

# Estudio y comparación de métodos de tokenización

Ricardo Quezada Figueroa\*, Laura Natalia Borbolla Palacios\*, Daniel Ayala Zamorano\*

\* Escuela Superior de Cómputo, Instituto Politécnico Nacional

Ciudad de México, México

qf7.ricardo@gmail.com, ln.borbolla.42@gmail.com, daz23ayala@gmail.com

**Resumen**—La tokenización consiste en el reemplazo de información sensible por valores sustitutos, llamados tokens, en donde el camino de regreso, del token a la información sensible, no es factible. En los últimos años este proceso se ha vuelto muy popular entre los comercios en línea, pues les permite descargar parte de las responsabilidades de seguridad adquiridas al manejar números de tarjetas de crédito en un tercero, proveedor de servicios de tokenización. Lamentablemente, existe una gran cantidad de desinformación alrededor de cómo generar los tokens, principalmente producida por las estrategias publicitarias de las empresas tokenizadoras, en donde cada una intenta convencer al comprador de que su sistema es el mejor, sin explicar realmente qué es lo que hacen para generar tokens. Uno de los mensajes más comunes entre la publicidad es que la criptografía y la tokenización son cosas distintas, y la segunda es mucho más segura. En este trabajo se explica a detalle en qué consiste la tokenización y cuál es su relación con la criptografía; se revisan y comparan los desempeños de los métodos más comunes para tokenizar; para terminar se concluye con una discusión alrededor de las ventajas y desventajas de cada uno.

**Palabras clave**—tokenización, criptografía, seguridad web

## I. INTRODUCCIÓN

Cuando el comercio a través de Internet comenzó a popularizarse, los fraudes de tarjetas bancarias se volvieron un problema alarmante: Visa y MasterCard reportaron, entre 1988 y 1998, pérdidas de 750 millones de dólares; según [13], en el año 2001 se tuvieron pérdidas de 1.7 miles de millones de dólares y, para 2002 aumentaron a 2.1. Como una medida para proteger los sistemas procesadores de pagos, las principales compañías de tarjetas de crédito publicaron un estándar obligatorio para todos aquellos que procesaran más de 20 000 transacciones anuales: el PCI DSS (en inglés, *Payment Card Industry Data Security Standard* [9]).

El PCI DSS cuenta con una gran cantidad de requerimientos, entre ellos, controles estrictos de acceso, monitoreo regular de redes, mantener programas de vulnerabilidades y políticas de seguridad de información, etcétera [2] [18]; por lo que, para negocios medianos y pequeños, resulta muy difícil obtener un resultado positivo. Además, a pesar de la publicación del estándar en 2004, han seguido existiendo grandes filtraciones de datos (*TJX* en 2006, *Hannaford Bros.* en 2008, *Target* en 2013, *Home Depot* en 2014, por mencionar algunos ejemplos).

Es en este contexto en el que surge el enfoque de la tokenización. Hasta antes de este momento, la idea era proteger la información sensible en donde quiera que se encontrara. Si los números de tarjetas de los usuarios se encontraban regados en diversas partes de un sistema, había que proteger todas esas partes. La tokenización consiste en concentrar la información

sensible en un solo lugar para hacer la tarea de protección más sencilla. Al momento de ingreso de un nuevo valor sensible, por ejemplo, la información bancaria de un usuario, se genera un token ligado a esa información: el token se usa en todo el sistema y la información sensible se protege en un solo lugar. Un posible adversario con acceso a los tokens no debe poder obtener la información sensible a partir de estos.

Una de las ventajas de la tokenización es que puede verse como un sistema autónomo, independiente del sistema principal. De esta manera se establece una separación de responsabilidades: el sistema principal se ocupa de la operación del negocio (por ejemplo, una tienda en línea) y el sistema tokenizador se dedica a la protección de la información sensible. Hoy en día varias compañías ofrecen servicios de tokenización que permiten que los comerciantes se libren casi por completo de cumplir con el PCI DSS. En la figura 1 se muestra una distribución bastante común para un comercio en línea: el sistema tokenizador guarda la información sensible en su base de datos y se encarga de realizar las transacciones bancarias.

Desde sus inicios, la tokenización se ha visto rodeada por una nube de desinformación: cada empresa usaba la palabra tokenización sin que existiera una definición formal, acordada por todos. En 2011, el PCI SSC (*Payment Card Industry Security Standard Council*) publicó unas guías de tokenización [8] con la idea de ayudar tanto a comerciantes en línea como a proveedores de servicios de tokenización a esclarecer qué es la tokenización y qué requerimientos debe de cumplir. Lamentablemente, el PCI no indicó cómo satisfacer los requerimientos, lo que dio pie a que muchas empresas se siguieran publicitando como tokenizadoras, sin que los usuarios pudieran saber qué es lo que pasa adentro de la caja negra.

Una de las consecuencias de esta desinformación es un mensaje bastante común entre las páginas de las empresas tokenizadoras: la tokenización y la criptografía son cosas distintas; la tokenización es una alternativa de la criptografía. Por ejemplo, lo único que *Shift4* dice con claridad sobre sus tokens es que se trata de valores aleatorios, únicos y alfanuméricos [15] [16]; para *Braintree*, la única manera de generar tokens es por métodos aleatorios [6]; para *Securosis* los tokens son valores aleatorios que nada tienen que ver con la criptografía [17]; *Jack Henry Card Processing Solution*, *Merchant Link* y *Bluepay TokenShield* no mencionan nunca cómo es que generan sus tokens. La mayoría de las soluciones tratan a sus métodos como secretos de compañía; esperan que el trato entre cliente y proveedor esté basado en la confianza

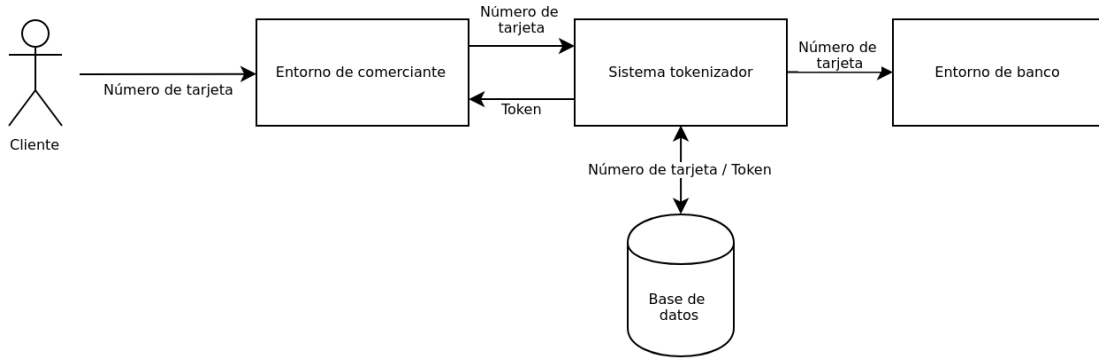


Figura 1. Arquitectura típica de un sistema tokenizador.

y no en la comparación de los propios métodos.

Como se verá por la exposición de algunos métodos tokenizadores, la tokenización es una aplicación de la criptografía y, como tal, debe ser analizada con la misma formalidad que las demás herramientas criptográficas. Por ejemplo, uno de los errores más comunes entre las empresas tokenizadoras es considerar a la generación de números aleatorios como fuera del campo de la criptografía.

En la sección dos de este trabajo se tocan algunos temas preliminares necesarios para presentar, en la sección tres, algunos de los métodos de tokenización. En la sección cuatro se presentan los resultados de las comparaciones de desempeño realizadas, y en la sección cinco se concluye con una discusión acerca de las ventajas y desventajas de cada método.

## II. PRELIMINARES

### II-A. Notación

Se denotarán a todas las cadenas de bits de longitud  $n$  como  $\{0, 1\}^n$ . Para cadenas  $x$  y  $y$  de símbolos sobre alfabetos arbitrarios, con  $|x|$  se denotará la longitud de la cadena y con  $x \parallel y$  la concatenación de ambas cadenas. El operador  $\oplus$  simboliza la combinación de los operandos, ya sea mediante sumas sin acarreo (equivalente del operador `xor` en cadenas binarias) o mediante suma global modular.

### II-B. Primitivas criptográficas

Un cifrado por bloques es un cifrado simétrico que se define por la función  $E : \mathcal{M} \times \mathcal{K} \rightarrow \mathcal{C}$  en donde  $\mathcal{M}$  es el espacio de textos en claro,  $\mathcal{K}$  es el espacio de llaves y  $\mathcal{C}$  es el espacio de mensajes cifrados. Tanto los mensajes en claro como los cifrados tienen una misma longitud  $n$ , que representa el tamaño del bloque [14].

Un cifrado que preserva el formato (en inglés *Format-preserving Encryption*, FPE) puede ser visto como un cifrado simétrico en donde el mensaje en claro y el mensaje cifrado mantienen un formato en común. Formalmente, de acuerdo a lo definido en [4], se trata de una función  $E : \mathcal{K} \times \mathcal{N} \times \mathcal{T} \times \mathcal{X} \rightarrow \mathcal{X}$ , en donde los conjuntos  $\mathcal{K}$ ,  $\mathcal{N}$ ,  $\mathcal{T}$ ,  $\mathcal{X}$  corresponden al espacio de llaves, espacio de formatos, espacio de *tweaks* y el dominio, respectivamente. El proceso de cifrado de un elemento del dominio con respecto a una llave  $K$ , un formato

$N$  y un *tweak*  $T$  se escribe como  $E_K^{N,T}(X)$ . El proceso inverso es también una función  $D : \mathcal{K} \times \mathcal{N} \times \mathcal{T} \times \mathcal{X} \rightarrow \mathcal{X}$ , en donde  $D_K^{N,T}(E_K^{N,T}(X)) = X$ .

Un generador determinístico de números aleatorios (DRBG, *Deterministic Random Bit Generator*) utiliza un mecanismo interno, generalmente una función hash o un cifrador por bloques, y un valor inicial llamado semilla, para producir bits de aspecto aleatorio. Dada la naturaleza determinística del proceso, a los números generados se les conoce como *pseudoaleatorios*. El método que utiliza como mecanismo interno a una función hash consiste en ir concatenando de forma consecutiva los valores hash derivados de la semilla e ir incrementando el valor de esta. El método basado en un cifrador por bloques utiliza el modo de operación de contador, en donde la semilla juega el papel de vector de inicialización. En ambos casos, la seguridad se basa en que la semilla sea un valor secreto [3].

## III. ALGORITMOS TOKENIZADORES

Como el enfoque de este artículo es ver a la tokenización como un servicio (figura 1), la interfaz para los procesos de tokenización y detokenización, desde el punto de vista de los usuarios del servicio, es sumamente simple: el proceso de tokenización es una función  $E : \mathcal{X} \rightarrow \mathcal{Y}$  y el de detokenización es simplemente la función inversa  $D : \mathcal{Y} \rightarrow \mathcal{X}$ , en donde  $\mathcal{X}$  y  $\mathcal{Y}$  son los espacios de números de tarjetas y tokens, respectivamente.

Los números de tarjetas bancarias cuentan con entre 12 y 19 dígitos, y se encuentran normados por el estándar ISO/IEC-7812 [12]. El último dígito se trata de un código de verificación calculado mediante el algoritmo de Luhn; este dígito hace que  $\text{ALGORITMODELUHN}(x) = 0$ . Para poder diferenciar a los números de tarjeta válidos de los tokens se establece que, para los tokens, el dígito verificador haga que  $\text{ALGORITMODELUHN}(x) = 1$ .

El PCI SSC (*Payment Card Industry Security Standard Council*) establece en sus guías de tokenización la siguiente clasificación para los algoritmos tokenizadores [8]:

- Métodos reversibles. Aquellos para los cuales es posible regresar al número de tarjeta a partir del token.

- Criptográficos. Ocupan un esquema de cifrado simétrico: el número de tarjeta y una llave entran al mecanismo de tokenización para obtener un token; el token y la misma llave entran al mecanismo de detokenización para obtener el número de tarjeta original.
- No criptográficos. Ocupan una base de datos para guardar las relaciones entre números de tarjetas y tokens; el proceso de detokenización simplemente es una consulta a la base de datos.
- Métodos irreversibles. Aquellos en los que no es posible regresar al número de tarjeta original a partir del token.
  - Autenticable. Permiten validar cuando un token dado corresponde a un número de tarjeta dado.
  - No autenticable. No permiten hacer la validación anterior.

La denominación *no criptográficos* resulta totalmente confusa, pues en realidad todos los métodos conocidos que caen en las categorías de arriba ocupan primitivas criptográficas. La segunda categoría (los irreversibles) carece de utilidad para aplicaciones que procesan pagos con tarjetas de crédito, pues la habilidad de regresar al número de tarjeta a partir de su token es uno de los requerimientos principales para los sistemas tokenizadores. Por lo anterior, en este trabajo se propone una clasificación distinta:

- Métodos criptográficos. Todos aquellos que ocupan herramientas criptográficas para operar.
  - Reversibles. Ocupan un esquema de cifrado simétrico: el número de tarjeta y una llave entran al mecanismo de tokenización para obtener un token; el token y la misma llave entran al mecanismo de detokenización para obtener el número de tarjeta original. El término *reversible* es porque se puede regresar al número de tarjeta sin ayuda de herramientas externas, como una base de datos.
  - Irreversibles. Ocupan herramientas criptográficas para generar el token de un número de tarjeta. Operan como funciones de un solo sentido: la única manera de regresar al número de tarjeta a partir de un token es mediante un ataque por fuerza bruta o mediante herramientas externas, como una base de datos.
- Métodos no criptográficos. Aquellos posibles métodos que no ocupen herramientas relacionadas con la criptografía; por ejemplo, un generador de números realmente aleatorio (TRNG, *True Random Number Generator*).

La clasificación de los *no criptográficos* solamente se propone para abarcar métodos de los cuales realmente se pueda decir que no se relacionan con la criptografía. En este trabajo no se presenta ningún método que clasifique en esa categoría.

A continuación se presentan algunos de los algoritmos tokenizadores más comunes. Al final de cada sección se explica en qué categoría cae según las dos clasificaciones anteriores.

### III-A. TKR

En [10] se analiza formalmente el problema de la generación de tokens y se propone un algoritmo que no está basado en cifrados que preservan el formato. Hasta antes de la publicación de este documento, los únicos métodos para generar tokens cuya seguridad estaba formalmente demostrada eran los basados en cifrados que preservan el formato.

El algoritmo propuesto usa un cifrado por bloques para generar tokens pseudoaleatorios y almacena en una base de datos la relación original de estos con los números de tarjetas. A continuación se muestran los procesos para tokenizar y detokenizar.

---

#### Algoritmo 1 Tokenización de TKR

---

```

1: función TKR-TOKENIZACIÓN( $x, k$ )
2:    $q \leftarrow \text{BUSCAR TARJETA}(x)$ 
3:   si  $q = 0$  entonces:
4:      $y \leftarrow \text{RN}(k)$ 
5:      $\text{INSERTAR}(x, y)$ 
6:   sino:
7:      $y \leftarrow q$ 
8:   fin si
9:   regresar  $y$ 
10: fin función

```

---



---

#### Algoritmo 2 Detokenización de TKR

---

```

1: función TKR-DETOKENIZACIÓN( $y$ )
2:    $q \leftarrow \text{BUSCAR TOKEN}(y)$ 
3:   si  $q = 0$  entonces:
4:     regresar error
5:   sino:
6:     regresar  $q$ 
7:   fin si
8: fin función

```

---

Las funciones `BUSCAR TARJETA`, `BUSCAR TOKEN` e `INSERTAR` sirven para interactuar con la base de datos. Lo único que queda por esclarecer es el contenido de la función generadora de tokens pseudoaleatorios, la función `RN`. El algoritmo de esta función se muestra en 3. Idealmente esta función debe regresar un elemento uniformemente aleatorio del espacio de tokens. La variable *cnt* mantiene un estado del algoritmo (mantiene su valor a lo largo de las distintas llamadas); el espacio de tokens contiene cadenas de longitud fija  $\mu$  de un alfabeto  $AL$  cuya cardinalidad es  $l$ ; el número de bits necesarios para enumerar a todo el alfabeto se guardan en  $\lambda = \lceil \log_2 l \rceil$ .

Existen varios candidatos viables para la función  $f$ : un cifrado de flujo, pues el flujo de llave de estos produce cadenas de aspecto aleatorio, o un cifrado por bloques con un modo de operación de contador. En la implementación de este trabajo se ocupa esta última opción.

Con la clasificación del PCI, este método cae, contradictoriamente, en los reversibles no criptográficos. Con la

**Algoritmo 3** Generación de tokens pseudoaleatorios en TKR

```

1: función TKR-RN( $k$ )
2:    $x \leftarrow F(k, cnt)$ 
3:    $x_1, x_2, \dots, x_m \leftarrow \text{CORTAR}(x, \lambda)$ 
4:    $t \leftarrow , i \leftarrow 0$ 
5:   mientras  $|t| \neq \mu$  hacer:
6:     si  $\text{ENTERO}(X_i) \leq l$  entonces:
7:        $t \leftarrow t \parallel \text{ENTERO}(X_i)$ 
8:     fin si
9:      $i \leftarrow i + 1$ 
10:  fin mientras
11:   $cnt \leftarrow cnt + 1$ 
12: fin función

```

clasificación propuesta en este trabajo se encuentra dentro de los criptográficos irreversibles.

**III-B. FFX (Format-preserving Feistel-based Encryption)**

Cifrado que preserva el formato presentado en [5] por Mihir Bellare, Phillip Rogaway y Terence Spies. En su forma más general, el algoritmo se compone de 9 parámetros que permiten cifrar cadenas de cualquier longitud en cualquier alfabeto; los autores también proponen dos formas más específicas (dos colecciones de parámetros) para alfabetos binarios y alfabetos decimales: A2 y A10, respectivamente. De aquí en adelante se hablará solamente de la colección A10.

FFX ocupa una red Feistel alternante junto con una adaptación de AES-CBC-MAC (usada como función de ronda) para lograr preservar el formato. La operación general del algoritmo se describe completamente por la operación de una red alternante:

$$\begin{aligned}
L_i &= \begin{cases} F_k(R_{i-1}) \oplus L_{i-1}, & \text{si } i \text{ es par} \\ L_{i-1}, & \text{si } i \text{ es impar} \end{cases} \\
R_i &= \begin{cases} R_{i-1}, & \text{si } i \text{ es par} \\ F_k(L_{i-1}) \oplus R_{i-1}, & \text{si } i \text{ es impar} \end{cases}
\end{aligned} \quad (1)$$

En la figura 4 se describe a la función de ronda. La idea general consiste en interpretar la salida de AES CBC MAC de forma que tenga el formato deseado. El valor de  $m$  corresponde al *split* en la ronda actual, esto es, la longitud de la cadena de entrada.

**Algoritmo 4** Función de ronda de FFX A10

```

1: función FFX-AES-CBC-MAC( $x, k, t$ )
2:    $a \leftarrow x \parallel t$ 
3:    $b \leftarrow \text{AES\_CBC\_MAC}(a, k)$ 
4:    $y' \leftarrow a[1..64], y'' \leftarrow a[65..128]$ 
5:   si  $m \leq 9$  entonces:
6:      $c \leftarrow y'' \text{ mód } 10^m$ 
7:   sino:
8:      $c \leftarrow (y' \text{ mód } 10^{m-9}) \times 10^9 + (y'' \text{ mód } 10^m)$ 
9:   fin si
10:  regresar  $c$ 
11: fin función

```

**Algoritmo** Cifrado  $BC_{F,s,b,w}(X,K,T)$ 

```

1.  $T_R \leftarrow T \text{ mód } 2^{32}$  y  $T_L \leftarrow (T - T_R)/2^{32}$ 
2.  $l \leftarrow \lceil b/2 \rceil$ 
3.  $r \leftarrow \lfloor b/2 \rfloor$ 
4.  $L_0 \leftarrow \sum_{j=0}^{l-1} X[j] \cdot s^j$ 
5.  $R_0 \leftarrow \sum_{j=0}^{r-1} X[j+l] \cdot s^j$ 
6. para  $i = 0$  hasta  $i = w - 1$ :
7.   si  $i$  es par:
8.      $L_{i+1} \leftarrow L_i \boxplus F_K((T_R \oplus i) \cdot 2^{f-32} + R_i) \pmod{s^l}$ 
9.      $R_{i+1} \leftarrow R_i$ 
10.  si  $i$  es impar:
11.     $R_{i+1} \leftarrow R_i \boxplus F_K((T_L \oplus i) \cdot 2^{f-32} + L_i) \pmod{s^r}$ 
12.     $L_{i+1} \leftarrow L_i$ 
13. para  $i = 0$  hasta  $i = l - 1$ :
14.    $Y_L[i] \leftarrow L_w \text{ mód } s$ 
15.    $L_w \leftarrow (L_w - Y_L[i])/s$ 
16. para  $i = l$  hasta  $i = r - 1$ :
17.    $Y_R[i] \leftarrow R_w \text{ mód } s$ 
18.    $R_w \leftarrow (R_w - Y_R[i])/s$ 
19.  $Y \leftarrow Y_L \parallel Y_R$ 

```

Figura 2. Cifrado interno BC.

Con la clasificación del PCI, este método cae en los reversibles criptográficos. Con la clasificación propuesta en este trabajo, se trata de un criptográfico reversible.

**III-C. BPS (Brier, Peyrin y Stern)**

Algoritmo de cifrado que preserva el formato capaz de cifrar cadenas formadas por cualquier conjunto de caracteres, descrito en [7] y cuyo nombre proviene de las iniciales de los apellidos de sus autores Eric Brier, Thomas Peyrin y Jacques Stern, aunque en el estándar [11], el NIST lo nombra como FF3.

BPS se conforma de 2 partes: un cifrado interno  $BC$  que se encarga de cifrar bloques de longitud fija, usando a su vez un cifrado por bloques  $F$ ; y un modo de operación especial, encargado de extender la funcionalidad de  $BC$  y permitir cifrar cadenas de un longitud de hasta  $max_b \cdot 2^{16}$  caracteres, donde  $max_b$  es la longitud máxima que puede tener una cadena para cifrarse con  $BC$ .

Este cifrado interno utiliza una red Feistel alternante y se define como  $BC_{F,s,b,w}(X,K,T)$ , donde:  $F$  es un cifrado por bloques con  $f$  bits de salida, como puede ser TDES o AES;  $s$  es la cardinalidad del alfabeto de la cadena a cifrar,  $b$  es su longitud,  $w$  es el número de rondas de la red Feistel,  $X$  es la cadena,  $K$  es una llave acorde al cifrado  $F$ , y  $T$  es un *tweak* de 64 bits.

El funcionamiento del cifrado  $BC$  es descrito en la figura 2.

Para cada bloque a cifrar, el cifrado  $BC$  debe instanciarse con una longitud de  $max_b = 2 \cdot \log_s(2^{f-32})$  caracteres, y cuando la longitud total del mensaje a cifrar no sea múltiplo de este valor, en el último bloque  $BC$  se tendrá que instanciar con una longitud igual a la de ese bloque.

El modo de operación de BPS es un variación de CBC, con la diferencia de que usa sumas modulares carácter por carácter en lugar de aplicar operaciones *xor*, además de que no emplea un vector de inicialización.

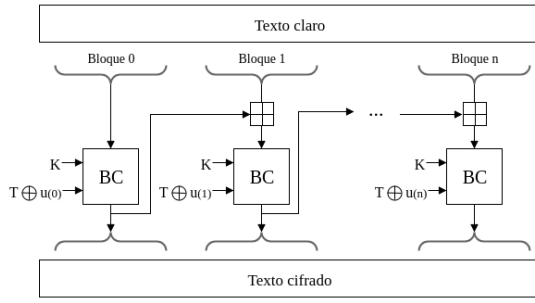


Figura 3. Modo de operación de BPS.

Otra característica de este modo de operación es que utiliza un contador  $u$  para aplicar un  $xor$  a los 16 bits más significativos de cada mitad del *tweak*  $T$  que utiliza BPS, por lo cual este se puede ver como una función  $u(n) = n \cdot (2^{16} + 2^{48})$ .

El funcionamiento del modo de operación se describe en la figura 3.

El PCI clasifica a este algoritmo dentro de los métodos de generación de tokens reversibles criptográficos, pero desde el punto de vista de este trabajo se tiene que es un método criptográfico reversible.

#### III-D. AHR (Algoritmo híbrido reversible)

En 2017, Longo, Aragona y Sala [1] propusieron un algoritmo híbrido reversible basado en un cifrador por bloques, una llave secreta y una entrada adicional. Las entradas del algoritmo son la parte del PAN a cifrar y una entrada adicional (por ejemplo, la fecha) que permite que se tengan varios tokens relacionados con la misma tarjeta.

El algoritmo necesita una función  $f$  pública que se encarga de poner el relleno en la entrada para obtener el bloque completo para el cifrador, pues, dada una cadena de longitud  $m$  regrese una de longitud  $n$ ; requiere también que solo se utilicen cifrados cuyo tamaño de bloque sea, mínimo, de 128 bits. Finalmente, como es un algoritmo reversible, se necesita una base de datos segura para almacenar los pares PAN-token.

Como se desea obtener un token que tenga el mismo número de dígitos que el PAN ingresado, se utiliza un método llamado *cycle-walking* para asegurarse de que el texto cifrado pertenezca al espacio del texto en claro.

A continuación se definen una serie de notaciones que se utilizarán en el algoritmo:

- $M$  Tamaño de bloque del cifrado por bloques que se usará.
- $l$  Longitud de la entrada. En este caso,  $13 \geq l \geq 19$ .
- $n$  Número de bits necesarios para representar a la entrada:  $n = \log_2(10^l)$ .
- $[y]_b^s$  Indica que  $y$  es menor que  $b^s$ :  $y < b^s$ .
- $\tilde{x}$  Representación de  $x$  en una cadena binaria cuando  $x$  es representado en su forma decimal y viceversa.

El primer paso es obtener el valor del bloque  $t$ ; es decir, concatenar los bytes más significativos de la salida de  $f(u, p)$  con la representación binaria de  $p$ . Después, se cifra el bloque

```

Algoritmo AHR( $p, u, k$ )
1.  $t = f(u, p) || [\tilde{p}]_b^s$ 
2.  $c = E(k, t)$ 
3. si  $(\tilde{c} \bmod 2^n) \geq 10^l$  entonces:
4.    $t = c$ 
5.   Regresar a 2.
6.  $token = [\tilde{c} \bmod 2^n]_{10}^l$ 
7. si  $comprobar(token) = \text{verdadero}$  entonces:
8.    $u = u + 1$ 
9.   Regresar al paso 1.
10. sino:
11.   regresar  $token$ 

```

Figura 4. Algoritmo híbrido reversible.

```

Algoritmo DRBG-tokenización( $n$ )
1.  $token \leftarrow$ 
2.  $cadena\_aleatoria \leftarrow drbg.generar(n)$ 
3. para  $i = 0$  hasta  $n - 1$ :
4.    $token[i] \leftarrow cadena\_aleatoria[i] \bmod 10$ 
5. regresar  $token$ 

```

Figura 5. Generación de tokens a partir de cadena binaria aleatoria.

$t$  con la llave  $K$  y se guarda en  $c$  la representación decimal de los últimos bits del bloque cifrado; aquí es donde se utiliza la caminata cíclica, pues si los dígitos de  $c$  son menores a los que le corresponden con  $p$ , se guarda  $t$  en  $c$  y se regresa al paso del cifrado. Finalmente, cuando se obtiene un token válido, se comprueba que no esté registrado en la base de datos (si lo está, se regresa al inicio, pero aumentando la entrada adicional  $u$  en 1) y se registra el nuevo par PAN-token. El pseudocódigo del algoritmo de tokenización se puede observar en 4.

#### III-E. Algoritmo basado en generador pseudoaleatorio

Probablemente este método es el más directo para generar tokens. La idea es producir una cadena binaria aleatoria con un DRBG e interpretarla para que tenga el formato de un token. El funcionamiento general es el mismo que en TKR (figura ??): la operación de tokenización primero verifica en la base de datos que el número de tarjeta no se encuentre ya asociado a un token; de ser este el caso, se regresa el token asociado previamente; en caso contrario se genera un nuevo token aleatorio, se guarda en la base de datos y se regresa. La detokenización es simplemente una consulta en la base de datos. En la figura 5 se muestra uno de los posibles métodos para generar un token a partir de una cadena binaria ( $n$  es la longitud que debe tener el token generado).

Según la clasificación del PCI, este método tendría que ser reversible no criptográfico; con la clasificación de este trabajo, se trata de un criptográfico irreversible. Al igual que con TKR, la denominación bajo la clasificación del PCI resulta contradictoria: los generadores pseudoaleatorios son posibles debido a herramientas criptográficas.

## IV. RESULTADOS DE COMPARACIONES DE DESEMPEÑO

Todos los resultados presentados en esta sección se llevaron a cabo en una computadora con las siguientes características:

- **Procesador:** Intel i5-7200U (2.5 GHz) de 4 núcleos.

Tabla I  
COMPARACIÓN DE TIEMPOS DE TOKENIZACIÓN.

Algoritmo	Tokenización ( $\mu$ s)	Detokenización ( $\mu$ s)
FFX	80	59
BPS	172	88
TKR	3379	885
AHR	3223	311
DRBG	2934	256

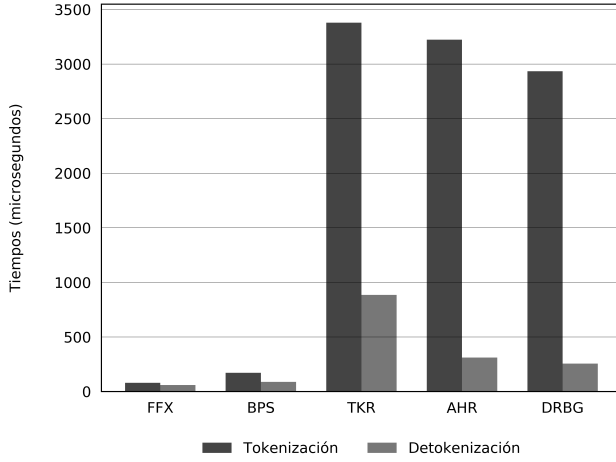


Figura 6. Comparación de tiempos de tokenización.

- **Sistema operativo:** Arch Linux, kernel 4.17.
- **Base de datos:** MariaDB 10.1.
- **Compilador:** GCC 8.1.1

En la tabla I y la figura 6 se muestran los resultados en tiempo de las ejecuciones de los algoritmos presentados en secciones anteriores.

## V. CONCLUSIONES

### REFERENCIAS

- [1] R. Aragona, R. Longo, and M. Sala. Several proofs of security for a tokenization algorithm. *Appl. Algebra Eng. Commun. Comput.*, 28(5):425–436, 2017.
- [2] U. C. Association. What is pci dss?, 2018.
- [3] E. Barker and J. Kelsey. Nist special publication 800-90a - recommendation for random number generation using deterministic random bit generators, 2015.
- [4] M. Bellare, T. Ristenpart, P. Rogaway, and T. Stegers. Format-preserving encryption. In M. J. J. Jr., V. Rijmen, and R. Safavi-Naini, editors, *Selected Areas in Cryptography, 16th Annual International Workshop, SAC 2009, Calgary, Alberta, Canada, August 13-14, 2009, Revised Selected Papers*, volume 5867 of *Lecture Notes in Computer Science*, pages 295–312. Springer, 2009.
- [5] M. Bellare, P. Rogaway, and T. Spies. The ffx mode of operation for format-preserving encryption. 2009.
- [6] Braintree. Tokenization secures cc data and meet pci compliance requirements, 2007.
- [7] E. Brier, T. Peyrin, and J. Stern. Bps: a format-preserving encryption proposal. 2010.
- [8] P. C. I. S. S. Council. Tokenization product security guidelines – irreversible and reversible tokens, 2015.
- [9] P. C. I. S. S. Council. Data security standard - version 3.2, 2016.

- [10] S. Diaz-Santiago, L. M. Rodríguez-Henríquez, and D. Chakraborty. A cryptographic study of tokenization systems. *Int. J. Inf. Sec.*, 15(4):413–432, 2016.
- [11] M. Dworkin. Nist special publication 800-38g - recommendation for block cipher modes of operation: Methods for format-preserving encryption, 2016.
- [12] I. O. for Standarization. *ISO/IEC 7812*. 5 edition, 2017.
- [13] J. S. Kiernan. Credit card and debit card fraud statistics, 2017.
- [14] A. Menezes, P. C. van Oorschot, and S. A. Vanstone. *Handbook of Applied Cryptography*. CRC Press, 1996.
- [15] S. Payments. The history of truetokenization, 2018.
- [16] S. Payments. Tried and true tokenization: The original tokenization solution for card data security, 2018.
- [17] Securosis. Understanding and selecting a tokenization solution, 2018.
- [18] S. Staff. The history of the pci dss standard: A visual timeline, 2013.