Sistemas Polialfabéticos

Este documento no pretende ser un curso exhaustivo, en el mejor de los casos, únicamente puede considerarse como un conjunto de notas complementarias en alguna asignatura. Por supuesto, todo documento es susceptible de mejora, cualquier sugerencia o comunicación de error u omisión será bienvenida.

Cifrado polialfabético

Las primeras aproximaciones al cifrado de mensajes consideraban una transformación fija para todos los elementos del alfabeto considerado para elaborar el mensaje a cifrar. La distribución no uniforme de los símbolos de un mensaje, sin importar el idioma, hace que estos sistemas de cifrado sean sensibles a un ataque por análisis de frecuencias.

El cifrado polialfabético se propuso como una vía de ocultar estas relaciones de frecuencia, proporcionando un método que permitía que un símbolo del mensaje se cifrara de varias formas posibles y que dos símbolos distintos se cifraran con el mismo símbolo en el critpograma. Estas características hacen, en principio, inviable un ataque por análisis de frecuencias de los símbolos del criptograma.

Cifrado Vigenèrè

Aunque no es el primer sistema de cifrado poliafabético descrito, el más conocido de ellos fue inicialmente atribuido a Blaise de Vigenèrè que lo describió en el siglo XVI aunque posteriormente, y pese a las (ligeras) diferencias entre las propuestas, posteriormente se le reconoció la autoría del método a Giovan Batista Belaso (en el siglo XIX). Debido a sus características, se consideró indescifrable hasta el siglo XIX.

Considerando un alfabeto de n símbolos, en un cifrado polialfabético la clave consiste en un conjunto de k valores de $\mathbb{Z}_n$, donde tanto k (el tamaño de la clave) como los valores de los desplazamientos deben permanecer secretos.

$$k = \{d_1, d_2, \ldots, d_k\}$$

El cifrado del símbolo que ocupa una cierta posición i considera el desplazamiento indicado por el elemento de la clave que ocupa posición congruente módulo k.

$$e_k(x_i) = x_i + k_{i \bmod k} \bmod n$$

y de forma análoga, el descifrado se consigue deshaciendo los desplazamientos indicados por la clave.

$$d_k(y_i) = y_i - k_{i \bmod k} \bmod n$$

Ejemplo 1. *Dado el mensaje sobre el alfabeto mayúsculas del español ($n = 27$ símbolos):*

`CONESTETEXTOINTENTAMOSILUSTRARUNCIFRADOPOLIALFABETICO`

y considerando la clave de tamaño 6:

`MCBGYM`

o lo que es lo mismo, el conjunto de desplazamientos $\{12, 2, 1, 6, 25, 12\}$, el cifrado del primer símbolo del criptograma se realiza considerando el valor en $\mathbb{Z}_{27}$ del símbolo 'C' y el primer desplazamiento:

$$'C' +' M' = 2 + 12 \bmod 27 = 14 =' \tilde{N}'$$

aplicando este proceso a los primeros símbolos del mensaje se obtiene:

`ÑQÑKQF`

una vez considerados todos los desplazamientos indicados en la clave, intentamos cifrar un símbolo cuya posición es congruente con la del primer símbolo (intentamos cifrar el septimo símbolo y $1 \equiv 7 \pmod 6$), por lo que necesitamos utilizar de nuevo el primer desplazamiento. En otras palabras, la clave se utiliza de forma cíclica mientras sea necesario cifrar nuevos símbolos. El cifrado del mensaje da como resultado:

`ÑQÑKQFPVFDRATOUKLFMÑPYGWGUUXYDGODÑDDMFPVNWTCMLYNPVJIN`

La clasificación de este método como polialfabético se basa en que, dada una clave de tamaño k, una forma de ver el cifrado es considerar que se utilizan k alfabetos de forma cíclica. La primera representación de este hecho es la *tabula recta* de Johannes Trithemius, representada en la Figura 1.

Una clave Vigenèrè considera k valores de un conjunto de talla igual a la del alfabeto, por lo que el espacio de claves del método Vigenèrè es n^k donde n denota el tamaño del alfabeto considerado para redactar los mensajes.

0	A B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z
1	B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z A
2	C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z A B
3	D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z A B C
4	E F G H I J K L M N Ñ O P Q R S T U V W X Y Z A B C D
5	F G H I J K L M N Ñ O P Q R S T U V W X Y Z A B C D E
6	G H I J K L M N Ñ O P Q R S T U V W X Y Z A B C D E F
7	H I J K L M N Ñ O P Q R S T U V W X Y Z A B C D E F G
8	I J K L M N Ñ O P Q R S T U V W X Y Z A B C D E F G H
9	J K L M N Ñ O P Q R S T U V W X Y Z A B C D E F G H I
10	K L M N Ñ O P Q R S T U V W X Y Z A B C D E F G H I J
11	L M N Ñ O P Q R S T U V W X Y Z A B C D E F G H I J K
12	M N Ñ O P Q R S T U V W X Y Z A B C D E F G H I J K L
13	N Ñ O P Q R S T U V W X Y Z A B C D E F G H I J K L M
14	Ñ O P Q R S T U V W X Y Z A B C D E F G H I J K L M N
15	O P Q R S T U V W X Y Z A B C D E F G H I J K L M N Ñ
16	P Q R S T U V W X Y Z A B C D E F G H I J K L M N Ñ O
17	Q R S T U V W X Y Z A B C D E F G H I J K L M N Ñ O P
18	R S T U V W X Y Z A B C D E F G H I J K L M N Ñ O P Q
19	S T U V W X Y Z A B C D E F G H I J K L M N Ñ O P Q R
20	T U V W X Y Z A B C D E F G H I J K L M N Ñ O P Q R S
21	U V W X Y Z A B C D E F G H I J K L M N Ñ O P Q R S T
22	V W X Y Z A B C D E F G H I J K L M N Ñ O P Q R S T U
23	W X Y Z A B C D E F G H I J K L M N Ñ O P Q R S T U V
24	X Y Z A B C D E F G H I J K L M N Ñ O P Q R S T U V W
25	Y Z A B C D E F G H I J K L M N Ñ O P Q R S T U V W X
26	Z A B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y

Figura 1: Tabula recta. Un cifrado con la clave $MCBGYM$ considera cíclicamente las filas 11, 2, 1, 6, 25 y 11 (desplazamientos sobre el alfabeto original) para el cifrado de los símbolos del mensaje

Criptoanálisis

Dado un criptograma obtenido mediante el cifrado Vigenèrè utilizando una clave cualquiera de tamaño k, es importante notar que los caracteres cuya posición es congruente módulo k se han cifrado utilizando la misma componente de la clave, esto es, se han cifrado con el mismo desplazamiento. Visto de otra forma, si descomponemos el mensaje en k subtextos, cada uno conteniendo los caracteres cuya posición en el criptograma es congruente módulo k, cada subtexto pueden considerarse un cifrados Caesar, pudiendo descifrarlos independientemente mediante un análisis de frecuencias. La recomposición de todos ellos dará como resultado el mensaje en claro.

Ejemplo 2. *Dado el siguiente criptograma obtenido mediante cifrado Vigenèrè:*

```
NDJPBLYWNO RXPXDIZOJT DEDXVXJWPR CÑÑCHMPNDX UYRXPEHNNH
XSBABTUJXX DABVXGTOMD ÑMLJQTASRV QJRHPLVXVN FNPOBBOCRE
WEWMDVYIJT YMCURXASMI UJJOFAVTWJ IIWAMEJPJW TRFXYYWPTT
ÑTCUHEDEOB LUECPFRVQY RXPMÑGCYJO XEWMJDJGTF BTMJIININI
JLZIÑANTDH THPLYMBOGX PXDIBLZIAB FBJGJQFEWI YOJHPNMIVN
NHPRKLBYQI YSKCNNIIQO BFJCRHTVRW DUPTGENXBY WEFAMTMJZX
TLRSJHISGN KGJFNWTSNX ÑMJGGEWVQU XTMRKIKÑJQ PRMIUJWIEU
UNJXTGMDKÑ VJIIWOCWNN UOMZKFQYRX ASRGMCAMOU KENNVYPPÑL
UCYIYRÑVXH XXDURLTUHO MVÑÑCCPMLA NTNHJJPCDI BCJWMSRXTY
ÑIXPUIBYWZ TRKVXGTXML
```

supongamos que, de algún modo, se sabe que el tamaño de la clave utilizada es 8. *Si consideramos la descomposición del mensaje en subtextos con caracteres que ocupan en el criptograma posición congruente módulo* 8, *el tercer subtexto (caracteres en posiciones,* 3, 11, 19, *etc.) es:*

```
JRJJHRXXTQRFRYRJIJWHERJJIZTGZJJNQIRPWZINJXJWTIURAVYXHPJJÑWT
```

*Un análisis de frecuencias sobre este subtexto da como resultado que el símbolo más frecuente es 'J' (*9 *en* $\mathbb{Z}_{27}$*). Asumiendo que el subtexto es suficientemente extenso para ser representativo, este símbolo corresponde al cifrado de 'E', se obtiene el tercer desplazamiento* $d_3 = 5$. *Esto permite* descifrar *el tercer subtexto deshaciendo el desplazamiento, obteniendo:*

```
ENEECNSSOMNANTNEDERCZNEEDUOBUEEIMDNLRUDIESERODPNVQTSCLEEJRO
```

que son los caracteres en posiciones, 3, 11, 19, *etc. del mensaje en claro.*

Análogamente, por ejemplo, el quinto subtexto (caracteres en posiciones, 5, 13, 21, *etc. del criptograma) es el siguiente*

```
BPDPPPBDMAPPWJAFWTTDPPXTNÑPPAFPPYQTGFTGTGMPEMWMAOPYDMLPMXTM
```

*En este caso, el análisis de frecuencias sobre este subtexto da como resultado que el símbolo más frecuente es 'P" (*16 *en* $\mathbb{Z}_{27}$*). De nuevo asumiendo que el subtexto es suficientemente extenso, este símbolo correspondería al cifrado de 'E', obtieniendo* $d_5 = 12$*.*

Procediendo de igual forma para cada subtexto, en caso que todos los textos sean contengan información suficiente, puede obtenerse que la clave es $k = \{9, 21, 5, 4, 12, 0, 10, 20\}$*. Se deja el descifrado de los sub textos y la recomposición del mensaje como ejercicio.*

Kasiski

Teniendo en cuenta que, una vez se conoce el tamaño de la clave, el cifrado Vigenèrè es sencillo de romper, a continuación se exponen dos aproximaciones distintas para obtener el tamaño de la clave.

La primera se debe a Friedrich Kasiski, oficial prusiano de infantería en el siglo X!X. El conocido como método Kasiski considera que, dada cualquier clave y cualquier fragmento de texto que aparece en un mensaje, debido a que la clave se considera de forma secuencial, la clave permite cifrar el fragmento de tantas formas distintas como tamaño tiene la clave.

Ejemplo 3. *Dada la clave Vigenèrè* $VIGENERE$*, esta permite cifrar un mismo texto de ocho (su longitud) formas distintas, ya que, independientemente de la longitud del texto, la clave puede considerarse de (en este caso) ocho formas distintas:*

```
FRAGMENTOQUEAPARECEENELMENSAJE
VIGENEREVIGENEREVIGENEREVIGENE
IGENEREVIGENEREVIGENEREVIGENEV
GENEREVIGENEREVIGENEREVIGENEVI
ENEREVIGENEREVIGENEREVIGENEVIG
NEREVIGENEREVIGENEREVIGENEVIGE
EREVIGENEREVIGENEREVIGENEVIGEN
REVIGENEREVIGENEREVIGENEVIGENE
EVIGENEREVIGENEREVIGENEVIGENER
```

De este modo, dos fragmentos iguales de criptograma de longitud suficiente, probablemente corresponden a fragmentos de mensaje en claro iguales que se han cifrado utilizando la clave de la misma forma, lo que implica que la distancia entre las posiciones en las aparecen estos fragmentos repetidos del criptograma es un múltiplo de la clave.

Mientras más cortos sean los fragmentos del criptograma que se consideran, mayor probabilidad existe de que estos no correspondan con el mismo fragmento en el mensaje original. No existe una regla para considerar un tamaño mínimo, en general, un fragmento de longitud cinco puede considearse representativo.

Ejemplo 4. *Considerando el siguiente criptograma Vigenèrè:*

```
AXKODVPKAJ SwriuvNJKS FOULJRKVRC TBBWILRXUL ISSWROYAAB
YRDLSHIDYW FLSJLAUÑÑÑ FZYDRSCDJJ EDSGRVNLJH GMRZSOCWSD
YOÑZQPZHLE PZPOSWCDDV IDKÑHLNHKD JHYLDRWKKV VCWLMSXOVE
FHPOIDFOGO YOFBRPJJES SWRWFTPSKÑ ZOÑZWXKFVP SHZDJHOSEV
WFAHPLEHQB UGRVPZOJHw riuvOFAHCM WOWAKPHOÑV MJKGRXDVJH
ÑGRCBYOSRH ADBPAHJHSZ SSWWSGVGJK QOQSIOELOS XDHLDHZDAW
VVJGWBJRIX BTWZÑVVDEL BGKFIOÑJEO YSÑCBVXIKP RCDVIDXHGF
MAWRUFÑÑBB JDJHYZTKAH VÑÑKBSESSW CDJTZWBLQF BRAHWXRAFY
IWZHACFJLB YWFFJYHOIÑ ÑGFBPWQLNL EHABKIRNUV OWKVÑDJLHS
OHZAMVOSXY VCBJLAUWÑV
```

en él se han señalado en minúscula los dos segmentos de mayor tamaño que aparecen repetidos. Teniendo en cuenta que el primero aparece en la posición 12 *y el segundo en la posición* 220*, de acuerdo con la hipótesis en la que se basa el método de Kasiski, la distancia entre estas posiciones* $220 - 12 = 208$ *es un múltiplo del tamaño de la clave.*

Índice de coincidencia

Otra aproximación para obtener el tamaño de la clave utilizada en un cifrado Vigenèrè tiene en cuenta una medida de dispersión de las frecuencia de aparición de los símbolos respecto a una distribución uniforme:

$$MD = \sum_{\forall i} \left(p_i - \frac{1}{n} \right)^2 = \sum_{\forall i} \left(p_i^2 - \frac{2p_i}{n} + \frac{1}{n^2} \right)$$

donde, teniendo en cuenta un alfabeto de 27 símbolos:

$$MD = \sum_{0 \leq i < 27} \left(p_i^2 - \frac{2p_i}{n} + \frac{1}{n^2} \right) = \sum_{0 \leq i < 27} \left(p_i^2 \right) - 0{,}037$$

Cada idioma tiene una distribución particular de aparición de los símbolos en un mensaje. En un texto donde los símbolos presentan una distribución própia del español:

$$IC = \sum_{0 \leq i < 27} \left(p_i^2 \right) = 0{,}072 \quad \Rightarrow \quad 0 \leq MD \leq 0{,}035$$

Intuitivamente, el *índice de coincidencia* (IC) mide la probabilidad de que escogidas dos posiciones distintas tomadas al azar en un mensaje, los símbolos que ocupan esas posiciones sean iguales. Considerando un criptograma, puede estimarse el IC utilizando la frecuencia de aparición de los símbolos:

$$IC \simeq \frac{\sum_{i=0}^{26} f_i(f_i-1)}{N(N-1)}$$

donde f_i denota el número de ocurrencias del caracter i-esimo del alfabeto en un criptograma de N símbolos. De este modo, en un texto en claro en español:

$$IC = MD + 0{,}037 \quad \Rightarrow \quad 0{,}037 \leq IC \leq 0{,}072$$

Ejemplo 5. *Considerando el criptograma Vigenèrè del ejemplo anterior:*

```
AXKODVPKAJ SWRIUVNJKS FOULJRKVRC TBBWILRXUL ISSWROYAAB
YRDLSHIDYW FLSJLAUÑÑÑ FZYDRSCDJJ EDSGRVNLJH GMRZSOCWSD
YOÑZQPZHLE PZPOSWCDDV IDKÑHLNHKD JHYLDRWKKV VCWLMSXOVE
FHPOIDFOGO YOFBRPJJES SWRWFTPSKÑ ZOÑZWXKFVP SHZDJHOSEV
WFAHPLEHQB UGRVPZOJHW RIUVOFAHCM WOWAKPHOÑV MJKGRXDVJH
ÑGRCBYOSRH ADBPAHJHSZ SSWWSGVGJK QOQSIOELOS XDHLDHZDAW
VVJGWBJRIX BTWZÑVVDEL BGKFIOÑJEO YSÑCBVXIKP RCDVIDXHGF
MAWRUFÑÑBB JDJHYZTKAH VÑÑKBSESSW CDJTZWBLQF BRAHWXRAFY
IWZHACFJLB YWFFJYHOIÑ ÑGFBPWQLNL EHABKIRNUV OWKVÑDJLHS
OHZAMVOSXY VCBJLAUWÑV
```

El índice de coincidencia de todo el criptograma es 0,0423264*, lo que indica que el criptograma no corresponde a un cifrado monoalfabético, o por permutación.*

Si dividimos el criptograma en dos bloques, el primero con los símbolos que ocupan posición congruente con 0 módulo 2 (posición par) y el segundo con los símbolos que ocupan posición congruente con 1 módulo 2 (posición impar), el IC de estos bloques devuelve los valores 0.0406256 y 0.0535734, que tampoco parece corresponder a cifrados monoalfabéticos.

Repetimos este proceso considerando tres bloques en lugar de dos, esto es, el de símbolos que ocupan posición congruente con 0, 1 y 2 módulo 3. Calculando de nuevo el IC obtenemos los valores 0.041238, 0.0449943 y 0.0400331, de nuevo lejos de lo que cabría esperar en un cifrado monoalfabético.

Repitiendo el proceso considerando más bloques, alfabetos de un cifrado monoalfabético, obtenemos:

k	IC
4	0,0498, 0,0573, 0,0567, 0,0652
5	0,0423, 0,0411, 0,0501, 0,0377, 0,0420
6	0,0350, 0,0509, 0,0396, 0,0579, 0,0396, 0,0479
7	0,0403, 0,0407, 0,0438, 0,0416, 0,0420, 0,0492, 0,0402
8	0,0613, 0,0818, 0,0841, 0,0841, 0,0999, 0,0642, 0,0707, 0,0786

Puede verse que, considerando 8 bloques, la mayoría de los IC calculados se ajustan a lo esperado en un cifrado monoalfabético, teniendo en cuenta que 8 es consistente con el filtrado Kasiski del tamaño de la clave, todo apunta a que este puede ser el tamaño de la clave. Se deja como ejercicio el análisis de los bloques y el descifrado del mensaje.