Mejorando la seguridad de un criptosistema OPE mediante la uniformización de los datos

Santi Martínez*, Daniel Sadornil†, Josep Conde*, Magda Valls* y Rosana Tomàs*
* Departament de Matemàtica, Universitat de Lleida
Email: {santi,jconde,magda,rosana}@matematica.udl.cat
† Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria
Email: sadornild@unican.es

***Resumen*—La cantidad de información almacenada en bases de datos crece constantemente. Una base de datos contiene múltiples registros divididos en varios campos. Algunos de éstos pueden contener información sensible, así que es necesario evitar que se acceda a ellos. Tradicionalmente, para proteger este tipo de información, se hace uso de la criptografía, pero la criptografía convencional tiene el problema de que, para consultas que necesitan acceder a un campo para todos los registros, se requiere descifrar el campo entero.**

La criptografía ordenable asegura que comparar datos cifrados produce el mismo resultado que comparar los datos originales, lo que permite ordenarlos sin descifrarlos. Por tanto, con este sistema se admiten búsquedas y consultas por intervalos en campos cifrados.

En este artículo proponemos un complemento, compatible con múltiples criptosistemas ordenables, consistente en transformar los datos de manera que se oculte su distribución estadística como paso previo al cifrado ordenable.

***Palabras clave*—bases de datos *(databases)*, criptografía ordenable *(order preserving encryption)*, criptografía simétrica *(symmetric key cryptography)*, distribución de probabilidad *(probability distribution)*, privacidad *(privacy)*, uniformización de datos *(data uniformization)*.**

I. Introducción

La criptografía permite esconder información sensible ante atacantes potenciales. Sin embargo, la información cifrada puede tener diferentes usos que requieran técnicas criptográficas específicas. En particular, este trabajo se centra en la privacidad y/o seguridad en bases de datos.

La seguridad en bases de datos es esencial para evitar un acceso no autorizado a información sensible. A título de ejemplo, en *The Toronto Star* [1] se explicó como un banco vendió un disco duro en eBay, olvidándose de borrar los datos en claro de centenares de clientes.

En bases de datos seguras, a veces se necesita permitir ciertas operaciones, o consultas, que requerirían que se descifrasen todos los campos necesarios para realizarlas, e.g. obtener los registros con un campo entre dos valores.

La criptografía ordenable (OPE, *Order Preserving Encryption*) permite hacer comparaciones de orden con datos cifrados, pues garantiza que éstos conservan el orden establecido entre datos en claro. Así, si un campo está cifrado de esta manera, las consultas de rango se pueden realizar eficientemente y asegurando que un atacante que tuviera acceso a la información almacenada en la base de datos no pueda obtener información de los datos en claro.

Consideremos una base de datos médica cifrada y supongamos que queremos saber el número de pacientes en un grupo de edad. Si el criptosistema usado no conserva el orden, esa consulta requerirá que cifremos cada uno de los valores del intervalo de edad y los comparemos con el campo correspondiente, o, alternativamente, que descifremos el campo de edad de todos los registros y los comparemos con los límites del intervalo. Si el algoritmo de cifrado no es determinista (o si en lugar de un entero el campo contiene un número real, como el nivel de azúcar en sangre) sólo la segunda alternativa es válida.

En cambio, si la base de datos usa OPE, sólo necesitamos cifrar los extremos del intervalo y comprobar cuantos registros tienen su campo cifrado de edad entre esos dos valores.

En esencia, un esquema OPE es una función estrictamente creciente del conjunto de datos en claro al conjunto de datos cifrados. Su seguridad recae en que dicha función, aun manteniendo el orden, parezca lo más aleatoria posible [2]. Esto asegurará que sólo aquellos con un conocimiento exacto de como calcularla (lo que está determinado por la clave del criptosistema) serán capaces de invertirla.

Los esquemas OPE son necesariamente simétricos, puesto que el conocimiento de la función de cifrado permite aproximar, hasta cualquier precisión, la función de descifrado. Nótese que, si un atacante con capacidad de cifrar valores arbitrarios quiere descifrar un valor concreto, podría realizar una búsqueda dicotómica de dicho valor, hasta llegar a una aproximación satisfactoria. Así, un hipotético OPE asimétrico sería automáticamente vulnerable a este ataque.

Los esquemas OPE han sido diseñados para entornos en que exista la posibilidad de que un intruso pueda acceder a la base de datos cifrada pero que no pueda cifrar ni descifrar valores arbitrarios.

Desde el punto de vista de un atacante, saber que un campo concreto ha sido cifrado con un esquema OPE proporciona una fuente útil de información si puede acceder a los datos almacenados. Si conociera los valores en claro de un conjunto de valores cifrados, podría crear una aproximación de la función de descifrado. Por lo que, si accediera a nuevos valores cifrados, podría usarla para aproximar los valores en claro correspondientes.

Por ejemplo, un atacante conoce x_1, x_2, y_1, y_2 y que $y_1 = Enc(x_1)$ y $y_2 = Enc(x_2)$; si obtuviera un valor cifrado y, con

ISBN: 978-84-9717-323-0

$y_1 < y < y_2$, entonces sabría que su descifrado, x, pertenece al intervalo (x_1, x_2). Además, podría deducir un valor x' cuya proximidad a x dependería de la predictibilidad de la función y la distancia entre los valores que supiera de antemano.

Por tanto, una función OPE debería minimizar este problema asegurando un alto nivel de impredecibilidad siendo lo más aleatoria posible [2]. Además, es deseable ocultar la distribución estadística de los datos cifrados.

En este artículo, proponemos la transformación de los datos a cifrar con un esquema OPE de manera que la entrada al criptosistema acabe siendo lo más uniforme posible. Consideramos un esquema OPE que cifra datos pertenecientes al intervalo $[0, 1]$, por lo que, antes de cifrar, convertiremos los datos desde la distribución inicial a una uniforme en $[0, 1]$. De forma similar, tras el descifrado del esquema OPE, convertiremos el valor obtenido, a un valor de la distribución inicial.

El resto del artículo está estructurado en las siguientes secciones: La sección II, expone algunas de las propuestas anteriores. En la sección III se propone la forma de escoger la distribución más apropiada para los datos. La sección IV explica cómo se complementa un sistema OPE concreto y como afecta la transformación a su eficiencia. Finalmente, las conclusiones se dan en la sección V.

II. Trabajos previos

Durante la última década, debido, en parte, al aumento de la preocupación por la privacidad de los datos preservando el análisis de los mismos, la criptografía ordenable ha experimentado un gran interés.

Bebek, en [3], realizó una primera propuesta donde proponía un método para cifrar un entero p añadiendo los p primeros valores de una secuencia pseudo-aleatoria segura de enteros positivos. Sin embargo, el coste de cifrar un valor p de n bits mediante este método es exponencial en n. Además, si μ es la media de la distribución de la secuencia pseudo-aleatoria, entonces $f(x) = \mu x$ aproxima la función de cifrado, y $f^{-1}(x) = x/\mu$, la de descifrado. Estas aproximaciones serán menos útiles si μ es cercano a 0 y la distribución tiene una desviación grande.

En [4], Ozsoyoglu et al. propusieron el uso de polinomios para el cifrado de enteros. Dichos polinomios no deben tener ningún extremo en el intervalo al que pertenecen los datos. Pero, el hecho de que algunos polinomios no dispongan de una fórmula explícita de su inversa, lleva a los autores a proponer la composición de varios polinomios fácilmente invertibles, de manera que el descifrado consista en aplicar las inversas en orden inverso. Para evitar desbordamientos de enteros, deciden controlar los coeficientes y usar logaritmos, lo que requerirá tratar con números en coma flotante y errores de precisión. Esto hace que la elección de la clave sea un proceso complejo. Además, descifrar es mucho más difícil que cifrar.

Agrawal et al. [5] propusieron la transformación de datos que siguieran cierta distribución estadística en datos que mantuvieran el orden y siguieran una distribución distinta, escogida de antemano. Para generar la función de cifrado, hacen uso de todos los datos a cifrar disponibles, así como una muestra de valores de la distribución objetivo. Por tanto, el tiempo de generación de la clave es lineal en el tamaño de la base de datos. Al cifrar, los datos se transforman en una distribución uniforme y, desde ésta, en la distribución objetivo. Para hacer esto, separan los datos en diversas particiones y, dentro de éstas, usan interpolación lineal. Si después de haber generado la clave se añadieran una gran cantidad de datos a la base de datos, podría ser necesario escoger una nueva clave y volver a cifrar la base de datos.

En [6], Lee et al. propusieron el esquema COPE (*Chaotic Order Preserving Encryption*). En este esquema, que reparte los datos en subconjuntos, se altera el orden de los mismos en función de la clave, por lo que no es un esquema OPE puro. El hecho de que haya que reordenar los subconjuntos para responder a una consulta afecta negativamente al coste.

Boldyreva et al. [2] presentaron una función OPE basada en el uso de un algoritmo de muestreo para la distribución hipergeométrica. Señalan el hecho de que, para funciones OPE con datos en claro y cifrados tomando valores enteros, el conjunto de salida es mayor que el de entrada (lo que permite que no haya dos valores en claro a los que corresponda el mismo valor cifrado). Así, una función de $\{1, ..., M\}$ a $\{1, ..., N\}$, con $M < N$, puede ser determinada unívocamente mediante la elección de un subconjunto (de cardinal M) del conjunto de salida que contendrá el cifrado de los valores del conjunto de entrada. Es más, basándose en las diferentes formas de hacer esta elección, proponen un criterio que debe cumplir una buena función OPE, recordando que, básicamente, la función, aun manteniendo el orden, tiene que ser lo más aleatoria posible.

En [7], se propuso un método OPE que cifraba datos pertenecientes al intervalo real $[0, 1]$. La función de cifrado se obtenía como composición de varias funciones básicas y el descifrado consistía en componer sus inversas en orden inverso. Cada función básica se definía por dos segmentos, el primero, desde el origen hasta el punto $P_{k_i} = (x_{k_i}, y_{k_i})$, y el segundo, de P_{k_i} al punto $(1, 1)$. La colección de todos los P_{k_i} constituía la clave del criptosistema. Para evitar funciones de mala calidad, la región donde se escogían los puntos de la clave estaba acotada.

En [8], se exponen dos metodologías para analizar la calidad de una función OPE. La primera de ellas se basa en la conversión de la función de cifrado en una secuencia que poder analizar como una señal de ruido. La segunda se basa en calcular las diferencias entre la función de cifrado y las aproximaciones que un atacante puede calcular a partir de un pequeño conjunto de puntos conocidos.

Más recientemente, se ha desarrollado un nuevo esquema OPE [9], que construye de forma iterativa una serie de puntos por los que pasará la función de cifrado. Inicialmente, la función es la identidad de $[0, 1]$ a $[0, 1]$. Así, en el primer nivel, se considera el rectángulo, cuyos lados son paralelos a los ejes, y con esquinas en $(0, 0)$ y $(1, 1)$, y se elige un punto en la diagonal descendiente. Este punto permite definir dos nuevos rectángulos con un vértice en el nuevo punto y otro en uno de los puntos extremos iniciales. En un segundo

nivel, se elige un punto en la diagonal descendiente de cada uno de los dos rectángulos, lo que permite definir cuatro rectángulos más pequeños. El proceso se repite hasta obtener la precisión deseada. Los puntos obtenidos constituyen la clave del criptosistema.

En este artículo, proponemos complementar un sistema OPE mediante la transformación de los datos a cifrar, de manera que oculte mejor la distribución estadística que éstos tuvieran inicialmente. El objetivo es que, antes de aplicar la función OPE, los datos sigan una distribución uniforme en $[0, 1]$.

III. Modelos de distribución de probabilidad de los datos

Nuestra propuesta consiste en complementar un sistema OPE mediante la ocultación de la distribución de probabilidad de los datos a cifrar. Por tanto, la transformación que proponemos no pretende servir como método de cifrado en sí mismo, sino como un preproceso que mejore la seguridad del sistema completo (para que los datos cifrados no conserven las propiedades de la distribución inicial).

Dicha ocultación de la distribución se realizará mediante la función de distribución acumulada (CDF, *Cumulative Distribution Function*). De esta forma, si X es una variable aleatoria continua con CDF F_X, entonces la variable aleatoria $P = F_X(X)$ tiene distribución uniforme en $[0, 1]$. Así, para recuperar la distribución original se usará la función cuantil F_X^{-1} (inversa de la CDF). En realidad, la transformación no necesita conocer la distribución real de los datos, pues es suficiente utilizar una que tenga el mismo soporte (i.e. que el intervalo para el que la función de densidad es no nula coincida en ambos casos). Esto se debe a que, para una variable X' con el mismo soporte que X, la variable $P' = F_X(X')$ también está distribuida entre 0 y 1, aunque ya no siga una distribución uniforme. Evidentemente, si conocemos la distribución de los datos a cifrar, lo ideal es utilizar la CDF y la cuantil de dicha distribución, pues es lo que dará mejor resultado.

Dependiendo del rango de valores que puedan tener los datos a cifrar consideramos tres grandes familias de distribuciones: con soporte finito, soporte infinito acotado inferiormente (o superiormente), y soporte infinito no acotado. Dentro de cada familia escogemos una distribución que, asumiendo un mínimo de información adicional, maximice la entropía.

Según el principio de máxima entropía, la distribución de probabilidad que mejor representa a una variable aleatoria es aquella en que, dadas unas ciertas condiciones, la desinformación es máxima. Así, una distribución de máxima entropía es aquella en que su entropía es al menos tan grande como la de cualquier otra distribución de su clase (donde una *clase* es el conjunto de distribuciones que cumplen una serie de restricciones). Por tanto, si de una distribución sólo se conocen unos pocos parámetros, la distribución a escoger es la que tenga máxima entropía para esos parámetros, lo que asegura que la distribución asume el mínimo de información adicional. Además, muchos sistemas tienden a seguir distribuciones de este tipo de manera natural.

En función del tipo de soporte hemos considerado las siguientes clases:

- Distribuciones con un mínimo y un máximo que conocemos. La distribución de máxima entropía para esta clase es la distribución uniforme continua.
- Distribuciones con mínimo y media conocidos. La distribución de máxima entropía para esta clase es la distribución exponencial.
- Distribuciones con media y varianza conocidas. La distribución de máxima entropía para esta clase es la distribución normal.

En la tabla I se muestran la CDF y la función cuantil de las distribuciones consideradas. Éstas son: la uniforme $\mathcal{U}(a, b)$, donde a y b son el mínimo y el máximo del soporte; la exponencial $\mathcal{E}(\lambda, \theta)$, donde θ es el mínimo, $\lambda = (\mu - \theta)^{-1}$ y μ es la media; y la normal $\mathcal{N}(\mu, \sigma^2)$, donde μ es la media y σ^2 la varianza.

Si tratamos con datos acotados sólo superiormente, cambiando el signo a cada valor, es posible usar también con ellos la distribución exponencial. Una vez obtenido su valor uniformizado, habrá que reflejar el resultado para no invertir el orden de los datos, i.e. $p = 1 - F_{\mathcal{E}(\lambda,\theta)}(-x)$.

III-A. *Determinación de la distribución*

Para poder determinar la distribución que se aproxima más adecuadamente a los datos a cifrar, necesitamos conocer información de los mismos. Pues, en función de que estén o no acotados inferiormente y/o superiormente, usaremos una distribución u otra. Si no conocemos esa información, pero tenemos una muestra con algunos de los datos a cifrar, trataremos de inferir la distribución a partir de éstos.

De hecho, si no se dispone de ninguna información de los datos a cifrar y ni siquiera tenemos una muestra que analizar, la opción más sensata es utilizar una distribución normal $\mathcal{N}(0, 1)$. Con ello nos aseguramos que cualquier valor que nos encontremos se transforme en un valor entre 0 y 1.

En lo sucesivo, se asume que disponemos únicamente de una muestra S con n valores $(x_1, x_2, \ldots x_n)$ de los datos a cifrar (donde puede haber valores repetidos). Con esta muestra, deberemos determinar cual de las distribuciones consideradas es la más adecuada.

En esta sección proponemos un método sencillo, basado en la cantidad de datos que hay en tres subintervalos iguales entre el mínimo y el máximo. Existen métodos alternativos, desde dividir el rango en una cantidad mayor de subintervalos (y proceder de manera similar), hasta estimar primero los parámetros de las distribuciones y luego hacer una prueba χ^2 de Pearson [10] para ver cuál se ajusta mejor.

Para el método de los tres subintervalos, primero buscamos el *mínimo*, m, y el *máximo*, M, de la muestra que tenemos. Ambos valores existirán siempre, incluso aunque la distribución no esté acotada, ya que S es finito. A continuación, partimos el intervalo $[m, M]$ en tres partes iguales, mediante los valores $l_1 = m + \frac{1}{3}(M - m)$ y $l_2 = m + \frac{2}{3}(M - m)$ y contamos cuantos datos hay en cada uno de los subintervalos:

Tabla I
FUNCIONES DE DISTRIBUCIÓN ACUMULADA Y CUANTIL

Distribución	CDF	Cuantil
$\mathcal{U}(a,b)$	$p = \begin{cases} 0 & \text{para } x < a \\ \frac{x-a}{b-a} & \text{para } a \le x < b \\ 1 & \text{para } x \ge b \end{cases}$	$x = a + p(b-a)$ para $0 \le p \le 1$
$\mathcal{E}(\lambda,\theta)$	$p = \begin{cases} 0 & \text{para } x < \theta \\ 1 - e^{-\lambda(x-\theta)} & \text{para } x \ge \theta \end{cases}$	$x = \begin{cases} \theta - \frac{\ln(1-p)}{\lambda} & \text{para } 0 \le p < 1 \\ +\infty & \text{para } p = 1 \end{cases}$
$\mathcal{N}(\mu,\sigma^2)$	$p = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$	$x = \begin{cases} -\infty & \text{para } p = 0 \\ \mu + \sqrt{2\sigma^2}\,\operatorname{erf}^{-1}(2p-1) & \text{para } 0 < p < 1 \\ +\infty & \text{para } p = 1 \end{cases}$

$c_1 = \#\{x \in S | x < l_1\}$, $c_2 = \#\{x \in S | l_1 \le x \le l_2\}$, $c_3 = \#\{x \in S | x > l_2\}$.

Nótese que, si la distribución sólo está acotada inferiormente, el máximo (cuando n crezca) tenderá a infinito, por lo que es razonable suponer que los datos se acumularán en el primer subintervalo (o en el tercero, si está acotada sólo superiormente). Asimismo, si la distribución no está acotada, podemos asumir que los datos tenderán a acumularse en el segundo subintervalo. En cambio, si el soporte de la distribución es finito, no se espera que los datos tiendan a acumularse tanto en ningún subintervalo.

Por tanto, si en el primer subintervalo hay tantos datos como en los otros dos juntos ($c_1 \ge c_2 + c_3$) usaremos la CDF de la exponencial, $p = F_{\mathcal{E}(\lambda,\theta)}(x)$, si es el central el que tiene tantos como los otros dos ($c_2 \ge c_1 + c_3$) usaremos la CDF de la normal, $p = F_{\mathcal{N}(\mu,\sigma^2)}(x)$, y, si es el tercero ($c_3 \ge c_1 + c_2$) usaremos también la exponencial con los cambios necesarios ($p = 1 - F_{\mathcal{E}(\lambda,\theta)}(-x)$). Finalmente, si ningún subintervalo domina a los otros dos, usaremos la CDF de la uniforme, $p = F_{\mathcal{U}(a,b)}(x)$.

Una vez escogida la distribución más adecuada para modelar los datos, deberemos estimar sus parámetros, lo que será diferente en cada caso particular.

III-B. Estimación de los parámetros

En esta sección proponemos los métodos de estimación de los parámetros de cada una de las distribuciones consideradas.

Tal como se ha indicado para la determinación de la distribución, si conocemos los valores exactos de los parámetros que queremos estimar, los usaremos directamente. Si no es el caso, podemos utilizar la muestra para tratar de hacer una estimación.

De entre los diversos tipos de estimadores se ha optado por utilizar estimadores insesgados [11], [12] de mínima varianza (UMVUE, *Uniformly Minimum-Variance Unbiased Estimator*). Sin embargo, para los parámetros que afectan al soporte de la distribución se ha optado por realizar algunos ajustes que disminuyan la posibilidad de encontrarnos con valores fuera de rango.

A continuación exponemos los estimadores que se han usado para las distinta distribuciones.

III-B1. Distribución uniforme: Ésta requiere dos parámetros a y b que son los valores extremos del soporte.

Notemos que todo valor menor o igual que $\hat{a}$ (la estimación de a) será interpretado como $\hat{a}$ y todo valor mayor o igual que $\hat{b}$ será interpretado como $\hat{b}$, por tanto, si el estimador de a tiene sesgo, es preferible que sea hacia la izquierda, i.e. $\hat{a} \le a$. De la misma manera, es preferible que $\hat{b} \ge b$.

Para determinar los valores $\hat{a}$ y $\hat{b}$ partiremos de los estimadores UMVUE, que son $\hat{a}' = m - \frac{M-m}{n-1}$ y $\hat{b}' = M + \frac{M-m}{n-1}$. Estos estimadores son insesgados, pero puede suceder que $\hat{a}'$ sea mayor que a o $\hat{b}'$ sea menor que b, por lo que les aplicaremos unas correcciones para disminuir dicha posibilidad.

En primer lugar, la media μ de la distribución debería coincidir con el valor medio de a y b. Por tanto, si la media muestral $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ no coincide con $\mu' = \frac{1}{2}(\hat{a}' + \hat{b}')$, modificaremos uno de los dos parámetros para forzar la coincidencia. Si $\bar{x} < \mu'$, disminuiremos el estimador de a: $\hat{a}'' = 2\bar{x} - \hat{b}'$, manteniendo $\hat{b}'' = \hat{b}'$. Si $\bar{x} > \mu'$, aumentaremos el estimador de b: $\hat{b}'' = 2\bar{x} - \hat{a}'$, manteniendo $\hat{a}'' = \hat{a}'$.

En segundo lugar, la varianza σ^2 de la distribución debería coincidir con $\frac{1}{12}(b-a)^2$. Si la varianza muestral $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ es mayor que ${\sigma^2}'' = \frac{1}{12}(\hat{b}'' - \hat{a}'')^2$, modificaremos ambos parámetros para forzar que coincidan. Para ello, ampliaremos el soporte en ambas direcciones de manera que no se altere el punto medio del intervalo, así, $\hat{a} = \hat{a}'' - \delta$ y $\hat{b} = \hat{b}'' + \delta$, donde $\delta = s\sqrt{3} - \frac{1}{2}(\hat{b}'' - \hat{a}'')$. Nótese que, si $s^2 < {\sigma^2}''$ no haremos esta última modificación, pues nos obligaría a reducir el soporte (lo que no es deseable), en tal caso, los estimadores serían $\hat{a} = \hat{a}''$ y $\hat{b} = \hat{b}''$.

III-B2. Distribución exponencial: Para determinar la distribución exponencial adecuada se requiere el valor de los parámetros λ y θ.

Para λ nos basaremos en que su valor es el inverso de la desviación estándar σ, i.e. $\lambda = \sigma^{-1}$. Por ello, usaremos un estimador UMVUE de este parámetro $\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)$ y luego lo invertiremos, $\hat{\lambda} = \hat{\sigma}^{-1}$. Como este parámetro no afecta al soporte de la distribución no haremos ninguna corrección adicional.

Para el parámetro θ, que corresponde a la cota inferior del soporte, es deseable que $\hat{\theta} \le \theta$, pues todo valor menor o igual que $\hat{\theta}$ será interpretado como $\hat{\theta}$. Por ello, escogeremos el menor de dos estimadores, el primero de ellos el UMVUE, que es $\hat{\theta}' = m - \hat{\sigma}/n$, y el segundo $\hat{\theta}'' = \bar{x} - \hat{\sigma}$ (que se basa en que la media μ de la distribución debería coincidir con $\theta + \sigma$). Así, $\hat{\theta} = \min(\hat{\theta}', \hat{\theta}'')$.

III-B3. Distribución normal: Esta distribución requiere los parámetros μ, la media, y σ^2, la varianza.

Ninguno de los dos parámetros afecta al soporte de la distribución, pues éste es infinito no acotado, por tanto, podemos utilizar los estimadores UMVUE en ambos casos.

Así, para μ usaremos la media muestral: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Y, para σ^2, la varianza muestral aplicando la corrección de Bessel: $\hat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Nótese que la corrección de Bessel se necesita porque para el cálculo de la varianza muestral usamos la media muestral $\bar{x}$. De hecho, si conociéramos la media poblacional μ (pero no la varianza), calcularíamos la varianza muestral usando μ (en vez de $\bar{x}$), en cuyo caso no usaríamos la corrección de Bessel, i.e. $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.

IV. Complementando un criptosistema concreto

El complemento propuesto en este artículo puede adaptarse a múltiples esquemas OPE. Con el fin de evaluar el sobrecoste que conlleva en el tiempo de cifrado y descifrado hemos decidido adaptarlo al criptosistema DOPE (*Diagonal-based Order Preserving Encryption*) [9] y realizar una serie de experimentos cuyos resultados mostramos en la tabla II.

El esquema DOPE cifra datos en $[0, 1]$, por lo que la adaptación es inmediata. Antes de cifrar, convertiremos los datos iniciales utilizando la función CDF de la distribución escogida mediante los criterios de la sección III. El resultado, que es un valor entre 0 y 1, se cifra mediante el esquema DOPE. Similarmente, después de descifrar, el valor obtenido se convierte a la distribución original mediante la función cuantil de la distribución correspondiente (ver tabla I).

Para la implementación se ha utilizado el criptosistema DOPE con pequeñas variaciones de seguridad y formato en los datos. Estos cambios han hecho que los tiempos de cifrado y descifrado sean mayores que los presentados en el artículo original.

Como se vio en las fórmulas de la tabla I, la conversión para la distribución uniforme es trivial. En el caso de la distribución exponencial, antes de cifrar habrá que calcular una exponenciación y, después de descifrar, un logaritmo. Asimismo, antes de cifrar un dato que siga una distribución normal habrá que calcular la función error complementaria erfc (pues la conversión es más rápida con ésta que con otras implementaciones de la CDF). Para cada una de estas tres funciones, se ha usado la implementación de la biblioteca estándar de C para valores de tipo `long double`. La función cuantil de la normal se ha implementado mediante el método de Newton-Raphson [13], por lo que consta de un bucle en el que, en cada iteración, se calculan la CDF (que, a su vez, calcula la función erfc) y la función de densidad (que requiere una exponenciación). Se ha escogido este método por ser de convergencia cuadrática.

Fijado un nivel de seguridad l, la clave del criptosistema DOPE es una lista de 2^l puntos (por los que pasa la función de cifrado). El cifrado (y el descifrado) requieren una búsqueda dicotómica para localizar la posición del valor a cifrar (o a descifrar) en la lista de abscisas (u ordenadas), por lo que su coste temporal es logarítmico en el tamaño de la lista, o, lo que es lo mismo, lineal en l.

La tabla II muestra los resultados para el criptosistema DOPE básico (sin cambio de distribución) y usando cada una de las tres distribuciones contempladas. Se ha usado un ordenador portátil con 2,4 GHz de CPU, 4 GB de RAM y el sistema operativo Debian GNU/Linux.

Se han realizado pruebas con un nivel de seguridad $14 < l < 19$, generando diez claves para cada valor de l. Con cada una de las claves se han cifrado y descifrado un millón de valores sin cambio de distribución, otro millón en el que los datos a cifrar seguían una distribución uniforme $\mathcal{U}$, otro más en que seguían una exponencial $\mathcal{E}$ y, finalmente, un millón más de datos que seguían una normal $\mathcal{N}$.

En la tabla II, las columnas T. cif. y T. des. indican el tiempo necesario para cifrar o descifrar un valor sin realizar ningún cambio de distribución. Las columnas T. cif. $\mathcal{U}$, $\mathcal{E}$ y $\mathcal{N}$ muestran el tiempo necesario para cifrar un valor que siga una de esas tres distribuciones, por lo que antes de usar la función de cifrado del esquema DOPE se debe uniformizar el valor mediante la CDF de la distribución. Las columnas T. des. $\mathcal{U}$, $\mathcal{E}$ y $\mathcal{N}$ muestran el tiempo necesario para descifrar un valor y luego convertirlo a su distribución original mediante la función cuantil de la distribución.

Podemos observar que el sobrecoste que conlleva cifrar o descifrar datos que siguen una distribución uniforme es de unos 10 ns y que éste es ligeramente mayor en el caso del descifrado. Si los datos siguen una exponencial, ambos sobrecostes son inferiores a 100 ns, siendo el del cifrado mayor que el del descifrado. Finalmente, en el caso de la normal, el sobrecoste al cifrado también es inferior a los 100 ns (incluso menor que el sobrecoste de la exponencial), pero para el descifrado, la implementación de la función cuantil ha causado un sobrecoste de más de 800 ns (y difícilmente podría ser mucho menor sin reducir la precisión).

Como es lógico, el sobrecoste debido al cambio de distribución es independiente del nivel de seguridad del esquema DOPE (y, por tanto, del tamaño de la clave), pues conlleva un proceso previo al cifrado o posterior al descifrado. Esto implica que, si se complementa cualquier otro criptosistema OPE, se deberían obtener sobrecostes similares.

V. Conclusión

En este artículo se ha propuesto un método de uniformización de datos como paso previo a la función de cifrado de un criptosistema ordenable. La finalidad de esta transformación es ocultar la distribución de probabilidad que siguen los datos en claro. Antes de cifrar datos que sigan una distribución conocida usaremos su función de distribución acumulada para convertirlos a una distribución uniforme en $[0, 1]$, y similarmente, después de descifrar usaremos la función cuantil para recuperar la distribución original.

También se ha propuesto un método para asignar una distribución de probabilidad a datos cuya distribución real no sea conocida. Se han considerado tres posibles casos: si los datos aparentan estar acotados en ambas direcciones,

Tabla II
RESULTADOS DE LA EXPERIMENTACIÓN

l	**T. cif.**	**T. des.**	**T. cif.** $\mathcal{U}$	**T. des.** $\mathcal{U}$	**T. cif.** $\mathcal{E}$	**T. des.** $\mathcal{E}$	**T. cif.** $\mathcal{N}$	**T. des.** $\mathcal{N}$
14	122,4 ns	122,7 ns	137,7 ns	135,1 ns	214,2 ns	195,6 ns	204,3 ns	976,2 ns
15	133,2 ns	136,2 ns	147,5 ns	148,9 ns	222,1 ns	207,0 ns	212,4 ns	1003,1 ns
16	195,1 ns	194,7 ns	204,4 ns	205,9 ns	285,2 ns	260,9 ns	265,9 ns	1074,3 ns
17	215,9 ns	212,7 ns	221,1 ns	223,3 ns	303,9 ns	285,1 ns	282,9 ns	1084,5 ns
18	236,9 ns	236,6 ns	244,5 ns	249,9 ns	326,1 ns	301,6 ns	307,1 ns	1122,9 ns
19	253,3 ns	253,4 ns	262,3 ns	275,0 ns	349,5 ns	328,7 ns	327,7 ns	1147,9 ns

se tratarán como si siguieran una distribución uniforme; si aparentan estar acotados sólo por un lado, se tratarán como si siguieran una exponencial (si la cota es superior requerirá una pequeña modificación); y, si aparentan no estar acotados, se tratarán como si siguieran una distribución normal.

Se han escogido estas distribuciones porque con ellas se pueden tratar datos definidos sobre cualquier tipo de intervalo, sea o no infinito. Además de eso, la uniforme es la distribución de máxima entropía de entre todas las que tienen soporte finito, la exponencial lo es entre todas las que tienen al menos una cota y tienen la misma media y la normal es la distribución de máxima entropía de entre todas las que tienen misma media y varianza.

La elección de la distribución y la estimación de sus parámetros se han realizado con el objetivo de disminuir la posibilidad de encontrarnos con valores fuera de su soporte. Nótese que, si el error supone asignar una distribución con soporte infinito a unos datos acotados o realizar una estimación de un máximo superior al real, dicho error, aunque no es deseable, no causará problemas.

Los métodos propuestos se han implementado y se ha comprobado qué sobrecoste representan en el tiempo de cifrado y descifrado de un criptosistema OPE.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por los proyectos MTM2010-21580-C02-01/02 y MTM2010-16051 del Gobierno de España y SGR2014-1666 de la Generalitat de Catalunya.

REFERENCIAS

[1] T. Hamilton, "Error sends bank files to eBay," The Toronto Star, September 15, 2003.

[2] A. Boldyreva, N. Chenette, Y. Lee, and A. O'neill, "Order-Preserving Symmetric Encryption," *Advances in Cryptology - EUROCRYPT 2009*, vol. LNCS, no. 5479, pp. 224–241, 2009, ISSN 0302-9743.

[3] G. Bebek, "Anti-tamper database research: Inference control techniques," Case Western Reserve University, Technical Report EECS 433, 2002, Final Report.

[4] G. Ozsoyoglu, D. A. Singer, and S. S. Chung, "Anti-Tamper Databases: Querying Encrypted Databases," in *17th Annual IFIP WG 11.3 Working Conference on Database and Applications Security*, 2003.

[5] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order Preserving Encryption for Numeric Data," in *ACM SIGMOD international conference on Management of data*, 2004, pp. 563–574.

[6] S. Lee, T.-J. Park, D. Lee, T. Nam, and S. Kim, "Chaotic Order Preserving Encryption for Efficient and Secure Queries on Databases," *IEICE Transactions on Information and Systems*, vol. E92-D, no. 11, pp. 2207–2217, 2009.

[7] S. Martínez, V. Mateu, R. Tomàs, and M. Valls, "Criptografía ordenable para bases de datos," in *XII Reunión Española sobre Criptología y Seguridad de la Información (RECSI)*, U. Zurutuza, R. Uribeetxeberria, and I. Arenaza-Nuño, Eds. Donostia–San Sebastián, España: Mondragon Unibertsitatea, September 2012, pp. 35–40, ISBN 978-84-615-9933-2.

[8] S. Martínez, J. M. Miret, R. Tomàs, and M. Valls, "Security Analysis of Order Preserving Symmetric Cryptography," *Applied Mathematics & Information Sciences (AMIS)*, vol. 7, no. 4, pp. 1285–1295, July 2013, ISSN 1935-0090.

[9] S. Martínez, J. M. Miret, R. Tomàs, and M. Valls, "Securing Databases by using Diagonal-based Order Preserving Symmetric Encryption," *Applied Mathematics & Information Sciences (AMIS)*, vol. 8, no. 5, pp. 2085–2094, September 2014, ISSN 1935-0090.

[10] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[11] V. G. Voinov and M. S. Nikulin, *Unbiased Estimators and Their Applications*, ser. Mathematics and Its Applications. Dordrecht, Netherlands: Kluwer Academic Publishers, 1993, vol. 1: Univariate Case.

[12] K. Lam, B. K. Sinha, and Z. Wu, "Estimation of parameters in a two-parameter exponential distribution using ranked set sample," *Annals of the Institute of Statistical Mathematics*, vol. 46, no. 4, pp. 723–736, 1994.

[13] J. Raphson, *Analysis Aequationum Universalis seu Ad Aequationes Algebraicas Resolvendas Methodus Generalis, & Expedita, Ex nova Infinitarum Serierum Methodo, Deducta ac Demonstrata*. London: Th. Braddyll, 1690.