Cómo identificar a usuarios de internet sin cookies

Cristian Hernández C.*
12 de junio de 2010

Resumen

Este informe muestra como los navegadores modernos de Internet son susceptibles a ser identificados sin que los navegantes lo sepan, por una “huella digital” a través de la información que transmiten a las páginas web sobre su versión y configuración. Además se presentan los conceptos usados y resultados obtenidos por el experimento Panopticlick, que con un simple algoritmo aspira a demostrar el método de rastreo por huellas digitales y su efectividad en un universo de usuarios conscientes de proteger su privacidad.

Introducción

Actualmente es sabido que muchos dispositivos tecnológicos poseen características sutiles, pero medibles que los hacen susceptibles de ser identificados unívocamente. Para eso se reconocen dichas singularidades que conformarían su “huella digital”, la que sería única para cada artefacto, pudiendo así distinguirlo del resto de su clase. Este mismo principio se cumple para los navegadores de Internet.

Es sabido que la forma más simple de identificar a un usuario por parte de un servidor web es a través del uso de las cookies (1). No obstante, si un usuario navega sin usar cookies, e incluso utilizando una proxy (2) para ocultar la IP, puede dificultar aún más su reconocimiento al navegar por Internet. Sin embargo, existen métodos que pueden superar todas estas barreras, y que efectivamente logran rastrear a los usuarios de los sitios de la red. Una herramienta útil para lograrlo es a través del uso de la “huella digital” del navegador. Los conceptos y la metodología utilizada para reconocer estas “huellas digitales” son presentadas a lo largo de este informe.

Huellas Digitales como amenaza a la privacidad en la Red

La manera más común de rastrear los navegadores de Internet (por “rastrear” nos referimos al hecho de asociar las actividades de un navegador en diferentes momentos, con algún sitios web) es a través de las cookies del protocolo http. Ya que estas son una conocida vulnerabilidad a la privacidad, muchos usuarios de Internet actualmente las bloquean, limitan o eliminan periódicamente. Por otro lado, el uso de las llamadas supercookies (3) por aplicaciones tan masificados en Internet (como Flash y Javascript) abren otra puerta que sigue exponiendo la privacidad de los usuarios.

En general, un usuario que busque evitar ser rastreado por la red, debe pasar por tres pruebas. La primera ya es difícil: encontrar una configuración apropiada que permita a los sitios usar cookies sólo para las características necesarias de interfaz con el usuario, pero que además prevenga otros potenciales usos de rastreo. La segunda es más difícil: aprender sobre todos los tipos de supercookies, y encontrar maneras de deshabilitarlos. Sólo una pequeña minoría de personas son capaces de pasar estas dos pruebas, pero los que pueden deben enfrentarse al tercer desafío: “las huellas digitales” de los navegadores de Internet.

Como mecanismo de rastreo usado hacia personas que limitan cookies, la “huella digital” posee también la astuta propiedad que es mucho más difícil de detectar que los métodos de las supercookies, ya que no deja evidencia alguna en el computador del usuario.

Huellas Digitales como Identificadores Globales

Cuando existe suficiente información dada a algún algoritmo de detección de huellas digitales para reconocer a un segmento único de los usuarios, estas huellas digitales pueden ser usadas esencialmente como un “Identificador Global” para esos usuarios. Tal identificador global puede ser pensado como semejante a una cookie que no puede ser borrada a menos que la configuración del navegador web cambie tanto, que sea suficiente como para anular la huella digital.

Aunque hayan incluso usuarios que no son identificados globalmente por un una huella digital en particular, hay maneras de hacerlos vulnerables a más métodos específicos y contextualizados de rastreo, implementados por el mismo algoritmo de reconocimiento de huellas digitales, si la huella se usa en combinación con otros datos.

Huellas Digitales enriquecidas

Algunos sitios web usan las supercookies LSO (4) de Adobe Flash como una manera de “regenerar” las cookies normales que han sido borradas por el usuario, o más discretamente para asocial las cookies ID anteriores del usuario con las nuevas cookies ID asignadas. Las huellas digitales poseen una amenaza similar de “regeneración de cookies”, incluso si las huellas digitales no son globalmente identificables. En particular, una huella digital que contenga no más de 15 a 20 bits de información sobre la identidad, en la mayoría de los casos va a ser suficiente para reconocer unívocamente al navegador, siempre que conozca su dirección IP, la subred a la que pertenece, o incluso solo con su Número de Sistema Autónomo (5).

Si el usuario borra sus cookies mientras continua con una dirección IP, una subred o un ASN que haya usado previamente, el “regenerador de cookies” podría, con alta probabilidad, enlazar las cookies nuevas con las anteriores. Finalmente, otro

uso para las huellas digitales es como medio para distinguir maquinas que están detrás de una única dirección IP (cuando se usa un proxy), incluso para aquellas maquinas que bloquean completamente las cookies.

Experimentando con Huellas Digitales

Un ejemplo en particular que identifica las huellas digitales de los navegadores de Internet, es el algoritmo implementado por el científico australiano Peter Eckersley de la EFF (Electronic Frontier Foundation) en el sitio web http://panopticlick.eff.org. El algoritmo recolecta información sobre características que los navegadores entregan a los servidores. Algunas son inferidas del simple contenido de las solicitudes http, y otras son recolectadas vía JavaScript AJAX (6). El algoritmo agrupa la información en ocho parámetros que conforman la huella digital del navegador, los cuales se presentan a continuación:

Variable	Fuente	Observaciones
Agente del usuario	Transmitida vía HTTP, solicitada por el servidor	Contiene la micro-versión del navegador, la versión del sistema operativo, el idioma, barras de herramientas y a veces otra información.
HTTP acepta encabezados	Transmitida vía HTTP, solicitada por el servidor	
Cookies habilitadas?	Inferida vía HTTP, solicitada por el servidor	
Resolución de la pantalla	Posteada vía JavaScript AJAX	
Zona horaria	Posteada vía JavaScript AJAX	
Plugins del navegador, versiones de los plugins y tipos MIME	Posteada vía JavaScript AJAX	Clasificados antes de recolectar. Microsoft Internet Explorer no ofrece maneras de enumerar los plugins; por lo que se usó el PluginDetect de la librería JavaScript para chequear los 8 plugins comunes en esa plataforma, además de código extra para estimar la versión del Adobe Acrobat Reader.
Tipos de letras del sistema	Applet Flash o Applet de Java, recolectada vía JavaScript AJAX	No clasificados.
Test parcial de supercookie	Posteada vía JavaScript AJAX	No se implementaron pruebas para cookies LSO de Flash, cookies de Silverlight, de bases de datos para HTML 5, o DOM globalStorage.

Tabla 1 – Componentes de las huellas digitales medidas en Panopticlick

Resultados del Experimento Panopticlick

Este experimento se realizó entre el 27 de enero y el 12 de febrero de 2010, y recolectó huellas digitales de un total de 470.161 navegadores operados por participantes informados sobre lo que se deseaba medir. Aunque podría decirse que las muestras de navegadores es un poco parcial, básicamente representa a la población de usuarios de Internet que prestan suficiente atención a la privacidad como para estar preocupado de estos detalles, como limitar las cookies, o quizá usar servidores proxy, y que en general están de acuerdo en la necesidad de evitar la posibilidad de rastreo y recolección de información de la mayoría de las actividades de los navegadores de Internet.

En estas muestras de usuarios conscientes, el 86.6 % de los navegadores vistos tuvieron instantáneamente una única huella digital, y sólo el 5.3% presentaban un anonimato frente las huellas digitales. Entre los navegadores que tenían habilitado o bien Adobe Flash o Java Virtual Machine, el 94.2% exhibieron instantáneamente una huella digital unívoca, y sólo el 4.8% tenían huellas digitales que fueron vistas exactamente dos veces. Sólo el 1% de los navegadores con Flash o Java presentaban un anonimato. Es decir, si se toma al azar una de las muestras, en el mejor de los casos sólo uno de cada 286.777 navegadores va a compartir la misma huella digital.

Cabe mencionar que durante el período del experimento, sólo un 37.4% de las huellas digitales exhibieron por lo menos un cambio en más de 24 horas. Desafortunadamente, encontraron que un simple algoritmo fue capaz de darse cuenta y seguir muchos de estos cambios en las huellas digitales. De hecho, el algoritmo era capaz de identificar correctamente una huella digital “progenitora” en el 99.1% de los casos.

Por otro lado los navegadores de los sistemas iPhone y Android son significativamente más difíciles de asignar una huella

digital que los navegadores de escritorio, ya que los smatphones no poseen la misma variedad de plugins presentes en los sistemas de escritorio. Por desgracia, los iPhones y Androids carecen de buenas opciones de control de las cookies (como sólo usar cookies durante cada sesión, o el uso de listas negras), así que esos usuarios son claramente rastreables por maneras más básicas que las huellas digitales.

Conclusiones

En este experimento se implementó un método en particular para identificar huellas digitales de los navegadores, por lo que, en general, hay muchas mediciones más que no fueron consideradas y que sin lugar a dudas mejorarían considerablemente los resultados obtenidos, reforzando la efectividad del método y la vulnerabilidad de la privacidad de los usuarios de Internet. Cabe señalar que hubo maquinas clonadas detrás de firewalls que resultaron ser resistentes al algoritmo usado, pero no pudieron resistir a huellas digitales que midan el sesgo del reloj (7) u otras características de hardware.

El uso de huellas digitales de los navegadores es una herramienta muy poderosa, por lo que deben ser consideradas en conjunto con las cookies, las direcciones IP y las supercookies, al momento de discutir sobre la privacidad de los usuarios y su rastreabilidad. Pese a que las huellas digitales no son particularmente estables en el tiempo, los navegadores revelan tanta información sobre su versión y configuración que igualmente siguen siendo rastreables. Por este motivo los desarrolladores de navegadores de Internet deberían considerar qué pueden hacer para disminuir la posibilidad de elaborar huellas digitales, particularmente a nivel API (8) de JavaScript.

*Universidad Técnica Federico Santa María Departamento de Electrónica.

Referencias:
1. Figueroa, F.: No hacen falta cookies para rastrearte online.
http://www.fayerwayer.com/2010/05/no-hacen-falta-cookies-para-rastrearteonline
2. Eckersley, P.: How unique is your Browser? http://panopticlick.eff.org/browseruniqueness.pdf
3. Corredera, L.: Cómo protegernos de las “supercookies”.
http://www.elreservado.es/news/view/220-noticias-espias/97-como-protegernosde-las-supercookies

notas:
1) Una cookie es un fragmento de información que se almacena en el disco duro del visitante de una página weba través de su navegador, a petición del servidor de la página. Esta información puede ser luego recuperada por el servidor en posteriores visitas.
2) El término proxy hace referencia a un programa o dispositivo que realiza una acción en representación de otro. Su finalidad más habitual es la de servidor proxy, que sirve para permitir el acceso a Internet a todos los equipos de una organización cuando sólo se puede disponer de un único equipo conectado, esto es, una única dirección IP.
3) Las supercookies son pequeños ficheros que una página Web remota crea en nuestros ordenadores sin nuestro permiso, parecidas a las cookies http, pero con nuevas características que las hacen mucho más “poderosas”.
4) Una cookie flash LSO (Local Shared Object, en español Objeto Local Compartido) es una colección de archivos tipo cookie almacenadas como archivo en computador del usuario. Las LSO son usadas por todas las versiones de Adobe Flash Player.
5) Un Sistema Autónomo (en inglés, Autonomous System, AS) es un conjunto de redes y routers que se encuentran administrados por una sola entidad (o en algunas ocasiones varias), y que cuentan con una política común de definición de trayectorias para Internet.
6) AJAX o Asynchronous JavaScript And XML (JavaScript asíncrono y XML), es una técnica de desarrollo web para crear aplicaciones interactivas que corren en el navegador del usuario, mientras en segundo plano mantiene una comunicación asincrónica con el servidor y le va enviando información de vuelta.
7) El sesgo de reloj se refiere a las diferencias en el tiempo (fecha y hora) mostrado en los relojes de las computadoras.
8) La API se refiere a la interfaz de programación de aplicaciones.