uc3m | Universidad Carlos III de Madrid

PRÁCTICA: ENTROPÍA

CURSO CRIPTOGRAFÍA Y SEGURIDAD INFORMÁTICA

Ana I. González-Tablas Ferreres
José María de Fuentes García-Romero de Tejada
Lorena González Manzano
Pablo Martín González
UC3M | GRUPO COMPUTER SECURITY LAB (COSEC)

HERRAMIENTAS

ENT. No disponible en el aula. (Descarga de http://www.fourmilab.ch/random/)

- Para Windows: Descomprímalo en una carpeta.
- Para Unix: Descomprímalo. Compílelo mediante *make y ejecútelo mediante el comando ./ent.*
- Ejecución:
    - Sitúese en la carpeta donde se encuentra el ejecutable. Si el archivo a analizar se encuentra en la misma carpeta: ent "NOMBRE DE ARCHIVO A ANALIZAR"

OPENSSL. Disponible en el aula (Linux), información para Windows: https://wiki.openssl.org/index.php/Binaries

- En Windows, para poder ejecutarlo desde cualquier ruta del sistema debe incluir la carpeta bin de OpenSSL dentro de la variable de entorno PATH. Utilice el comando: *set PATH=%PATH%;"PATH DONDE INSTALE OPENSSL"/bin*

INTRODUCCIÓN

En criptografía, una de las características que se pide a todo algoritmo criptográfico es que la salida que ofrece el mismo sea lo más aleatoria posible. Una salida no aleatoria puede facilitar el criptoanálisis, exponiendo potenciales debilidades que podrían ser aprovechadas por terceros. Desgraciadamente, no existe una definición exacta del término aleatoriedad, por lo que **nunca se puede saber con certeza si una serie de datos se puede considerar como aleatoria**. Para paliar este problema, a lo largo de los años se han propuesto diferentes test que miden empíricamente la aleatoriedad de una serie de datos. Si bien **estos test no pueden asegurar con certeza absoluta la aleatoriedad**, si pueden detectar series que aunque lo parezcan, no lo son.

En esta práctica vamos a analizar la aleatoriedad de diferentes ficheros (cifrados y sin cifrar) y series de datos pseudo-aleatorias. Además, vamos a comprobar las

consecuencias que tienen características como la aleatoriedad en operaciones como la compresión. Finalmente identificaremos ficheros con alta entropía (número de bits de información) y que sin embargo están alejados de ser aleatorios (ejemplo: fichero jpg).

Ejemplo de salida de la batería de tests ENT:

ENT performs a variety of tests on the **stream of bytes** in *infile* (or standard input if no *infile* is specified) and produces output on the standard output stream. Example:

Entropy = 7.980627 bits per character. (max. 8)

Optimum compression would reduce the size of this 51768 character file by 0 percent.

Chi square distribution for 51768 samples is 1542.26, and randomly would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 125.93 (127.5 = random).

Monte Carlo value for Pi is 3.169834647 (error 0.90 percent).

Serial correlation coefficient is 0.004249 (totally uncorrelated = 0.0).

Observaciones:

El resultado negativo (no pasa alguna de las pruebas) de una batería de pruebas estadísticas sobre un fichero descarta aleatoriedad, pero el resultado positivo (todos las pruebas superadas) no garantiza nada.

Entropía por byte:

Fuente de mensajes: 2^8 posibles bytes. Entropía máxima: $\log_2 2^8 = 8$.

Índice de compresión:

Mide la capacidad de eliminar redundancia.

Chi Cuadrado:

Este test es muy sensible y es el que más se utiliza. Computa un valor para el conjunto de bytes del fichero y establece un porcentaje que indica con qué probabilidad un fichero realmente aleatorio excedería dicho valor. Si el porcentaje es mayor que 99% o menor que 1%, la secuencia es casi seguro no aleatoria. Si el porcentaje está entre el 99% y el 95% o entre el 1% y el 5%, la secuencia es

sospechosa. Porcentajes entre el 90% y el 95% o entre el 5% y el 10% indican que la secuencia es casi sospechosa. El óptimo porcentaje sería de un 50%.

Ejemplo:
51768 bytes – histograma de frecuencias - Grados de libertad: 2^8 - Diferencia con una distribución uniforme: 1542.26 – **Porcentaje: 0,01 de que una secuencia aleatoria de bytes excediera este valor** (equivale a la probabilidad de que un fichero aleatorio hubiera dado este valor).

Conclusión: es poco probable que el fichero sea aleatorio.

Media:

Se calcula la media. Los valores posibles son desde 0 a 255. Si la distribución es uniforme, la media debe ser próxima a 127.5.

Montecarlo para el cálculo de Pi:

Se traza un cuadrado. Se inscribe un círculo. Se toman coordenadas como secuencias de 6 bytes para generar una coordenada X y una Y. Se ven cuántas caen dentro del círculo y cuántas fuera. Esto da una aproximación de la superficie del círculo. Y dado que la superficie es ($radio^2 \cdot Pi$), podremos hacer la siguiente regla de tres:

la cantidad de puntos dentro es al total de puntos dibujados como
la superficie del círculo es a la superficie del cuadrado.

Y podremos calcular la superficie del círculo con la siguiente fórmula:

superficie.círculo = puntos.dentro × superficie.cuadrado / total.puntos

Converge (al valor de pi) muy lentamente: necesita secuencias largas para aproximarse bien al valor de Pi.

Coeficiente de correlación en serie:

Es una medida estadística de la correlación de un byte con el siguiente. Sólo mide secuencias de un byte. Valores máximos: -1, 1. Valores óptimos: cercanos a 0.

EJERCICIOS

Ejercicio 1 :

Descargue los siguientes ficheros a la carpeta de ENT. Si no puede descargar estos ficheros, puede sustituirlos por otros del mismo tipo:

⇒ **Tipo doc:**
https://d9db56472fd41226d193-1e5e0d4b7948acaf6080b0dce0b35ed5.ssl.cf1.rackcdn.com/spectools/docs/wd-spectools-word-sample-04.doc

⇒ **Tipo c:**
hhttps://www.sanfoundry.com/c-program-replace-line-text-file/
(copiar el primer programa en un fichero y poner extensión .c)

⇒ **Tipo jpeg:** http://www.stallman.org/IMG_5884.JPG

⇒ **Tipo gif:** http://www.ritsumei.ac.jp/~akitaoka/cogwhee1.gif

⇒ **Tipo bmp:** http://www.websiteoptimization.com/secrets/web-page/6-4-balloon.bmp

a) Ejecute ENT sobre estos ficheros. Describa, interprete y analice cada uno de los resultados obtenidos.

b) A la vista del análisis presentado en el apartado anterior, y atendiendo a la naturaleza de los ficheros (e.g. si son textuales, imágenes, con/sin estructura, con/sin compresión de contenidos), ¿son razonables los resultados de la aleatoriedad de los ficheros?

Solución:

a) Tras ejecutar ENT sobre los ficheros, el resultado es el siguiente:

```
[redacted]\Downloads\random>ent.exe wd-spectools-word-sample-04.doc
Entropy = 4.206582 bits per byte.

Optimum compression would reduce the size
of this 71680 byte file by 47 percent.

Chi square distribution for 71680 samples is 5041318.18, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 45.4639 (127.5 = random).
Monte Carlo value for Pi is 3.816842458 (error 21.49 percent).
Serial correlation coefficient is 0.540488 (totally uncorrelated = 0.0).
```

```
[redacted]\Downloads\random>ent.exe cfile.c
Entropy = 4.271430 bits per byte.

Optimum compression would reduce the size
of this 2030 byte file by 46 percent.

Chi square distribution for 2030 samples is 59510.13, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 68.1798 (127.5 = random).
Monte Carlo value for Pi is 4.000000000 (error 27.32 percent).
Serial correlation coefficient is 0.526674 (totally uncorrelated = 0.0).
```

```
[redacted]\Downloads\random>ent.exe IMG_5884.JPG
Entropy = 7.976906 bits per byte.

Optimum compression would reduce the size
of this 1344317 byte file by 0 percent.

Chi square distribution for 1344317 samples is 43454.94, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 129.2505 (127.5 = random).
Monte Carlo value for Pi is 3.115455341 (error 0.83 percent).
Serial correlation coefficient is 0.003607 (totally uncorrelated = 0.0).
```

```
[redacted]\Downloads\random>ent.exe cogwhee1.gif
Entropy = 7.985225 bits per byte.

Optimum compression would reduce the size
of this 21602 byte file by 0 percent.

Chi square distribution for 21602 samples is 431.11, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 128.9456 (127.5 = random).
Monte Carlo value for Pi is 3.102222222 (error 1.25 percent).
Serial correlation coefficient is 0.013769 (totally uncorrelated = 0.0).
```

```
[redacted]\Downloads\random>ent.exe 6-4-balloon.bmp
Entropy = 6.898953 bits per byte.

Optimum compression would reduce the size
of this 75088 byte file by 13 percent.

Chi square distribution for 75088 samples is 763343.67, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 82.0634 (127.5 = random).
Monte Carlo value for Pi is 3.384689148 (error 7.74 percent).
Serial correlation coefficient is 0.026482 (totally uncorrelated = 0.0).
```

A la vista de los resultados se puede afirmar que ningún fichero es aleatorio, pues no superan todos los test. Sin embargo, los algoritmos de codificación .jpg, .gif, .bmp sí que presentan mejores resultados pues pasan (o están próximos a pasar) algunos de los test.

b) Los resultados son razonables porque era de esperar que los ficheros de texto (.doc y .c) superen un menor número de test por ser más redundantes (entre otras cosas). En cambio, los ficheros de imágenes suelen tener una entropía mayor.

Ejercicio 2:

a) Utilizando el manual de la aplicación (https://www.openssl.org/docs/man1.0.2/), investigue y explique qué generan los siguientes comandos:

   ⇒ *openssl rand -out r1000 -rand FILE -base64 1000*

   ⇒ *openssl rand -out r1000000 -rand FILE -base64 1000000*

   FILE puede corresponderse con cualquier tipo de fichero, por ejemplo "CA.pl"

b) Ejecute los comandos anteriores. Ejecute ahora ENT sobre los ficheros resultantes. Explique los resultados de la ejecución de ENT y efectúe conclusiones sobre la aleatoriedad del contenido de los ficheros. Para ello puede ayudarse consultando:
   https://www.openssl.org/docs/man1.0.1/crypto/rand.html

Solución:

a)

   ⇒ *openssl rand [options] num: Genera un fichero aleatorio de num bits.*

   ⇒ *Options: [-out r1000 -rand FILE] El fichero de salida es r1000 y la semilla para el generador de números seudoaleatorios se coge del fichero FILE (o de cualquier otro fichero)*

   Con la misma semilla se generan resultados distintos.

b) El resultado de ejecutar ENT sobre los ficheros generados es el siguiente (nótese que el FILE escogido ha sido CA.pl. Si escoge otro fichero los resultados podrían variar, aunque no deberían hacerlo sustancialmente).

```
[redacted]\Downloads\random>ent.exe r1000
Entropy = 6.020157 bits per byte.

Optimum compression would reduce the size
of this 1378 byte file by 24 percent.

Chi square distribution for 1378 samples is 4178.20, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 82.7837 (127.5 = random).
Monte Carlo value for Pi is 4.000000000 (error 27.32 percent).
Serial correlation coefficient is 0.107646 (totally uncorrelated = 0.0).

[redacted]\Downloads\random>ent.exe r1000000
Entropy = 6.044390 bits per byte.

Optimum compression would reduce the size
of this 1375004 byte file by 24 percent.

Chi square distribution for 1375004 samples is 3958522.06, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 83.3086 (127.5 = random).
Monte Carlo value for Pi is 4.000000000 (error 27.32 percent).
Serial correlation coefficient is 0.116514 (totally uncorrelated = 0.0).
```

Como se puede apreciar, no todos los test se pasan satisfactoriamente (e.g compresión está lejos de ser 0). Por tanto, los ficheros no son aleatorios, aunque el comando openSSL está pensado para la generación de ficheros de contenido aleatorio.

Ejercicio 3:

a) Comprima el fichero ".doc" del ejercicio 1, calcule la entropía y compárela con dicho ejercicio.

b) Cifre el fichero ".doc" del ejercicio 1 con OpenSSL de la siguiente forma

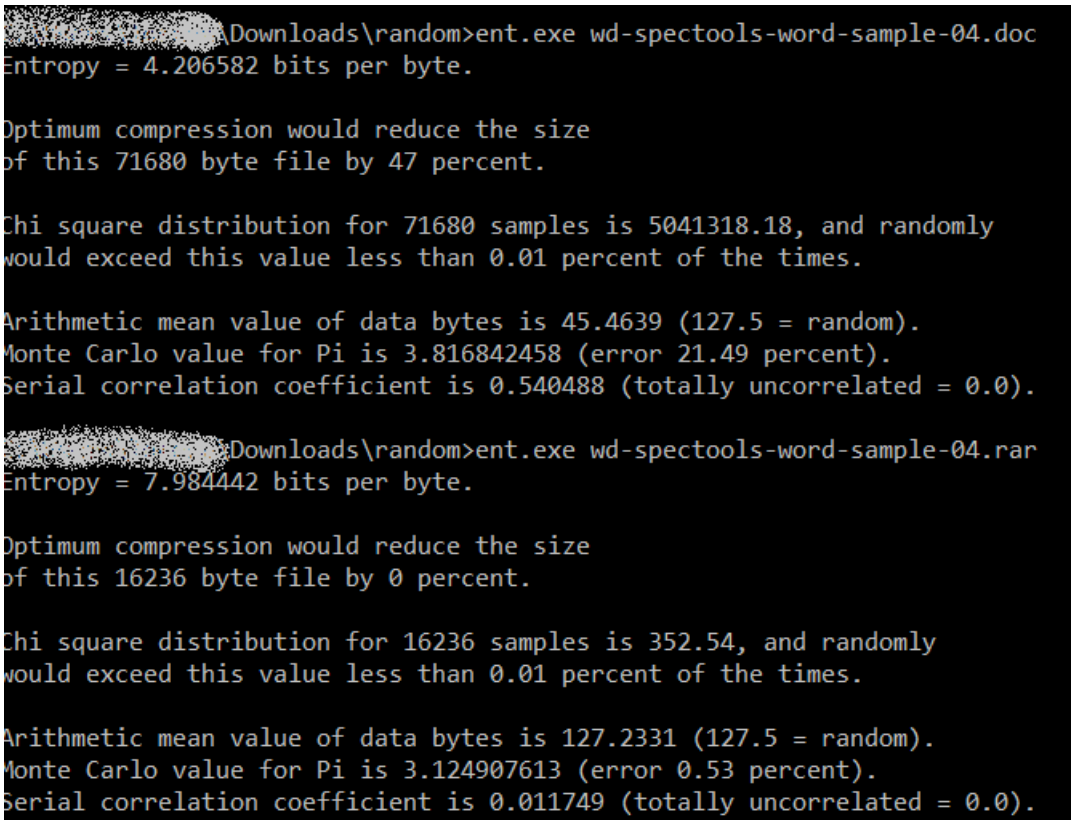
openssl enc -aes-256-cbc -salt -in FILE.doc -out FILE_ENCRYPTED.doc

Analice los resultados que ofrece ENT sobre el fichero *"FILE_ENCRYPTED.enc"* y compárelos con la entropía del fichero original.

c) Comprima el fichero *FILE.enc* con Winzip, Winrar o 7zip. ¿Hay variación de tamaño? Explique la causa que lo motiva. Calcule la entropía sobre este nuevo fichero (cifrado) y compárela con el fichero original y con el generado en el apartado "a" (cifrado).

Solución:

a) Los resultados de ejecutar ENT en el fichero .doc original y el comprimido son los siguientes:



Es posible apreciar que la mayoría de los test se pasan satisfactoriamente en el fichero comprimido. De modo que la compresión ha aumentado la entropía. Sin embargo, el test de la Chi square no se ha superado (50 es el óptimo) y por tanto, no se considera un fichero aleatorio.

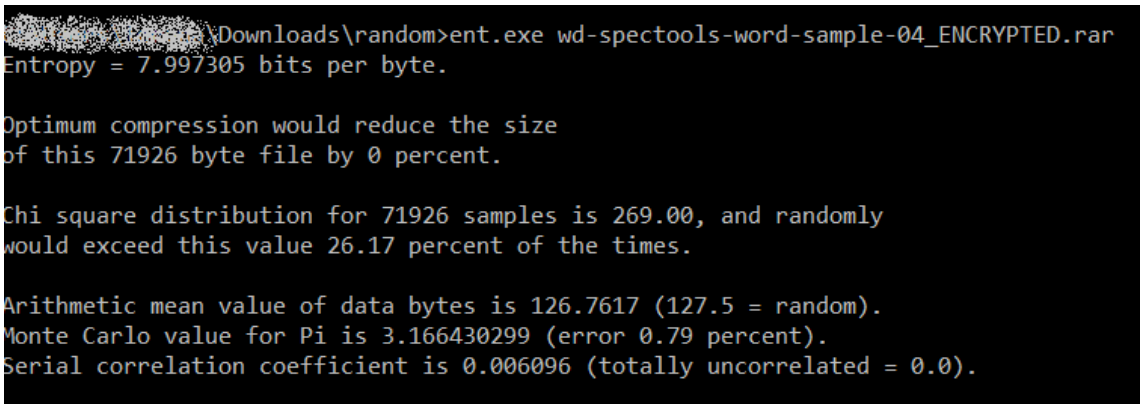
b) El fichero ha sido cifrado con la palabra "Seguridad". El resultado de ENT sobre el fichero en claro y cifrado es el siguiente:

```
[redacted]\Downloads\random>ent.exe wd-spectools-word-sample-04.doc
Entropy = 4.206582 bits per byte.

Optimum compression would reduce the size
of this 71680 byte file by 47 percent.

Chi square distribution for 71680 samples is 5041318.18, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 45.4639 (127.5 = random).
Monte Carlo value for Pi is 3.816842458 (error 21.49 percent).
Serial correlation coefficient is 0.540488 (totally uncorrelated = 0.0).

[redacted]\Downloads\random>ent.exe wd-spectools-word-sample-04_ENCRYPTED.doc
Entropy = 7.997409 bits per byte.

Optimum compression would reduce the size
of this 71712 byte file by 0 percent.

Chi square distribution for 71712 samples is 257.52, and randomly
would exceed this value 44.40 percent of the times.

Arithmetic mean value of data bytes is 126.9033 (127.5 = random).
Monte Carlo value for Pi is 3.162985274 (error 0.68 percent).
Serial correlation coefficient is 0.004573 (totally uncorrelated = 0.0).
```

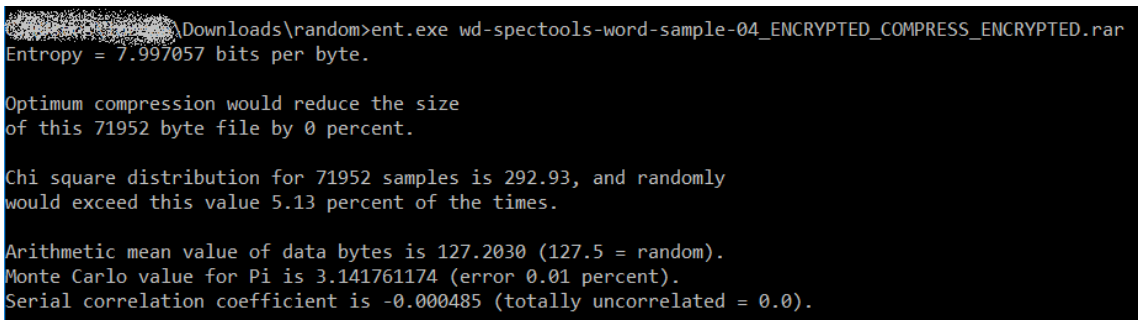
Se puede comprobar que la mayoría de los test han mejorado sustancialmente tras el cifrado. Dado que el test de Chi square está un tanto alejado del valor deseado (50.00) y la media aritmética está en 126.9 (aunque muy carca del valor deseado, 127.5), se descartaría la aleatoriedad. No obstante, en este caso la aleatoriedad podría ser cuestionable.

c) El comprimir el fichero cifrado (con winrar en este caso), los resultados son muy similares o incluso peor, como apunta el test de la Chi square. Esto puede ocurrir debido a que un fichero cifrado debe tener una alta entropía y no tendría porqué mejorarse tras su compresión.

En cuanto al tamaño, la compresión no lo ha afectado porque si un fichero cifrado tiene alta aleatoriedad, no puede comprimirse. De hecho, el tamaño ha aumentado ligeramente por los datos incluidos por la aplicación compresora.



El fichero comprimido ha sido cifrado nuevamente y el resultado tras pasar ENT es el siguiente:



Se puede apreciar en los resultados que, dado que el fichero ha pasado por un doble cifrado, hay test como la media aritmética que ya sí se han superado. Sin embargo, el test de la Chi square ha empeorado y ahora no se ha superado. En resumen, los resultados son muy similares entre ambos ficheros.