中国科学技术大学 University of Science and Technology of China | 网络空间安全学院 School of Cyber Science and Technology

作品类别： ✔软件设计 ☐硬件制作 ☐工程实践

《密码学导论》课程大作业作品设计报告

作品题目：以多个字符串为密钥的维吉尼亚密码变体

团队名称：吉奥万巴蒂斯塔贝拉索队

团队人员：李梓豪

2024 年 6 月 6 日

基本信息表

作品题目：以多个字符串为密钥的维吉尼亚变体密码

作品内容摘要：

将普通维吉尼亚密码的密钥字符串升级为包含多个字符串的字符串库，在加密中随机选取其中的字符串进行加密而形成的维吉尼亚变体密码，其相比普通维吉尼亚密码有一定安全性的提升；同时仍存在一些不足之处，从攻击者的角度，作者设计了一个用于检验密文中单字母、双字母和三字母出现频率的程序，对多个密文进行分析。

关键词（五个）：

维吉尼亚密码 频率统计攻击 随机数

团队成员（按在作品中的贡献大小排序）：

序号	姓名	学号	任务分工
1	李梓豪	PB22000130	程序设计、代码实现、后续思考

1.作品功能与性能说明

维吉尼亚密码是一种简单的多表代换密码。其被破译的核心是通过重合指数法、Kasiski 等办法找到密钥的长度，然后对每一位置上的数据进行频率统计攻击等针对单表代换密码的攻击方法。针对这一缺陷，可以将原有的密钥的单个字符串扩展成为多个字符串（数量不超过 36）的集合，从而产生一种维吉尼亚密码的升级版本。简而言之，它相当于多个普通维吉尼亚密码以随机的顺序混合而成。

如上述，此维吉尼亚密码的改进版本避免了单一密钥字符串反复加密带来的弱点从而使得普通的重合指数法、Kasiski 失效；从频率统计分析来看单字母、双字母、三字母的各个出现频率都更加均匀；另外由于随机数的存在，即使明文、密钥相同，每次运行产生的密文也不一样。相比普通维吉尼亚密码，缺点有：1.密钥复杂得多，包含多个字符串且顺序不能混淆；2.密文更长（相比普通维吉尼亚密码的密文多了一个 c2 字符串的长度，其长度约为原密文的 $1/E(l_i)$， $E(l_i)$表示密钥字符串的平均长度）3.若攻击者能获得同样明文和密钥经过多次加密形成的不同密文，则有可能恢复出密钥每个字符串的长度，进而用类似破译维吉尼亚密码的方法进行破译

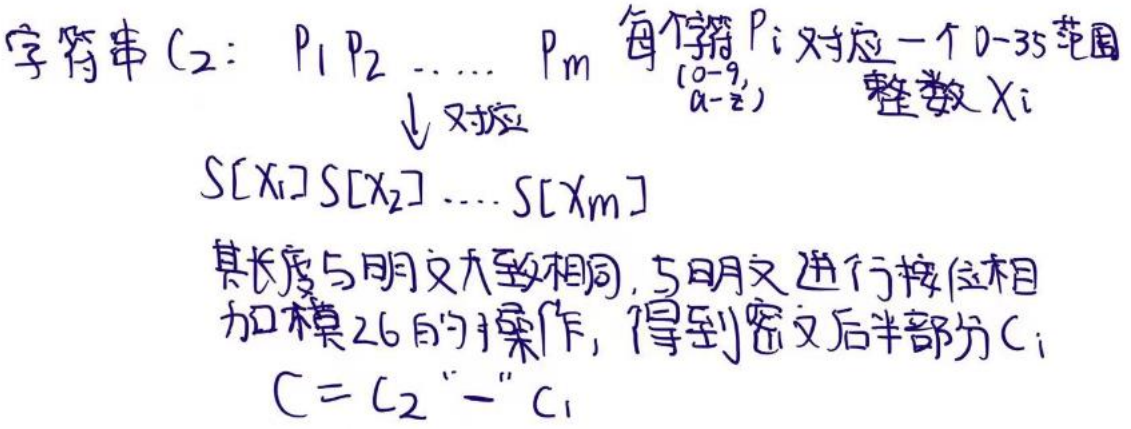
以恢复每个字符串长度为核心的攻击简要思路如下：

根据 c2 字符串中出现的字符及数量可以确定每个字符串被用到的次数，同时密文‘-’后的长度已知即明文长度已知，由此可以列出一个 n 元 1 次不等式（这里的 n 为密钥包含的字符串数）。当可以获得多次加密得到的密文时，则可以列出多个不等式组成的不等式组。得到的密文数量越多，求该不等式组得到的解越少；得到可能的每个字符串长度后，将明文密文分割即可得到多个单表代换密码，再分别进行频率统计攻击即可。

2.设计与实现方案

2.1 实现原理

将所有的密钥字符串编号 0-9，a-z；加密时，每次产生一个随机数（范围为字符串个数）从而随机选取一个密钥字符串对明文进行加密，该密钥字符串（设其长度为 l_i）加密完明文的 l_i 个长度后就进行更换，随机选取下一个密钥字符串进行加密。用一个新字符串（记为 c2）记录选取的所有密钥字符串的顺序，新密文将 c2 与密文连起来，中间用一个横杠连接，则由密文和密钥可以完全恢复出明文。



2.2 参考文献

中国科学技术大学 24 春《密码学导论》课件《Crypt01-古典密码学》

2.3 运行结果

明文选择了英文诗歌《未选择的路》共 565 个字母；为更好测试程序性能，将原明文重复 180 遍得到新的明文，长度 101700，密文长度在 110800 左右；decrypt.py 程序对密文进行解密，得到《未选择的路》原文重复 180 遍（即长度 101700 的新明文），实现了基本的加密解密功能

另外，由于该长度 101700 的新明文由一段文字重复多遍而成，其频率统计特性更加突出，相比正常的 10e5 长的英文更易受攻击

2.4 技术指标

单字母、双字母、三字母的频率统计特性；加密、解密算法运行速度

3.系统测试与结果

3.1 测试方案

为测试其对抗频度攻击的性能，编写 frequency.py，分别记录密文各个字母出现的概率、连着两个字母出现最多的 100 个组合、以及连着三个字母出现最多的 100 个组合。

3.2 功能测试

根据 3.1 的方案来测试该密码抵御频度攻击的功能：

先看单字母的频率统计特性：（同样的明文加密三次得到三个不同的密文）

a	0.03974	0.03903	0.03976
b	0.03865	0.03745	0.03701
c	0.03624	0.03696	0.03658
d	0.03837	0.0383	0.03806
e	0.04097	0.04105	0.04065
f	0.03857	0.03761	0.03815
g	0.03871	0.03939	0.03935
h	0.03679	0.0377	0.03762
i	0.0387	0.03972	0.03807
j	0.03954	0.03886	0.03954
k	0.03903	0.03894	0.04024
l	0.04195	0.04257	0.04053
m	0.03489	0.03532	0.03487
n	0.03524	0.0343	0.03547

o	0.04482	0.04526	0.04483
p	0.0361	0.03636	0.03773
q	0.03583	0.03509	0.03695
r	0.03522	0.03607	0.03509
s	0.04188	0.04305	0.04297
t	0.04054	0.03957	0.03932
u	0.03484	0.03403	0.03463
v	0.03839	0.03833	0.0386
w	0.04385	0.04256	0.0422
x	0.03566	0.03639	0.03648
y	0.03774	0.03808	0.03727
z	0.03772	0.038	0.03801

观察发现每个字母出现频率都在 0.0385 上下，较之单表代换密码和普通维吉尼亚密码更加平均；但是仍不是完全平均，例如从三次加密来看，每次 o 频率都在 0.0045 左右，为 26 字母最频繁出现，说明当密钥和明文一定的时候密文仍然会呈现一定的单字母统计规律。

为什么出现这样的单字母频率特性？本密码其实相当于多个维吉尼亚密码的随机混合，而维吉尼亚密码在给定密钥字符串的情况下密文也是会呈现出频率统计特性。当明文足够长时，每个密钥被选到的频率比较接近，整体的频率统计特性可以近似为是所有单个的维吉尼亚密码加密的频率统计的平均。这个也可以作为攻击该密码时的一个突破点。

双字母、三字母的频率统计特性如下：

ow	239	os	235	jw	247
wt	231	ab	234	sv	234
wo	229	wl	233	jt	230
jw	229	iy	226	wl	222
da	228	le	225	fs	219
kw	227	za	223	wh	219
ov	223	jw	223	eq	219
tl	222	oo	221	vg	217

kn	220	ik	221	ow	217
bl	218	cj	220	rj	217

mff	36	mff	33	qzx	32
taa	35	lwj	32	jve	31
ovt	33	lxs	30	zjw	30
jve	32	fkw	30	abi	29
ksk	30	siy	30	omf	29
rjr	29	tfg	30	xcj	29
abh	28	tez	29	qss	29
egd	28	rjr	29	lwj	29
mwo	27	gtf	28	uza	28
phc	27	ovt	28	boo	28

相较于 10e5 的明文密文长度，这些双字母三字母组合的出现频率不算很高；但是三次加密中，jw,ow,wl 均在双字母出现频率 top10 中出现 2-3 次；mff,ovt,lwj 在三字母出现频率 top10 中出现 2 次，这样事件出现的概率远高于完全随机字母的字符串的频率统计，说明 jw，ow,wl 很可能是 of，an，as 等常见单词刚好经过同一密钥字符串的同一位置加密，mff,ovt,lwj 则可能对应常见英文单词中的 the,and 等等。

3.3 性能测试

理论的加密、解密算法均为 O（n）,n 为明文长度；明文长度为 10e5 左右时，运行加密或解密平均约需要 0.045 秒；长度为 10e6 左右时，时间平均约需要 0.45 秒。运行速度较快。

3.4 测试数据与结果

明文 m.txt(真正用于加密的明文是它的内容重复 180 遍而成)；
密钥 key.txt；密文 c.txt；由密文解密得到的明文 mnew.txt

4.应用前景

该密码作为简单多表代换密码的升级版本，可以对一些对安全性要求不是很高的消息进行加密；该密码优点是加密解密思路简单、计算量少（与普通维吉尼亚密码大致相同），因此适合在无法接触计算机、只能进行人力运算的条件下对信息进行加密解密。

5.结论

通过将普通维吉尼亚密码的密钥升级为多个字符串可以有效提高其抗频度攻击的功能，而在加密解密的计算速度上几乎保持一致；缺点是密钥更加复杂、密文长度增加，以及仍然可以通过频度攻击、还原密钥长度等方式对该密码进行有效攻击。