

CRIPTOGRAFÍA PARA INGENIER@S

Class4crypt

© Jorgeramió 2022

Aula virtual de
criptografía
aplicada

Diapositivas
utilizadas en las
clases grabadas
de Class4crypt

Módulo 4 Teoría de la información en la criptografía

Dr. Jorge Ramió Aguirre © 2022



Attribution-NonCommercial-
NoDerivatives 4.0 International
(CC BY-NC-ND 4.0)



Dr. Jorge Ramío Aguirre

*El ingenio es intrínseco al ser humano,
solo hay que darle una oportunidad
para que se manifieste.*

<https://www.criptored.es/cvJorge/index.html>

Class4crypt

Tu aula virtual de criptografía aplicada

Módulo 1. Principios básicos de la seguridad

Módulo 2. Matemáticas discretas en la criptografía

Módulo 3. Complejidad algorítmica en la criptografía

➡ **Módulo 4. Teoría de la información en la criptografía**

Módulo 5. Fundamentos de la criptografía

Módulo 6. Algoritmos de criptografía clásica

Módulo 7. Funciones hash

Módulo 8. Criptografía simétrica en bloque

Módulo 9. Criptografía simétrica en flujo

Módulo 10. Criptografía asimétrica

Class4crypt

Módulo 4. Teoría de la información en la criptografía

- 4.1. Cantidad de información e incertidumbre
- 4.2. Entropía de la información y codificador óptimo
- 4.3. Ratio y redundancia del lenguaje
- 4.4. Secreto perfecto y distancia de unicidad
- 4.5. Métodos de difusión y confusión en la criptografía

Lista de reproducción del módulo 4 en el canal Class4crypt

https://www.youtube.com/playlist?list=PLq6etZPDh0kvx_KxYPDn14wzSzx2qZM0p

Class4crypt c4c4.1

Módulo 4. Teoría de la información en la criptografía

Lección 4.1. Cantidad de información e incertidumbre

4.1.1. Definiciones de información y teoría de la información

4.1.2. La figura de Claude Shannon

4.1.3. Cantidad de información asociada a un mensaje

4.1.3.a. Análisis en función de su extensión (visión subjetiva)

4.1.3.b. Análisis en función de su utilidad (visión subjetiva)

4.1.3.c. Análisis en función de su probabilidad (visión objetiva)

4.1.4. Definiciones de incertidumbre y de cantidad de información

Class4crypt c4c4.1 Cantidad de información e incertidumbre
<https://www.youtube.com/watch?v=ejhZLLD8G6k>

Definiciones relacionadas con información

- Información

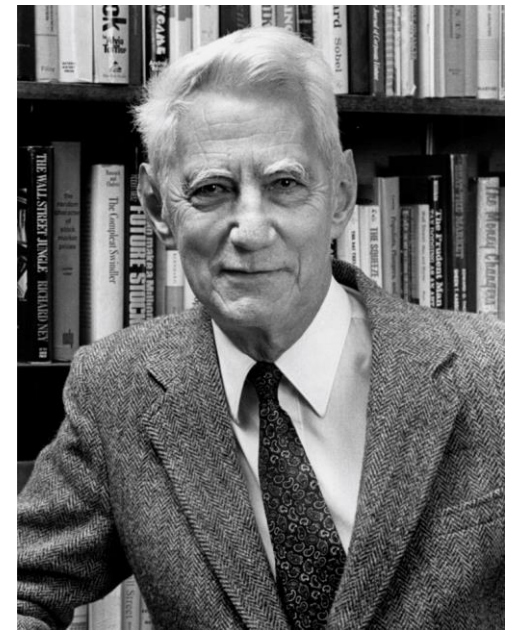
- Es el conjunto de datos o mensajes inteligibles, creados con un lenguaje de representación y que debemos proteger ante las amenazas del entorno, durante todo su ciclo de vida (creación, almacenamiento, transmisión, gestión y eliminación) usando, por ejemplo, técnicas criptográficas

- Teoría de la información

- La teoría de la información mide la **cantidad** de información que contiene un mensaje, a través del número **medio** de bits necesario para codificar todos los posibles mensajes mediante un codificador **óptimo**
- Cosas que deben aclararse en la definición anterior:
 - ¿Cantidad de información? ¿Número medio de bits? ¿Codificador óptimo?

Claude Elwood Shannon

- Matemático, ingeniero y criptólogo, reconocido como el padre de la teoría de la información
 - A mathematical theory of communication, Bell System Technical Journal 27 (1948)
 - Communication theory of secrecy systems, Bell System Technical Journal 28 (1949)
 - Introduce el concepto de unidad de información shannon, que actualmente se denomina bit, **binary digit**
 - Demuestra que las fuentes de información (el telégrafo, la radio, el teléfono, personas hablando, textos, etc.) pueden medirse, es decir, que la información es cuantificable, y que existe una velocidad máxima de transferencia o capacidad de canal, y sienta las bases para la corrección de errores, la supresión de ruidos y la redundancia del lenguaje
 - Varios de estos temas serán fundamentales en la criptografía



1916 - 2001

Representación de la información

- La información puede estar representada por
 - Números
 - 15/01/21 corresponde a la fecha 15 de enero del año 2021
 - 04:15:33 se interpreta como 4 horas, 15 minutos y 33 segundos
 - Símbolos
 - H_2O corresponde a la fórmula química del agua
 - Estos emoticonos nos indican estados de ánimo: 😊 ☹
 - Lenguaje de códigos o letras
 - “Te envío un fuerte abrazo” (personas)
 - El código fuente de un programa (máquinas)

Todo es información y
sería ideal cuantificarla...

Cantidad de información de un mensaje

- La información puede analizarse desde 3 perspectivas
 - a) En función de la **extensión** del mensaje recibido
 - b) En función de la **utilidad** del mensaje recibido
 - c) En función de la **probabilidad** de recibir un mensaje
- Este último enfoque orientado a la ingeniería y que fue usado por Claude Shannon en su estudio, es el que aquí nos interesa
 - Si un mensaje es poco probable, diremos que contiene una alta cantidad de información y, por el contrario, si un mensaje es muy probable, diremos que su cantidad de información es baja
 - Pero esto habrá que cuantificarlo para poder medirlo...

Cantidad de información según extensión

- Ante una pregunta cualquiera, una respuesta precisa y más extensa nos entregará, subjetivamente, una mayor información sobre el tema en particular. Diremos así que estamos ante una mayor “cantidad de información”
- **Pregunta: ¿Hace mucho calor allí? (en una playa de España en particular en agosto)**
 - Respuesta 1: Sí
 - Respuesta 2: Sí, hace mucho calor, y cuando no hay viento la temperatura puede superar con facilidad los 35 grados a la sombra
- ¿Dónde pensamos que hay una mayor cantidad de información?
 - En la respuesta 2 (subjetivamente)
 - Hay más cantidad de información en la respuesta 2 porque al ser más extensa y precisa, nos entrega más datos y nos define de mejor manera la situación

Cantidad de información según utilidad

- Ante una pregunta cualquiera, una respuesta más útil y clara nos dejará con la sensación subjetiva de haber recibido una mayor “cantidad de información”
- Pregunta: ¿Hace mucho calor allí? (en una playa de España en particular en agosto)
 - Respuesta 1: Sí, sobre los 30 grados normalmente
 - Respuesta 2: Depende. Si el viento es del sur, que siempre es más cálido que el que viene del norte, es normal que la temperatura suba algo
- ¿Dónde pensamos ahora que hay una mayor cantidad de información?
 - En la respuesta 1 (subjetivamente)
 - Ahora hay más cantidad de información en la respuesta 1, porque es más precisa y entrega una solución a nuestra pregunta. La segunda respuesta es muy amplia, pero no nos dice nada que sea de utilidad para aclarar nuestra duda

Cantidad de información según probabilidad (1/2)

- Una respuesta que sea poco probable por lo inesperada, nos dará una sensación objetiva de que contiene una mayor “cantidad de información”
- Pregunta: ¿Hace mucho calor allí? (en una playa de España en particular en agosto)
 - Respuesta 1: Normalmente sí
 - Respuesta 2: No, nada. Lo normal es que la temperatura no supere los 15 grados y por la noche suele bajar unos 10 grados. Incluso es posible que algún día nieve
- ¿Dónde pensamos en este caso que hay una mayor cantidad de información?
 - En la respuesta 2 (objetivamente)
 - La segunda respuesta es totalmente inesperada, y por lo tanto era muy poco probable escucharla. Lo sorprendente de la misma, nos deja la sensación de que hemos recibido mucha información, aunque tengamos la certeza de que la información recibida no sea la correcta, ni cierta, ni tampoco la deseada

Cantidad de información según probabilidad (2/2)

- La respuesta que sea menos probable porque hay muchos más estados posibles de la misma, nos dará la sensación de contener una mayor “cantidad de información”
- **Pregunta: ¿Dónde le daría alegría a su cuerpo Macarena según la canción?**
 - Respuesta 1: En una país de la península ibérica
 - Respuesta 2: En una capital de provincia de España
 - Respuesta 3: En el número 3 de la calle Sierpes de Sevilla
- ¿Dónde habría aquí una mayor cantidad de información?
 - En la respuesta 3 (objetivamente)
 - Porque al ser más extenso el número de calles (> 4.500) en la ciudad de Sevilla que el de las 50 capitales de provincia de España, y esto último mayor que los 4 países que forman la península ibérica, la respuesta 3 tiene asociada una mayor **incertidumbre** (muchos estados -aquí direcciones de calles- posibles)

Conjeturas subjetivas sobre incertidumbre

- Ante varios mensajes posibles, en principio todos equiprobables, aquel que tenga una menor probabilidad de aparición en su grupo será el que contenga una mayor cantidad de información
- Es decir, hay una mayor incertidumbre ante ese mensaje con baja probabilidad frente a otro que tenga una alta probabilidad
- A mayor incertidumbre, mayor cantidad de información
 - En el ejemplo anterior, podríamos suponer que todas las calles de Sevilla (estados respuesta posibles) eran equiprobables. Entonces, hay una gran incertidumbre sobre qué calle será ($1/4.500$) con respecto, por ejemplo, a las capitales de provincia ($1/50$) ... pero esto habrá que demostrarlo

Estados de una variable

- Sea X una variable aleatoria con n estados posibles
- Con $X = x_i$ una ocurrencia i -ésima

Es decir $X = \{x_1, x_2, x_3, \dots, x_{n-1}, x_n\}$

Con probabilidades $p_1 = p(x_1), p_2 = p(x_2), \dots, p_n = p(x_n)$

Como $0 \leq p_i \leq 1$ para $i = 1, 2, \dots, n$

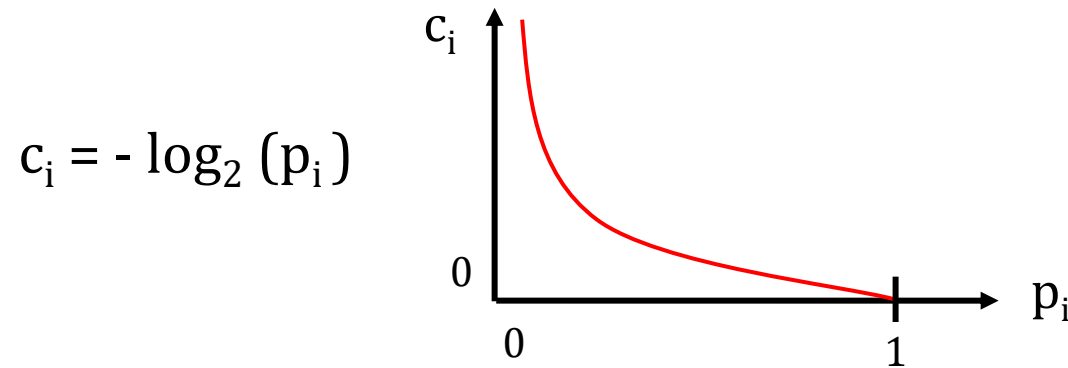
Entonces

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

- Como asignamos probabilidades de ocurrencia menores que 1 a estos n estados, lógicamente la suma de todas será igual a la unidad

Definición de cantidad de información

- Definiremos c_i a la cantidad de información del estado i , y se representará como el logaritmo en base dos de la probabilidad de que ocurra el estado i -ésimo



- Logaritmo $p(x_i) = 1 \Rightarrow$ no hay incertidumbre: $c_i = 0$
 $p(x_i) = 0 \Rightarrow$ máxima incertidumbre: $c_i \rightarrow \infty$
- Signo $p(x_i) < 1 \Rightarrow \log p(x_i)$ será negativo
- Base 2 Fenómeno binario \Rightarrow dos estados (un bit)

Grados de indeterminación (incertidumbre)

- En una bolsa hay dos papeles con círculos, en otra bolsa hay dos papeles con cuadrados y en una tercera bolsa hay dos papeles con triángulos; en todas hay uno negro y uno blanco
- Sacamos a ciegas un papel de cada bolsa y tenemos la combinación que indica la flecha

Combinación 1 ○ □ △

Combinación 2 ○ □ ▲

Combinación 3 ○ ■ △

Combinación 4 ○ ■ ▲



Combinación 5 ● □ △

Combinación 6 ● □ ▲

Combinación 7 ● ■ △

Combinación 8 ● ■ ▲

- Si hay equiprobabilidad, entonces $p(x_i) = 1/8$
- ¿Qué cantidad de información tiene cada uno de los estados?

$$c_i = \log \frac{\text{Grado de indeterminación previo}}{\text{Grado de indeterminación posterior}}$$

Solución a la incertidumbre con pistas

Combinación 1	○	□	△
Combinación 2	○	□	▲
Combinación 3	○	■	△
Combinación 4	○	■	▲
Combinación 5	●	□	△
Combinación 6	●	□	▲
Combinación 7	●	■	△
Combinación 8	●	■	▲



- Como $p(x_i) = 1/8$ entonces la incertidumbre inicial $I_i = 8$
- Si se van dando pistas, como resultado la incertidumbre irá bajando
 1. Las figuras no son del mismo color: I_i baja de 8 a 6 (descartamos combinaciones 1 y 8)
 2. El círculo es blanco: I_i baja de 6 a 3 (descartamos combinaciones 5, 6 y 7)
 3. Hay dos figuras blancas: I_i baja de 3 a 2 (descartamos combinación 4)
 4. El cuadrado es negro: I_i baja de 2 a 1 (descartamos combinación 2)
- Se acaba la incertidumbre pues la solución es la combinación 3

Solución matemática a la incertidumbre

- Usaremos la expresión logarítmica de c_i de momento sin definir la base, se hará después
- Pista 1. Las figuras no son del mismo color. La incertidumbre I_i baja de 8 a 6
$$c_{i1} = \log (8/6) = \log 8 - \log 6$$
- Pista 2. El círculo es blanco. La incertidumbre I_i baja de 6 a 3
$$c_{i2} = \log (6/3) = \log 6 - \log 3$$
- Pista 3. Hay dos figuras blancas. La incertidumbre I_i baja de 3 a 2
$$c_{i3} = \log (3/2) = \log 3 - \log 2$$
- Pista 4. El cuadrado es negro. La incertidumbre I_i baja de 2 a 1 y desaparece
$$c_{i4} = \log (2/1) = \log 2 - \log 1$$
- Todas las magnitudes se pueden sumar: $c_i = c_{i1} + c_{i2} + c_{i3} + c_{i4} = \log 8 - \log 1 = \log 8$
- ¿Qué base habrá que usar para el logaritmo?

Base del logaritmo en la incertidumbre

- Sean I_i la incertidumbre inicial e I_f la incertidumbre final
- $c_i = \log(I_i/I_f) = \log I_i - \log I_f$
- La cantidad de información tiene como unidad de medida la de un fenómeno de sólo dos estados, un fenómeno binario
- Luego: $c_i = \log_b(2/1) = \log_b 2 - \log_b 1 = \log_b 2$
- Si la incertidumbre desaparece cuando $\log_b 2 = 1$, entonces la base b debe ser 2
- Precisamente a esta unidad se le llama bit, binary digit (shannon)
- En el ejemplo anterior $c_i = \log_2 8 = 3$ (sólo 3 preguntas cambian la incertidumbre por certeza)
 - Pregunta 1: ¿Está entre la opción 1 y la 4? \Rightarrow Respuesta Sí
 - Pregunta 2: ¿Está entre la opción 1 y la 2? \Rightarrow Respuesta No
 - Pregunta 3: ¿Es la opción 4? \Rightarrow Respuesta No
 - La solución es la opción 3 $c_i = -\log_2(p_i) = -\log_2(1/8) = -0 + \log_2 8$ $c_i = \log_2 8 = 3$

Cuánta falta nos hace hoy Claude Shannon

- “La información es poder”



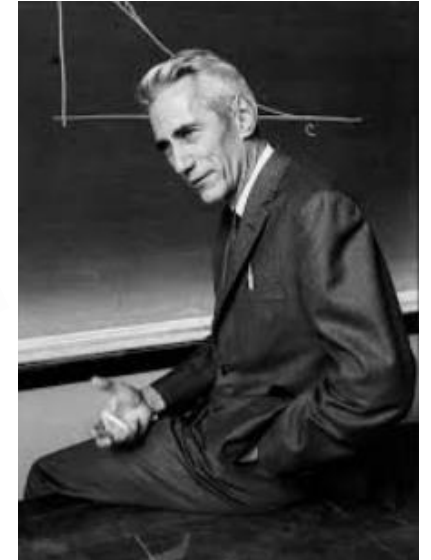
Francis Bacon
1561 - 1626



Thomas Hobbes
1588 - 1679



Leviatán



Claude Shannon
1916 - 2001

- ¿Has pensado cómo manejan hoy nuestra mente con noticias falsas para fines políticos y económicos?
- ¡Qué bien nos vendría ahora un estudio como el de Shannon para cuantificar y medir la **desinformación**!



Conclusiones de la Lección 4.1

- Claude Shannon, uno de los investigadores más importantes del siglo XX, hace un inmenso aporte a la criptografía con su trabajo en teoría de la información y los sistemas secretos. Mi opinión personal: convierte a la criptografía en una ciencia
- Definición básica: la teoría de la información mide la cantidad de información que contiene un mensaje, a través del número medio de bits necesario para codificar todos los posibles mensajes mediante un codificador óptimo
- La información puede analizarse desde tres perspectivas: la extensión del mensaje (vista subjetiva), su utilidad (vista subjetiva) y su probabilidad (vista objetiva)
- Esta última es la que nos interesa para poder cuantificar y medir la información
- La cantidad de información es la relación entre el grado de incertidumbre inicial I_i e incertidumbre final I_f que se tiene sobre un mensaje: $c_i = \log_2 (I_i/I_f) = -\log_2 (p_i)$
- A mayor incertidumbre de un mensaje, mayor es la cantidad de información

Lectura recomendada

- A Goliath Amongst Giants, Claude E. Shannon, Bell Labs
 - <https://www.bell-labs.com/claude-shannon>
- A Mathematical Theory of Communication, C. E. Shannon, The Bell System Technical Journal, Vol. 27, 1948
 - <http://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>
- Communication Theory of Secrecy Systems, C. E. Shannon, The Bell System Technical Journal, Vol. 28, 1949
 - <https://www.cs.virginia.edu/~evans/greatworks/shannon1949.pdf>
- shannon (unidad), Wikipedia
 - [https://es.wikipedia.org/wiki/Shannon_\(unidad\)](https://es.wikipedia.org/wiki/Shannon_(unidad))
- Criptografía y Seguridad en Computadores, Capítulo 3: Teoría de la información, Manuel Lucena, versión 5-0.1.4, noviembre 2019
 - <http://criptografiayseguridad.blogspot.com/p/criptografia-y-seguridad-en.html>

Class4crypt c4c4.2

Módulo 4. Teoría de la información en la criptografía

Lección 4.2. Entropía de la información y codificador óptimo

4.2.1. Definición de entropía

4.2.2. Propiedades de la entropía

4.2.3. Codificador óptimo

4.2.4. Codificación mediante el método de Huffman

4.2.5. Ejemplos de entropía de mensajes

Class4crypt c4c4.2 Entropía de la información y codificador óptimo

<https://www.youtube.com/watch?v=ZvJ0bFAhT9w>

De la clase anterior...

La teoría de la información mide la cantidad de información que contiene un mensaje, a través del **número medio de bits** necesario para codificar todos los posibles mensajes mediante un **codificador óptimo**

Definición de entropía

- La entropía de un mensaje X , que se representa por $H(X)$, es el **valor medio ponderado** de la **cantidad de información** de los diversos estados del mensaje
- Es una medida de la incertidumbre media acerca de una variable aleatoria y el número de bits de información, es decir, la cantidad de información media que ésta y sus símbolos contiene

$$H(X) = - \sum_{i=1}^k p(x_i) \log_2 p(x_i)$$

- ¿Cómo se llega a esta ecuación?



Recordando la cantidad de información

- Sabemos que un fenómeno de k estados equiprobables tendrá un grado de indeterminación igual a k
- Y que la probabilidad de que se dé uno de esos estados será
 - $p = 1/k$
- Vimos en la lección anterior que la cantidad de información c_i era
 - $c_i = \log_2 (k/1)$
 - Que puede representarse de esta otra manera
 - $c_i = \log_2 [1/(1/k)] = \log_2 1 - \log_2 (1/k)$
 - Así, podemos concluir que: $c_i = -\log_2 p$ (estados equiprobables)

Deduciendo la ecuación de la entropía

- Si ahora cada uno de estos estados tiene una probabilidad distinta p_i , diremos que la entropía $H(X)$ será igual a la suma ponderada (proporcional a esa probabilidad p_i) de la cantidad de información
 - $H(X) = - p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_{k-1} \log_2 p_{k-1} - p_k \log_2 p_k$
- Que puede expresarse como se había definido previamente

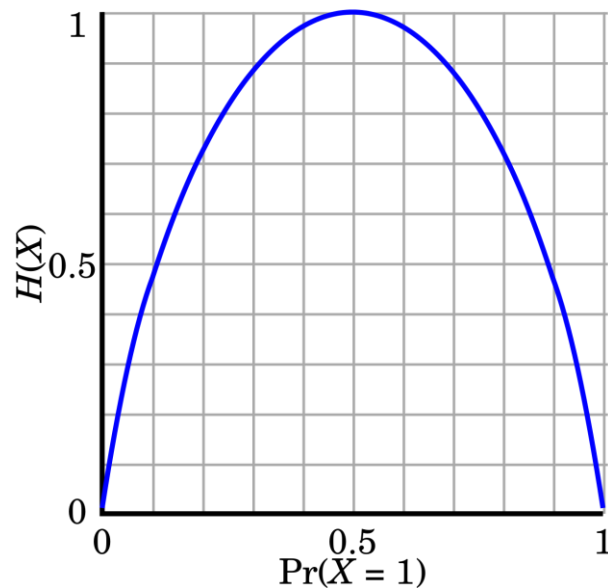
$$H(X) = - \sum_{i=1}^k p(x_i) \log_2 p(x_i)$$

- Nota: aunque la ecuación parece bastante lógica, su obtención no resulta tan inmediata

Propiedades de la entropía

- La entropía será siempre positiva y se anula si y sólo si la probabilidad de un estado es igual a 1 y la del resto de estados es igual 0. La demostración es obvia
- La entropía será máxima, es decir habrá una mayor incertidumbre sobre el mensaje, cuando exista una equiprobabilidad en todos los estados de la variable X. Por lo tanto, para una variable de n estados se cumplirá que $H(X)_{\text{máx}} = \log_2 n$
- La demostración empírica (que haremos a continuación) es muy fácil. No obstante, la demostración matemática de esta ecuación no es sencilla
- Si hay n estados equiprobables, entonces $p_i = 1/n$
- Entonces $H(X) = - \sum p_i \log_2 p_i = - n(1/n) \log_2 (1/n) = - (\log_2 1 - \log_2 n) = 0 + \log_2 n$
- Por lo que la entropía máxima para una variable X de n estados será $H(X)_{\text{máx}} = \log_2 n$

Curva típica de la entropía con 2 estados



Ref.: Brana, Wikimedia Commons

$$\begin{aligned} H(X) &= - \sum p_i \log_2 p_i = - 1/2 \log_2 (1/2) - 1/2 \log_2 (1/2) \\ H(X) &= - 0,5 (\log_2 1 - \log_2 2) - 0,5 (\log_2 1 - \log_2 2) \\ H(X) &= - 0,5 (0 - 1) - 0,5 (0 - 1) = 0,5 + 0,5 = 1 \end{aligned}$$

- Ensayo de Bernoulli. Experimento binario en donde la variable aleatoria X puede tomar solo los valores 0 o 1
- Se supone éxito a la probabilidad de obtener en el experimento el resultado 1
- Si $p(x=1) = 0,5$ entonces la entropía será máxima y su valor igual a 1
- Existe equiprobabilidad entre el 0 y el 1 y, por tanto, tenemos la mayor incertidumbre
- Si $p(x=1) < 0,5$ la entropía disminuye
- Si $p(x=1) > 0,5$ la entropía disminuye

Codificación óptima y mínimo de bits

- Un codificador óptimo es aquel que para codificar un mensaje X usa el menor número posible de bits
- Introduciendo el signo negativo dentro de la ecuación de la entropía

$$H(X) = \sum_{i=1}^k p(x_i) \log_2 [1/p(x_i)]$$

- En esa ecuación, la expresión $\log_2 [1/p(x_i)]$ representará el número necesario de bits para codificar el mensaje X en un codificador óptimo
- Por ejemplo, el código ASCII no es un código óptimo porque usa 8 bits para representar a sus 256 caracteres (letras, números, signos, símbolos, etc.), sin importarle si su uso es muy frecuente o poco frecuente en un archivo

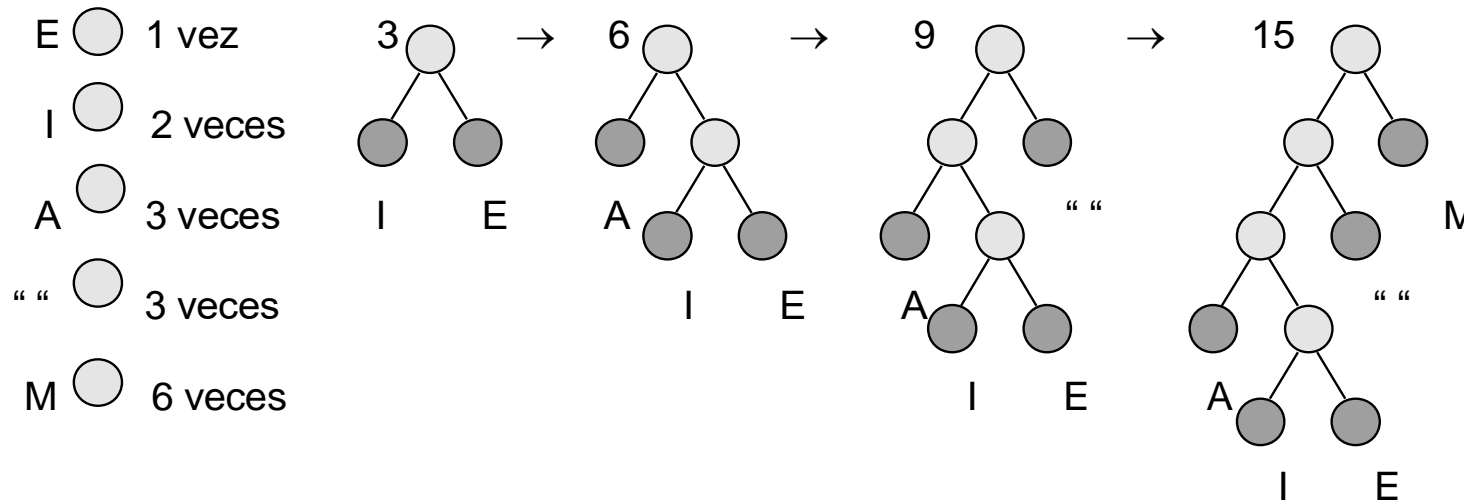
Codificación por el método de Huffman



1925-1999

- Para el texto **MI MAMA ME MIMA** de 15 caracteres, se va creando un árbol con las frecuencias observadas como el que se muestra abajo. Un desplazamiento hacia la derecha es un 1 y un desplazamiento hacia la izquierda es un 0

Letra Frecuencia Ocurrencias →



Se codificará con 33 bits:

1 0010 01 1 000 1 000 01 1
0011 01 1 0010 1 000

Una media de $33/15 = 2,2$
bits por carácter

Esto era la media de bits en
la definición de entropía →

Para codificar en ASCII haría
falta $15 \cdot 8 = 120$ bits

- Luego codificamos así: **M** = 1, " " = 01, A = 000, I = 0010, E = 0011

Entropía de MI MAMA ME MIMA (1/2)

- Probabilidades de cada uno de los 5 caracteres del texto X
 - M aparece 6 veces, luego $p(M) = 6/15$
 - El espacio en blanco aparece 3 veces, luego $P(" ") = 3/15$
 - A aparece 3 veces, luego $P(A) = 3/15$
 - I aparece 2 veces, luego $P(I) = 2/15$
 - E aparece 1 vez, luego $P(E) = 1/15$
- Usando la ecuación de la entropía
 - $H(X) = \sum p(x) \log_2 [1/p(x)]$
 - $p(M) \log_2 [1/p(M)] + p(" ") \log_2 [1/p(" ")] + p(A) \log_2 [1/p(A)] + p(I) \log_2 [1/p(I)] + p(E) \log_2 [1/p(E)]$

Entropía de MI MAMA ME MIMA (2/2)

- Reemplazando valores:
- $H(X) = 6/15 \cdot \log_2(15/6) + 3/15 \cdot \log_2(15/3) + 3/15 \cdot \log_2(15/3) + 2/15 \cdot \log_2(15/2) + 1/15 \cdot \log_2(15/1)$
- Y resolviendo:
 - $= 0,4 \cdot \log_2(2,5) + 0,2 \cdot \log_2(5) + 0,2 \cdot \log_2(5) + 0,133 \cdot \log_2(7,5) + 0,066 \cdot \log_2(15)$
 - $= 0,4 \cdot 1.322 + 0,2 \cdot 2.322 + 0,2 \cdot 2.322 + 0,133 \cdot 2.907 + 0,066 \cdot 3.907$
 - $= 0,5288 + 0,4644 + 0,4644 + 0,3867 + 0,2579 = 2,1022$
- $H(X) = 2,1022$
- Un valor que se acerca mucho al número medio de bits encontrado en la transparencia anterior haciendo la media $33/15 = 2,2$

Entropía y la seguridad de las claves

- En un sistema criptográfico, la seguridad o fortaleza del mismo debe residir solamente en la clave (principio de Kerckhoffs que veremos en otra clase)
- Si en un algoritmo de cifra, por ejemplo AES, podemos introducir por teclado una clave en formato ASCII o en formato hexadecimal, ¿qué es preferible?
- Es recomendable que sea en hexadecimal, porque su entropía (es decir el desorden, equiprobabilidad) será mayor y, por tanto, más difícil de adivinar
 - Una clave de 16 bytes en AES (128 bits) introducida por teclado en ASCII sólo permite caracteres imprimibles, algo más de 200 de los 256 caracteres del ASCII extendido. Por ejemplo, no podemos usar bytes desde posición 00000000 hasta 00011111 (códigos de control) y difícilmente algunos códigos de la zona alta
 - Además, tenderemos a poner cadenas de caracteres que sean fáciles de recordar
 - Y el hexadecimal permite bytes desde 00000000 = 0x 00 hasta 11111111 = 0x FF

Conclusiones de la Lección 4.2

- La entropía $H(X)$ de un mensaje X es el valor medio de bits con el cual dicho mensaje se codifica utilizando una codificación óptima
- También puede decirse que la entropía es la incertidumbre media acerca de una variable aleatoria X , en este caso la información
- La entropía será positiva y sólo se anula si un estado tiene probabilidad 1 y el resto 0
- La entropía será máxima cuando los n estados de la variable información X sean equiprobables $p_i = 1/n$, y vendrá dada por $H(X)_{\text{máx}} = \log_2 n$
- La codificación según método de Huffman, en el que los caracteres más frecuentes del mensaje se codifican con una corta longitud de bits y los menos frecuentes con longitudes mayores, nos entrega la entropía del mensaje
- Para ello, realizamos la división del número de bits utilizados en dicha codificación entre el número de caracteres del mensaje

Lectura recomendada (1/2)

- Entropía (información), Wikipedia
 - [https://es.wikipedia.org/wiki/Entrop%C3%ADa_\(informaci%C3%B3n\)](https://es.wikipedia.org/wiki/Entrop%C3%ADa_(informaci%C3%B3n))
- Shannon entropy calculator
 - <http://www.shannonentropy.netmark.pl/calculate>
- Codificación Huffman
 - <https://riptutorial.com/es/algorithm/example/23995/codificacion-huffman>
- Huffman Tree Generator
 - <http://huffman.ooz.ie/>
- Huffman Coding
 - <https://people.ok.ubc.ca/ylucet/DS/Huffman.html>

Lectura recomendada (2/2)

- Códigos y tablas de uso frecuente en criptografía, CLCript 00, Jorge Ramió
 - https://www.criptored.es/download/Codigos_y_tablas_de_uso_frecuente_en_criptografia.pdf
- Criptografía y Seguridad en Computadores, Capítulo 3 Teoría de la información, Manuel Lucena, versión 5-0.1.4, noviembre 2019
 - <http://criptografiayseguridad.blogspot.com/p/criptografia-y-seguridad-en.html>

Class4crypt c4c4.3

Módulo 4. Teoría de la información en la criptografía

Lección 4.3. Ratio y redundancia del lenguaje

4.3.1. Recordando el concepto de entropía $H(X)$

4.3.2. La ratio absoluta del lenguaje R

4.3.3. La ratio real del lenguaje r

4.3.4. Redundancia del lenguaje D

4.3.5. Debilidades de la cifra clásica por la redundancia del lenguaje

4.3.6. Relación entre la redundancia del lenguaje y la compresión alcanzada en los programas zip

Class4crypt c4c4.3 Ratio y redundancia del lenguaje
<https://www.youtube.com/watch?v=ho7aNkaGQNo>

Recordando la definición de entropía

- La entropía de un mensaje X es el valor medio ponderado de la cantidad de información de los estados de un mensaje
- Medida de la incertidumbre media acerca de una variable aleatoria y de la cantidad de información que ésta y sus símbolos contiene

$$H(X) = \sum_{i=1}^{i=k} p(x_i) \log_2 [1/p(x_i)]$$

- Se anula si sólo si un estado tiene $p(x_i) = 1$ y los demás $p(x_i) = 0$
- Es máxima si $p(x_i)$ la misma para todos sus estados (equiprobable)
- Entonces $H(X)_{\text{máx}} = \log_2 p$
- Para una clave aleatoria de 128 bits
 $H(X)_{\text{máx}} = \log_2 [1/(1/128)]$
 $H(X)_{\text{máx}} = \log_2 128 = 7 \text{ bits}$

Entropía del lenguaje mod 27 y ASCII

- Para las letras mayúsculas del alfabeto `ABCDEFGHIJKLMNOPQRSTUVWXYZ`
 - $H(X)_{\text{máx}} = \log_2 27 = 4,75$ bits
 - Para un texto con varias centenas de letras, por ejemplo el primer párrafo de El Quijote con 958 letras:
 - $H(X)_{\text{típica}} = 4,04$ bits (y se mantiene en 4,04 aunque el texto sea más largo)
- Para los 107 caracteres más el espacio, que típicamente vemos en un teclado `°1234567890'¡ª!".$%&/()=?¿\|@#~¬qwertyuiop`+QWERTYUIOP^*€[]asdfghjklñ´çASDFGHJKLÑ¨Ç{}<zxvcvbnm,.->ZXCVBNM;:_`
 - $H(X)_{\text{máx}} = \log_2 108 = 6.75$ bits, siendo $H(X)_{\text{máxASCII}} = \log_2 256 = 8$ bits
 - Para ese primer párrafo de El Quijote en ASCII, ahora con 1.204 caracteres
 - $H(X)_{\text{típica}} = 4,22$ bits (5 primeros párrafos y 6.815 caracteres: sube a 4,27)

Ratio absoluta del lenguaje R

- Si codificamos un mensaje cualquiera, letra a letra y suponiendo además una equiprobabilidad entre esas letras, se obtiene la ratio absoluta del lenguaje R, que es igual a la entropía $H(X)$
 - $R = H(X)$ bits/letra
- Si las 27 letras mayúsculas del alfabeto español o las 26 letras del alfabeto inglés fuesen equiprobables, la ratio absoluta sería
 - $R_{\text{mod } 27} = \log_2 n = \log_2 27 = 4,75489$ bits/letra (es común indicar 4,75)
 - $R_{\text{mod } 26} = \log_2 n = \log_2 26 = 4,70044$ bits/letra (entropía algo más baja)
- Y para el código ASCII extendido de 256 caracteres equiprobables
 - $R_{\text{mod } 256} = \log_2 n = \log_2 256 = 8$ bits/letra (entropía más alta, pero no tanto)

Ratio real del lenguaje r

- Sabemos que esa equiprobabilidad entre letras no se cumple nunca, por lo tanto la ratio real en un mensaje de texto será bastante menor
- La ratio real es el número de **bits de información** en cada letra para mensajes con una longitud igual a N letras. Es decir, según la definición de entropía
 - $r = H(X)/N$ bits/letra
- Como las letras que aparecen en un texto no tienen igual probabilidad, su frecuencia de aparición es distinta, los lenguajes están muy estructurados, hay bloques muy frecuentes de dos letras, de tres letras, de cuatro letras, e incluso de hasta cinco y seis letras, se puede asignar código a grupos de letras, la ratio real de un texto será mucho más baja que la ratio absoluta R
 - $1,2 < r < 1,5$ bits/letra (valor estimado)

Características de la ratio real r

- Si la ratio real r está entre 1,2 y 1,5 bits por letra, ¿quiere decir que puedo codificar todos los mensajes con letras mayúsculas usando sólo 2 bits?
 - Obviamente, no. Con dos bits sólo podremos codificar hasta cuatro letras o caracteres, representados de forma binaria por 00, 01, 10 y 11
 - Esos 1,2 o 1,5 bits significan los **bits de información** que contiene cada letra, en un texto donde se manifiestan las características del lenguaje, como lo es por ejemplo la redundancia que veremos en esta clase
- Un ejemplo de la redundancia del lenguaje lo tenemos si en un texto más o menos conocido quitamos las vocales. Sigue siendo posible entenderlo
 - **NNLGRDLMNCHDCYNMBRNQRCRDRM** Este texto será fácil...
 - **CNDSBTSSLPDRMSCDFCRHSTCTRLTRS** ¿Y este otro texto?

Significado y alcances de la ratio real r

- No es necesario codificar la información letra a letra
- Se podría codificar en bloques de 2 letras, en bloques de 3 letras, etc., según el estudio realizado por Claude Shannon
- Así se llega a este valor de ratio real estimada $1,2 < r < 1,5$
- ¿Qué significa esto en la práctica?
- Para un alfabeto de L elementos, por ejemplo 27 letras (ABCD ... WXYZ), con una entropía máxima igual a $H(X)_{\text{máx}} = \log_2 27 = 4,75$ bits/letra
 - Existirán $2^{R \cdot N}$ mensajes posibles de longitud N (con y sin sentido)
 - Para $N = 5$: desde AAAAA, AAAAB, AAAAC, ..., hasta ZZZZX, ZZZZY, ZZZZZ
 - Pero sólo habrá $2^{r \cdot N}$ mensajes que tengan sentido
 - Para $N = 5$: ABACO, CARTA, NADAR, PERRO, SIMIL, RATIO, SABOR, ZUECO, ...

Ejemplo demostrativo de la ratio r (1/3)

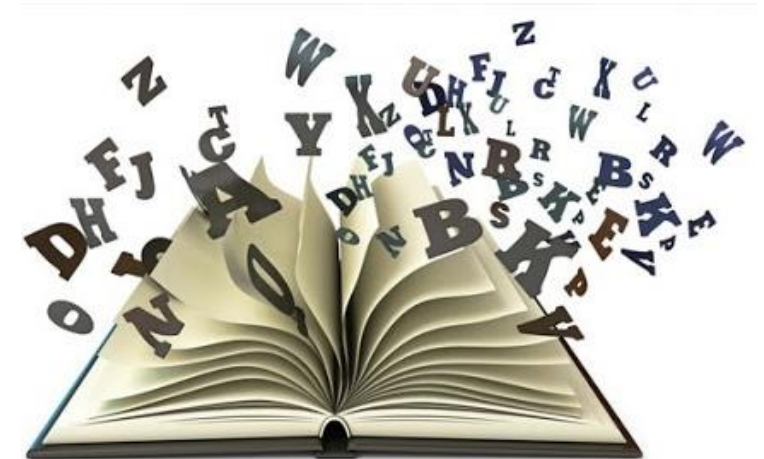
- Vamos a suponer un alfabeto muy simple para que el resultado nos entregue menos de 100 palabras y estas puedan mostrarse en una transparencia
 - Sea nuestro alfabeto un subconjunto del castellano que consta solamente de 5 letras: A, E, O, S y N
 - Supondremos que este alfabeto es “más o menos” representativo del lenguaje dado que las frecuencias características de esas letras suman el 50%
 - A = 12,5 %
 - E = 13,7 %
 - O = 8,7 %
 - S = 8,0 %
 - N = 6,7 %
- Las frecuencias típicas del lenguaje se verán en una próxima diapositiva



Vamos a buscar palabras de 4 letras con este alfabeto y que tengan sentido

Ejemplo demostrativo de la ratio r (2/3)

- Pregunta: ¿Cuántos mensaje de longitud 4 existen y cuántos con sentido?
- Como $R = \log_2 5 = 2,3219$
 - Existirán $2^{R*4} = 2^{2,3219*4} = 625$ palabras de longitud 4 con y sin sentido
 - Obviamente $625 = 5^4$ mensajes (alfabeto con 5 letras y la longitud es 4)
- Como la ratio real r estará entre 1,2 y 1,5 entonces cabe esperar x mensajes con sentido de longitud 4 dentro de estos valores
 - $2^{1,2*4} < x < 2^{1,5*4}$ es decir, $27 < x < 64$
 - Valor medio de $x = (27 + 64)/2 = 45$
- Con la ayuda de un diccionario, buscamos palabras de longitud 4 que contengan esas 5 letras AEOSN en cualquier orden y forma

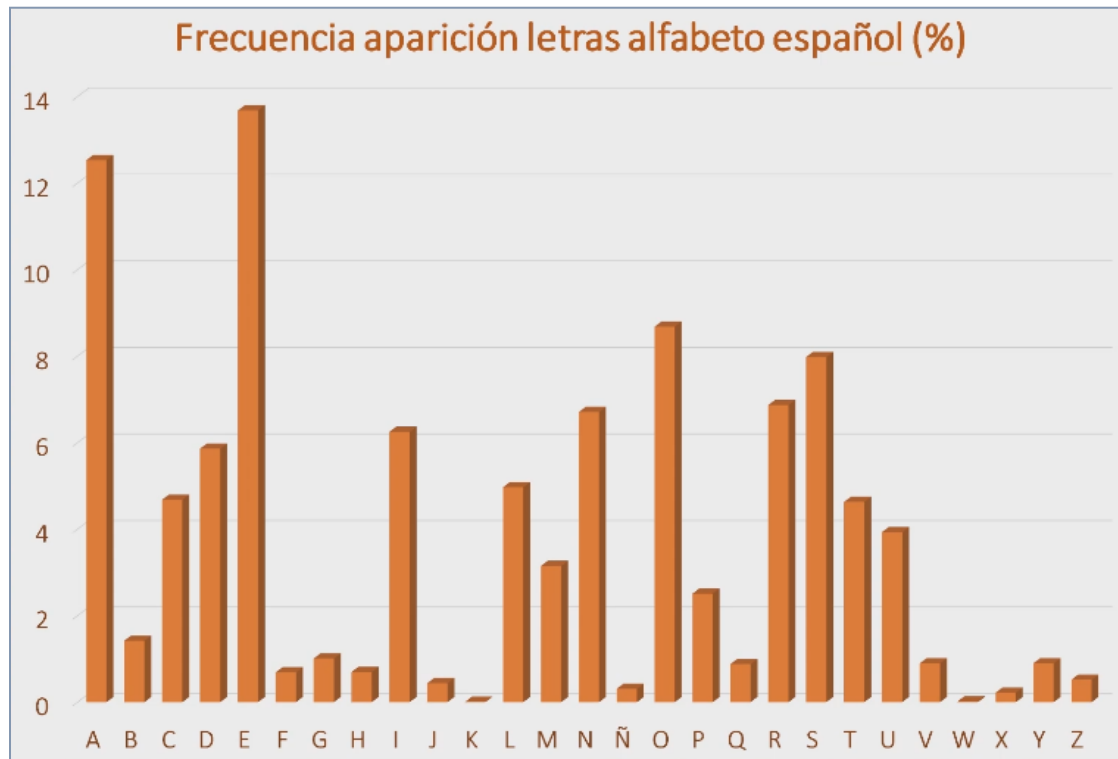


Ejemplo demostrativo de la ratio r (3/3)

- Mensajes de longitud 4 con sentido para un alfabeto de 5 letras (A, E, O, S, N), que representan el 50% de las frecuencias características de las 27 letras del alfabeto en mayúsculas del castellano
 - ANAS, ANEA, ANON, ANOS, ANSA, ASAN, ASAS, ASEA, ASEE, ASEN, ASEO, ASES, ASNA, ASNO
 - EASO, ENEA, ENEO, ENES, ESAS, ESES, ESOS
 - NANA, NANO, NAOS, NASA, NASO, NENA, NENE, NEON, NEOS, NONA, NONO
 - ONAS, OSAN, OSAS, OSEA, OSEN, OSES, OSOS
 - SANA, SANE, SANO, SEAN, SEAS, SENA, SENO, SESO, SOSA, SOSO
- “Casualmente” 😊 han aparecido estas 49 palabras, un valor muy cercano a esa media igual 45 de la diapositiva anterior

La redundancia del lenguaje

- El estudio de Shannon demuestra que es la estructura del lenguaje (siendo diferente para cada lenguaje) la que produce esta redundancia



- Existen diferencias en la frecuencia de aparición de cada una de las letras dentro de un texto
- Si el texto tiene varios cientos o, mejor aún, miles de letras, hay una distribución característica de sus frecuencias como se muestra en la figura, sobresaliendo estas 9 letras: E, A, O, S, R, N, I, D, C (frec. > 5,0 %)

Frecuencia de las letras en el lenguaje

- Frecuencias del lenguaje (español) obtenidas del software Criptoclásicos v2.1, Herramientas – Estadísticas del Lenguaje – Tabla frecuencias monogramas

A 12,53 %	B 1,42 %	C 5,68 %
D 5,86 %	E 13,68 %	F 0,69 %
G 1,01 %	H 0,70 %	I 6,25 %
J 0,44 %	K 0,01 %	L 4,97 %
M 3,15 %	N 6,71 %	Ñ 0,07 %
O 8,68 %	P 2,51 %	Q 0,88 %
R 6,87 %	S 7,98 %	T 4,63 %
U 3,93 %	V 0,90 %	W 0,02 %
X 0,22 %	Y 0,90 %	Z 0,52 %

- Las 9 letras (E, A, O, S, R, N, I, D, C) con frecuencias mayores que un 5,0% suman un 74,24 %
- Es decir, sólo 9 de las 27 letras del alfabeto (un tercio) aportan el 75% de las frecuencias de un texto amplio
- Las palabras ENRISCADO, CONSIDERA, INCREADOS y DISECARON usan estas 9 letras más frecuentes

Redundancia del lenguaje y cifra clásica

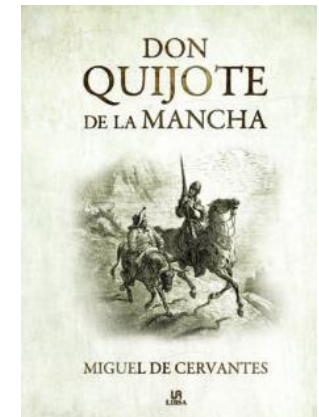
- No sólo hay redundancia por unas letras más frecuentes que otras. Existen bloques de dos letras (EN, ES, ...), de tres letras (ADO, IDA, ...), de cuatro letras (ANDO, LADO, ...) y hasta de cinco letras (IENDO, MENTE, ...) que son muy frecuentes en un texto
- Existe, además, una estructuración típica de frases y oraciones con sentido en nuestro lenguaje
- Todo esto dará pistas al criptoanalista para atacar un sistema. Y nuestra misión es crear algoritmos que sean seguros y que eviten estos ataques
- Esta información servirá para romper la gran mayoría de los sistemas de cifra clásica, porque la redundancia del lenguaje se sigue manifestando en el criptograma. Lógicamente, no sucederá lo mismo en la cifra moderna

Calculando la redundancia del lenguaje

- La redundancia del lenguaje mod 27, representada por D , es la diferencia entre la ratio absoluta $R = 4,75$ y la ratio real r , que se estima entre 1,2 y 1,5
 - $D = R - r$
 - Como $1,2 < r < 1,5$
 - $D_1 = R - r_1 = 4,75 - 1,2 = 3,55$ y $D_2 = R - r_2 = 4,75 - 1,5 = 3,25$
 - Por lo tanto $3,25 < D < 3,55$
- ¿Qué significa esto?
 - Que el número de bits extras (*bits redundantes*) al codificar un mensaje suponiendo un alfabeto de 27 letras, será en media $(3,55 + 3,25)/2 = 3,4$
 - Para 27 letras, deberíamos codificar con 5 bits ya que $2^5 = 32$ y $2^4 = 16$
 - $D_2/R = 68,42 \% < \text{Redundancia del Lenguaje} < 74,73 \% = D_1/R$

Archivos zip y redundancia en txt mod 27

- Se copian los capítulos 1 y 2 del libro El Quijote (18.156 letras) en un archivo txt, convirtiendo todo el texto a mayúsculas, es decir módulo 27
- Puede usar Criptoclásicos v2.1, algoritmo del César cifrando con $b = 0$
 - El archivo Quijote1.txt tiene 18.204 bytes
 - El archivo Quijote1.zip tiene 7.978 bytes
 - Como $7.978/18.204 = 0,4383$
 - El archivo txt se reduce al 43,83 %
 - La reducción ha sido de un 56,17 %
 - Se han quitado $18.204 - 7.978 = 10.226$ bytes redundantes
 - Redundancia = $10.226/18.204 = 56,17 \%$
 - Entropía (número medio de bits con un codificador óptimo)
 - $7.978 * 8 \text{ bits} = 63.824 \text{ bits} / 18.156 \text{ letras} = 3,5 \text{ bits/letra}$ (de los 8 bits del ASCII)



Se trata sólo de un ejemplo ilustrativo

Archivos zip y redundancia en txt ASCII

- Si copiamos los 5 primeros capítulos de El Quijote en un archivo de texto con ASCII extendido de 256 caracteres, aparecen 47.457 caracteres sin espacios y 57.887 caracteres con espacios (10.430 espacios, un 18%)
 - El archivo Quijote2ASCII.txt tiene 59.514 bytes
 - El archivo Quijote2ASCII.zip tiene 22.052 bytes
 - Se han quitado $59.514 - 22.052 = 37.462$ bytes redundantes
 - Redundancia = $37.462 / 59.514 = 62,9 \%$
- Aunque hay muchos caracteres nuevos en el texto, por ejemplo minúsculas, á, é, í, ó, ú, etc., la redundancia ha aumentado porque el espacio en blanco con esa frecuencia tan alta del 18% hace bajar mucho la entropía del documento
- Sin espacio en blanco, la redundancia es: $(49.119 - 20.726) / 49.119 = 57,8 \%$
- No aplica a archivos con formato: docx, jpg, pdf, mp4, etc. (+ o - comprimidos)

Nuevamente, se trata tan sólo de un ejemplo ilustrativo

Conclusiones de la Lección 4.3 (1/2)

- La ratio absoluta del lenguaje R supone que las letras sean equiprobables y se codifiquen una a una. Por lo tanto, su valor coincide con el de la entropía
- Para mod 27 se cumple que $R = 4,75$ bits
- Existirán 2^{R*N} mensajes con y sin sentido de longitud N
- La ratio real del lenguaje r es mucho más baja porque las letras no son equiprobables y además no es necesario codificar sólo letra a letra
- Para mod 27 se cumple que $1,2 < r < 1,5$ bits/letra (estimado)
- En mod 27, si se codifican las 27 letras con 5 bits, la ratio real nos indica que de esos 5 bits usados hay 3,6 bits redundantes y sólo 1,4 bits de información
- Existirán aproximadamente 2^{r*N} mensajes con sentido de longitud N

Conclusiones de la Lección 4.3 (2/2)

- Siempre se cumplirá que $2^{r*N} \ll 2^{R*N}$ (en realidad será muchísimo menor)
 - Texto 10 letras mod 27: $2^{1,35*10} = 11.585$ y $2^{4,75*10} = 199.032.864.766.430$
 - DAMEUNBESO, NILOSUEÑES, QUELASTIMA, ESLOQUEHAY, NOME GUSTAS, YTUTAMPOCO, HASTALUEGO, HASTANUNCA... ☺ y máximo 32.000 textos
- El lenguaje es muy redundante al haber letras que se repiten más que otras, e incluso lo mismo para grupos de dos o más letras
- La redundancia se define como $D = R - r$
- Redundancia para un alfabeto de 27 letras: $3,25 < D < 3,55$ bits/letra
- Redundancia para un alfabeto de 27 letras: $68,42 \% < D < 74,73 \%$
- La estructura y redundancia del lenguaje facilita el criptoanálisis de varios algoritmos de cifra clásica por un análisis de frecuencias en el criptograma

Lectura recomendada

- Shannon entropy
 - <https://planetcalc.com/2476/>
- RapidTables, Logarithm Calculator online
 - https://www.rapidtables.com/calc/math/Log_Calculator.html
- Ratio de entropía, Wikipedia
 - https://es.wikipedia.org/wiki/Ratio_de_entrop%C3%ADa
- Biblioteca Virtual Miguel de Cervantes
 - <http://www.cervantesvirtual.com/obra-visor/el-ingenioso-hidalgo-don-quijote-de-la-mancha--0/html/>
- Criptoclásicos v2.1
 - https://www.criptored.es/software/sw_m001c.htm
- Criptografía y Seguridad en Computadores, Capítulo 3 Teoría de la información, Manuel Lucena, versión 5-0.1.4, noviembre 2019
 - <http://criptografiayseguridad.blogspot.com/p/criptografia-y-seguridad-en.html>

Class4crypt c4c4.4

Módulo 4. Teoría de la información en la criptografía

Lección 4.4. Secreto perfecto y distancia de unicidad

4.4.1. Secreto de un sistema criptográfico

4.4.2. Cifrado con secreto perfecto

4.4.3. Cifrado sin secreto perfecto

4.4.4. Modelo de cifrador aleatorio

4.4.5. Distancia de unicidad

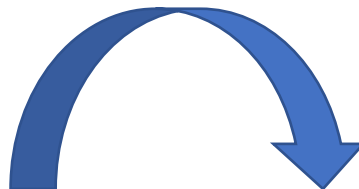
Class4crypt c4c4.4 Secreto perfecto y distancia de unicidad
https://www.youtube.com/watch?v=4JXy85IV_3k

Secreto de un sistema criptográfico

- Shannon define el secreto de un criptosistema como la incertidumbre del mensaje en claro M conocido el criptograma C , en cuya cifra se ha usado una clave K
- Tenemos estos espacios: $M = \{M_1, M_2, \dots M_n\}$, $C = \{C_1, C_2, \dots C_n\}$, $K = \{K_1, K_2, \dots K_n\}$
- Y en ellos se cumple que: $\sum p(M_i) = 1$, $\sum p(C_i) = 1$, $\sum p(K_i) = 1$
- Sea $p(M)$ la probabilidad de enviar un mensaje M . Si tenemos n mensajes M_i equiprobables, entonces $p(M_i) = 1/n$
- Sea $p(C)$ la probabilidad de recibir un criptograma C . Si cada uno de los n criptogramas C_i tienen igual probabilidad, entonces $p(C_i) = 1/n$
- Sea $p_M(C)$ la probabilidad de que a partir de un texto en claro M_i se obtenga un criptograma C_i
- Y sea $p_C(M)$ la probabilidad de que una vez recibido el criptograma C_i éste provenga de un texto en claro M_i

Cifrado con secreto perfecto (1/3)

Un sistema tiene secreto perfecto si el conocimiento del texto cifrado no nos proporciona ninguna información acerca del mensaje. Es decir, cuando la probabilidad de acierto al recibir el elemento $i+1$ es la misma que en el estado i


$$\text{Secreto perfecto} \Rightarrow p(M) = p_c(M)$$

La probabilidad p de enviar un mensaje M con texto en claro $p(M)$ o, en otras palabras, la probabilidad a priori, será igual a la probabilidad p de que una vez conocido un criptograma C , éste se corresponda a un mensaje M cifrado con la clave K . Esta última, ahora conocida como la probabilidad a posteriori, es $p_c(M)$

Cifrado con secreto perfecto (2/3)

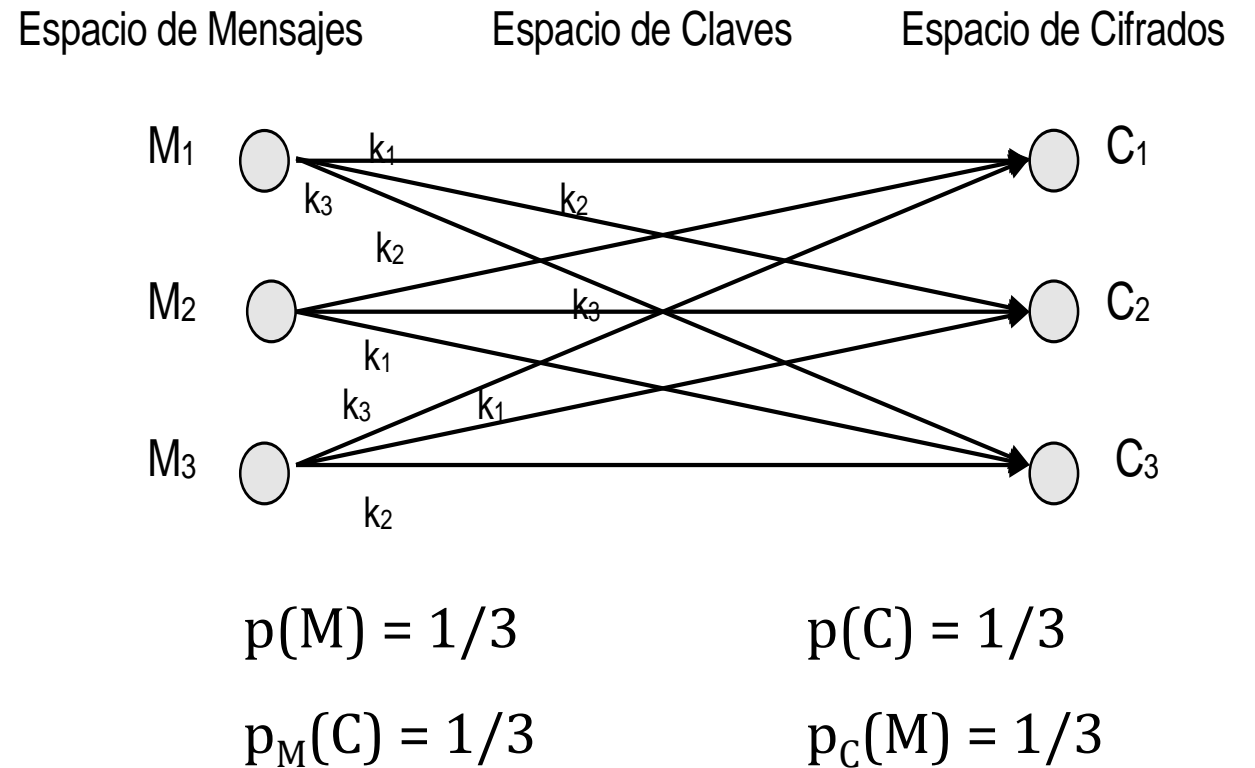
- La probabilidad p de cifrar un mensaje en claro M usando una clave K y que se convierta en un criptograma C será $p_M(C)$
- Luego, como M debe haberse cifrado con alguna clave K

$$p_M(C) = \sum_{k=1}^{k=n} p(K) \quad \text{donde } C = E_K(M) \quad (E = \textit{encrypt}, \text{ con clave } K)$$

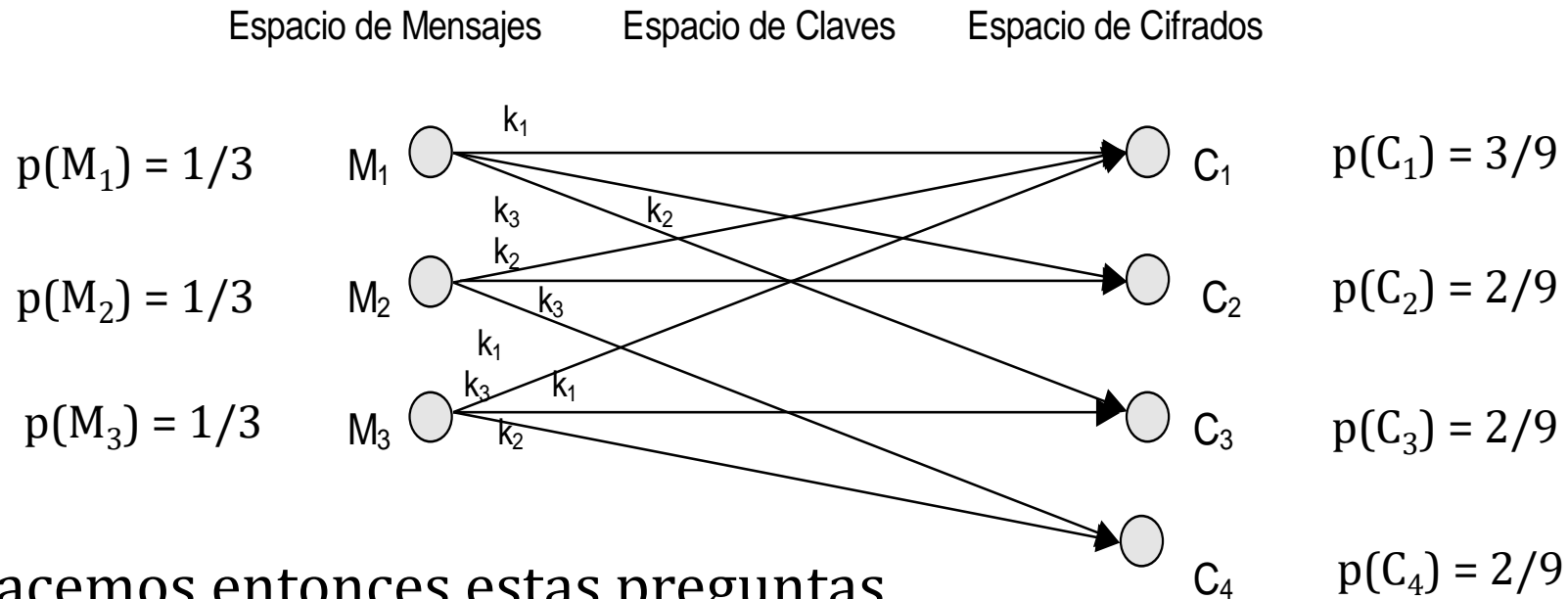
- Es decir, deberá cumplirse que: $\exists k_j / E_{k_j}(M_i) = C_i$
- Esto nos indica que, para lograr un secreto perfecto, el espacio de claves debe ser mayor o al menos de igual tamaño que el espacio de mensajes
- Algo en teoría imposible, excepto en el denominado cifrado de Vernam

Cifrado con secreto perfecto (3/3)

- La condición necesaria y suficiente del secreto perfecto es que para cualquier valor de mensaje M , se cumpla que la probabilidad de recibir el criptograma C , resultado de la cifra de M con una clave K , sea la misma que recibir ese mismo criptograma C , resultado ahora de la cifra de un mensaje M' cifrado con una clave K'
- Es decir, $p_M(C) = p(C)$ para todo valor de M



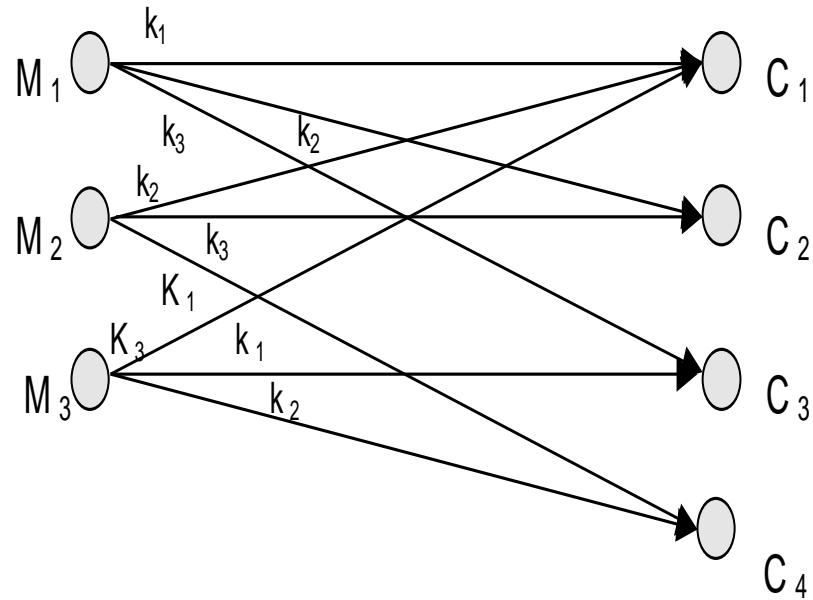
Cifrado sin secreto perfecto (1/2)



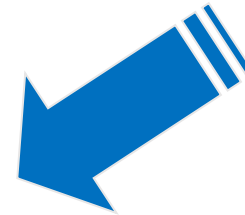
- Nos hacemos entonces estas preguntas
- ¿Cuál es la probabilidad de que un mensaje M_i cifrado con una clave K se convierta en un criptograma C_i , es decir $[P_{M_i}(C_i)]$, y la de que un criptograma C_i sea el resultado de la cifra de un mensaje M_i con clave K , es decir $[P_{C_i}(M_i)]$?



Cifrado sin secreto perfecto (2/2)



$P_{C1}(M_1) = 1/3$	$P_{C1}(M_2) = 1/3$	$P_{C1}(M_3) = 1/3$
$P_{C2}(M_1) = 1/2$	$P_{C2}(M_2) = 1/2$	$P_{C2}(M_3) = 0$
$P_{C3}(M_1) = 1/2$	$P_{C3}(M_2) = 0$	$P_{C3}(M_3) = 1/2$
$P_{C4}(M_1) = 0$	$P_{C4}(M_2) = 1/2$	$P_{C4}(M_3) = 1/2$



$p_{M1}(C_1) = 1/3$	$p_{M1}(C_2) = 1/3$	$p_{M1}(C_3) = 1/3$	$p_{M1}(C_4) = 0$
$p_{M2}(C_1) = 1/3$	$p_{M2}(C_2) = 1/3$	$p_{M2}(C_3) = 0$	$p_{M2}(C_4) = 1/3$
$p_{M3}(C_1) = 1/3$	$p_{M3}(C_2) = 0$	$p_{M3}(C_3) = 1/3$	$p_{M3}(C_4) = 1/3$

Definición de distancia de unicidad

- Se entenderá por distancia de unicidad a la cantidad de N letras en el texto cifrado o criptograma, mínima necesaria para que se pueda intentar con ciertas expectativas de éxito un criptoanálisis para romper la clave
- Este valor se obtiene cuando la equivocación de esa clave $H_c(K)$, o entropía condicional de la clave, se acerca a cero o tiende a anularse
- Dada la redundancia del lenguaje, a medida que se tenga un criptograma más largo la tarea de ataque del criptoanalista se va reduciendo
- Esto porque el atacante puede confiar más en los resultados que obtiene al aplicar las estadísticas y características del lenguaje sobre el criptograma
- Se buscará entonces el tamaño mínimo de N letras del criptograma que permita esperar que la solución de la clave K buscada sea única
- Para ello, se supondrá un esquema de cifrador aleatorio como el que se describe en las diapositivas siguientes

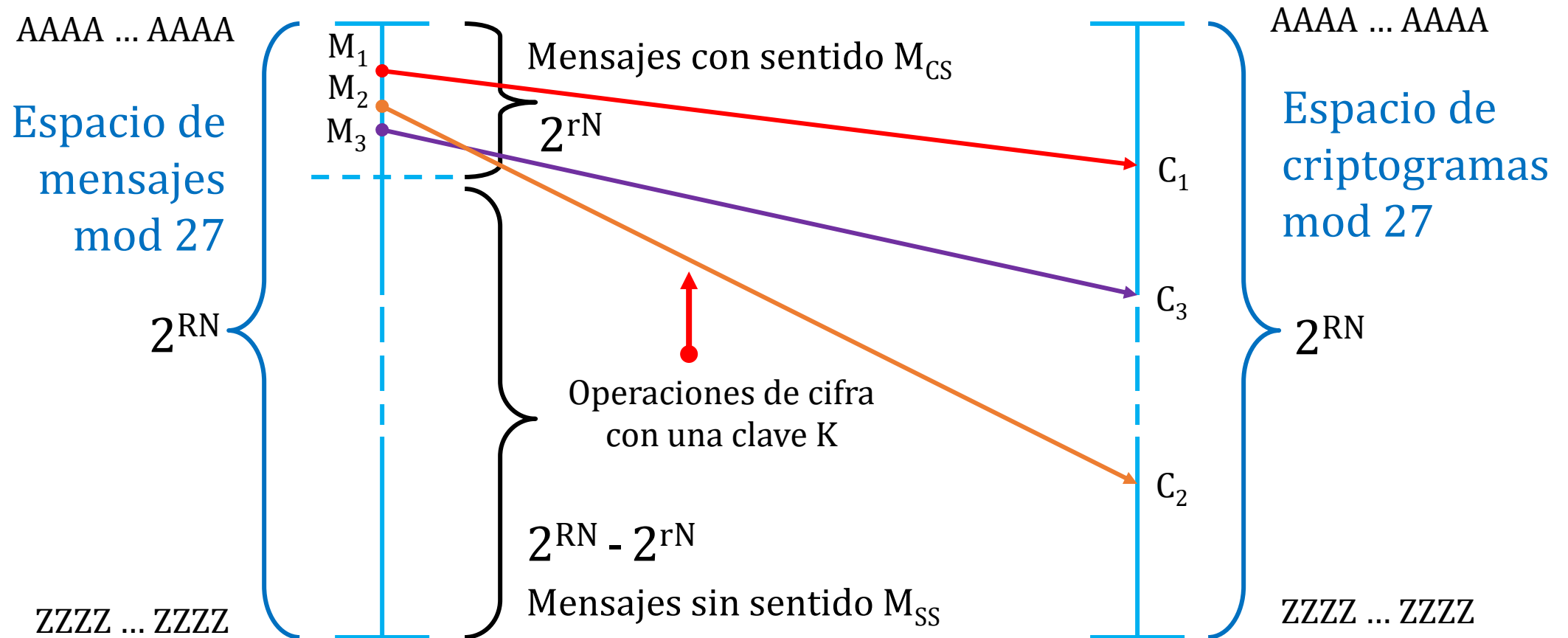
Parámetros del cifrador aleatorio (1/2)

- Existirán 2^{RN} mensajes posibles de longitud N
- Existirán 2^{rN} mensajes de longitud N con sentido
- El espacio de mensajes de longitud N se dividirá en:
 - Espacio de los mensajes con sentido: $M_{CS} = 2^{rN}$
 - Espacio de los mensajes sin sentido: $M_{SS} = 2^{RN} - 2^{rN}$
- En este escenario, los 2^{rN} mensajes con sentido M_{CS} van a ser equiprobables, siendo esta probabilidad $p(M_{CS}) = 1/2^{rN} = 2^{-rN}$
- El resto $(2^{RN} - 2^{rN})$ son mensajes sin sentido M_{SS} , no generados porque hablamos de texto, con una probabilidad nula $p(M_{SS}) = 0$

Parámetros del cifrador aleatorio (2/2)

- Además, existirán $2^{H(K)}$ claves equiprobables
- En donde $H(K)$ es la entropía de la clave
- Cada una de ellas con una probabilidad $p(K) = 1/2^{H(K)} = 2^{-H(K)}$
- Con cada una de estas claves se cifrarán todos los mensajes con sentido M_{CS} , dando lugar a 2^{RN} criptogramas posibles C con una longitud N
- A diferencia de los mensajes, como es lógico los criptogramas C obtenidos serán todos equiprobables, todos son “sin sentido”
- Con estos datos se obtiene el esquema de un cifrador aleatorio que se muestra en la siguiente diapositiva

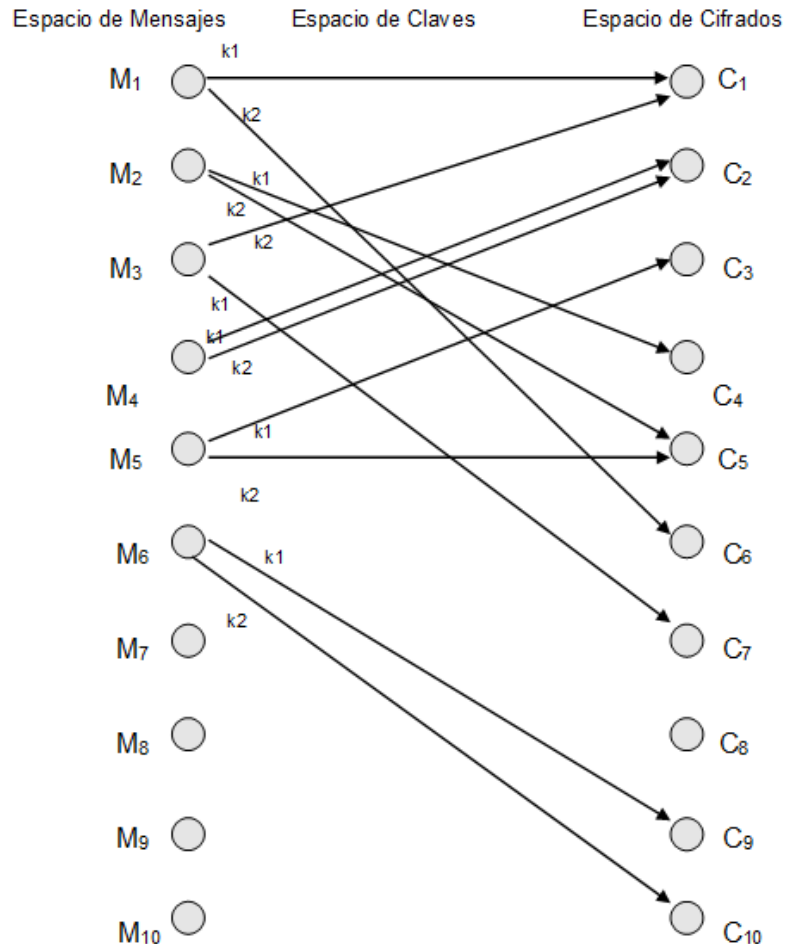
Cifrador aleatorio mensajes de longitud N



Veamos los escenarios de este modelo de cifra para sólo dos claves: k_1 y k_2



Escenarios cifrador aleatorio para k_1 y k_2



- Una solución verdadera **SV**, será aquella en la que un criptograma está asociado sólo a un texto en claro con sentido y que ha sido cifrado con una única clave k_i
 - Una solución falsa **SF** será cualquier otro resultado de cifra diferente al anterior
 - Soluciones verdaderas **SV** en el esquema mostrado
 - $C_3 = E_{k_1}(M_5)$ $C_4 = E_{k_1}(M_2)$ $C_6 = E_{k_2}(M_1)$ $C_7 = E_{k_1}(M_3)$
 - $C_9 = E_{k_1}(M_6)$ $C_{10} = E_{k_2}(M_6)$
 - Soluciones falsas **SF** en el esquema mostrado
 - $C_2 = E_{k_1}(M_4)$ $C_2 = E_{k_2}(M_4)$ $C_5 = E_{k_2}(M_2)$ $C_5 = E_{k_2}(M_5)$
 - $C_1 = E_{k_1}(M_1)$ $C_1 = E_{k_2}(M_3)$
1. Una solución falsa **SF** obvia será la del criptograma C_2
 2. Una solución falsa **SF** débil será la del criptograma C_5
 3. Una solución falsa **SF** fuerte será la del criptograma C_1

Cálculo de la distancia de unicidad (1/2)

- Para cada solución verdadera **SV** de un texto M cifrado con una clave k del espacio de claves $2^{H(K)}$, existirán otras $(2^{H(K)} - 1)$ claves con la misma probabilidad de entregar una solución falsa **SF**
- Sea q la probabilidad de obtener un mensaje con sentido M_{CS} al descifrar (o criptoanalizar) un criptograma C
- $q = 2^{rN} / 2^{RN} = 2^{(r - R)N} = 2^{-DN}$ (Redundancia $D \approx 3,4$ en módulo 27)
- Luego, **SF** = $(2^{H(K)} - 1) q = (2^{H(K)} - 1) 2^{-DN} = 2^{H(K) - DN} - 2^{-DN}$
- Si 2^{-DN} puede despreciarse por pequeño, entonces **SF** $\approx 2^{H(K) - DN}$
 - O lo que es lo mismo, $H(K) - DN = \log_2 \text{SF}$

Cálculo de la distancia de unicidad (2/2)

- La solución $SF = 2^{H(K)} - DN = 0$ es imposible porque sólo se llegaría a ella de forma asintótica y con un valor de N infinito
- Se aceptará entonces que haya como máximo una sola solución falsa SF y, por tanto, nos acercamos a la solución única
- Si $SF = 2^{H(K)} - DN = 1$, entonces $H(K) - DN = 0$
- Por lo tanto $N = H(K)/D$ será la distancia de unicidad
- Este valor es sólo de referencia para comparar los diferentes sistemas de cifra clásica. Para romper una cifra será necesario contar al menos con un tamaño de criptograma 10 veces mayor a N , siendo habitual una cantidad mucho mayor de texto cifrado

Conclusiones de la Lección 4.4

- Shannon define el secreto de un sistema de cifra como la incertidumbre del mensaje en claro M conocido el criptograma C
- Tanto los mensajes M , como las claves K y los criptogramas C tendrán una probabilidad p asociada y tendrán, además, su correspondiente espacio
- Un sistema tiene secreto perfecto si el conocimiento del texto cifrado no proporciona ninguna información acerca del mensaje. Se logra si el espacio de claves es al menos de igual tamaño que el espacio de mensajes
- La distancia de unicidad es la cantidad de N letras del criptograma mínima necesaria para intentar un criptoanálisis con ciertas expectativas de éxito
- Definiendo un esquema de cifrador aleatorio, se llega a que esa distancia de unicidad (un indicador sólo para comparación) viene dada por $H(X)/D$

Lectura recomendada

- Communication Theory of Secrecy Systems, C. E. Shannon, The Bell System Technical Journal, Vol. 28, 1949
 - <https://www.cs.virginia.edu/~evans/greatworks/shannon1949.pdf>
- Distancia de unicidad
 - https://es.wikipedia.org/wiki/Distancia_de_unicidad
- Shannon's Theory of Secrecy Systems, Eli Biham - May 3, 2005 (diapositivas de clase)
 - <http://www.cs.technion.ac.il/~cs236506/04/slides/crypto-slides-02-shannon.2x2.pdf>
- Criptografía y Seguridad en Computadores, Capítulo 3 Teoría de la información, Manuel Lucena, versión 5-0.1.4, noviembre 2019
 - <http://criptografiayseguridad.blogspot.com/p/criptografia-y-seguridad-en.html>

Class4crypt c4c4.5

Módulo 4. Teoría de la información en la criptografía

Lección 4.5. Métodos de difusión y confusión en la criptografía

4.5.1. Importancia de la difusión y la confusión en la criptografía

4.5.2. Método de difusión

4.5.3. Obtención de difusión mediante operaciones de permutación

4.5.4. Método de confusión

4.5.5. Obtención de confusión mediante operaciones de sustitución

4.5.6. Cifradores de producto

Class4crypt c4c4.5 Métodos de difusión y confusión en la criptografía
<https://www.youtube.com/watch?v=S1-RK6YcXfk>

Métodos de difusión y confusión

- Claude Shannon propone usar dos métodos en los algoritmos de cifra para que ésta cumpla con dos objetivos básicos, a saber:
 1. Difuminar la redundancia del lenguaje en el criptograma
 2. Dificultar el descubrimiento de la clave usada en la cifra
- Estos métodos son la difusión y la confusión. Propuestos en un informe secreto en 1946, finalmente fueron de dominio público en 1949 en su artículo *Communication Theory of Secrecy Systems*
- Aplicando estos métodos, se convierte un texto en claro en un criptograma que tiene una apariencia aleatoria y que difumina las características del lenguaje, muy difícil de romper por un atacante

Difusión y confusión según Shannon

Communication Theory of Secrecy Systems[★]

By C. E. SHANNON

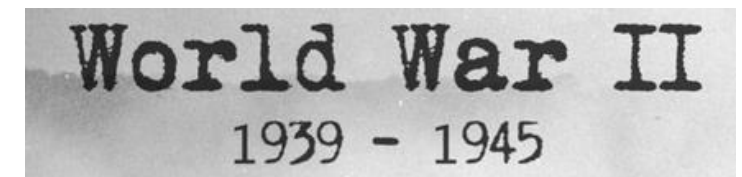
Two methods (other than recourse to ideal systems) suggest themselves for frustrating a statistical analysis. These we may call the methods of *diffusion* and *confusion*. In the method of diffusion the statistical structure of M which leads to its redundancy is “dissipated” into long range statistics—i.e.,

1 INTRODUCTION AND SUMMARY

Página 708

The problems of cryptography and secrecy systems furnish an interesting application of communication theory¹. In this paper a theory of secrecy systems is developed. The approach is on a theoretical level and is intended to complement the treatment found in standard works on cryptography². There, a detailed study is made of the many standard types of codes and ciphers, and of the ways of breaking them. We will be more concerned with the general mathematical structure and properties of secrecy systems.

[★] The material in this paper appeared in a confidential report “A Mathematical Theory of Cryptography” dated Sept.1, 1946, which has now been declassified.



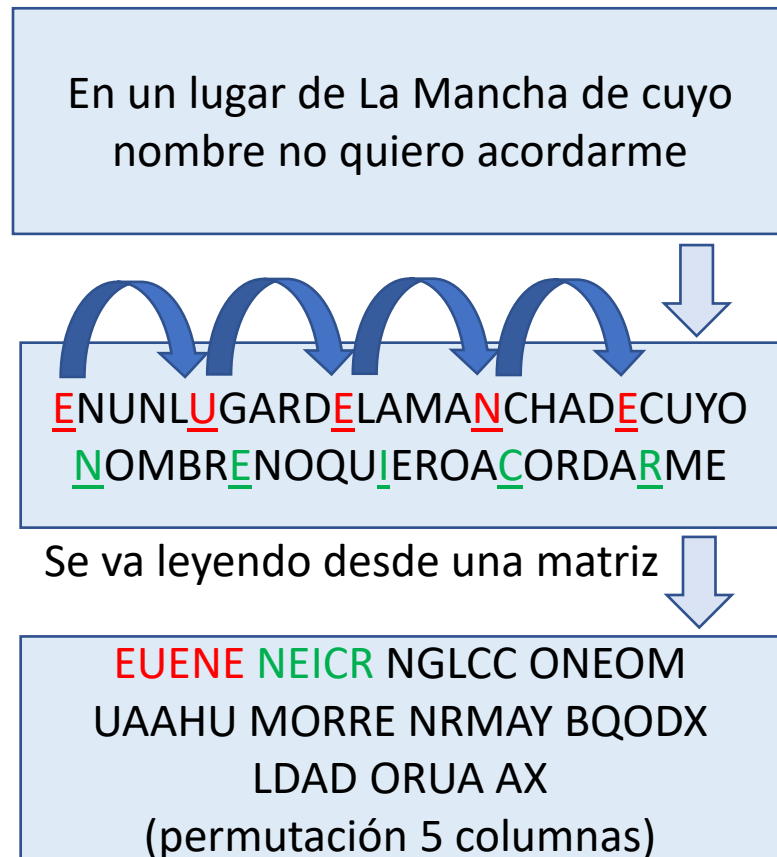
El método de difusión y permutaciones

- La difusión pretende difundir las características del texto en claro en todo el criptograma, ocultando así la relación entre el texto en claro y el texto cifrado
- Para lograr la difusión se aplicarán al texto en claro operaciones de permutación o transposición de letras
- De esta manera, las mismas letras del texto en claro aparecerán dispersas o desordenadas en todo el criptograma
- En la criptografía moderna, esta permutación operará sobre bits o bytes, como sucede con las permutaciones de bits en clave y texto en claro del DES y en la función ShiftRows sobre bytes del AES

Cifra por permutación mod 27

Criptoclásicos v2.1

URL en lectura recomendada



- En el ejemplo, se cifra mediante técnica de permutación por 5 columnas, en la que se va formando el criptograma leyendo el texto el claro de 5 en 5 letras desde una matriz formada ex profeso, obteniendo: $E+5 = U+5 = E+5 = N$, etc., incluyendo relleno si fuera necesario
- Al aplicar una permutación al texto en claro, el criptograma tiene las mismas letras que las del texto en claro, pero ubicadas en posiciones diferentes

El método de confusión y sustituciones

- El método de confusión pretende confundir al atacante, de manera que no le sea fácil poder establecer una relación sencilla entre el criptograma y la clave usada para el cifrado del texto en claro
- Para lograr la confusión, se aplicarán al texto en claro operaciones de sustitución de una letra por otra letra
- De esta manera, las letras del texto en claro aparecerán cambiadas en el criptograma por otras letras diferentes
- En la criptografía moderna, esta sustitución operará sobre bits y bytes, como sucede con los 48 bits de entrada de las cajas S del DES que se convierten en 32 y en la función SubBytes del AES

Cifra por sustitución mod 27

Criptoclásicos v2.1

URL en lectura recomendada

En un lugar de La Mancha de cuyo
nombre no quiero acordarme

ENUNLUGARDEELAMANCHADECUYO
NOMBRENOQUIEROACORDARME

Se sustituyen las letras

GOWON WICTF GNCÑC OEJCF
GEWAQ OQÑDT GOQSW KGTQC
EQTFC TÑG
(desplazamiento +2 espacios)

- En el ejemplo se ha cifrado mediante la técnica de sustitución, realizando un desplazamiento al texto en claro de dos letras hacia la derecha, como se muestra en estos dos alfabetos
 - ABCDEFGHIJKLMNOPQRSTUVWXYZ
 - CDEFGHIJKLMNOPQRSTUVWXYZAB
- Al aplicar una sustitución al texto en claro, lógicamente en el criptograma no aparecerán las mismas letras que las que había en ese texto en claro

Apariencia aleatoria del criptograma

- Dado que un texto cifrado debe tener una apariencia aleatoria, deberá eliminarse cualquier relación estadística entre el texto en claro y su texto cifrado
- Esto se logra si se usan técnicas de permutación o de sustitución
- Sin embargo, si se aplican ambas técnicas de forma independiente, no resulta suficiente para cifrar un texto de una manera segura

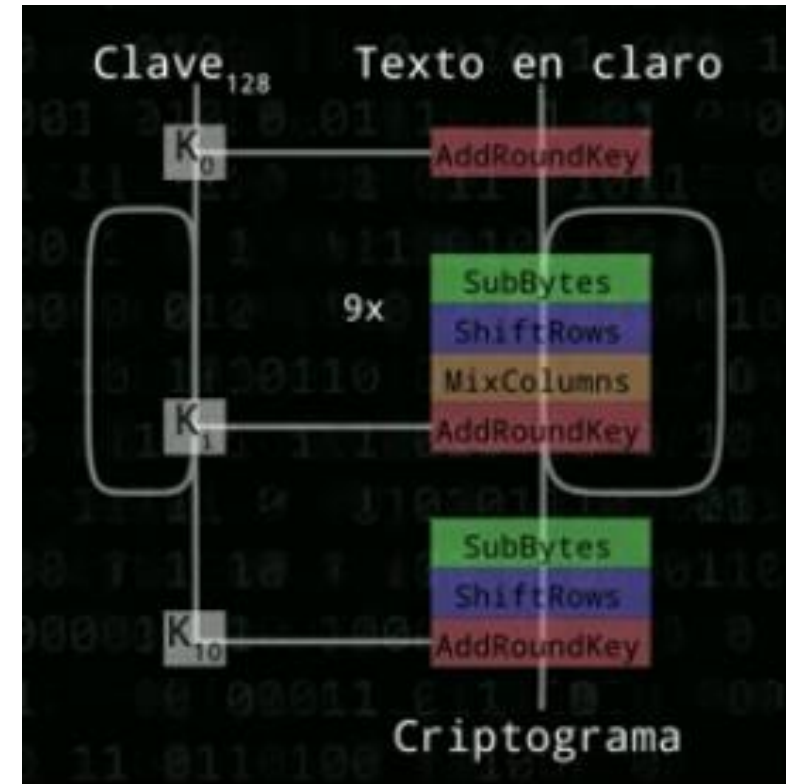
Lección 2. Sistemas de cifra con clave secreta



Lección 2 Sistemas de cifra con clave secreta,
proyecto intypedia (Criptored)
Dr. Fausto Montoya Vitini
Consejo Superior de Investigaciones Científicas CSIC
<https://www.youtube.com/watch?v=46Pwz2V-t8Q>

Sustitución + permutación = producto

- El uso combinado de las técnicas de sustitución y permutación, ofusca y dispersa la estructura estadística del texto en claro en el texto cifrado, dando así fortaleza a la cifra
- Los algoritmos que usan de forma simultánea técnicas de sustitución y permutación, son los denominados cifradores de producto. Un ejemplo lo tenemos en el algoritmo AES de cifra moderna simétrica en bloque

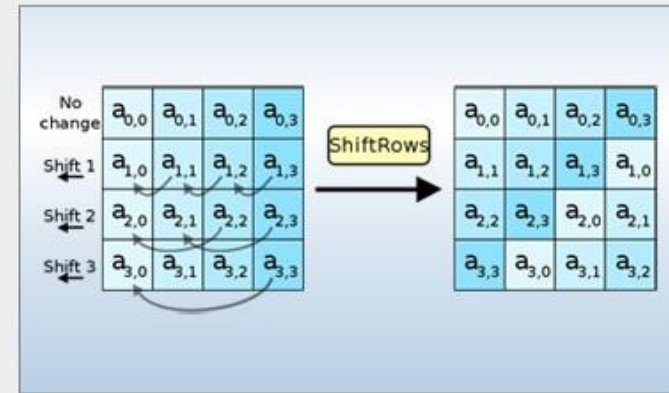
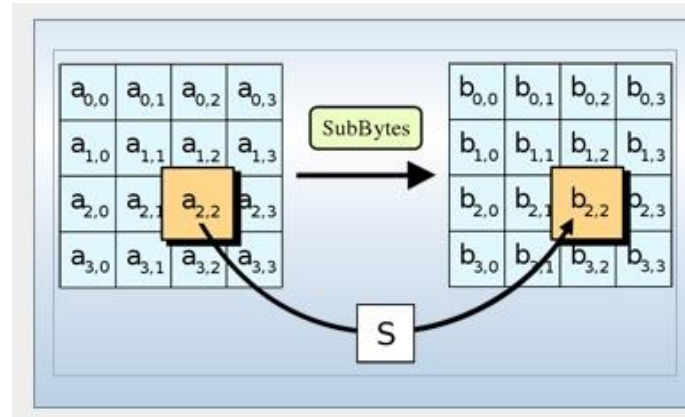


Píldora Thoth 30 ¿Cómo se cifra con el algoritmo AES?
<https://www.youtube.com/watch?v=tzj1RoqRnv0>

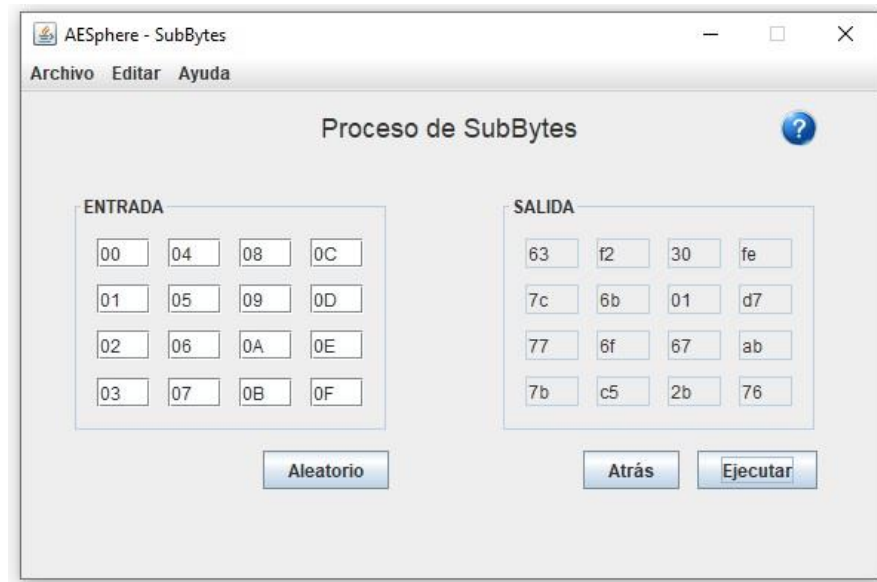
Cifrador de producto: AES



Sustitución



Permutación



Estudiados los tres pilares de la criptografía

- Con esta lección c4c4.5. terminamos estos 4 módulos iniciales de Class4crypt, que se han dedicado a una introducción y a los tres pilares de la criptología
 1. Módulo 2: [La teoría de los números](#)
 - Estudio de las matemáticas discretas, operaciones aritméticas en grupos y cuerpos que permiten el cifrado y el descifrado ([6 lecciones](#))
 2. Módulo 3: [La teoría de la complejidad algorítmica](#)
 - Estudio de la clasificación de los problemas como computacionalmente tratables o tipo P e intratables o tipo NP ([4 lecciones](#))
 3. Módulo 4: [La teoría de la información](#)
 - Estudio de la cantidad de información contenida en los mensajes y claves, así como su entropía y la redundancia del lenguaje ([5 lecciones](#))

Conclusiones de la Lección 4.5

- El método de la difusión pretende difundir o difuminar las características y estructura del lenguaje en todo el criptograma
- Para lograr la difusión se emplean técnicas del cifrado por permutaciones
- El método de la confusión pretende dificultar el ataque al no permitir que se establezca una relación sencilla entre el criptograma y la clave
- Para lograr la confusión se emplean técnicas de cifrado por sustituciones
- Los algoritmos que usan las técnicas de sustitución y permutación de forma conjunta se conocen como cifradores de producto
- Esto permite que el criptograma tenga una apariencia lo más aleatoria posible
- Los algoritmos en bloque DES y AES son ejemplos de cifradores de producto

Lectura recomendada

- Communication Theory of Secrecy Systems, C. E. Shannon, The Bell System Technical Journal, Vol. 28, 1949
 - <https://www.cs.virginia.edu/~evans/greatworks/shannon1949.pdf>
- Introducción a la seguridad informática y criptografía clásica, Lección 3: Conceptos básicos de la criptografía, Apartado 2, Técnicas usadas para la cifra, MOOC Crypt4you, Jorge Ramió, 2016
 - <https://www.criptored.es/crypt4you/temas/criptografiaclassica/leccion3.html#apartado2-1>
- Criptoclásicos v2.1, Juan Contreras Rubio, dirección Jorge Ramió, julio 2020
 - https://www.criptored.es/software/sw_m001c.htm
- Lección 2 Sistemas de cifra con clave secreta, proyecto intypedia, Fausto Montoya, 2010
 - <https://www.criptored.es/intypedia/docs/es/video2/GuionIntypedia002.pdf>
- Criptografía y Seguridad en Computadores, Capítulo 3 Teoría de la información, Manuel Lucena, versión 5-0.1.4, noviembre 2019
 - <http://criptografiayseguridad.blogspot.com/p/criptografia-y-seguridad-en.html>