数学与密码分析

中国人民解放军保密委员会保密通信

数学与密码分析

Mathematics and Cryptanalysis

付茂泉　郭廷清　邵先锋
李　乔　　陈建明　译

一九八八年十月

Library of Congress Cataloging-in-Publication Data

Cryptology Yesterday, today, and tomorrow.
1. Cryptography. 2. Ciphers. I. Deavours, Cipher
A. Z103.C76 1987 652'.8 87—14396
ISBN 0—89006—253—6

ARTECH HOUSE, INC.
685 Canton Street
Norwood,MA 02062

International Standard Book Number:0-89006-253-6
Library of Congress Catalog Card Number:87-14396
10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Cryptology Yesterday, today, and tomorrow.
1. Cryptography. 2. Ciphers. I. Deavours, Cipher A. Z103.C76 1987 652'.8 87—14396
ISBN 0—89006—253—6

ARTECH HOUSE, INC.
685 Canton Street
Norwood,MA 02062

International Standard Book Number:0-89006-253-6
Library of Congress Catalog Card Number:87-14396
10 9 8 7 6 5 4 3 2 1

出 版 说 明

《数学与密码分析》译自1987年美国Artech House出版的《Cryptology Yesterday, Today and Tomorrow》一书的第三章：Mathematics and Cryptanalysis。

该书第一章主要讲述密码的历史，文献和人物评议；第二章主要介绍机械密码及其解法；第三章分析并列举破译不同密码体制所需的密文长度，讨论了密码体制的复杂度进而提出用序列的复杂度作为密码性能的测度，介绍了密码的自动分析（即用计算机自动脱密）方法等有关密码理论的敏感问题。为使有关单位和人员及时了解国外数学与密码分析的研究情况，我们将该书第三章单独翻译成册，供热心于数学与密码分析的读者学习参考。

本书由空军第六研究所搜集、提供英文原稿并承担全部译校工作。其中，付茂泉同志翻译第1、4、5、7篇；邵先锋同志翻译第2、10篇；郭廷清同志翻译第3、6、9篇；李香同志翻译第8篇；陈建明同志翻译第11篇。最后由付茂泉、郭廷清同志校阅了全部译稿。

我们谨向提供原稿、参加译、校工作的单位和个人表示衷心地感谢。

在本书的出版、发行工作中，得到空六所第七研究室的大力支持，在此表示谢意。

由于时间仓促，本书错误之处难免，欢迎批评指正。

中国人民解放军保密委员会保密通信办公室

1988年10月

目　录

密码分析中的唯一解点

Cipher A. Deavours

C. 仙农，信息论之父，由于1945年他的《密码学的数学理论》一书的出版，又奠定了密码分析的数学基础(后来重版时改用了较少说明性的书名《保密系统的通信理论》〔1〕)。在这篇文章中，仙农用特有的明快与直率的文体陈述了密码编码学与密码分析学的要点。任何对密码学基础感兴趣的人都可以从该文中受益，而且文中大部分只要求最低程度的数学知识。

文中很有意思的一节是论述唯一解点（unicity point）的概念。一种密码的唯一解点就是电文的长度——超过这个长度，用已知的系统脱密就变成唯一的过程。若电文小于唯一解点距离，通常有多种脱密，而且所谓的密码分析者无法从这些可用的脱密中选择出正确的脱密。因此，即使假定密文截收者完全了解所用的加密系统（包括专用密钥），如果所截收的电文量少于唯一解点所需要的量，没有单义解是可能的。

对一种随机密码（random cipher），其唯一解点可以用下面的公式估计：

$$U = H(K)/D$$

式中H(K)是给定系统中可能的密钥数的对数，D是源消息中每个字母的多余度。英文的多余度约为1.11位数或78%。这个公式很简单，但其精确度可增加或减少到这样的程度：按仙农的观点，加密系统是随机的。要在这里详细讨论究竟怎样构造一种随机密码会需要很多的篇幅。不过，两个最严格的条件

是：每个加密密钥被等可能地使用，并且用随机选择密钥脱密一份电文等可能地产生某一可能的源消息。（对实际应用来说，源消息（source message）意指任何任意的明文字符串，有意义的或无意义的，其长度与密文相等）。本文的附录A中讨论了一种有关的密码模型。单表代替和移位密码并不是随机密码的令人满意的例子，但是，即使在这些密码中，唯一解点公式也只用作实际唯一解点的下界〔2〕。更复杂的密码，如普莱费尔（Playfair）密码，三叉（trifid）密码，N元代替密码以及密钥相当长的多表密码通常都使随机密码达到令人满意的程度。接近随机值0.38的重合指数在许多情况下都是合适的统计试验。

仙农推导的唯一解点公式是用信息论的术语表达的，而且还可以用一种更简单的方法〔2〕。假定明电文由字母表中字符组成，就有26^N种可能的长度为N的源消息。通常把这个数写作指数形式：

$$26^N = 10^{(\log 26)N} = 10^{1.41N}$$

长度为N的字符串中，大多数是毫无意义的字母堆集，但有些字符串可组成有效的英文明文段。对相当长的字符串、已经找到的有效的英文明文段数约为$10^{.30N}$。从经典的统计方法类推，常量·30被叫做该语言每个字母的熵（entropy）。如果我们随机地选择一个加密密钥并脱密一份N个字符的密报，我们可能得到一份可识别的英文消息或得到一批N个不连贯的字母。即使这份消息是有效的英文，我们也不能假定它就是那份预定的消息，因为我们可能选择了错误的密钥。得到有意义的消息的机会即概率P平均应为有意义的消息数除以可能的源消息总数，所以

$$P = 10^{.30N}/10^{1.41N} = 10^{-1.11N}$$

一般最后一项写成10^{-DN}，这里的D被称做该语言用数字表示的多余度（redundancy）。对于英文，D值1.11表示英文约有1.11/log26 = .78 = 78%是多余的。

假设这种密码有10^H个密钥，我们把每个密钥都试一遍，会得到多少份有意义的消息呢？至少有一份有意义的消息（正确的那份）；用其余10^H-1个密钥给出大约$10^{H}\cdot P$种以上有意义的消息。因此，假脱密的期望数为

$$E = 10^{-1.11N}\ (10^H - 1)\ = 10^{-1.11N+H} - 10^{-1.11N}$$

当N适当大时，右边第二项很小，一般可忽略不计。如果指数$-1.11N+H \gg 0$，则E很大；若$-1.11N+H \ll 0$，则E很小。$-1.11N+H=0$或$N=H/1.11$的那个点把少数解的范围与多数解的范围分开，因此这个点被恰当地称做唯一解点。图1表示周期5的维吉尼亚密码在假定消息源有26^5种密钥和不同的多余度情况下的唯一解点线$\log E = -DN+H$。消息源的多余度越少，其唯一解距离就越远（$\log E = 0$）。

实际上，在应用于密码前，先把消息源编码可以压缩它的多余度。可以用简单的缩写达此目的，例如，REPORT RECEIVED可以缩写成RPRT RCVD。

为了说明仙农的结论，我们将计算适用于几种经典密码的某些样本的唯一解点，假定英文为消息源语言。对使用长度为P的字母表密钥的维吉尼亚密码，其密钥可有26^P种选择方式。因为字母表序列是规则的，在那个方向上不存在自由选择。因此

$$H\ (K)\ = \log 26^P = P\log 26 = 1.41P$$

所以

$$U = 1.41P/1.11 = 1.27P\text{个字母}$$

对这个结论，我们做如下解释。假定截获一份已知用维吉尼亚密表加密的密报、提出一种假设解，产生有效的电文。如

果假设周期为P，则原密报中出现的字符多于1.27P个表示所提供的假设解可能是唯一的。密报中出现的字符少于1.27P个，使我们不得不排除假设解为可能解集中的唯一解。

如果我们把上述系统中的密码字母表混合起来，其密钥数增加到26！·26^P，因为字母表可以有26！种不同的混排方式(A有26种可能的代替，B有25种，等等。) 26^P个长度为P的可能的密钥片语中，任何一个都可用于这些排列中的每一个排列。这时，唯一解点移到

$$U = \log(26! \cdot 26^P) = 23.97 + 1.27P\text{个字母}$$

注意，混合那条字母序列，把唯一解点延长到一定的距离，这个距离与密钥片语的长度"无关"（Independent）并且对用长密钥片语的系统或长密文的保密性没有什么影响。下面给出另外一些唯一解点的抽样数据。

密码类型	唯一解点（用字母数表示）
维吉尼亚，密钥长度P	1.27P
维吉尼亚，混合密表，密钥长度P（沼泽Ⅱ）	23.97＋1.27P
维吉尼亚，混合密表与混合明表	47.94＋1.27P
序列密码，密钥长度P（沼泽Ⅳ）	23.97N
连续使用的N个无关混合密表	23.97N＋1.27P
N个无关的混合密表，密钥长度P，密钥由M个不同字符组成	.90logM＋23.97N
随机双码代替密	1460.61
普莱费尔密表	22.69
四方密表（两个混合密表）	45.38
随机N码代替密	.90log(26^N)！

每个字母有N个代替的多名码代替　$(\log 26N!/N!^{26})/1.11$

表中最后一项结果使用的密钥数可按下述方法求出。因为每个字母有N个代替，一共有26N种代替。"A"的N种代替可从$26N!/N!25N!$种方式中选择（每次从26N种组合中选取N种）。"B"的N个代替则可以从$25N!/N!24N!$种方式中选择，等等。因此，脱密时供选择的密钥总数为

$$\frac{26N!}{N!25N!}\cdot\frac{25N!}{N!24N!}\cdots\frac{N!}{N!O!}=\frac{26N!}{(N!)^{26}}$$

多名码代替是很有意思的，因为在多名码代替中，人们要加密和产生唯一的密报，不仅需要知道代替密钥，而且需要知道代替表的使用顺序，而脱密只需要知道代替表。上面的结论基于这样的假定：如同用同一个代替表代替每一个字母产生一个单表代替一样，这里所用的密钥不受进一步的限制。对于很大的N，可以用Stirling 因子近似法(factorial approximation)，求出的唯一解点值约为$U=33.14N$。实际上，这个结论告诉我们如何设计一种唯一解点总是长于电文长度的多名码代替密码。例如，加密一份500个字母的电文，我们需要$U=33.14N>500$即$N>15$以免出现唯一解。

普莱费尔密码的唯一解点对大多数人好象都太短。谁能够破译一份只有23个字母的普莱费尔密码？初学者解这种密码的方法是只用双码频率表而不连续构造密钥方表（keysquare），显然是徒劳无功的，因为这种方法需要大约1400个字符，远长于通常的电文长度。

破译这样一种接近唯一解点的有趣的例子可在1936年11—12月版的《陆军通讯兵公报》（Signal Corps Bulletin）中找到。A. Monge——一位美陆军二等兵破译了一位英国将军提供的一份30个字母的复杂的普莱费尔密报：

BUFDA GNPOX IHOQY TKVQM PMBYD AAEQZ。Monge获得的明文使这份密报的意思变得十分清楚。虽然这份密报的长度接近唯一解点，但Monge正确地假定其密钥方表属于那种密钥字混合型。把自己限制到这种不完全的混合字母表就更进一步缩短了唯一解点。举例来说，如果密钥方表的最末一排可以假定为VWXYZ，其唯一解点不会落在16.56个字母以外。因此，Monge的状况比第一次出现要好。（根据Monge的解，好象没有有关资金转移的记录。）

正如大家都知道的，证明一个解是唯一的（这就是唯一解点要做的）和实际上找到这个解（破译密报）是两码事。唯一解点研究仅指可以落入敌手而不致泄露明文的同样地键控电文的总量。

作为一个例子，我们研究在D·卡恩的名著《破译者》〔3〕中描述的盟军在第二次世界大战中使用的SYKO（西科）战地密码。一种SYKO体制由32张互无联系的乱序密表组成，每张表37个字符(26个字母，10个数字和一个间隔“-”)。顺序使用这些密表产生一个周期为32的多表密码。我们得到

$$U = \log(37)!^{32}/1.11 = 1244\text{字符}$$

每张密表大约39个字符。根据Kerckhoff的重迭法破译这种体制也要求每张密表39个字符。在这种情况下，唯一解点和求解需要的电文量是很接近的。

唯一解点研究的用途，在对用混合密钥即Vernam密钥密码的研究中提供了另一个例子。一种推移序密钥的维吉尼亚密表是这种体制的一个很简单的例子。如果起始密钥长度为7，推移序指数为1，则密钥反复前的总周期$7\times 26 = 182$个字符。这种密码的唯一解点不是$1.27P = 1.27(182) = 231.4$个字符，如同前面的表中推定的那样。因为密钥是混合密钥，起始密钥有

26^7种选择，推移指数有26种（1，2，3，…，26），构成总数为$26^7 \cdot 26$种可能的密钥。这就产生一个$\log 26^7 \cdot 26/1.11 = 10.20$个字符的唯一解点。真是惊人的减少！

一般说来，如果起始密钥是P_1个字母表字符后紧接P_2个字母表字符的派生密钥，那么，密钥选择数为$26^{P1} \cdot 26^{P2} = 26^{P1+P2}$种。因此，这种混合密钥体制的唯一解点是

$$U = \log 26^{P1+P2}/1.11 = 1.27(P_1 + P_2)$$

这个有趣的结果表明，Vernam式的混合密钥使唯一解点增加的长度取决于各个密钥的和，尽管密钥长度取决于各个密钥的积。可见Vernam式的混合密钥在改善一种密码体制的保密性方面作用是微乎其微的。Bryant Tuckerman用计算机调查了Vernam—Vigenere密码〔4〕，揭示出Vernam密钥构成的同样的不一致的特性。仙农的著作清楚地描述了为什么Tuckerman正好得到了他得出的结论。现在回到我们前面求E即求假脱密的期望数的公式上来，我们可能获取某些有意义的信息。假定一种滚动密钥的维吉尼亚密码用正规的英文明文作密钥。经验证明：只要密报有足够的长度，几乎所有使用这种体制的密报都是可破的。仅根据这一点，人们就可以估计出英文的多余度。长度为N个字符的连贯的英文明文串数约为

$$10^{RN}$$

这里的R是英文的多余度，N是存在于明文串中的字符数。只有当N超过15个左右的字符，这个公式才成立。对一种滚动密钥密码，上面这个数也是长度为N的密钥数。总的说来，求E的公式就变成

$$E = 10^{-DN+RN} + 10^{-DN} = 10^{(-D+R)N}$$

对于足够大的N，如果这种密码是唯一可破的，那么，随着N增大，E必定接近于零。只有指数$-D+R$小于零时才出现

这种情形。因为$D = \log 26 - R$，我们有

$$-D + R = -D + \log 26 - D = -2D + \log 26 < 0$$

或

$$D > \log 26/2 = \cdot 71$$

用百分数表示，·71的多余度约为$\cdot 71/\log 26 = 50\%$。我们得出结论：如果所描述的这种滚动密钥密码是唯一可破的，那么英文的多余度至少必须为50%。

回到我们原先求E的公式并代入$R = .30$和$D = 1.11$的标准值，我们发现，$N > 1$时，$E < 1$。换句话说，一个滚动密钥的维吉尼亚密码的解是唯一的即使只接收到一个字符。这个结论显然是错误的。问题出在原来的公式上，它是用英文的多余度1.11进行计算的。实际上，只有在计算英文的熵时，所用的字母组多于$N = 20$时才达到这个值。D和R都是N的函数，在求E的公式中，一个较好的方法应是用根据英文熵计算出的适当的N值的D和R。附录B指出如何能推导出这类近似公式。把这个近似公式用在求E的方程中，结果形成唯一解点曲线，见图Ⅰ。这条曲线表明，滚动密钥的维吉尼亚密码的唯一解点大约为8个字符。作者以为这个结果与经验一致。

大多数熟悉唯一解点理论的人都觉得其结果过于严格，并且觉得实际上要影响一份给定密报的解所需要的字符数远大于规定的最小数目。毫无疑问，这种感觉来源于用某种方法分析一种特定的密码得到的经验。被带入密文的消息源的多余度是多数密码分析方法的基础。一种给定的方法对出现在密报中的多余度利用得越少，脱密所需要的密报文就越多。根据这些想法，人们可以定义一个有效唯一解点（effective unicity point）。

以单表代替为例，（这并不是随机密码的一个令人满意的例子，但用它容易说明我们的方法。）单表代替的唯一解点约为

$$U = \log 26!/1.11 = 24\text{字符}$$

（由于W·F·Friedman相应的估计量为25个字符，这是一个相当令人满意的估计量。）如果人们打算只用字母频率来破译单表加密的电文，那么，用来计算唯一解点的有效多余度是0.20位数。这是因为根据单字母频率计算出的英文熵约为1.21；因而，其多余度为$\log 26 - 1.21 = .20(14\%)$。这样

$$U_{eff} = \log 26!/.20 = 133\text{字符}$$

相应地，只用双码组频率数据，$U_{eff} = 65$字符；用三码组频率数据$U_{eff} = 55$字符。当我们用8码组频率数据时，有效解约需38个字符。在实际解密过程中是这样做的：用一阶熵知识获得一个立脚点，当暴露出整个明文插入码（Patches)时再迅速扩大到用高阶熵知识。这就说明了为什么只需要25个字符，也说明了为什么通常在一个合理的时间量内，达成破译所需的字符数远远多于其最小字符数。

一般说来，加密结果要进行"扩散"。这种扩散处理迫使一份电文的侵入者要先截获大量的加密资料才能够重新推出源消息中的多余度。一个简单的启发式的论据将说明这一点。假定用密钥字KING对片语"ofthe"进行维吉尼亚加密。英文明文的熵被定义为

$$R = \lim_{N\to\infty} -\sum_{i=1}^{26^N} P_i^N \log P_i^N / N$$

其中P_i^N是英文N码组出现的概率。在密码中，片语"ofthe"可能表现为YNGNO，WSZRM，BLDPR，或UPBVK，这取决于这个片语与密钥字KING的交叉位置。这四个5码组在收到的电文中，每一组出现的概率均为P/4，假如P是"ofthe"出现在明电文中的概率。另外，我们可假设密文片语"OFTHE"出现

的概率为零，而且明文片语“yngno”，“wsgrm”，“bldpr”和“upbvk”出现的概率也为零，因为它们不是英文中连贯的片语。

一个周期为P的多表密码，在明文中出现的概率为P_i^N的一给定N码组可产生P个N码组的密文。每一组出现的概率均为P_i^N/P。如果N非常大而且这个密文N码组在明文中出现的概率为零，那么，原来有N码组明文在密文中的期望出现概率也为零。

因此，如果我们计算消息源的熵，一串形如

$$-P_i^N \log P_i^N/N$$

的项要被替换。如果是根据密文来计算熵，被替换的项形如

$$-P(P_i^N/P)\log(P_i^N/P)/N$$

总计被替换的项数为

$$-(P_i^N/P)\log(P_i^N/P)/N$$

这两个值逐项的差为

$$-P_i^N \log P/N$$

所有这类差的总和为

$$(\sum_i P_i^N \log p)/N = (\log P/N)(\sum_i P_i^N) = \log P/N$$

最末一项可看作是多表密码的一个“熵扩散因子”(entry diffusion factor)。因为当$N\to\infty$且P固定时，$\log P/N\to 0$，可见如果截获足够的电文，就可以用密文而不是用明文来计算整个的多余度。这是很重要的，因为用来分析电文的大多数统计资料最终是以密报中出现的多余度为基础的。

因为源消息中的多余度被带进密报中，这给密码分析者提供了基本的攻击手段，所以任何减少源消息中出现的多余度的方法最终将使将来的侵入者的问题复杂化。英文的估计多余度为78%；这就意味着大多数英文原文可删去78%，而且删除后

的原文可以清清楚楚地恢复原样。（当然，并不是"任何"原文都可删去78%，原文中某些部分比其他部分含有更多的意义。）如果充分利用这种方法，将把英文的多余度减少为零，从而使唯一解点扩大到无限大。具有零多余度的电文加密后绝不能唯一地脱密，不管截获多少电文。人们可以计算出原文删除的百分比不同时，将出现的多余度的百分比的变化情况。

令P表示被删除的原文的百分率。"删除后"长度为N的英文电文总数等于"删除前"长度为N′的英文电文的总数，这里N和N′的关系由方程$N' - PN' = N$表示。由于原来的电文总数是$10^{RN'}$（R = 英文的熵 = .30），长度为N的新的电文数为

$$10^{(R/(1-P))N}$$

删除后的有效熵由R变为R/(1 − P)。通过计算

$$(\log 26 - (R/(1-P))/\log 26$$

可以求出多余度的新的百分比。图Ⅲ表示明文删除的百分比与剩下的多余度百分比的关系曲线。

举例来说，所有的元音字母A，E，O，U和Y（当作为元音字母出现时）通常都可从英文中删去而不会冒丢失原义的危险；这种删除使原文平均缩短40%。因此，删除后的有效熵变为.30/.6 = .5或用多余度表示约为65%。

图Ⅲ回答了下面这样一个有趣的问题：为了使唯一解变得不可能，滚动密钥维吉尼亚密码的密钥和明文究竟必须删除多少？仅仅删除元音字母是不够的，因为我们前面的结论已经表明，需要少于50%的多余度。参见图Ⅲ，我们看到，把多余度的百分比降到50%或更低，必须同时把密钥和明文都删除58%左右。

假如读者想试一试自己在多余度减少的密码方面的技能，下面给出一份滚动密钥维吉尼亚密码的密报，其中密钥和明文