Fingerprinting automático de contenidos digitales inspirado en las secuencias de ADN

David Megías
Universitat Oberta de Catalunya,
Internet Interdisciplinary Institute (IN3),
Estudis d'Informàtica, Multimèdia i Telecomunicació,
Rambla del Poblenou 156,
08018 Barcelona
Email: dmegias@uoc.edu

Josep Domingo-Ferrer
Universitat Rovira i Virgili,
Department d'Enginyeria Informàtica i Matemàtiques,
Cátedra UNESCO de Privacidad de Datos,
Av. Països Catalans 26,
43007 Tarragona
Email josep.domingo@urv.cat

***Resumen*—La distribución *multicast* de contenidos no es adecuada para el comercio electrónico, dado que produce exactamente la misma copia del contenido, de manera que los culpables de una distribución ilegal no pueden ser identificados. Por otro lado, la distribución *unicast* requiere una conexión para cada comprador, pero permite incrustar un número de serie diferente para cada usuario, lo que permite identificar a un distribuidor ilegal. La distribución por pares (P2P) proporciona una tercera opción que puede combinar algunas de las ventajas del *multicast* y el *unicast*: por un lado, el vendedor sólo necesita establecer conexiones *unicast* con unos pocos compradores-semilla, quienes se encargan de las futuras distribuciones del contenido; por otro lado, si se utiliza un mecanismo apropiado de fingerprinting, los distribuidores ilegales todavía pueden ser identificados. En este artículo se propone un esquema de fingerprinting inspirado en las secuencias de ADN que permite identificar a los redistribuidores, al mismo tiempo que se preserva el anonimato de la mayoría de los compradores honestos.**

I. Introducción

El fingerprinting de contenidos digitales [2] es una técnica viable para la protección de los derechos de autor en la compra-venta de estos contenidos a través de Internet. Básicamente, las técnicas de fingerprinting consisten en incrustar una marca imperceptible en el contenido adquirido (que puede ser audio, fotografías o vídeo) que identifique al comprador. En este caso, la marca incrustada es diferente para cada comprador, si bien el contenido debe ser perceptualmente idéntico en todos los casos. En caso de que el comprador distribuya el contenido de manera ilegal, la marca incrustada permitirá su posterior identificación y podrá acarrear acciones legales en su contra. Además, los esquemas de fingerprinting se clasifican en tres tipos [3]: simétricos, asimétricos y anónimos. En el primer tipo, es el vendedor quien incrusta la marca en el contenido, de manera que el comprador no puede ser acusado formalmente de distribución ilegal, puesto que el mismo vendedor podría haber ejercido este comportamiento indeseado. En el caso del fingerprinting asimétrico el vendedor no tiene acceso a la copia marcada, pero sí se puede obtener la marca en caso de distribución ilegal, permitiendo la identificación del comprador malicioso. En el tercer y último tipo, además de la asimetría, el comprador conserva su anonimato, y por tanto no se le puede vincular a la compra de un contenido concreto, a no ser que participe en una distribución ilegal.

Básicamente, todos los esquemas de fingerprinting propuestos hasta la actualidad son centralizados. Es decir, la distribución del contenido se realiza de forma *unicast*, desde el vendedor a cada comprador, lo que encarece el proceso considerablemente, tanto en forma de tiempo de cómputo como de ancho de banda de comunicaciones. Esta cuestión se agrava con el uso de esquemas asimétricos y anónimos, puesto que los protocolos para garantizar la asimetría y el anonimato suelen ser muy costosos y requerir muchos recursos, ya sea en forma te tiempo de CPU, de comunicaciones o de almacenamiento. Idealmente, una distribución en forma de difusión (*multicast*) sería más atractiva, dado que una emisión se distribuye de manera simultánea a un grupo de receptores, evitándose así el establecimiento de una distribución individual para cada receptor. No obstante, la distribución en difusión no permite enviar copias diferentes a cada usuario, tal y como exigen los esquemas de fingerprinting. Parece, por lo tanto, que la distribución *unicast* es la más adecuada para los esquemas de fingerprinting pese a los elevados costes que implica. Estos costes acabarían repercutiendo en los compradores para que el vendedor pueda mantener su margen de beneficio.

Una alternativa intermedia son los sistemas de distribución por pares (P2P) en que los receptores de un contenido se convierten en distribuidores para otros. Este modelo puede verse como una mezcla de las características de los sistemas *unicast* y *multicast*. La distribución P2P de todo tipo de contenidos se ha popularizado en los últimos años con el incremento del ancho de banda de las comunicaciones domésticas. BitTorrent [5], Kademlia [12] o eDonkey2000 [11] son algunos ejemplos de protocolos P2P para el intercambio de archivos a nivel privado. Cabe destacar que la distribución P2P no se limita a usuarios privados en el ámbito doméstico,

Este trabajo está financiado parcialmente por el Ministerio de Economía y Competitividad a través de los proyectos TSI200765406-C03-01/03 "E-AEGIS", TIN2011-27076-C03-01/02 "CO-PRIVACY" y CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES"; y por la Generalitat de Catalunya a través de la subvención 2009 SGR 1135. El segundo autor está financiado parcialmente a través de un premio ICREA-Acadèmia de la Generalitat de Catalunya y dirige la Cátedra UNESCO de Privacidad de Datos, pero los puntos de vista expresados en este artículo no comprometen a UNESCO.

sino que algunas compañías están empezando a utilizar este sistema para las descargas de sus productos (por ejemplo [15]). El uso de un esquema de distribución P2P permite al vendedor establecer sólo un pequeño número de conexiones *unicast* con un conjunto de M compradores-"semilla" que se convierten en nuevas fuentes del contenido para otros compradores. El contenido puede, finalmente, alcanzar un conjunto N de compradores con $M \ll N$. La cuestión que aborda este artículo es cómo llevar a cabo esta compra-venta en un entorno P2P tal que los redistribuidores ilegales sigan siendo identificables.

El artículo propone un esquema de distribución P2P, que requeriría un software específico para esta aplicación y no uno genérico, en que el vendedor crea únicamente un conjunto de M copias semilla del contenido y las hace llegar a M compradores semilla. Todas las copias subsiguientes se generan a partir de las M copias semilla. La copia obtenida por cada comprador es una combinación de las copias proporcionadas por sus fuentes (progenitores). El fingerprint de cada comprador se construye como una secuencia binaria formada por combinación de las secuencias de sus fuentes, de manera análoga a como las secuencias de ADN de los seres vivos se construyen combinando las secuencias de ADN de sus progenitores. El esquema propuesto, que conlleva ahorro en ancho de banda y tiempo de cómputo para el vendedor, todavía permite localizar a los distribuidores ilegales usando un algoritmo de localización del fingerprint similar a los tests de ADN de las investigaciones forenses.

Por otro lado, en el esquema propuesto los fingerprints de los compradores no se registran de ninguna forma, de manera que los compradores preservan su privacidad (anonimato) a no ser que se produzca una redistribución ilegal, en cuyo caso el fingerprint de la copia distribuida debe ser vinculado a un usuario particular. Para ello, es necesario que el software de distribución P2P guarde un registro de cada transacción entre compradores; alternativamente, se puede crear un registro externo de transacciones.

El algoritmo de localización de los distribuidores ilegales se inicia con un "test de ADN" del fingerprint extraído de la copia distribuida ilegalmente con los fingerprints de los M compradores semilla. Un simple test de correlación indica cuál de los compradores semilla es el antepasado más probable del distribuidor ilegal, lo que permite seguir buscando entre los "hijos" de éste. De esta manera se va realizando una búsqueda a través del grafo de distribución de contenidos hasta que se identifique al distribuidor ilegal o algún comprador rechace participar en el test de ADN, en cuyo caso será acusado de haber sido la fuente de la distribución. Si el test de ADN se realiza a través de un esquema de computación segura multipartito [4], [7], el fingerprint exacto de los compradores honestos no necesita ser revelado, aunque su anonimato no será garantizado en el sentido que deberán colaborar en la cadena de tests de ADN y, por lo tanto, desvelar que han adquirido ese contenido en particular. No obstante, las simulaciones que presentamos en el artículo señalan que el porcentaje de compradores implicados en los tests de ADN no sólo es bajo, sino que decrece monótonamente a medida que aumenta el número de compradores.

Finalmente, cabe destacar que el esquema propuesto es intrínsecamente asimétrico y anónimo, en el sentido que el vendedor no tiene acceso a ninguna copia marcada del contenido salvo las de los M compradores semilla, quienes podrían ser simplemente compradores ficticios, es decir, diferentes réplicas del contenido colocadas en diferentes nodos de Internet por el propio vendedor para iniciar el proceso.

El resto del artículo se estructura de la manera siguiente. La sección II introduce el concepto de los fingerprints automáticos inspirados en las secuencias de ADN. En la sección III se presenta el esquema de distribución propuesto. La sección IV presenta el algoritmo para localizar a los redistribuidores ilegales del contenido. En la sección V se presentan algunos resultados de simulación y la sección VI finaliza el artículo con las conclusiones más relevantes del trabajo realizado.

II. Fingerprints inspirados en el ADN

Esta sección presenta el concepto de fingerprints binarios análogos a las secuencias de ADN y generados automáticamente. Los términos que se usan a lo largo de este artículo están basados en los que se usan en genética para las cadenas de ADN y la herencia. Las definiciones de estos términos en el contexto de este artículo se introducen a continuación.

Secuencia de ADN: en la naturaleza, el ADN es una molécula formada por una secuencia ordenada de nucleótidos, donde cada nucleótido es una de las siguientes cuatro moléculas más pequeñas: adenina (A), citosina (C), guanina (G) y timina (T). Aunque la molécula de ADN está formada por dos cadenas entrelazadas, los nucléotidos están siempre emparejados A-T y C-G en las dos cadenas, de manera que la estructura del ADN es redundante. El ADN es una molécula en forma de doble hélice en que cada cadena es la complementaria de la otra. Por lo tanto, una de las cadenas contiene toda la información necesaria para construir la otra.

En este artículo, los fingerprints se construyen formando una secuencia binaria cuyos bits pueden ser considerados como los equivalentes de los nucleótidos en las secuencias de ADN reales. Aunque en las secuencias de ADN reales los nucleótidos pueden interpretarse como símbolos de dos bits (dado que existen cuatro nucleótidos diferentes), la analogía todavía es perfectamente válida con nucleótidos de un solo bit.

Gen: un segmento de ADN que codifica una determinada proteína –y por tanto tiene cierto impacto en la herencia y en la bioquímica del ser vivo– se denomina gen. De la misma forma, un segmento del fingerprint inspirado en ADN se denomina "gen" en este artículo. Pese a que los genes reales tienen longitudes diferentes, en este artículo se utilizan genes de longitud fija por simplicidad y sin pérdida de generalidad.

Por otro lado, en la naturaleza no todos los segmentos de la secuencia de ADN forman parte de genes. Sin embargo, en este artículo, todos los "nucleótidos" (bits) de los fingerprints pertenecen a alguno de los genes de la secuencia.

Apareamiento y herencia: en la naturaleza, los genes de un ser vivo son básicamente una combinación de los genes de sus

progenitores (aunque algunos procesos como la mutación y el cruce pueden producir fragmentos de ADN que son diferentes en un descendiente respecto a ambos progenitores).

Análogamente, en este artículo, cuando un comprador obtiene su copia de un contenido distribuido en la red P2P usando algún software concreto, el fingerprint de su copia será una combinación de los genes de las fuentes del contenido (que se denominan "padres" o "progenitores" por seguir con la analogía biológica). En este caso, el número de padres de un contenido no tiene que ser exactamente dos como en el mundo natural. Por lo tanto, el proceso de apareamiento en el escenario propuesto debe ser entendido de manera generalizada y no limitada a dos progenitores.

En esta propuesta, los fingerprints pueden entenderse como "generados automáticamente" a partir de los fingerprints de los progenitores. A pesar de esta "generación automática de los fingerprints", las secuencias obtenidas todavía son válidas para identificar a los distribuidores ilegales, de igual modo a como los restos de ADN pueden usarse en investigaciones criminales para identificar al sospechoso de un delito.

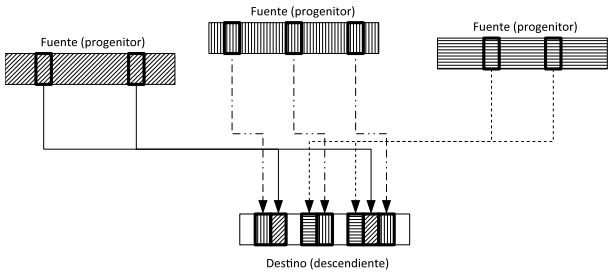

Figura 1: Subida/bajada del contenido (apareamiento)

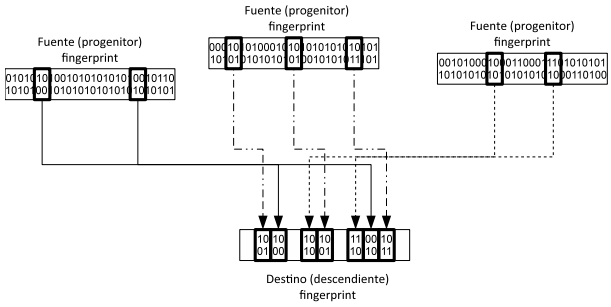

Figura 2: Construcción automática de los fingerprints

Mutación y cruce: Diferentes tipos de alteraciones pueden afectar a las moléculas de ADN provocando la modificación de algunos de los nucleótidos de sus secuencias. Estos cambios pueden afectar a un único nucleótido o a un segmento de la secuencia de ADN. Básicamente, el cruce ocurre cuando las dos cadenas complementarias de ADN se recambian durante la replicación del ADN y la mutación se refiere a diferentes errores de tipo aleatorio durante la replicación.

La mutación y el cruce proporcionan mecanismos que permiten a la secuencia de ADN de un descendiente incluir genes que son diferentes de los de sus progenitores. Si se permite que un comprador pueda obtener su copia de un único progenitor, la mutación (y el cruce) deberá usarse para producir una versión diferente del fingerprint, como se requiere en este tipo de aplicaciones (dos compradores deben tener siempre fingerprints diferentes). Nótese que, aunque los fingerprints se han definido como una única cadena binaria, puede pensarse que existe un complementario implícito que es su negación y usarse éste para simular cruces.

Test de relación de ADN: en las investigaciones del mundo real, los tests de relación de ADN se realizan para probar o descartar un parentesco entre dos o más individuos. Teniendo en cuenta las propiedades del apareamiento y la herencia, los parientes de sangre comparten segmentos más largos de sus secuencias de ADN en comparación con los que no lo son.

El equivalente a estos tests de relación de ADN en el esquema de fingerprinting propuesto es una función para calcular la correlación de una secuencia binaria con otra. Dado que los antepasados y los descendientes del escenario de fingerprinting distribuido comparten varios genes, existe una correlación medible entre sus fingerprints binarios.

III. Distribución P2P de contenidos con fingerprints inspirados en ADN

La base de la distribución P2P es que los contenidos se distribuyen desde unos usuarios a otros. Al recibir fragmentos del contenido, los usuarios de destino se convierten en fuentes para otros. De esta forma, un archivo se obtiene al acoplar los fragmentos obtenidos de diferentes fuentes. Este proceso de subida y bajada de archivos de diferentes fuentes se muestra en la Figura 1. En esta figura, el destino obtiene fragmentos de tres fuentes diferentes que se acoplan para formar el contenido.

III-A. Requerimientos del algoritmo de incrustación

Para extraer los fingerprints descritos en la sección II en un contenido distribuido usando un esquema P2P, es necesario disponer de un esquema de incrustación para insertar el fingerprint de los M compradores semilla. En principio es suficiente con incrustar fingerprints generados aleatoriamente para los compradores semilla siempre que la correlación entre los M fingerprints sea pequeña.

El esquema de incrustación debe satisfacer las siguientes condiciones. Primero, la marca incrustada debe ser una secuencia binaria distribuida a lo largo de todo el contenido (archivo). Además, el fingerprint debe quedar separado en piezas que se incrustarán en diferentes bloques (o fragmentos) que serán distribuidos por el software P2P. Cada una de estas piezas (de longitud fija) contendrá un gen completo del fingerprint. Por ejemplo, si el software P2P fragmenta los contenidos en bloques de 32 kilobytes (kB), cada gen del fingerprint deberá estar incrustado en uno de estos fragmentos y la extracción del fingerprint debe ser robusta contra este tipo de fragmentación en unidades de 32 kB, siempre que el principio y el final de cada fragmento se respeten. Este proceso se ilustra en la Figura 2. Es importante puntualizar que esto no siempre es posible con sistemas de incrustación no orientados a bloques. Un ejemplo de esquema de incrustación

orientado a bloques que podría usarse en este escenario se presenta en [13]. Segundo, incluso si las diferentes copias de los contenidos obtenidos por diferentes compradores no son idénticas bit a bit (dado que el fingerprint incrustado será diferente para cada uno de ellos), estas versiones diferentes deben ser idénticas "perceptualmente", porque los contenidos deben tener la misma calidad para todos los compradores. Por tanto, si el esquema de distribución P2P utiliza una función hash para la indexación de los contenidos, será necesario aplicar un hash perceptual [6] que proporcione valores idénticos para las diferentes versiones del contenido. Una función hash estándar que produzca valores diferentes aunque se modifique un único bit del archivo no será útil en una aplicación de este tipo.

Si se cumplen estas dos condiciones, el fingerprinting ocurre de manera automática a medida que los compradores obtienen sus contenidos de diferentes fuentes. No hay ningún sobrecoste adicional para incrustar la marca.

El fingerprinting automático descrito de esta forma requiere que haya más de una fuente para cada comprador. En caso de que exista una única fuente, el fingerprint sería idéntico para la fuente y el nuevo comprador. Esta limitación se podría resolver usando la mutación y el cruce. Sin embargo, si se usan mutación y cruce, el fingerprinting de la nueva copia ya no sería automático, puesto que habría que modificar varios (bits) genes, cosa que implicaría reincrustar información.

III-B. Coprivacidad en la relación progenitor-descendiente

Si se puede forzar a que haya al menos dos progenitores para cada comprador, esta es la solución más simple y efectiva, puesto que, en este caso, el fingerprinting sería automático y no haría falta incrustar nueva información en las transacciones entre compradores. Afortunadamente, el interés egoísta de un comprador hijo es el de obtener su archivo de más de una fuente y el interés egoísta de cada comprador progenitor es el de no distribuir el contenido completo a un comprador hijo.

Si un hijo obtiene su contenido de un único progenitor, en caso de no usar mutación/cruce, su fingerprint será el mismo que el del progenitor. En caso de que el progenitor distribuya el contenido ilegalmente, el descendiente se arriesga a ser injustamente acusado de redistribución. Obtener el contenido de más de un progenitor es una solución más simple y automática de evitar este riesgo que la mutación/cruce. Del mismo modo, si un progenitor envía todo el contenido a un único hijo, este descendiente heredará el fingerprint completo del progenitor. Por lo tanto, si el descendiente distribuyese el contenido ilegalmente, el progenitor se arriesgaría a ser acusado injustamente de redistribución. Dividir el contenido entre diferentes hijos es la mejor opción para un progenitor.

Esta situación en la que la mejor estrategia para preservar la propia privacidad consiste en actuar de tal manera que se protege la privacidad de los demás se denomina coprivacidad [9], [10]. Expresado en términos de la teoría de juegos, el vector de estrategias $(\text{múltiples hijos}, \text{múltiples padres})$ es un equilibrio de Nash entre progenitores y descendientes.

La propiedad de coprivacidad asegura que siempre que un comprador hijo pueda obtener el contenido de más de un padre, lo hará. También asegura que un padre no estará interesado en pasar el contenido completo a un único descendiente. Esta última condición se puede forzar fácilmente en el software de distribución P2P. Por ejemplo, si un progenitor envía el contenido a un descendiente, se puede cerrar la conexión tan pronto como se supere un cierto umbral (por ejemplo un 50 o un 60 %) de contenido enviado. El software también puede bloquear cualquier intento posterior de establecer una conexión del mismo hijo con el mismo padre para el mismo contenido.

III-C. Protocolo de distribución P2P

Para arrancar el sistema, el vendedor debe producir unas pocas semillas del contenido marcado. La aproximación propuesta pasa porque el vendedor produzca un pequeño número M de versiones del contenido con fingerprints pseudoleatorios diferentes usando algún esquema de incrustación que satisfaga las condiciones indicadas en las secciones anteriores. Estos M compradores podrían ser reales o ficticios, y los compradores de la segunda generación contactarían con ellos para obtener nuevas copias del contenido. El propio vendedor o algún tipo de autoridad confiable almacenarían la asociación de estos primeros M fingerprints con las identidades (o quizá algún pseudónimo) de los primeros M compradores. Una vez el sistema se haya arrancado, cualquier transacción futura ocurrirá sin ninguna necesidad de ejecutar el esquema de incrustación de la marca (suponiendo que no se usarán ni la mutación ni el cruce). Además, todos los fingerprints desde el comprador $M+1$ hasta el último (N) son completamente anónimos (conocidos únicamente por el propio comprador) y no guardan ningún tipo de relación con las identidades de los compradores. Cabe notar que este método para conseguir el fingerprinting anónimo es mucho más simple que los métodos de fingerprinting anónimo propuestos en la literatura [14], [3], [1], [8], que requieren el uso de algún tipo de protocolo criptográfico en *cada* transacción. Solamente el software de distribución P2P guardará un registro de las transacciones realizadas en caso de que sea necesario utilizarlo para tests de relación de ADN futuros.

Algoritmo 1 (Distribución P2P):

1. Para $i := 1$ hasta M, el vendedor genera la copia i-ésima del contenido con un fingerprint aleatorio incrustado.
2. Para $i := M+1$ hasta N, el comprador i-ésimo obtiene su copia del contenido acoplando los fragmentos obtenidos de un conjunto S_i de progenitores tal que $S_i \subseteq \{B_1, \cdots, B_{i-1}\}$ y $|S_i| > 1$, donde $|\cdot|$ es la cardinalidad de un conjunto y B_j hace referencia al j-ésimo comprador.

IV. Identificación de culpables de redistribución

Esta sección muestra como el método de fingerprinting propuesto puede usarse para identificar a los culpables de redistribución ilegal. Suponiendo que el esquema de incrustación es robusto y suficientemente seguro para que los usuarios maliciosos no puedan eliminar fácilmente sus fingerprints sin dañar el contenido haciéndolo inservible (que es la hipótesis habitual de los esquemas de marcado, [2]), el método siguiente

puede usarse por parte de una autoridad para identificar la fuente de una copia distribuida ilegalmente:

Algoritmo 2 (Identificación):

1. El fingerprint f de la copia distribuida ilegalmente X_f se obtiene usando el sistema de extracción de la marca.
2. El conjunto de test inicial T_0 se construye con los M compradores de las versiones semilla del archivo.
3. Sea $i := 0$.
4. Se realiza un test de relación de ADN con cada fingerprint de los compradores del conjunto actual T_i. Este paso se realiza calculando una simple correlación entre cadenas binarias. Dado el fingerprint f a ser localizado y el de test f' extraído de la copia $X_{f'}$ correspondiente a un comprador en T_i, ambas de longitud L, la correlación $C(f, f')$ entre f y f' puede calcularse como sigue:

$$C(f, f') = \frac{1}{L} \sum_{j=1}^{L} (-1)^{f_j \oplus f'_j}, \tag{1}$$

donde f_j y f'_j son, respectivamente, los bits j-ésimos de f y f', y $\oplus$ hace referencia a la suma exclusiva (XOR). Evidentemente, proporcionar X_f al comprador y esperar que éste retorne el valor correcto de $C(f, f')$ tiene la debilidad de que un comprador malicioso podría retornar una correlación muy baja. Una mejor opción consiste en que el comprador proporcione su contenido marcado $X_{f'}$ a una autoridad, sin que el comprador tenga acceso a X_f. Cabe notar que, si un comprador culpable intenta alterar aleatoriamente $X_{f'}$, esto convertirá al contenido en inservible (por la hipótesis de marcado de [2]), y la alteración será detectada. Si el comprador teme ser acusado injustamente por la autoridad, podría preferir no revelar $X_{f'}$; en este caso, el comprador y la autoridad podrían usar un esquema multipartito de computación segura para el cálculo de la expresión (1).

5. Existen tres posibles resultados del paso anterior:
   a) Uno o más compradores en T_i rechazan colaborar con la autoridad en el cálculo de sus correlaciones con f. En ese caso, los compradores que lo rechacen será acusados de redistribución ilegal o de incumplimiento de contrato.
   b) Uno de los compradores en T_i proporciona una correlación $C(f, f') = 1$. En ese caso, este comprador es acusado de ser el culpable de la distribución.
   c) En cualquier otro caso, el comprador de T_i que proporcione la máxima correlación con f será escogido como el antepasado más probable del comprador de la copia distribuida ilegalmente. En este caso, se construye un nuevo conjunto T_{i+1} de compradores con los descendientes (hijos) de este comprador, excluyendo todos aquellos para los cuales su correlación ya haya sido analizada (recordemos que un comprador tendrá varios progenitores). Estos hijos pueden ser obtenidos consultando el registro del software de distribución P2P usado por el antepasado más probable del comprador de la copia distribuida ilegalmente. Una vez el construido el nuevo conjunto T_{i+1}, se reasigna $i := i + 1$ y se regresa al paso 4 de este algoritmo.

Nótese que los compradores que quieran usar la plataforma deberán firmar un acuerdo legal que especificará las condiciones de uso de la misma y, por lo tanto, se comprometerán a participar en los tests de relación de ADN en caso de ser requeridos para ello.

Aunque escoger la máxima correlación funcionará en la mayoría de las ocasiones, no se puede descartar que, eventualmente, se obtenga una correlación mayor para un comprador que no sea un antepasado del comprador de la copia distribuida ilegalmente. Por ejemplo, un descendiente del distribuidor ilegal podría tener, como otro antepasado, un nodo del grafo que sí sea antepasado del distribuidor ilegal. Esto produciría una correlación alta, pero la cadena desde el comprador seleccionado no conduciría al distribuidor ilegal. En esta situación, sería necesario aplicar *backtracking* en el algoritmo de localización descrito. Una subred del grafo sería analizada hasta que todos los nodos del subgrafo sin descendencia hayan sido examinados. Cuando toda una subred haya sido explorada, el elemento de T_i con la segunda máxima correlación sería escogido como candidato a antepasado del distribuidor ilegal. Cuando todos los elementos de un conjunto T_i hayan sido examinados sin éxito (es decir, sin haber podido acusar a ningún comprador), el procedimiento volvería al conjunto T_{i-1}. Cabe destacar que el *backtracking* se produce en un número muy reducido de las simulaciones realizadas.

V. Resultados de simulación

Esta sección presenta un conjunto de experimentos de simulación para ilustrar las propiedades del sistema propuesto. En todas las simulaciones, se han usado fingerprints de 4096 bits divididos en 128 genes de 32 bits cada uno.

Cuadro I: Número medio y porcentaje de tests de relación de ADN a compradores no semilla

Gener.	Pobl.	Núm. medio de tests ADN		Backtrack.
		1 simulación	100 simulaciones	(100 sim.)
$k = 2$	$N = 20$	3,40 (34,0 %)	3,71 (37,1 %)	0 %
$k = 3$	$N = 40$	6,93 (23,1 %)	7,29 (24,3 %)	0 %
$k = 4$	$N = 80$	12,26 (17,5 %)	11,69 (16,7 %)	0,6 %
$k = 5$	$N = 160$	18,99 (12,7 %)	17,05 (11,4 %)	1,2 %
$k = 6$	$N = 320$	24,31 (7,8 %)	23,76 (7,7 %)	2,7 %

Las simulaciones consisten en producir varias generaciones de compradores usando un crecimiento exponencial. En cada generación se producen tantos compradores como la población total hasta ese momento. La primera generación tiene M compradores (las semillas); la segunda, otros M; la tercera $2M$, y así sucesivamente. El número de semillas escogido es $M = 10$ y cada comprador obtiene su copia de entre dos y cuatro progenitores diferentes. De esta forma, el número medio de progenitores para los nodos no semilla es de tres. Los resultados se muestra en el Cuadro I y en la Figura 3. Con esta configuración, la probabilidad de que dos compradores tuviesen el mismo fingerprint sería de 10^{-128}.

Los resultados del Cuadro I muestran los valores para una simulación y la media para 100 simulaciones usando 100 valores diferentes para la semilla del generador de números pseudoaleatorios para reducir los posibles sesgos de una sola simulación. Se puede comprobar que no hay diferencias importantes entre 1 y 100 simulaciones. La última columna representa el número medio de compradores que requieren *backtracking* en las 100 simulaciones. Como era de esperar, a medida que la red crece de tamaño, más compradores necesitan *backtracking*, pero su porcentaje es siempre pequeño.

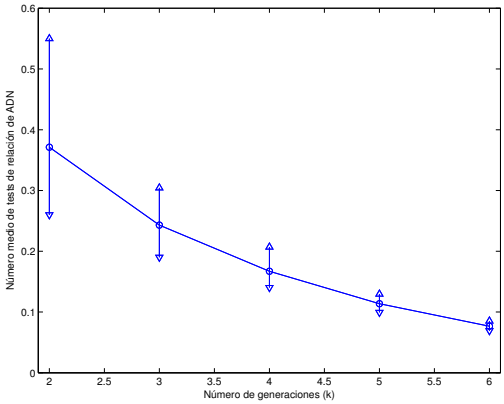

Figura 3: Fracción media de compradores no semilla afectados por tests de relación de ADN. Las líneas verticales representan intervalos max-min.

La Figura 3 muestra los intervalos para el valor medio de la fracción de compradores no semilla involucrados en tests de relación de ADN a medida que el número de generaciones aumenta. Para cada número de generaciones, la línea vertical representa un intervalo, con el triángulo de arriba mostrando la máxima fracción y el triángulo de abajo indicando la mínima fracción. Estos valores medios de las fracciones corresponden a los porcentajes indicados en el Cuadro I para las 100 simulaciones. Puede comprobarse que la fracción de compradores no semilla afectados por un test de relación de ADN disminuye a medida que el número de generaciones (y por tanto la población de compradores) crece: cuantos más compradores hay en el sistema, mayor es la probabilidad de permanecer completamente anónimos y no tener que participar en un test de relación de ADN. Sin embargo, a medida que la población crece, también podría crecer el número de redistribuciones ilegales, por lo que se requeriría de más tests de ADN para investigarlos, lo que también podría implicar una mayor probabilidad para un comprador no semilla de participar en un test de relación de ADN y perder su anonimato. Estos resultados de simulación se han ratificado mediante un análisis teórico que no se incluye por motivos de espacio.

VI. Conclusiones

Se ha presentado un esquema de fingerprinting inspirado en las secuencias de ADN y diseñado para funcionar en un esquema de distribución P2P. El esquema propuesto permite al vendedor localizar los culpables de una redistribución ilegal del contenido. Además, el vendedor sólo conoce, como mucho, los fingerprints de los compradores semilla, pero no los del resto de compradores (la gran mayoría). De hecho, el vendedor desconoce incluso las identidades de los compradores no semilla. Cuando se debe localizar al culpable de una distribución ilegal, sólo una pequeña fracción de compradores honestos deben renunciar a su anonimato y proporcionar sus copias, de manera que el esquema puede calificarse como cuasi-anónimo.

Como trabajo futuro se plantea la necesidad de dotar al sistema de resistencia frente a la confabulación de diversos compradores maliciosos que puedan juntarse para eliminar sus fingerprints. También puede trabajarse en mejorar el algoritmo de localización, intentando reducir los casos de *backtracking* y el número de tests de relación de ADN que afecten a compradores honestos.

Referencias

[1] Y. Bo, L. Piyuan, and Z. Wenzheng. An efficient anonymous fingerprinting protocol. In *Computational Intelligence and Security*, LNCS 4456, Springer, pp. 824-832, 2007.

[2] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. In *Advances in Cryptology-CRYPTO'95*, LNCS 963, Springer, pp. 452-465, 1995.

[3] J. Camenisch. Efficient anonymous fingerprinting with group signatures. In *Asiacrypt 2000*, LNCS 1976, Springer, pp. 415-428, 2000.

[4] D. Chaum, I. Damgård, and J. van de Graaf. Multiparty computations ensuring privacy of each party's input and correctness of the result. In *Advances in Cryptology-CRYPTO'87*, LNCS 293, Springer, pp. 87-119, 1988.

[5] B. Cohen. The BitTorrent Protocol Specification. 2008. Available at http://www.bittorrent.org/beps/bep_0003.html.

[6] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography*. Burlington MA: Morgan Kaufmann, 2008.

[7] I. Damgård, Y. Ishai, and M. Krøigaard. Perfectly secure multiparty computation and the computational overhead of cryptography. In *EUROCRYPT 2010*, LNCS 6110, Springer, pp. 445-465, 2010.

[8] J. Domingo-Ferrer. Anonymous fingerprinting based on committed oblivious transfer. In *Public Key Cryptography-PKC 1999*, LNCS 1560, Springer, pp. 43-52, 1999.

[9] J. Domingo-Ferrer. Coprivacy: towards a theory of sustainable privacy. In *Privacy in Statistical Databases-PSD 2010*, LNCS 6344, Springer, pp. 258-268, 2010.

[10] J. Domingo-Ferrer. Coprivacy: an introduction to the theory and applications of co-operative privacy. *SORT-Statistics and Operations Research Transactions*, vol. 35, special issue: Privacy in statistical databases, pp. 25-40, 2011.

[11] O. Heckmann and A. Bock. The eDonkey 2000 Protocol. KOM Technical Report 08/2002, Ver. 0.8. Department of Electrical Engineering & Information Technology & Department of Computer Science. Darmstadt University of Technology (Germany). 2002.

[12] P. Maymounkov and D. Mazières. Kademlia: a peer-to-peer information system based on the XOR metric. In *IPTPS 2002-First International Workshop on Peer-to-Peer Systems*, LNCS 2429, Springer, pp. 43-65, 2002.

[13] D. Megías, J. Serra-Ruiz, and M. Fallahpour. Efficient self-synchronised blind audio watermarking system based on time domain and FFT amplitude modification. *Signal Processing*, 90(12):3078-3092, 2010.

[14] B. Pfitzmann and M. Waidner. Anonymous fingerprinting. In *Advances in Cryptology-EUROCRYPT'96*, LNCS 1233, Springer, pp. 88-102, 1997.

[15] H. Xie, Y. R. Yang, A. Krishnamurthy, Y. G. Liu, and A. Silberschatz. P4P: provider portal for applications. *SIGCOMM Comput. Commun. Rev.*, 38(4):351-362, 2008