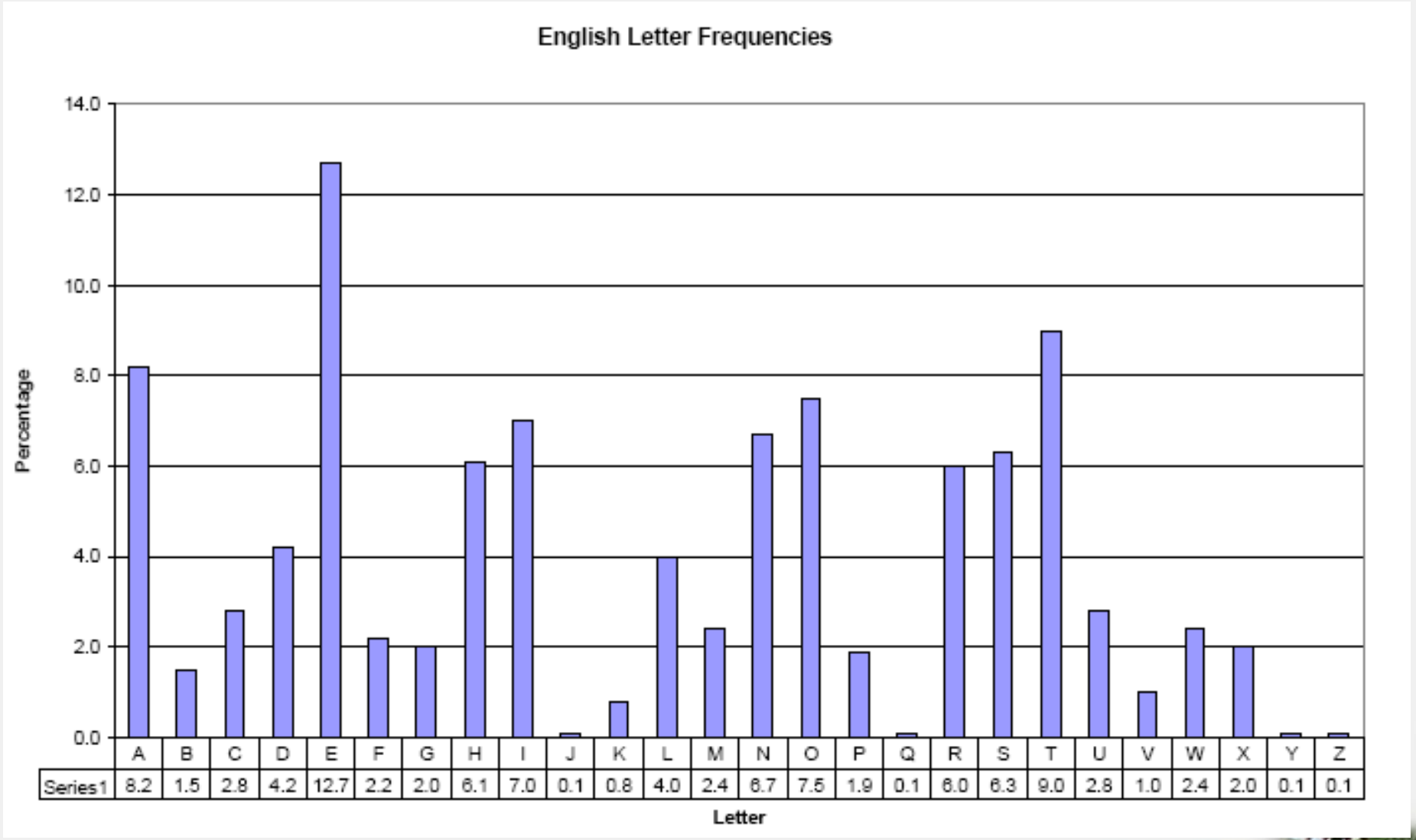
$$d_\pi(y)=\pi^{-1}(y)$$

这里 π^{-1}代表置换 π 的逆置换。

事实上,在代换密码的情形下,我们也可以认为$\mathcal{P}$和$\mathcal{C}$是 26 个英文字母。在移位密码中使用$\mathbb{Z}_{26}$是因为加密和解密都是代数运算。但是在代换密码的情形下,可更简单地将加密和解密过程直接看做是一个字母表上的置换。

语言的冗余性和密码攻击

- 人类的语言是有冗余性的
- 比如从“th lrd s m shphrd shll nt wnt”中我们可以大概猜出些什么
- 字母使用的频率是不一样的
- 英文字母E是使用最频繁的，然后是T, R, N, I, O, A, S等
- 有些字母使用得很少，如Z, J, K, Q, X
- 这样可以得到英文字母使用频率分布表
- 同时，统计双字母组合和三字母组合的使用频率也是非常有用的


English Letter Frequencies
Percentage
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Series1 8.2 1.5 2.8 4.2 12.7 2.2 2.0 6.1 7.0 0.1 0.8 4.0 2.4 6.7 7.5 1.9 0.1 6.0 6.3 9.0 2.8 1.0 2.4 2.0 0.1 0.1
Letter

密码分析

- 单字母代换的密码都无法抗击字母频率攻击
- 例1.10 利用仿射密码中获得的密文：

FMXVEDKAPHFERBNDKRXRSREFMORUDS
DKDVSHVUFEDKAPRKDLYEVLRHHRH

第一步，分析频率。

表 1.2　密文出现字母频率统计

字母	频率	字母	频率
A	2	N	1
B	1	O	1
C	0	P	2
D	7	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

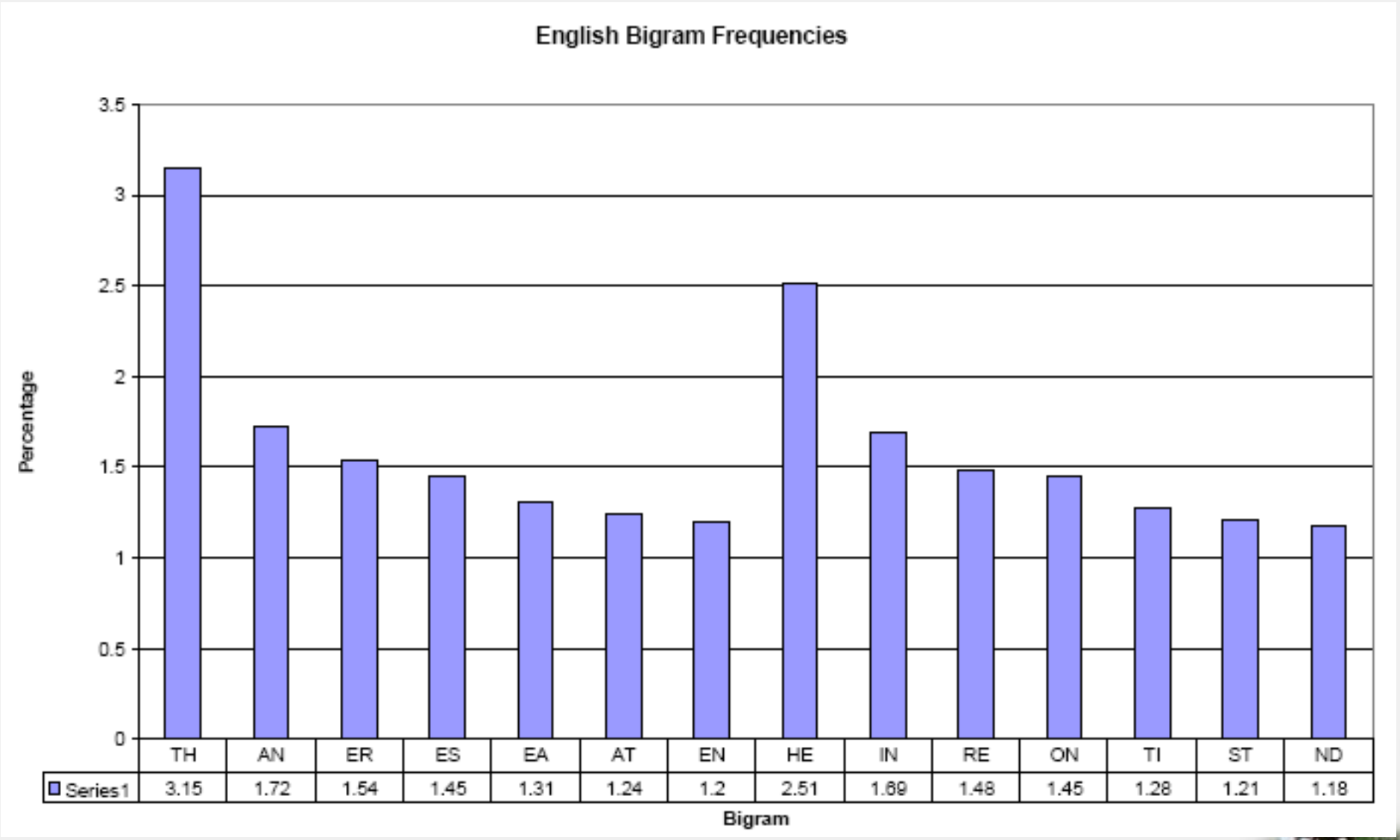
第二步，猜测a,b。多次猜测

这里虽然只有 57 个字母，但它足以分析仿射密码，最大频率的密文字母是：R(8 次)，D(7 次)，E、H、K(每个 5 次) 和 S、F、V(各 4 次)。首先，我们可以猜想 R 是 e 的加密而 D 是 t 的加密，因为 e 和 t(分别) 是两个出现频率最高的字母。以数字表达即为 $e_K(4) = 17$ 和 $e_K(19) = 3$，因为 $e_K(x) = ax + b$，这里 a 和 b 是未知的，所以我们有如下的关于两个未知数的线性方程组：

$$4a + b = 17$$
$$19a + b = 13$$

这个方程组有惟一解 $a = 6, b = 19$(在 $\mathbb{Z}_{26}$)，但这是一个不合法的密钥，因为 $\gcd(a, 26) = 2 > 1$。所以我们的猜想肯定是不正确的。

我们再猜测 R 是 e 的加密，而 E 是 t 的加密，继续使用上述的方法，得到 $a = 13$，这也是一个不合法的密钥。再试一种可能性：R 是 e 的加密，H 是 t 的加密，则有 $a = 8$，这也是不合法的。继续进行，我们猜测 R 是 e 的加密，K 是 t 的加密，这样可得 $a = 3, b = 5$，首先它至少是一个合法的密钥，下一步工作就是检验密钥 $K = (3,5)$ 的正确性。如果我们能得到有意义的英文字母串，则可证实是有效的。


English Bigram Frequencies
Percentage
3.5
3
2.5
2
1.5
1
0.5
0
TH AN ER ES EA AT EN HE IN RE ON TI ST ND
Series1 3.15 1.72 1.54 1.45 1.31 1.24 1.2 2.51 1.69 1.48 1.45 1.28 1.21 1.18
Bigram

对抗频率分析的办法

- 多名或同音代替密码（映射是一对多）
- **多表代替密码**
- 多字母代替密码

多表代换密码

多表代换密码：以一系列(两个以上)代换表依次对明文消息的字母进行代换的加密方法。

明文字母序列：$m=m_1 m_2\ldots$，

代换序列：$\pi=(\pi_1, \pi_2,\ldots)$为代换序列。

密文字母序列：$c=E_k(m)=\pi(m)=\pi_1(m_1)\pi_2(m_2)\ldots$

- 非周期多表代变换密码，π为非周期的无限序列。这类密码，对每个明文字母都采用不同的代换表(或密钥)进行加密，称作是一次一密钥密码(One-time Pad Cipher)。
- 周期多表代换密码，π为周期序列，重复地使用，

代换序列：$\pi=\pi_1\pi_2\ldots\pi_d\pi_1\pi_2\ldots\pi_d\ldots$

密文：$c=E_k(m)=\pi(m)=\pi_1(m_1)\pi_2(m_2)\ldots\pi_d(m)\pi_1(m_{d+1})\ldots\pi_d(m_{2d})$

当$d=1$时就退化为单表代换。

多表代换密码

1. 维吉尼亚密码Vigenére cipher

- 1858年法国密码学家Blaise de vigenere所发明。
- 移位代换表：$\boldsymbol{\pi}=\pi_1\pi_2\ldots\pi_d$，由$d$个字母序列给定的密钥
- $$\mathbf{k}=(k_1, k_2, \ldots, k_d)\in \boldsymbol{Z}_q^{\,d}$$
- $k_i(i=1, \ldots, d)$确定明文第$i+td$个字母(t为正整数)的移位次数，即 $$c_{i+td}=E_{ki}(m_{i+td})\equiv m_{i+td}+k_i \bmod q$$
- 称$\boldsymbol{k}$为**用户密钥**(user key)或**密钥字**(key word)，其周期地延伸就给出了整个明文加密所需的**工作密钥**(working key)。

维吉尼亚密码

密码体制 1.4　维吉尼亚密码

设 m 是一个正整数。定义 $\mathcal{P}=\mathcal{C}=\mathcal{K}=(\mathbb{Z}_{26})^m$，对任意的密钥 $K=(k_1,k_2,\cdots,k_m)$，定义

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

和

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

以上所有的运算都是在 $\mathbb{Z}_{26}$ 上进行。

维吉尼亚密码

- 例令q=26，密钥字k=beads，即周期d=5，则有

Plaintext:	the man and the woman retrieved the letter from the post office
Key:	bea dsb ead sbe adsbe adsbeadsb ean sdeads bead sbe adsb eadbea
Ciphertext:	VMF QTP FOH MJJ XSFCS SIMTNFZXF YIS EIYUIK HWPQ MJJ QSLV TGJKGF

- 其中，同一明文字母t在不同的位置上被加密为不同的字母*V, M, Y和U*。
- 维吉尼亚密码是用d个凯撒代换表周期地对明文字母加密。

Vigenére cipher-破译

加密过程中可以观察到一个很重要的信息：**密钥的重复部分与明文中的重复部分的连接，在密文中也产生一个重复部分！**也就是说，如果字母在明文中重复，那它总是用密钥词的相同部分进行加密，这样密文也包含重复的字符串。

当然某段密文 的两次出现也可能是偶然的，而不一定是用相同密钥加密相同明文序列导致的。但当信息足够长时，就会有大量重复的密文序列出现．通过计算重复密文序列间距的最大公约数，就很有可能猜测出密钥的长度，从而最终获得密钥和明文内容

Kasiski测试法

- 两个相同的明文段将加密成相同的密文段。若他们的位置间距为d，则$d=o(\mathrm{mod}\ m)$.
- 反过来，如果在密文中观察到两个相同的长度为3的密文段，那么将给攻击者带来很大的方便，因为他们实际上对应了相同的明文串。
- 测试法：搜索长度至少为3的相同的密文段，记下其离起始点的距离。假如得到如下几个距离，$d_1,d_2,\ldots,$那么可以猜测m为他们最大公因子的因子。

Kasiski 测试法

- Kasiski测试法是由Charles Babbage 和 Friedrich Kasiski 分别发现的。
- 密文中的重复性可以暗示出密钥长度，如果两个相同明文序列之间的距离是密钥长度的整数倍，那么产生的密文序列也是相同的

Consider the following concrete example with the password `beads` (spaces have been added for clarity):

Plaintext:	`the man and the woman retrieved the letter from the post office`
Key:	`bea dsb ead sbe adsbe adsbeadsb ean sdeads bead sbe adsb eadbea`
Ciphertext:	`VMF QTP FOH MJJ XSFCS SIMTNFZXF YIS EIYUIK HWPQ MJJ QSLV TGJKGF`

Note that the word `the` is mapped sometimes to `VMF`, sometimes to `MJJ` and sometimes to `YIS`. However, it is mapped *twice* to `MJJ`, and in a long enough text it is likely that it would be mapped multiple times to each of the possibilities. The main observation of Kasiski is that the distance between such multiple appearances (except for some coincidental ones) should be a multiple of the period length. In the above example, the period length is 5 and the distance between the two appearances of `MJJ` is 40 (8 times the period length). Therefore, the *greatest common divisor* of all the distances between the repeated sequences should yield the period length t.

重合指数法： 1920年William Friedman提出

如果考虑来自26个字母表的完全随机文本，则每个字母都以相同的概率1/26出现，假定另一个随机文本放在第一个的下面，在上下位置出现相同字母a的概率为$(1/26)^2$在两个随机文本的上下位置找到任意两个相同字母总的概率为 $26(1/26)^2 = 1/26 = 0.0385$

但实际上，由于英文字母出现的概率是不同的，设字母a，b，c，…，z出现的概率分别为 $p_0, p_1, p_2, \cdots p_{25}$

这样找到两个相同字母的概率为 $\sum_{i=0}^{25} p_i^2 = 0.065$

这个值比随机文本的概率大得多，称为重合指数。

定义 设一个语言由*n*个字母构成，每个字母出现的概率 $p_i, 1 \le i \le n$，则重合指数是指其中两个随机元素相同的概率，记为

$$CI = \sum_{i=1}^{n} p_i^2$$

这样对于一个完全随机的文本*CI*=0.0385，与一个有意义的英语文本*CI*=0.065，差异是比较明显的。

实际分析中，重合指数的利用体现在几个方面。

- ✓ 如果密文的重合指数较低，那么就可能是多表替代密码。维吉尼亚密码将密文分行，每行是单表替代密码。
- ✓ 在单表替代时，明文的字母被其它字母代替，但不影响文本的统计属性，即加密后密文的重合指数仍不变，*CI*（明文）= *CI*（密文），由此可以判断文本是用单表还是用多表替代加密的。
- ✓ 如果密钥长度（即密文分行的列数）正确，同一行密文有相同字母的概率接近0.065；如果密钥长度不对，则概率将大大小于0.065，显得更随机，由此得到密钥长度（可与Kasisiki测试的结果对比）。
- ✓ 重合指数的估算能用于分析两个不同密文，比如接收到两段文本*C*1，*C*2，如果它们用同样的方式加密，则 $CI(C_1) \approx CI(C_2)$

实际密文长度有限，从密文中计算的重合指数值总是不同于理论值，所以通常用*C*I的估计值*CI*′，以字母出现的频度近似表示概率，则

$$CI' = \sum_{i=1}^{m} C_{x_i}^2 / C_L^2 = \sum_{i=1}^{m} x_i(x_i - 1) / L(L-1)$$

重合指数法

假设我们使用维吉尼亚密码加密的密文串为 $y = y_1 y_2 \cdots y_n$。将串 y 分割为 m 个长度相等的子串，分别为 $y_1, y_2, \cdots, y_m$，这样可以以列的形式写出密文，组成一个 $m \times (n/m)$ 矩阵。矩阵的每一行对应于子串 $y_i, 1 \leqslant i \leqslant m$。换言之，我们有如下形式：

$$
\begin{aligned}
y_1 &= y_1 y_{m+1} y_{2m+1} \cdots \\
y_2 &= y_2 y_{m+2} y_{2m+2} \cdots \\
&\vdots \quad \vdots \quad \vdots \\
y_m &= y_m y_{2m} y_{3m} \cdots
\end{aligned}
$$

如果 $y_1, y_2, \cdots, y_m$ 按如上方法构造，则 m 实际上就是密钥字的长度，每一个 $I_c(y_i)$ 的值大约为 0.065。另一方面，如果 m 不是密钥字的长度，那么子串 y_i 看起来更为随机，因为它们是通过不同密钥以移位加密方式获得的。易知，对一个完全的随机串，其重合指数为：

$$I_c \approx 26(1/26)^2 = \frac{1}{26} = 0.038$$

值 0.065 和 0.038 的差别是比较大的，按这种方法通常可以确定密钥字的长度（或确信一个利用 Kasiski 方法猜测的密钥字长）。

已知使用维吉尼亚密码加密获得如下密文：

CHREEVOAHMAERATBIAXXWTNXBEEOPHBSBQMQEQERBW
RVXUOAKXAOSXXWEAHBWGJMMQMNKGRFVGXWTRZXWIAK
LXFPSKAUTEMNDCMGTSXMXBTUIADNGMGPSRELXNJELX
VRVPRTULHDNQWTWDTYGBPHXTFALJHASVBFXNGLLCHR
ZBWELEKMSJIKNBHWRJGNMGJSGLXFEYPHAGNRBIEQJT
AMRVLCRREMNDGLXRRIMGNSNRWCHRQHAEYEVTAQEBBI
PEEWEVKAKOEWADREMXMTBJJCHRTKDNVRZCHRCLQOHP
WQAIIWXNRMGWOIIFKEE

首先使用 Kasiski 测试法。在密文中密文串 CHR 共出现在 5 个位置，开始位置分别为 1、166、236、276 和 286，其距离分别为 165、235、275 和 285。这三个整数的最大公约数为 5，故我们猜测密钥字的长度很可能为 5。

我们再使用重合指数法确认这一猜测。当 $m = 1$ 时，重合指数为 0.045；当 $m = 2$ 时，两个重合指数分别为 0.046 和 0.041；当 $m = 3$ 时，为 0.043、0.050 和 0.047；当 $m = 4$ 时，为 0.042、0.039、0.046 和 0.040；当 $m = 5$ 时，可获得的值分别为 0.063、0.068、0.069、0.061 和 0.072。这些值为密钥字的长度为 5 提供了强有力的证据。