Identificación forense de contenido digital mediante variaciones de hashes

Gustavo Daniel Presman[*]

[*]Ingeniero Electrónico certificado internacionalmente en Informática Forense. Perito informático. Docente del Postgrado de Seguridad Informática de la Universidad Nacional del Salvador. Docente del TCP en Derecho Informático de la Universidad Nacional de San Luís. Docente de la Maestría en Seguridad Informática de la Universidad Nacional de Buenos Aires.

Abstract: El análisis de hashes es una actividad rutinaria que el perito informático realiza con el objeto de identificar objetos conocidos e indubitables en una investigación forense informática. Con la creciente disponibilidad de herramientas que permiten intervenir el contenido binario de un archivo digital en una época caracterizada por el incipiente incremento de la criminalidad informática, la dificultad de analizar grandes volúmenes de información almacenada electrónicamente en búsqueda de determinados contenidos específicos se hace cada vez más compleja. El notorio aumento de los delitos de pornografía infantil en Internet a un nivel que podría considerarse epidémico está impulsando el desarrollo de nuevas tecnologías de análisis forense informático para asistir a las investigaciones y procesos judiciales. En este trabajo desarrollaré algunas variantes de los algoritmos de hashes que intentan encontrar solución al problema de la alteración maliciosa de contenido digital de un archivo permitiendo su rápida identificación. También presentaré algunas técnicas y herramientas novedosas para este cometido que ya están disponibles o próximas a ser liberadas.

Palabras Claves: Delitos Informáticos – Comparación unívoca – Hash - Pericias Informáticas – Análisis Forense Informático

Introducción:

La utilización tradicional del cálculo del *hash* de un archivo, a fin de determinar su unicidad es una de las técnicas fundamentales del análisis forense informático que se emplea de rutina en pericias informáticas donde se requiera determinar la existencia de un archivo determinado a través de un archivo de muestra considerado como indubitable.
El *hash* es una función matemática que genera una clave de longitud fija por medio de la aplicación de algún algoritmo de cálculo.
La función de hash por lo tanto genera una clave unívoca y unidireccional Conceptualmente se interpreta que una modificación en el contenido binario de un archivo por mínima que sea altera en forma drástica el valor de hash resultante, asumiendo que si dos archivos producen resultados diferentes de hash es por que el contenido de los mismos es diferente también.
Desde el punto de vista matemático, la cantidad de valores de hash diferentes depende de la longitud de la clave generada por el algoritmo de hash.

El uso del Hash en el Análisis Forense Informático

El uso del Hash en una investigación forense informática o peritaje informático es habitual en los siguientes casos:

1- Identificación de objetos conocidos: Determinar cual de los objetos existentes en un archivo de evidencia constituyen objetos de una aplicación conocida es de gran utilidad para reducir el universo de objetos a procesar en un caso y acelerar los tiempos de respuesta, toda vez que dichos archivos conocidos pueden excluirse tanto de las búsquedas como del procesamiento posterior. Resulta fundamental contar para ello con bases de datos de hashes indubitables respecto del contenido. Estas bases de datos (*hash sets*), pueden ser creadas por el propio investigador a partir de la disponibilidad de los archivos en cuestión, compartidas entre peritos o descargadas de fuentes confiables.

2- Identificación de objetos Notables: En esta categoría se incluyen todos esos archivos que pueden ser de interés para una investigación en particular. El caso típico es el de una pericia en la que se disponen de archivos de evidencia

provenientes de varias computadoras, incluso de diferentes usuarios sobre las que se intenta probar la vinculación de las actividades entre ellas, a través de la existencia de idénticos archivos en las diferentes computadoras. Posiblemente una de las situaciones más frecuentes con las que se enfrentan los peritos informáticos de diversas fuerzas de la ley es la investigación de pornografía infantil en Internet, la cual es auxiliada a través de la INTERPOL [1] facilitando las bases de datos de hashes correspondientes a imágenes de pornografía infantil, de esta manera pueden efectuar una comparación unívoca sin necesidad de disponer de las imágenes.

Una función de hash permite la obtención de una clave de longitud fija dado el objeto original y siendo esta irreversible es decir que no resulte posible llegar al objeto original a partir del hash.
Matemáticamente resulta posible que existan claves resultantes iguales para objetos diferentes, ya que el rango de posibles claves es mucho menor que el de posibles objetos a considerar.

En el caso de usar la función de hash MD5 la longitud de la clave es de 128 bits mientras que la función de hash SHA-1 produce una clave de 160 bits de longitud , estas dos funciones de hash son las de mayor aceptación en Informática Forense.
Desde el punto de vista criptográfico un Hash no tendría validez si dos mensajes diferentes produjeran el mismo valor de hash, a esta situación se la denomina colisión. Una investigación publicada en el año 2004 [2] demostró que la función de Hash MD5 posee colisiones y lo mismo ocurre con la función de Hash SHA-1.
A pesar de estos hallazgos de la comunidad criptográfica, las funciones de hash MD5 y SHA-1 siguen siendo utilizadas en forma rutinaria para establecer una "huella digital" e identificar los objetos de una investigación forense informática o autenticar una imagen forense con el objeto de preservar la evidencia durante todo el proceso.
Esto se debe a que resulta imposible modificar el contenido de un objeto de manera que conserve el hash original. De hecho la investigación que prueba las colisiones del MD5 es desarrollada a partir de bloques de datos muy específicos que impactan sobre el algoritmo, situación que podríamos denominar "de laboratorio".

[1] INTERPOL International Police - **http://www.interpol.int/Public/Children/Default.asp**.

[2] Wang Xianyan, Feng Dengguo, Lai Xuejia, Yu Hongbo. "Collisions for Hash Functions. MD4, **MD5** Haval-128 and RIPEMD." CRYPTO '04; Revised August 17, 2004

Una aplicación utilizada para el cálculo de la función de hash es el comando de línea **dcfldd** que se encuentra presente en todas las distribuciones forenses de Linux [3].
A continuación veremos como se produce el cambio drástico en el valor de hash alterando solamente un carácter de un documento de texto, para lo cual procedo a calcular el hash MD5 de la primer hoja de este trabajo y luego a quitarle la primer letra al título y recalcular el valor del hash. Las siguientes son las imágenes de estos documentos con sus respectivos valores de hash MD5:

11a300153318 7ab2312427632bc9ddb5

76cc36b8125b88c3e0a6d374da154af9

[3] Tool for Hashing dcfldd - **http://dcfldd.sourceforge.net/**

Es posible observar en el ejemplo mostrado que el valor de hash cambia drásticamente a pesar de tratarse de un cambio mínimo en el contenido del archivo.

Sin lugar a dudas, el cálculo del hash de los objetos existentes en un archivo de evidencia es una herramienta fundamental para el investigador forense o perito informático, pero resulta claro que si el sospechoso produjo alguna modificación insignificante en el contenido que aun permita la distribución del archivo sin perdida relevante de su contenido, esta técnica no resultará apropiada para identificarlo.

A continuación desarrollaré alguna de las técnicas existentes que permiten aproximarse a la resolución de este problema.

Hashes Parciales

La idea básica del cálculo de hashes parciales consiste en subdividir el archivo en porciones a las que se le calcula por separado la función de Hash, de manera tal que un archivo modificado parcialmente mantendrá coincidencias solas con aquellas porciones del archivo original que no hayan sido cambiadas. Se puede establecer un porcentaje de coincidencias, según el tipo de investigación que se esté realizando, para establecer un grado de similitud entre ambos archivos.

Esta técnica ya es utilizada por el software EnCase Forensic [4] para autenticar parcialmente los archivos de evidencia en formato E01, de manera que resulta posible aislar aquellos segmentos cuyo valor de hash parcial, que en ese caso se calcula con un algoritmo CRC de 32 bits, no coincide con el obtenido al momento de la recolección de la evidencia, pudiéndose mantener la integridad del resto de los segmentos que no han sufrido variación del hash parcial.

Seguidamente podemos observar, para la primer hoja del presente trabajo, los hashes parciales de cuatro fragmentos del documento y sus correspondientes variaciones al efectuar la eliminación del primer carácter del título.

[4] Guidance Software EnCase Forensic - **http://www.guidancesoftware.com/computer-forensics-ediscovery-software-digital-evidence.htm**

d19c26a8d8b63c178fbf7396c76ec55f

13eeb4fb431fa57989d2d6567b05200d

9560b2e9fdfbe92d494cc21d40e9948c

160424e2fdb2364f86dc72aff38c08a5

02f1a2c45de5d322f7b06639df059ae1

13eeb4fb431fa57989d2d6567b05200d

9560b2e9fdfbe92d494cc21d40e9948c

160424e2fdb2364f86dc72aff38c08a5

Teniendo la cuenta que el archivo fue dividida en cuatro partes, podemos concluir de la comparación de los hashes parciales, que el segundo archivo presenta una similitud del 75 % con el original.

Naturalmente que si subdividiéramos el archivo en mayor cantidad de segmentos y consecuentemente tuviéramos mayor cantidad de hashes parciales, seguramente el grado de similitud entre el archivo modificado y el original sería mayor.

Según lo explicado para obtener un mayor número de hashes parciales se requiere de un mayor tiempo de procesamiento de los objetos de un archivo de evidencia, resultará entonces una solución de compromiso según el tiempo y los recursos informáticos disponibles.

Hashes Rodantes

El método de hashes parciales presenta algunas dificultades para su utilización en archivos en los que la variación de los mismos no consiste en una modificación sino en la eliminación parcial de su contenido digital.
Una evolución del método de hashes parciales son los hashes rodantes. Conceptualmente el archivo se sigue subdividiendo en segmentos pero los hashes rodantes se utilizan para identificar las fronteras de cada uno de ellos, desplazándose como si fuese una ventana.
Inicialmente se calcula el hash de la ventana y a medida que se va desplazando la misma se recalcula este valor adicionando el resultado de la función de hash aplicada a los bytes que se incorporan a la ventana y substrayendo el valor de los bytes que se dejan fuera.
Un trabajo publicado por Jesse Kornblum [5] establece el siguiente paralelo entre el contenido binario de un archivo y el contenido genético de materia viva:

> *"...Los archivos con un solo bit diferente son prácticamente idénticos y comparten una gran homología. Inspirándose en la genética, dos cromosomas son homólogos si tienen secuencias idénticas de genes en el mismo orden. Del mismo modo, dos archivos de computadora pueden tener secuencias homólogas si tienen secuencias de bits idénticos en el mismo orden. Los dos archivos son idénticos excepto en un conjunto de inserciones, modificaciones y supresiones de datos..."*

Propone entonces la utilización combinada del hash tradicional con el rodante en lo que denomina Hash contextual CPTH *(Contextual piecewise triggered hash*).
Una utilidad de código abierto que permite su cálculo es la herramienta ssdep [6]. Existe también una implementación comercial del hash contextual es la que se encuentra disponible en Forensic Toolkit de Accesdata [7] y que se denomina Hash difuso (*Fuzzy Hashing*).

[5]Identifying almost identical files using context triggered piecewise hashing – Jesse Kornblum – Digital Investigation Journal ·S (2006)

[6]Tool for Fuzzy Hashing ssdeep - **http://ssdeep.sourceforge.net/**

[7]AccessData FTK - **http://www.accessdata.com/forensictoolkit.html**

Mapeo de Hashes en bloques

Las variaciones al método tradicional de hashing que se han presentado, en todos los casos requieren de disponer del archivo lógico que se requiere comparar. Para el caso en que el sospechoso haya borrado el mismo, sería necesario efectuar previamente una búsqueda en el espacio no utilizado del disco de manera de poder acceder a un archivo lógico que si bien ha perdido todos sus metadatos asociados con las fechas MAC y el propio nombre del archivo, aún presentaría contenido digital susceptible de comparación.

Si la situación es aun más compleja y los archivos en el espacio no utilizado, poseen su encabezamiento sobrescrito o se encuentran muy fragmentados en el sistema de archivos, entonces la recuperación del archivo lógico no sería posible y por consiguiente los métodos propuestos anteriormente fallarían.

Una alternativa propuesta para la identificación forense de fragmentos de contenido binario que se encuentran en el espacio no utilizado (unallocated) o en el espacio descuidado (slack space) es el Mapeo de Hashes en bloque que consiste en efectuar, para el archivo indubitable, un cálculo de hashes tradicionales pero de cada uno de los bloques que el mismo ocupa en el disco, ese resultado es almacenado en una matriz o mapa de hashes en bloque.

A continuación el investigador seleccionará el dominio de búsqueda entre el espacio no utilizado, el espacio sin particionar y el slack space y se calcula el hash de cada bloque de esta selección el cual se compara con el mapa del archivo indubitable estableciendo cuales son los bloques del dominio de búsqueda que tienen coincidencia Este método ha sido desarrollado por Guidance Software y se implementa a través de un script en EnCase Forensic.

Al ser un procedimiento de cálculo bloque a bloque, el algoritmo es de procesador intensivo, ofreciendo resultados en tiempos de procesamiento muy extensos, pero resultando una metodología original para identificar segmentos de archivos que se encuentran diseminados en espacios físicos del medio de almacenamiento, fuera del control del sistema de archivos.

Alternativa para la identificación de imágenes

Las imágenes son un contenido binario de vital importancia en investigaciones de delitos informáticos . La identificación de imágenes ligeramente modificadas pero que conservan el espíritu y la idea original es uno de los desafíos de la Informática Forense de los próximos años.

Hany Farid, un profesor del Dartmouth College [8] especializado en análisis forense

[8] Hany Farid CV&Bio - **http://www.cs.dartmouth.edu/farid/**

de imágenes, desarrolló con el soporte de Microsoft Corporation una tecnología de identificación de imágenes que permite el reconocimiento de fotografías que han sido ligeramente modificadas. Esta tecnología que Microsoft denomina PhotoDNA [9] y que ha donado al centro nacional de niños desaparecidos y explotados de los Estados Unidos (NCMEC) consiste en redimensionar la fotografía a un tamaño, color y resolución estándar, como paso previo al calculo tradicional del hash, de esta manera la fotografía podría ser identificada sin importar el tamaño, resolución y profundidad de colores que la misma tenga.

Esta tecnología se encuentra en fase experimental, conservándose en secreto los detalles de la misma para evitar que el algoritmo sea difundido, facilitándose las técnicas antiforenses sobre el mismo.

Consideraciones Finales

El incremento de la criminalidad informática, especialmente en aquellos delitos relacionados con la pornografía infantil en Internet, sumado a la facilidad de descargar y utilizar herramientas que modifican el contenido de un archivo digital pero mantienen su esencia, hace imperativo el desarrollo de nuevas técnicas de comparación binaria que permitan identificar esta circunstancia asistiendo a la justicia en estas investigaciones.

[8] Microsoft PhotoDna - **http://www.microsoftphotodna.com/**