Privacidad

Alexandre Viejo Galicia
Jordi Castellà-Roca

PID_00183949

Universitat Oberta de Catalunya

Índice

Introducción

La tecnología es una pieza clave en el engranaje, que hace avanzar a la sociedad en conjunto. No obstante, la tecnología aplicada a la información consigue potenciar a los individuos en particular.

Aquellos que en el pasado empleaban una cantidad ingente de horas adquiriendo conocimiento de libros y otros documentos, con el inicio de la era de Internet conocieron un nuevo mundo donde la cantidad de información disponible se multiplicaba y la facilidad de acceso a ella alcanzaba cotas difícilmente imaginables anteriormente. En este nuevo escenario, los ordenadores de todo el mundo están conectados entre sí. Computadoras de personas anónimas, de pequeñas empresas, grandes corporaciones y Gobiernos se comunican y comparten sus conocimientos. Internet permite a individuos que jamás se conocerían en persona, compartir lo que ven y lo que saben.

Internet se ha convertido en el gran escaparate para cualquiera que quiera darse a conocer al mundo. Es el lugar idóneo para las empresas en busca de publicidad directa, para los exhibicionistas y también para aquellos *voyeurs* que entienden la red como su televisión a la carta en donde pasar sus ratos de ocio.

La digitalización masiva de documentos antiguos, la utilización de dicho formato para generar los documentos actuales y la interconexión de los ordenadores ha creado información inalterable y disponible para todos. Históricamente, la privacidad de las personas se ha basado en que éstas controlaban su propia información. No obstante, mantener ese control en un mundo dominado por el formato digital y los ordenadores conectados es prácticamente imposible. Una vez que un cierto dato se hace público, puede ser replicado y recolocado en cualquier otro lugar de la red, sin verse alterada su calidad o sin que ésta pueda variar con el tiempo.

Este escenario ha propiciado que los ordenadores se conviertan en una herramienta indispensable para interactuar con el mundo. Foros, blogs, redes sociales... todas esas aplicaciones permiten a los individuos saciar su curiosidad aunque ese camino también les lleva a compartir su propio conocimiento.

Finalmente, aquellos que se creían solo consumidores de ese gran teatro de variedades han descubierto que ellos también son actores principales en una gran obra global. Han comprobado que su rastro en Internet aumenta a medida que interactúan con la red y que sus propios datos personales, intereses u opiniones pueden ser adquiridos fácilmente, tan fácilmente como ellos encuentran información referente a otros individuos.

No obstante, no solo los individuos anónimos pueden acceder a todos esos datos, las empresas y los Gobiernos han aprendido a explotar toda esta cantidad de información personal disponible y proporcionada por las propias personas. De esta forma han desarrollado nuevas formas de obtener beneficios e influir en las masas. En muchos casos, la privacidad ya es cosa del pasado. Recuperarla, es una de las tareas del presente.

Objetivos

En los materiales didácticos asociados a este modulo el estudiante encontrará los contenidos necesarios para alcanzar los siguientes objetivos:

1. Entender qué es la huella digital de un usuario, qué contiene y cómo se crea.

2. Conocer los mecanismos utilizados habitualmente en Internet para identificar y rastrear a un usuario sin que éste se percate.

3. Conocer los atributos y datos personales que forman un perfil de usuario y entender los distintos métodos utilizados para generarlos.

4. Entender cómo las empresas y organizaciones pueden utilizar los perfiles de usuario para aumentar sus beneficios.

5. Conocer la definición de privacidad en el marco de las tecnologías de la información y entender las implicaciones de las políticas de privacidad aplicadas por las empresas de Internet que tratan con datos personales.

6. Poseer nociones básicas sobre las técnicas disponibles para preservar la privacidad de los usuarios conociendo sus virtudes, inconvenientes y sus distintos ámbitos de aplicación.

1. La huella digital

En la actualidad, la mayoría de los habitantes de las sociedades desarrolladas interactúan de forma habitual con el mundo digital (por ejemplo, TV, teléfonos móviles, Internet, sensores, RFID, etc.). La huella digital de una cierta persona está formada por el rastro que dejan dichas interacciones.

La huella digital es esencial para proporcionar personalización, marketing dirigido, reputación digital y varios servicios basados en el medio social.

Más allá de proporcionar servicios o anuncios personalizados, la importancia de la huella digital respecto a la reputación de los individuos no debe ser pasada por alto. La utilización de las huellas digitales en los procesos de selección de personal es una realidad. En este sentido, Coutu y otros (2007) explica en su estudio que en el pasado, el CV de una persona era totalmente controlado por ésta y se basaba únicamente en lo que dicha persona comunicaba a la empresa interesada. Por el contrario, en la actualidad, el CV de un individuo está formado por los primeros diez ítems que aparecen en Google.

La utilización de la huella digital no se limita a los departamentos de recursos humanos de las empresas. Una serie de encuestas y análisis realizados en el 2007 por Pew Internet sugieren que el 47% de las personas han buscado información sobre sí mismas en Internet. Este tipo de búsqueda se la conoce como "búsqueda de vanidad" o de "ego". Más allá de la búsqueda de información propia, el mismo estudio indica que más de la mitad de todos los usuarios adultos de Internet han utilizado un motor de búsqueda para seguir las huellas de otra persona. Un 11% de estos usuarios buscaban información acerca de alguien a quien estaban pensando contratar, no obstante, llama la atención que un 19% de los usuarios buscaba información sobre compañeros de trabajo, colegas o competidores. Respecto a la información adquirida, un 72% se centraba en los datos de contacto, un 37% en logros profesionales e intereses, un 33% en perfiles de redes sociales o profesionales, un 31% en fotos y finalmente, un 28% buscaba también información respecto a antecedentes personales.

Lectura recomendada

Para consultar las encuestas y análisis realizados en el 2007 por Pew Internet podéis consultar el siguiente artículo:

M. Madden; S. Fox; A. Smith; J. Vitak (2007). "Digital Footprints: Online identity management and search in the age of transparency".

Otros resultados interesantes de la encuesta de Pew Internet indican que el 10% de los usuarios de Internet tienen un trabajo que les obliga a promover o publicitar su reputación digital. Adicionalmente, un 20% de los adultos estadounidenses dicen que sus empresas tienen políticas especiales para promover

que los empleados se publiciten correctamente en Internet. Finalmente, cabe destacar que las personas concluyen que un 90% de la información que localizan sobre sí mismos es precisa.

Ante estos datos, la importancia de la huella digital en la sociedad actual queda libre de toda duda. A continuación, las preguntas que deben ser resueltas son: ¿qué contiene dicha huella? y ¿cómo se crea?

1.1. El contenido de la huella digital

Comúnmente, se considera que la huella digital de una persona la forman sus visitas a ciertas páginas web, sus búsquedas, sus perfiles en redes sociales o similares, correos, blogs, posts en foros, etc. No obstante, la huella digital va mucho más allá. En este sentido, el Discovery Channel ofrece una aplicación que ilustra cómo una persona deja un rastro digital en muchas de sus acciones cotidianas.

Como ejemplos de esta situación podemos destacar las siguientes interacciones y el rastro que dejan:

- **Leer noticias por Internet**: el sitio web de noticas puede identificar al usuario (por ejemplo, mediante un identificador de usuario/contraseña, *cookies* o IP) y extraer cierta información, como sus horarios, las noticias que le interesan, los anuncios que selecciona, etc. Estos datos ayudan a proporcionar un servicio personalizado al usuario pero también permiten conocer sus intereses y actividades, lo cual puede ser un problema para este.

- **Ir en coche al trabajo**: los coches más modernos son capaces de almacenar datos relacionados con la velocidad, frenos, utilización de cinturones de seguridad, etc. Los coches equipados con GPS son capaces de almacenar las rutas seguidas. Esta información es útil en caso de accidente, pero también puede ser utilizada en su contra (por ejemplo, las aseguradoras pueden utilizar este conocimiento para rechazar a ciertos conductores).

Venta de datos de conducción

Según recogió el diario alemán *AD*, el fabricante del GPS TomTom vendió los datos de conducción de sus usuarios a la policía de los Países Bajos.

- **Pagar con tarjeta de crédito**: los bancos guardan registros completos de los movimientos realizados con las tarjetas que emiten. Las tarjetas ofrecen comodidad a los compradores, pero los datos de transacciones pueden ser utilizados por terceros para conocer los hábitos de compra de las personas. Dichos datos también pueden ser robados en ataques informáticos a las bases de datos de los bancos.

- **Pagar un peaje**: incluso pagando con dinero en metálico (el cual se considera un sistema de pago anónimo), las cámaras situadas en los peajes guardan las imágenes de los vehículos y sus matrículas. Estos sistemas evitan que conductores deshonestos se salten los peajes, no obstante, tam-

bién permiten localizar a los conductores honestos y analizar hasta cierto punto sus hábitos de conducción y rutas.

- **Aparcar en un parking público de pago**: la utilización de cámaras de seguridad para garantizar el pago del servicio implica una situación similar al del pago del peaje.

- **Utilizar un tag RFID para identificarse en el trabajo**: la utilización de estos dispositivos evita que personas no autorizadas entren en entornos de trabajo restringidos. No obstante, los *tags* RFID pueden contener una gran cantidad de información personal del propietario. Dicha información puede ser robada por un atacante que se encuentre en el radio de acción del *tag*. Adicionalmente, estos dispositivos permiten a la empresa conocer los hábitos del trabajador en el lugar de trabajo.

- **Consultar el correo electrónico**: muchos sistemas de correo utilizados habitualmente (por ejemplo, G-mail/Googlemail, Hotmail, etc.) almacenan y analizan los mensajes enviados y recibidos. A partir de esta información las empresas proporcionan anuncios personalizados al usuario. No obstante, indirectamente también pueden tener acceso a una gran cantidad de información personal de los usuarios comprometiendo la confidencialidad de sus mensajes.

- **Acceder a una red wireless de un lugar público**: la información transmitida durante la utilización de Internet mediante este sistema puede ser almacenada por el proveedor de servicios. Adicionalmente, estos sistemas son más susceptibles de ser atacados por *hackers* que podrían acceder a dicha información.

- **Llamar por teléfono**: las compañías telefónicas almacenan la hora y número de cada llamada. La agencia de seguridad nacional de EE. UU. (NSA) puede monitorizar las llamadas internacionales sin informar a los usuarios. Adicionalmente, ciertos *hackers* pueden acceder a la base de datos de la compañía telefónica y acceder a toda esta información.

- **Utilizar una red social**: la información publicada en estos sitios web es almacenada y analizada por las empresas que gestionan dichos sistemas. Estas empresas utilizan los datos adquiridos con fines comerciales. La información que los usuarios publican en estas redes es generalmente muy personal e incluyen números de teléfono, direcciones, estado civil, opiniones sobre productos, actividades realizadas, viajes, etc.

Escuchas telefónicas

En marzo del 2005 el primer ministro de Grecia confirmó que su teléfono había sido pinchado, al igual que el del alcalde de Atenas y otros 100 dignatarios de alto rango. Las escuchas también se realizaron a empleados de la Embajada Americana. Una detallada descripción del caso fue publicado por Vassilis Prevelakis y Diomidis Spinellis en la revista *IEEE Spectrum* (vol. 44 núm. 7) en julio del 2077 con el título "The Athens Affair".

- **Utilizar un sistema de mensajería instantánea**: las empresas que proporcionan estos servicios almacenan y analizan las conversaciones realizadas. Generalmente, los datos adquiridos son utilizados con fines comerciales.

- **Enviar una consulta a un buscador de Internet**: el motor de búsqueda puede identificar al usuario (por ejemplo, mediante usuario/contraseña, *cookies* o dirección IP) y extraer sus intereses a partir de las consultas que el usuario realiza. La información extraída generalmente se utiliza con fines comerciales.

Configuración de los navegadores

La configuración de los navegadores puede utilizarse para identificar usuarios de forma casi unívoca (Eckersley, 2010).

Como se puede observar, la mayoría de interacciones que una persona realiza en su día a día deja un cierto rastro que forma su huella digital. El tamaño de dicha huella variará en función del número e importancia de sus interacciones, no obstante, se puede inferir la facilidad con la que se puede extraer información personal de los individuos. En este módulo nos centraremos en los problemas de privacidad asociados a las interacciones que se producen en Internet.

1.2. Cómo se crea la huella digital en Internet

La huella digital de una persona se puede construir de forma activa o pasiva. En cualquier caso, cabe destacar que son las acciones del individuo las que dejan al descubierto sus trazas. La diferencia entre construcción activa y pasiva se basa en si dichos rastros se dejan de forma consciente (de forma activa) o inconsciente (de forma pasiva).

La **creación activa de una huella digital** se fundamenta en acciones que el usuario hace deliberadamente. Desde la llegada de la Web 2.0, la presencia en la red de aplicaciones sociales en las cuales los usuarios generan un perfil con fotos, opiniones y todo tipo de datos personales (por ejemplo, trabajo, situación civil, número de teléfono, etc) ha crecido de forma muy significativa. En cualquier caso, la huella digital creada de forma activa es responsabilidad del propio usuario y éste puede aprender a controlarla para su propio beneficio.

Lectura recomendada

El siguiente artículo analiza la utilización de los perfiles públicos por parte de las empresas para seleccionar a sus trabajadores:

J. Terry (7, febrero, 2008). "Leaving a digital footprint: Online activities follow students to job interviews, professional world". *The State News*.

La **creación pasiva de una huella digital** se fundamenta en elementos prácticamente invisibles para el usuario como el *web caching* y las *cookies*. Estos elementos son transparentes para el usuario que incluso puede no conocer su existencia y por lo tanto, se pueden considerar más peligrosos desde el punto de vista de la privacidad.

Web 2.0

El término Web 2.0 está comúnmente asociado con aplicaciones web que facilitan compartir información, la interoperabilidad y la colaboración en la World Wide Web (WWW).

Ejemplos de la Web 2.0 son las comunidades web, servicios web, redes sociales, wikis, blogs, etc.

El *web caching* se basa en la utilización de cachés web. Este tipo de caché almacena todo tipo de información accesible mediante el *browser*. Su motivación es reducir el ancho de banda consumido, la carga de los servidores y el retardo en la descarga. Un caché web almacena copias de los documentos que pasan por él, de forma que las subsiguientes peticiones pueden ser respondidas por el propio caché. La caché web se puede situar en diversos puntos de la arquitectura de comunicaciones: puede localizarse en el propio *browser* del usuario (llamado *browser-caché*) o puede estar situado en un *proxy* a cargo del propio servidor (o servidores) web, a los cuales el usuario intenta acceder. En este último caso hablaríamos de un *proxy-caché* capaz de almacenar el tráfico web del usuario y poner a disposición del propietario del servidor una gran cantidad de información del individuo en cuestión.

Uso de *cookies*

Un ejemplo sencillo del uso de *cookies* ocurre cuando el usuario intenta conectarse a cierta página de noticias en la cual está registrado, y no necesita recordar su identificador/contraseña (*login/password*) puesto que el servidor accede a la *cookie* guardada en el ordenador y obtiene dichos datos sin la interactuación de la persona.

Las *cookies* guardan información de los usuarios en su propio ordenador. Las páginas web utilizan estos contenedores de datos para reconocer al usuario que se conecta frecuentemente a dicha página y conocer sus preferencias y requisitos.

Este sistema de reconocimiento automático facilita la interacción de los usuarios, no obstante, las *cookies* pueden almacenar información personal que puede resultar peligrosa desde el punto de vista de la privacidad. En este sentido, estos contenedores pueden almacenar cualquier dato que el usuario haya introducido en un formulario (por ejemplo, la dirección de su casa) y, obviamente, pueden guardar datos más comunes como dirección de correo electró-

nico, proveedor de servicios de Internet, etc. La extensa implantación de esta tecnología, su sencillez de ejecución y su capacidad de acción sin requerir la atención del usuario convierte a las *cookies* en un punto controvertido en cuanto a los peligros de privacidad y la huella digital de los usuarios.

Ved también

La generación y utilización de *cookies* se comentan con detalle en el subapartado 1.2.1. Posteriormente, se dedica el subapartado 1.2 a detallar algunas tecnologías con características similares a las *cookies* que podrían usarse como sustituto.

1.2.1. Http *cookies*

Bases de datos sobre individuos

La compañía RapLeaf construye bases de datos sobre individuos utilizando sus actividades en redes sociales, historial de compras y otras interacciones con la web.

Los datos recogidos se utilizan para campañas de publicidad dirigida.

Las *cookies* son utilizadas como un mecanismo de comunicación persistente entre los sitios web que las originan y los *browsers* de los visitantes. La persistencia de las *cookies* se traduce en que son un simple fragmento de texto que se almacena en el disco duro del ordenador del usuario y permite al servidor web identificar al usuario cada vez que se conecta, conocer sus preferencias de usuario, realizar funciones de "cesta de la compra" y cualquier otra actividad que pueda realizarse almacenando un fragmento de texto en un ordenador. Cabe destacar que el sistema de identificación por *cookies* es la forma más sencilla y efectiva de identificar a un mismo usuario que vuelve a un sitio web debido a que la extendida utilización de direcciones IP dinámicas descarta este método para dicho cometido, y requerir al usuario que introduzca un *login/password* puede ser un inconveniente para la persona.

Como ya se ha dicho, las *cookies* son simplemente fragmentos de texto. No son software, lo que implica que no pueden ser programadas, no pueden llevar *virus*, *spyware* o *malware* en general. No obstante, sí pueden ser utilizadas por software dañino para conocer las actividades del usuario en Internet. Hay que recordar que las *cookies* indican las páginas a las cuales accede el usuario, el *login/password* que utiliza, sus preferencias y ciertos datos personales que haya podido introducir en algún momento. De igual forma, las *cookies* pueden ser robadas por un *hacker* que acceda a la computadora del usuario.

En resumen, las *cookies* no perjudican a la privacidad del usuario directamente pero pueden ser utilizadas con ese fin.

Funcionamiento básico de las *cookies*

Como ya se ha indicado anteriormente, las *cookies* son generadas por el servidor. Esto puede realizarse mediante CGI *scripting*. Posteriormente, la *cookie* es colocada en el lado del cliente. En una posterior conexión entre el cliente y el servidor, el servidor web será capaz de obtener todas las *cookies* que se encuentran en el lado del cliente.

Atributos de las *cookies*

Las *cookies* están formadas por ciertos atributos, destacamos los más relevantes desde el punto de vista de la privacidad:

1) ***Domain***: un servidor solo puede acceder a *cookies* generadas por otro servidor dentro del mismo dominio.

Ejemplo

Los servidores de bali.vacation.com y mexico.vacation.com pueden acceder a las *cookies* generadas por servidores de vacation.com.

2) ***Path***: este atributo funciona de forma similar al dominio al limitar la visibilidad de las *cookies* basándose en el *path* de la URL.

Ejemplo

En este sentido una *cookie* donde se especifique *domain= bali.vacation.com* y *path=account* provocará que a la *cookie* solo se acceda desde páginas situadas o que cuelgan de la URL: *bali.vacation.com/account*. Este atributo puede dejarse inactivo para evitar esta restricción.

3) ***Expires*** **y** ***Max-Age***: este atributo indica al *browser* cuándo eliminar una cierta *cookie*. La fecha marca exactamente el día y la hora de expiración. Existe una variación (RFC 2965) en la cual se puede especificar el tiempo de vida de la *cookie* en segundos desde que fue recibida del servidor. Las *cookies* marcadas para eliminación se destruyen al cerrar el *browser*.

4) ***Secure***: este atributo indica al *browser* que la *cookie* en cuestión sólo puede ser utilizada en conexiones cifradas. El servidor debe situar una *cookie* de este tipo mediante un canal seguro.

5) ***HttpOnly***: este atributo indica al *browser* que la *cookie* en cuestión sólo puede ser utilizada por el protocolo HTTP. Esto evita que *scripts* situados en el lado del cliente accedan a las *cookies* y, por lo tanto, que puedan ser robadas por ataques basados en *cross-site scripting*.

Cross-site scripting

El *cross-site scripting* generalmente abarca cualquier ataque que permita ejecutar código de "scripting", como VBScript o JavaScript, en el contexto de otro sitio web.

Estos errores se pueden encontrar en cualquier aplicación que presente información en un navegador web. El problema está en que usualmente no se validan correctamente los datos de entrada.

Creación de la huella digital mediante *cookies*

A continuación se muestran una serie de pasos que indican cómo los servidores web pueden utilizar las *cookies* para guardar un registro de las páginas visitadas por el usuario:

1) El usuario visita portal.com, una página gestionada por advts.com, y selecciona con el ratón en un *banner* correspondiente a shoe.com, también gestionado por advts.com.

Lectura recomendada

En la siguiente página web podéis encontrar más información respecto a ataques basados en *cross-site scripting*: "The Cross-Site Scripting (XSS) FAQ".

2) El servidor advts.com coloca una *cookie* en el *browser*: *portal.com::zapatos.com* y direcciona al usuario a shoe.com pasando la información correspondiente a shoe.com.

3) El usuario visita un *banner* de vacaciones.com situado en zapatos.com, y el sitio vacaciones.com también está gestionado por advts.com.

4) El servidor advts.com obtiene la *cookie* del *browser* y la actualiza a *portal.com::zapatos.com::vacaciones.com*, entonces direcciona el usuario a vacation.com.

Problemas de privacidad asociados al uso de *cookies*

Ejemplo

Siguiendo el ejemplo del subapartado anterior, supongamos que advts.com introduce en la *cookie* un identificador único (por ejemplo, "1234") para el usuario que accede a zapatos.com. Dentro de este sitio web, el usuario compra unas zapatillas de *trekking*. Posteriormente, pincha con el ratón en un *banner* de vacaciones.com. Este último sitio está dentro del dominio de advts.com y, por lo tanto, puede acceder a la *cookie* del *browser* del usuario. A partir de la *cookie*, se obtiene el identificador "1234" asignado y puede contactar con zapatillas.com para conocer qué zapatos ha comprado el usuario "1234" en el sitio web. Tras conocer que el usuario ha obtenido unas zapatillas de *trekking*, vacaciones.com puede ofrecer al usuario paquetes de vacaciones enfocados al montañismo.

En este pequeño ejemplo hemos podido comprobar la importancia de la información que desprendemos al interactuar con la web y cómo las *cookies* ayudan a que los distintos servidores sigan nuestro rastro y obtengan beneficios de ello. En este caso, el problema de privacidad analizado está asociado a que todas las páginas consultadas estaban situadas dentro del mismo dominio y por tanto, todos los servidores involucrados podían recuperar la *cookie* del usuario. Si el usuario accediese a un servidor externo a ese dominio, dicho servidor no podría acceder a la *cookie* ni al rastro del usuario.

Se puede ver, por lo tanto, la utilización de dominios como una medida de control de acceso a la información del usuario. No obstante, esta medida de seguridad puede ser fácilmente evitada: aunque las *cookies* sólo se envían al servidor que las definió o a otro situado dentro del mismo dominio, una página web puede contener imágenes u otros componentes almacenados en servidores de otros dominios. Las *cookies* que se crean durante las peticiones de estos componentes se llaman ***cookies* de terceros**.

Analizadores de tráfico

La utilización de analizadores de tráfico se ha demostrado efectiva para obtener el identificador de usuario y la contraseña (*login/password*) de los usuarios de Facebook mediante la captura de sus *cookies* que se transmiten mediante conexiones HTTP no cifradas.

Con la utilización de *cookies* a terceros, una entidad puede realizar un seguimiento de los usuarios a través de todas las páginas donde haya colocado imágenes o componentes similares.

Finalmente, cabe destacar que en una conexión HTTP no cifrada, las *cookies* viajarán en claro junto con el resto de información transmitida. Un atacante que intercepte mediante un *sniffer* dicha conexión podrá tener acceso a la información contenida en las *cookies*. Esto representa un problema de privaci-

dad importante que puede ser resuelto utilizando seguridad a nivel de capa de transporte (TLS) para cifrar la conexión entre *browser* y servidor. El protocolo https sería un ejemplo del tipo de conexión que solucionaría esta situación.

1.2.2. Alternativas a las *cookies*

Existen alternativas validas a las *cookies* a la hora de identificar a un usuario (o navegador) en particular. Sin embargo, estas alternativas generalmente no son tan fiables y, por lo tanto, acaban convirtiendo a las *cookies* en la opción preferida a la práctica.

Local shared objects (*flash cookies*)

Las *flash cookies* son porciones de datos que los servidores web que utilizan Adobe Flash guardan en los ordenadores de los usuarios. Generalmente, dichos servidores utilizan las *flash cookies* para almacenar preferencias de usuario, no obstante, son también utilizadas para obtener la información de navegación de los usuarios evitando los controles que éstos hayan podido aplicar sobre las clásicas http *cookies*. Ryan Singel (2009) explica que, en esas fechas, más de la mitad de los sitios web más relevantes utilizaban *flash cookies* para seguir usuarios y almacenar información sobre ellos. De todos ellos, solo cuatro sitios web mencionaban su utilización en sus políticas de privacidad.

Con la configuración por defecto, el cliente *flash* no solicita al usuario ningún permiso para almacenar *flash cookies* en su disco duro. Además, la configuración inicial de la aplicación puede almacenar hasta 100 Kb de información en el dispositivo. De forma similar a las http *cookies*, una flash *cookie* no puede ser leída por servidores fuera del dominio de la entidad que la creó.

Dirección IP

Anteriormente, hemos hecho una breve referencia a la utilización de direcciones IP como técnica poco fiable para realizar un seguimiento de usuarios. Este método principalmente se basa en almacenar la dirección IP del ordenador que solicita las páginas. Esto es así debido al propio protocolo IP. Cada solicitud que recibe el servidor tiene la dirección IP del ordenador en el que se ejecuta el navegador (o de su servidor *proxy* en caso de usarse). El servidor puede guardar esta información, independientemente del uso o no de *cookies*.

El problema principal de esta medida es su escasa fiabilidad a la hora de identificar unívocamente a un cierto usuario. Los ordenadores y *proxies* utilizados pueden estar compartidos por varios usuarios o el mismo ordenador puede tener asignadas diferentes direcciones IP en diferentes sesiones (caso típico de asignación dinámica de IP). Además, cabe destacar que esta técnica solo proporciona identificación, no puede sustituir el uso de *cookies* para el almacenaje de preferencias y aplicaciones similares.

Identificación mediante *browser*

Peter Eckersley (2010) ha investigado el nivel de *singularidad* de las "huellas" de los navegadores a partir de los datos de configuración que éstos transmiten bajo demanda a los distintos sitios web. En particular, ha analizado las huellas de una gran cantidad de navegadores distintos que se conectaron a un determinado sitio web (http://panopticlick.eff.org). De los resultados obtenidos se concluyó que la huella de un navegador (su configuración) contiene al menos 18,1 bits de entropía. Esto implica que al elegir un navegador al azar, en las mejores circunstancias solo uno entre 286.777 navegadores compartirá la misma configuración. En el caso de navegadores que utilizan Flash o Java, el grado de singularidad aumenta. En dichos casos, los navegadores contienen 18,8 bits de información que puede ser utilizada para identificarlos unívocamente. Como resultado de esto, en el estudio realizado el 94,2% de los navegadores con Flash o Java utilizaban una configuración diferente del resto.

Mediante la observación de los visitantes que volvían repetidas veces a la página web, el autor fue capaz de estimar cómo de rápido las huellas de los navegadores cambiaban con el tiempo. Se concluyó que las configuraciones de los navegadores cambian rápidamente, pero también se demostró que incluso las heurísticas más simples eran capaces de identificar un mismo navegador que había sido modificado (generalmente actualizado). En este sentido, los resultados del estudio indican que las heurísticas utilizadas identificaban correctamente a un navegador "actualizado" con una probabilidad del 99,1%. El falso positivo ocurría con una probabilidad del 0,86%.

URL (query string)

Otra técnica para seguir la navegación de un cierto usuario que no acepta *cookies* consiste en incrustar información en la URL. Normalmente se usa para este fin la *query string* que es parte de la URL, pero también se pueden utilizar otras partes.

Formato de la *query string*

La *query string* es la parte de una URL que contiene los datos que se deben pasar al servidor web para que genere la página que el navegador solicita.

Una URL típica que contenga una *query string* tendría el siguiente formato: http://server/path/program?query_string.

Este método consiste en que el servidor web añade *query strings* a los enlaces de la página web que contiene, a la hora de servirla al navegador. Cuando el usuario sigue uno de esos enlaces "modificados", el navegador devuelve al servidor la *query string* añadida.

Las *query strings* utilizadas de esta manera son muy similares a las *cookies*: básicamente, las dos tecnologías se basan en el uso de porciones de información definidos por el servidor y devueltas por el navegador del usuario posteriormente. Sin embargo, existen diferencias: dado que una *query string* es parte de una URL. Si la URL es reutilizada posteriormente, se estará enviando al servidor la misma porción de información. Si, por ejemplo, las preferencias de un

usuario están codificadas en la *query string* de una URL, y el usuario envía esa URL a otro usuario por algún medio, esas preferencias serán utilizadas también por este otro usuario.

Además, incluso si el mismo usuario accede a la misma página dos veces, no hay garantía de que se utilice la misma *query string* en las dos ocasiones. Un ejemplo de esta situación se da cuando el mismo usuario llega a la misma página partiendo de dos orígenes distintos: una proveniente de otra página del mismo servidor web y la otra de un buscador. En este caso, las respectivas *query strings* serán normalmente diferentes, mientras que en el caso de haber utilizado *http cookies*, éstas hubiesen sido idénticas.

Otra desventaja de las query strings está relacionada con la seguridad: almacenar en una *query string* información que identifica una sesión permite o simplifica varios ataques contra la seguridad de los usuarios. Por ejemplo, permite a un atacante realizar una fijación de sesión, forzando al usuario a trabajar con una identificación o sesión en particular elegida por dicho atacante.

Autenticación HTTP

El protocolo HTTP incluye mecanismos de autenticación (por ejemplo, *digest access authentication*) que permiten acceder a una página web sólo cuando el usuario ha facilitado un nombre de usuario y contraseña correctos. Una vez que se han introducido los credenciales, el navegador las almacena y las utiliza para acceder a las páginas siguientes, sin pedirlas de nuevo al usuario.

Desde el punto de vista del usuario, la aplicación de estas técnicas tienen un efecto similar al de la utilización de *http cookies*: el nombre de usuario y la palabra clave sólo se piden una vez, y a partir de entonces el usuario obtiene acceso a las páginas del servidor. Internamente, el nombre de usuario y la contraseña se envían al servidor con cada petición del navegador. Esta información puede ser utilizada para seguir al usuario durante su navegación de forma efectiva.

Cabe destacar que se aplican mecanismos de expiración de sesión a estos métodos. Por ejemplo, una sesión en particular normalmente expira tras un periodo de inactividad determinado, quedando así invalidada para una posterior recuperación.

Adicionalmente a la medida anterior, se debe considerar la aplicación de técnicas criptográficas al envío del nombre de usuario y contraseña puesto que, de ser enviados en claro, podrían ser capturados por un analizador de tráfico (*sniffer*).

2. Perfiles de usuario

En este apartado primero resumiremos qué elementos forman un perfil de usuario. Posteriormente explicaremos varios entornos donde esos elementos pueden ser adquiridos debido a su publicación activa o pasiva por parte de los propietarios. Finalmente, se introducirán varios escenarios donde se muestra la importancia de los perfiles de usuarios y cómo las empresas los pueden explotar.

2.1. Atributos de los perfiles de usuario

A continuación resumiremos los diferentes elementos que forman un perfil de usuario:

- **Atributos personales**: son los datos personales típicos: nombre, edad, sexo, dirección, etc.
- **Intereses**: se refiere a los temas preferidos o de interés del usuario, como por ejemplo: futbol, coches, etc.
- **Opiniones**: las que pueda tener el usuario sobre ciertos temas, por ejemplo, cine, música, etc.
- **Topología de amistades**: los amigos del usuario y sus identidades.
- **Localización**: los lugares que frecuenta el usuario, las rutas habituales que sigue para ir al trabajo, a casa, etc.

2.2. Creación de los perfiles de usuario

Los elementos que forman parte de un perfil de usuario pueden ser obtenidos en las distintas interacciones activas o pasivas de este con varias aplicaciones disponibles en Internet. A continuación analizaremos a nivel general las aplicaciones más comunes y la información personal que se puede extraer de ellas.

2.2.1. Creación activa: aplicaciones sociales

Este subapartado se ocupa de las aplicaciones que generan una huella digital del usuario de forma activa, es decir, el propio usuario con sus interacciones publica de forma consciente su información personal y la identifica como suya. Existe una gran cantidad de aplicaciones sociales que explotan la Web 2.0. A continuación se analizan las categorías más relevantes.

Redes sociales

Krishnamurthy y Wills (2010) han realizado un análisis de la accesibilidad y disponibilidad de ciertos atributos personales en doce redes sociales (por ejemplo, Facebook, LinkedIn, etc). Específicamente, los elementos que buscaban eran:

- foto personal,
- localización,
- sexo,
- nombre,
- amigos,
- actividades,
- edad,
- escuelas,
- lugar de trabajo,
- fecha de nacimiento,
- código postal,
- dirección de correo electrónico,
- número de teléfono y
- dirección física.

En la tabla siguiente se muestra el grado de disponibilidad de cada atributo (situado en cada fila de la tabla). La primera columna indica el número de redes sociales donde el atributo en cuestión se encuentra disponible para todos los usuarios de la red y el propietario no puede restringir el acceso. El atributo en cuestión podría ser accesible incluso por individuos externos a la red social. La segunda columna muestra el número de redes sociales donde el atributo se encuentra disponible a los usuarios mediante la configuración de privacidad por defecto, pero en este caso el usuario puede restringir su acceso a voluntad. La tercera columna muestra el número de redes sociales donde el atributo en cuestión puede ser cumplimentado por los usuarios, pero por defecto su valor no se muestra a todo el mundo. La cuarta columna muestra el recuento de las redes sociales donde el atributo en cuestión no forma parte del perfil del usuario y por lo tanto, su información no se encuentra disponible.

Grado de disponibilidad de atributos personales en las redes sociales

Atributos	Nivel de disponibilidad			
	Siempre disponible	Disponible por defecto	No disponible por defecto	Nunca disponible
Foto personal	9	2	1	0
Localización	5	7	0	0
Sexo	4	6	1	0
Nombre	5	6	1	0
Amigos	1	10	1	0

Atributos	Nivel de disponibilidad			
	Siempre disponible	Disponible por defecto	No disponible por defecto	Nunca disponible
Actividades	2	8	0	2
Edad	2	5	4	1
Escuelas	0	8	1	3
Trabajo	0	6	1	5
Cumpleaños	0	4	7	1
Código postal	0	0	10	2
Correo	0	0	12	0
Teléfono	0	0	6	6
Dirección	0	0	4	8

Consumer Reports en el 2010 publicó una encuesta realizada en 2.000 hogares estadounidenses respecto a qué información personal publicaban en Facebook. En ella se ofrece un punto de vista complementario al mostrado en la tabla anterior. En la tabla siguiente se muestra el tanto por ciento de individuos sobre el total de hogares encuestados que publican abiertamente un cierto atributo en la red social Facebook.

Lectura complementaria

Consumer Reports National Research Center (2010). "Annual State of the Net Survey". *Consumer Reports* (vol. 75, núm. 6).

Usuarios que publican ciertos atributos en las redes sociales

Atributo	Porcentaje
Nombre completo	84%
Fotos personales	63%
Correo	51%
Fecha nacimiento	42%
Fotos de familiares	24%
Nombres de familia y amigos	19%
Trabajo	17%
Nombres de familiares	16%
Dirección de casa	7%
Número de móvil	7%
Número de teléfono fijo	4%

Además de atributos personales, las redes sociales también almacenan las relaciones de amistad entre usuarios y sus intereses en cuanto a música, películas, deportes, etc.

Por último, algunas redes sociales también incluyen información sobre la ubicación de los usuarios en tiempo real.

Blogs, microblogs y foros

Este tipo de aplicaciones (por ejemplo, Twitter, Google Blog) proporcionan un conjunto de atributos personales que es similar al que ofrecen las redes sociales. Sin embargo, en este caso, obtener los intereses de los usuarios puede ser una tarea más compleja, puesto que generalmente el usuario no puede seleccionarlos de forma explícita.

En los blogs y microblogs (y aplicaciones similares), una gran parte de la información personal (como intereses u opiniones) aparece en los mensajes de los usuarios: si un determinado usuario habla en su blog acerca de la navegación, otros usuarios entienden que está interesado en barcos, regatas y, posiblemente, otros deportes que se desarrollen en el mar.

Sin embargo, para un ordenador puede ser más difícil entender los intereses y opiniones que se ocultan en el texto de un blog. El tratamiento de datos de texto y la interpretación de su semántica es una tarea compleja. De hecho, la semántica es un rasgo inherentemente humano que se define mediante un consenso social (Sanchez, Isern, Millán, 2010). En consecuencia, la interpretación semántica de datos en formato texto se basa en evidencias encontradas en una o varias fuentes de conocimiento construidas de forma manual. La idea detrás de este proceso es imitar el razonamiento humano usando el conocimiento implícito o explícito.

Un posible método para lograr esto es el uso del conocimiento estructurado modelado en forma de taxonomías o más generalmente ontologías.

> Las ontologías hacen referencia a la formulación de un exhaustivo y riguroso esquema conceptual dentro de uno o varios dominios dados; con la finalidad de facilitar la comunicación y el intercambio de información entre diferentes sistemas y entidades. Las ontologías se construyen para ser procesadas por las máquinas y por lo tanto, pueden ser aplicadas en este campo que nos ocupa.

En este sentido, las ontologías se han utilizado con éxito en áreas relacionadas con la extracción de información de recursos textuales (Sanchez, Isern, Millán, 2010).

Iniciativas como la web semántica (Berners-Lee, Hendler, Lassila, 2001), han propiciado la creación de muchas ontologías que abarcan desde un ámbito general hasta temas concretos. Gracias a la utilización de estas fuentes de conocimiento, es posible mapear ciertas palabras encontradas en textos generados

por usuarios (por ejemplo, "deportes acuáticos") a los conceptos ontológicos relacionados (por ejemplo, deportes acuáticos: "los deportes que implican una actividad física en el agua").

Adicionalmente, las herramientas de extracción de información de recursos textuales son capaces de obtener la opinión de los usuarios respecto a ciertos temas. Por ejemplo, si un usuario está publicando su opinión sobre una cierta película en Twitter, este tipo de esquemas pueden analizar si la calificación del usuario es positiva o negativa.

Compartición de contenidos multimedia

Este tipo de aplicaciones (por ejemplo, Youtube, Picasa, etc.) se basan en el intercambio de contenidos multimedia (por ejemplo, vídeos, fotos, etc.). En este escenario, la información personal es revelada principalmente debido al uso de metadatos (o *tags*) vinculados a cada archivo multimedia.

> El etiquetado social es una manera informal de asignar etiquetas definidas por el usuario a los distintos elementos compartidos. En lugar de clasificar el contenido publicado de acuerdo con las directrices de clasificación bibliográficas, los usuarios definen sus propios términos de manera informal basándose únicamente en las asociaciones que evoca el elemento a clasificar.

Grootveld y otros (2008) elaboraron un estudio sobre la contribución de las etiquetas sociales, los metadatos de profesionales y los metadatos generados automáticamente en el proceso de recuperación de información de vídeos. En este trabajo, 194 participantes etiquetaron un total de 115 vídeos, mientras que otros 140 participantes realizaron búsquedas en la colección de vídeos para obtener respuestas a ocho preguntas. Los resultados obtenidos muestran que en el contexto actual las etiquetas sociales proporcionan un proceso de recuperación efectivo, mientras que los metadatos generados de forma automática no lo consiguen.

Kakogianni y Soderberg (2010) han proporcionado una categorización de las variables utilizadas por los usuarios en el proceso de etiquetaje. Con el fin de investigar qué elementos utilizan los usuarios para etiquetar, los autores consultaron las etiquetas más populares de Flickr. Todas las etiquetas consultadas se pueden dividir en categorías de acuerdo con el análisis de la imagen en cuestión. La tabla siguiente muestra esta categorización.

Categorías más utilizada para etiquetar los contenidos multimedia

	Categoría	Etiqueta
Quién	Personas	familia, amigos, yo, bebé, chica

	Categoría	Etiqueta
	Animales	gato, perro, animales, pájaro, pájaros
	Edificios	casa, iglesia, museo
	Objetos	flores, agua, flor, comida, coche
	Colores	verde, negro, azul, color, rojo
	Paisaje	sol, cielo, puesta de sol, nubes, playa
Qué	Actividades	viaje, vacaciones, excursión, travesía
	Eventos	boda, fiesta, cumpleaños
Cuándo	Temporada	verano, invierno, primavera, otoño
	Vacaciones	navidad, halloween
	Momento del día	noche, día, atardecer
Dónde	Continente	Europa, Australia, Asia
	País	Japón, Italia, Francia, USA, China
	Lugar específico	parque, jardín, zoo, casa
	Características geográficas	playa, nieve, calle, ciudad, mar
Acerca de	Abstracto	arquitectura, arte, moda, amor

Los autores concluyen que hay una forma de pensamiento similar entre las personas al describir una fotografía. Por supuesto, el número de etiquetas que se pueden utilizar son infinitas, no obstante, a la práctica, los elementos utilizados tienden a centrarse en esas categorías.

De forma adicional, las aplicaciones de intercambio de fotos y vídeos generalmente permiten a los usuarios compartir grandes cantidades de datos de localización para indicar en qué lugar se generaron dichos elementos.

Mensajería instantánea

Los usuarios de estas aplicaciones generalmente rellenan un formulario con algunos atributos personales. Estos atributos son similares a los utilizados en las redes sociales y, por lo tanto, los resultados analizados anteriormente son aplicables en este subapartado.

Este tipo de aplicaciones por lo general ofrecen una lista de contactos al usuario y muestran sus respectivos estados (por ejemplo, en línea, fuera de línea, ocupado, etc.) con el fin de facilitar la comunicación. Esta lista de contactos es una representación directa de las relaciones de amistad de cada usuario y funciona de forma similar a la subred de amigos que se utiliza en las redes sociales.

Adicionalmente, hoy en día las aplicaciones de mensajería instantánea se integran en los móviles y dispositivos siendo capaces de proporcionar datos de localización.

También cabe destacar que mediante estas herramientas es posible extraer los intereses y las opiniones de los usuarios de las conversaciones en formato texto. Estas herramientas utilizan métodos similares a los explicados en el subapartado dedicado a blogs y microblogs. Sin embargo, la complejidad del proceso dependerá de cómo se gestionan dichos mensajes.

Gestión de mensajes

La complejidad del proceso dependerá de cómo se gestionan los mensajes. Por ejemplo, si son almacenados en el servidor durante mucho tiempo (por ejemplo, Skype), el proceso de extracción de información se podrá realizar de forma más sencilla que si los mensajes sólo se almacenan en los ordenadores de los usuarios.

Mundos virtuales y juegos multijugador masivos en línea

Los usuarios de este tipo de juegos rellenan un formulario con algunos atributos personales. Al ser un proceso similar al de las redes sociales, se esperan unos resultados similares en cuanto a la información que contienen estos servicios.

Los mundos virtuales y juegos en línea masivos hasta cierto punto se pueden considerar representaciones gráficas de una red social y por lo tanto, pueden contener datos muy similares a los de una red social estándar (por ejemplo, intereses, topología de amistades, etc.). Es muy frecuente que estas aplicaciones contengan comunicación textual entre los usuarios, de la cual se puedan extraer también opiniones o intereses.

Por lo tanto, el problema de estos entornos no es la inexistencia de información, sino la forma de adquirirla. Por ejemplo, en un mundo virtual los contactos realizados entre usuarios pueden ser considerados como las relaciones de amistad de las redes sociales. No obstante, obtener estas relaciones puede ser muy complejo dependiendo de la estructura del juego analizado. Por el contrario, en las redes sociales esta información aparece perfectamente reflejada y gestionada. Respecto a opiniones e intereses, adquirirlos dependerá también de la capacidad de almacenar y recuperar los mensajes generados por los usuarios.

2.2.2. Creación pasiva: navegación y motores de búsqueda

Dentro de los métodos para crear de forma pasiva la huella de los usuarios nos centraremos en dos muy comunes y prácticamente inherentes al uso de Internet: la navegación entre páginas web y la utilización de motores de búsqueda.

Historial de navegación

En la obtención de los patrones de navegación de los usuarios, estos no intervienen directamente. Por lo tanto, consideramos que esta información se obtiene de los usuarios de forma pasiva y en muchas ocasiones sin su propio conocimiento.

A partir de las páginas visitadas por un usuario, la información básica de su perfil que se obtiene son sus intereses. El proceso de adquirir estos datos se fundamenta en el uso de mecanismos para identificar unívocamente al usuario que accede a distintas páginas web. Una vez el usuario es identificado y las páginas consultadas almacenadas, se puede asociar la temática de dichas páginas con los intereses del usuario analizado.

Ved también

Los distintos mecanismos existentes para adquirir datos (por ejemplo, *http cookies*) han sido detallados en profundidad en el subapartado 1.2 de este módulo.

Motores de búsqueda

Cuando un usuario quiere buscar un cierto término en un motor de búsqueda (por ejemplo, Google), teclea las palabras clave de la consulta en una barra de búsqueda. Entonces, el motor aplica técnicas de recuperación para seleccionar y clasificar los resultados. Después de esto, el usuario evalúa la lista de páginas mostrada y obtiene la información.

Junto con este proceso, el motor de búsqueda construye un perfil del usuario en función de las consultas que realiza.

Ejemplo

Google afirma que sus servidores registran automáticamente las solicitudes formuladas por los usuarios incluyendo la consulta, su dirección IP, su tipo de navegador, el idioma del navegador, la fecha y hora de la solicitud y una o más *cookies* que pueden identificar unívocamente el navegador del usuario.

Una vez el usuario es identificado y todas las consultas que realiza registradas, es posible obtener la temática de las consultas realizadas debido a que generalmente ya se trata de palabras clave que se refieren a algún tema en concreto. En el caso de ser consultas más elaboradas, sería posible extraer su temática aplicando medidas similares a las explicadas en el subapartado 2.2.1 dedicado a blogs y microblogs. En cualquier caso, mediante este proceso es posible conocer los intereses de los usuarios y, de hecho, los motores de búsqueda lo utilizan para proporcionar búsquedas personalizadas.

Un ejemplo de esto ocurre cuando un usuario ha realizado consultas anteriores sobre temas relacionados con la química y realiza una nueva consulta sobre "mercurio". El motor de búsqueda puede utilizar los intereses del usuario extraídos de consultas anteriores para mostrar respuestas relacionadas con el elemento químico antes que respuestas relacionadas con el planeta.

Como se puede observar, este proceso no cuenta con la aprobación directa del usuario y por lo tanto, esta aplicación participa en la creación pasiva de la huella digital del usuario.

2.3. Explotación de los perfiles de usuario

A continuación se introducen y comentan varios escenarios donde los perfiles de usuario son (o pueden ser) utilizados por distintas empresas para obtener beneficios.

2.3.1. Explotación de los intereses

Los intereses de los usuarios son muy valiosos. Se pueden utilizar para proporcionar **publicidad dirigida**. Por ejemplo, las compañías de automóviles muestran publicidad relacionada con aquellas personas que tienen interés en el mundo del automóvil.

Otro ejemplo en este sentido sería el de Facebook. Esta red social obtiene beneficios mediante la personalización de los anuncios que muestra a los usuarios. Cuanta más información muestran los usuarios en sus perfiles, más personalizada es la publicidad que reciben (Wortham, 2010).

Google AdWords

Un ejemplo del uso de los intereses para mejora de la publicidad seria Google AdWords. Esta tecnología es un método que utiliza Google para hacer publicidad patrocinada. Los anuncios que muestra el motor de búsqueda están directamente relacionados con las consultas realizadas por el usuario. Google cobra a cada empresa por cada clic realizado sobre su anuncio.

2.3.2. Explotación de las opiniones

La aplicación de modelos para agregar las opiniones de los distintos colectivos permite obtener información extremadamente importante sobre sus comportamientos y proporciona a las empresas la capacidad de prever tendencias futuras. Además, recolectar grandes cantidades de opiniones sobre ciertos productos en particular resulta de ayuda para diseñar campañas de marketing y publicidad (Adamic, Leskovec, Huberman, 2006).

Como demostración de esta situación, Asur y Huberman (2010) analizaron la capacidad de prever beneficios respecto a los distintos estrenos cinematográficos utilizando los mensajes publicados en Twitter. La primera parte del estudio se basó en entender cómo se construyen las expectativas y la atención sobre una película en particular. La segunda parte se centró en cómo las opiniones de las personas tanto positivas como negativas se propagaban e influenciaban a los demás. Las conclusiones de dicho trabajo fueron las siguientes:

- Las opiniones adquiridas en entornos sociales son indicadores efectivos del comportamiento de las masas en el mundo real.

- La ratio de mensajes relacionados con una cierta película puede ser utilizado para construir un modelo de predicción de los beneficios que esta obtendrá. Además, el modelo generado es más efectivo que el estándar utilizado generalmente por la industria del cine (*Hollywood stock exchange*) para dicho propósito.

Los mercados financieros son otro escenario donde las opiniones de los usuarios de aplicaciones sociales son utilizadas para prever los beneficios obtenidos por las distintas empresas y el éxito de sus productos.

El software diseñado para el análisis lingüístico es utilizado para extraer los sentimientos del mercado. En este sentido, la agencia Dow Jones creó un diccionario de 3.700 palabras que indican cambios en el sentimiento. Ejemplos de estas palabras serían: *fortaleza*, *ganador*, *riesgo* o *conspiración*. Estos programas analizan el contexto de las frases y detectan su tendencia para posteriormente avisar de comportamientos masivos.

Bloomberg

La agencia Bloomberg monitoriza noticias y publicaciones en Twitter y alerta en el caso de que exista una gran cantidad de individuos enviando mensajes respecto a un cierto tema (por ejemplo, "Apple").

Otro escenario donde se han demostrado eficaces las opiniones de los usuarios es en la organización de viajes y vacaciones. La firma PhocusWright (2011), encargada de realizar investigaciones de marketing orientado a viajes, ha confirmado que los entornos sociales y las opiniones que generan producen una enorme influencia en los hábitos de compra de este tipo de productos.

En este sentido, los investigadores concluyen que los usuarios generalmente otorgan mucha importancia a las valoraciones proporcionadas por amigos o individuos similares a ellos mientras que la publicidad proporcionada a través de los medios habituales recibe una atención menor.

2.3.3. Explotación de la localización

Los servicios de localización se han convertido en una herramienta de marketing muy destacada para pequeñas empresas que dependen del tráfico de clientes, como restaurantes o bares (Pattison, 2010). El crecimiento de estos servicios se fundamenta en su capacidad para explotar los nuevos dispositivos móviles que poseen una excelente conectividad y que han sido adoptados por una gran cantidad de personas.

Los servicios de localización pueden jugar muchos papeles. Ofrecen al cliente herramientas para relacionarse, programas de puntos, guías de ciudades o valoraciones de sitios en particular. Una de sus aplicaciones más interesantes para las pequeñas empresas comentadas anteriormente es que permiten adquirir datos de los clientes que se encuentran dentro de su zona de influencia y presentarles ofertas, vales descuento, etc. Por ejemplo, si un usuario se encuentra a pocos metros de distancia respecto a cierto restaurante, este puede recibir un aviso en su dispositivo móvil alertando del restaurante y de una cierta oferta adaptada a sus gustos personales.

Básicamente, estas aplicaciones permiten a los negocios conectar con las personas y fidelizar a los clientes.

2.3.4. Explotación de los perfiles de forma global

En los anteriores subapartados se han introducido casos centrados en la explotación de una parte del perfil. No obstante, existen escenarios donde los beneficios se obtienen de la adquisición de todo el perfil completo y la utilización de los diversos elementos que la componen dependiendo del uso que le quiera dar la empresa adquiriente.

La utilización de perfiles para mejorar las estrategias de contratación de empleados es una realidad en la actualidad. Además, se constata que el uso de estas nuevas técnicas ayuda a reducir los costes en el proceso de contratación.

Lectura recomendada

(Mayo, 2010). "Social savvy recruiters utilising social media in their recruitment strategy". *Personnel Today*.

Las compañías de seguros también utilizan los perfiles de los usuarios para obtener evidencias de fraude y reducir posibles pérdidas económicas en este sentido (Li, 2011). La tendencia incluso va un poco más lejos y las compañías del sector están estudiando la utilización de los perfiles para fijar precios personalizados para cada cliente (Beattie, Stagg-Macey, 2010). En este caso, los datos más relevantes a extraer de los perfiles serían estilos de vida y problemas médicos.

La compra de perfiles de usuario a las empresas que ofrecen servicios sociales es significativamente común: por ejemplo, AOL recibe 1.000 peticiones de datos para ser utilizados en delitos penales o civiles (Hansell, 2006). Facebook recibe entre 10 y 20 peticiones al día. Cabe destacar que los perfiles pueden ser solicitados tanto por empresas como por gobiernos interesados en cuestiones de seguridad nacional.

3. Definición y políticas de privacidad

En este apartado se introduce el concepto de privacidad y sus implicaciones respecto a la adquisición y explotación de datos personales mediante los métodos explicados anteriormente. También se detallan los dos niveles de privacidad que se tienen en cuenta en el diseño de herramientas para la protección de la privacidad. Finalmente, se analiza de forma general las políticas de privacidad que se aplican a los datos personales de los usuarios recolectados por las empresas.

3.1. Definición de privacidad

> La privacidad es un derecho humano fundamental. Este concepto engloba otras ideas relacionadas, como la dignidad humana y la libertad de expresión. La privacidad se ha convertido en uno de los derechos humanos más importantes de la edad moderna (Rotenberg, 2000).

La privacidad es reconocida en diversas regiones y culturas de todo el mundo. Es un derecho protegido por la Declaración Universal de los Derechos Humanos, el Pacto Internacional de Derechos Civiles y Políticos, y por muchas otras organizaciones internacionales y tratados regionales de derechos humanos. Casi todos los países del mundo incluyen el derecho a la privacidad en su constitución. Como mínimo, esas leyes contemplan los derechos de inviolabilidad del domicilio y el secreto de las comunicaciones. Además, las constituciones redactadas más recientemente incluyen derechos específicos para el acceso y control de la información personal por parte del propio individuo. En el caso de los países en los cuales la constitución no menciona explícitamente el derecho a la privacidad, los tribunales han encontrado este derecho reflejado en otras disposiciones (Laurant, 2003).

Se pueden diferenciar cuatro aspectos diferentes respecto a la privacidad:

- **Privacidad de la información**. Este aspecto implica el establecimiento de normas que rigen la recolección y tratamiento de datos personales tales como datos económicos, médicos y registros gubernamentales.

- **Privacidad corporal**. Este aspecto abarca la protección física de las personas contra procedimientos invasivos como análisis genéticos, pruebas de drogas, etc.

- **Privacidad de las comunicaciones**. Este aspecto incluye la seguridad y privacidad del correo, teléfono, correo electrónico y cualquier otra forma de comunicación.

- **Privacidad territorial**. Este aspecto implica el establecimiento de límites a la intrusión en los entornos domésticos y otros escenarios, como el lugar de trabajo o el espacio público. Esto incluye búsquedas, videovigilancia y controles de identidad.

La huella digital tratada en este módulo se enfoca principalmente en su creación, utilización y control en el ámbito de Internet. Por lo tanto, solo los aspectos relacionados con la privacidad de la información y la privacidad de las comunicaciones se aplican en los escenarios considerados.

La privacidad de las comunicaciones se proporciona en su mayor parte aplicando medidas de seguridad que limiten los ataques a las transmisiones. Más concretamente, las soluciones asociadas a este aspecto estarían relacionadas con la utilización del hardware adecuado, actualizaciones de software y herramientas específicas de seguridad informática.

En el caso de la privacidad de la información, este aspecto está claramente relacionado con el problema del manejo de datos de carácter personal facilitados tanto por los propios usuarios (creación activa de la huella digital) como obtenidos de forma invisible para el usuario (creación pasiva de la huella digital). Por lo tanto, este apartado se centra en este último aspecto de la privacidad.

Ved también

En el apartado 4 se ofrecen posibles soluciones para mitigar el problema de la explotación de datos personales.

En cuanto al diseño de esquemas que proporcionen privacidad de la información, se pueden definir dos niveles de privacidad: **anonimato** y **no-enlazabilidad**. Un sistema mantiene el anonimato de los usuarios cuando impide que su identidad se haga pública. No-enlazabilidad es un nivel más fuerte que anonimato y se refiere al hecho de que las diferentes interacciones de un mismo usuario con un cierto sistema no puedan ser relacionadas entre sí. La no-enlazabilidad impide el seguimiento de usuarios y la creación de perfiles.

Para mostrar la diferencia entre los dos niveles, se pueden introducir dos medidas de privacidad sencillas, una para cada nivel. Para proporcionar anonimidad, el uso de seudónimos puede ser suficiente. El uso de un seudónimo permite al usuario que sus actividades digitales se asocien a una identidad falsa proporcionando privacidad a su identidad real. No obstante, el uso de seudónimos no proporciona no-enlazabilidad dado que todas las actividades que se realicen con ese seudónimo se pueden relacionar entre ellas. Si bajo alguna circunstancia, una tercera entidad es capaz de descubrir la identidad real bajo el seudónimo, ésta también será capaz de enlazar todas las actividades realizadas bajo dicho seudónimo con la identidad real del usuario. Para evitar

esta situación, un sistema que proporcione no-enlazabilidad debería cambiar frecuentemente los seudónimos utilizados por el usuario siguiendo algún mecanismo que impida que puedan ser relacionados entre sí.

3.2. Políticas de privacidad: quién es el propietario de la información personal

En este subapartado trataremos las políticas de privacidad de las empresas que recopilan datos personales de los usuarios y que construyen su huella digital tanto de forma activa como pasiva. Estas políticas son muy relevantes para decidir la propiedad de los usuarios sobre sus propios datos y la forma en la que pueden ser utilizados.

3.2.1. Políticas respecto a la creación activa de la huella digital

Respecto a la creación activa de la huella digital, los usuarios de redes sociales generalmente creen que la información que publican (por ejemplo, el contenido de su perfil, fotos, etc.) es de su propiedad y control. No obstante, esto no siempre es así.

En el 2009, Facebook realizó un cambio en sus términos de uso. Uno de los nuevos párrafos añadidos decía literalmente lo siguiente.

> "Las siguientes secciones sobrevivirán al término de su cuenta en los Servicios de Facebook: conducta prohibida, contenido de usuario, su política de privacidad, créditos de regalos, propiedad, derechos de propiedad, licencias, peticiones, disputas de usuario, quejas, indemnizaciones, renuncia general, limitación de responsabilidad, terminación y cambios en el servicio de Facebook, servicio de arbitraje, ley de administración, lugar y jurisdicción y otros".

Anteriormente a este cambio, si el usuario decidía cerrar su cuenta, el contenido que hubiese publicado también desaparecía, pero con dicho cambio, el contenido continuaba perteneciendo a Facebook incluso después de cerrar la cuenta, así que la empresa podía vender las fotos del usuario o utilizarlas para publicidad sin que dicho usuario recibiese dinero a cambio. La ola de críticas que la modificación supuso obligó a la empresa a restaurar los términos originales, no obstante, esta situación deja patente la importancia de los términos de uso de las aplicaciones sociales que generalmente los usuarios aceptan sin leer.

Las condiciones de uso de las distintas redes sociales varían de forma significativa entre sí.

LinkedIn

> LinkedIn, por ejemplo, exige a sus usuarios a conceder a la empresa una "licencia no exclusiva, irrevocable, mundial, perpetua, ilimitada, transferible, transmisible y que proporciona a la empresa derecho a copiar, modificar parcialmente, mejorar, distribuir, publicar, eliminar, retener, agregar, usar y comercializar de cualquier manera conocida ahora o en el futuro sin ningún tipo de consentimiento, previo aviso o compensación alguna para el usuario o para terceros".

Twitter

Twitter por su parte "no reclama derechos de propiedad intelectual sobre el material proporcionado por los usuarios" y añade que los usuarios "pueden eliminar su perfil en cualquier momento mediante la supresión de su cuenta y que dicha acción también elimina cualquier texto e imágenes que los usuarios hayan almacenado en el sistema".

Como conclusión, los usuarios de este tipo de aplicaciones deben revisar los términos de uso y políticas de privacidad de las empresas que ofrecen estos servicios con el objetivo de conocer cómo su privacidad será tratada en el futuro y las implicaciones que el contenido que hagan público tendrá en su huella digital.

3.2.2. Políticas respecto a la creación pasiva de la huella digital

Respecto a la **creación pasiva de la huella digital**, a modo general se puede estudiar la política de privacidad aplicada por Google a sus distintos servicios (por ejemplo, motor de búsqueda, calendario, blog, correo electrónico, localización, mapas, etc.), los cuales utilizan distintos métodos de identificación de usuarios (por ejemplo, *http cookies*, credenciales de acceso) y adquieren múltiples datos personales en ciertos casos sin la colaboración activa de los usuarios.

En las condiciones de servicio de Google, concretamente en el apartado dedicado a información personal y privacidad, se explica que "el usuario de los servicios ofrecidos por la empresa acepta el uso de sus datos de acuerdo con las políticas de privacidad de Google".

Las políticas de privacidad de Google se basan en cinco principios que describen la forma en que Google trata la privacidad y la información de los usuarios en todos sus productos:

1) Utilizar la información para ofrecer a los usuarios productos y servicios valiosos.

2) Desarrollar productos que reflejen prácticas y estándares de privacidad firmes.

3) Recopilar información personal de forma transparente.

4) Ofrecer a los usuarios alternativas significativas para proteger su privacidad.

5) Supervisar de forma responsable la información que almacenamos.

Se puede observar que los puntos 1 y 3 hacen referencia explícita a la adquisición de datos personales y su utilización para aumentar el valor de los productos y servicios aplicados. La ambigüedad de dichos principios permite a Google explotar los datos personales para obtener beneficios tanto para el usuario como propios.

Respecto al tipo de información recolectada por Google, su política de privacidad explicita los siguientes elementos:

- **Información que proporciona el usuario**: al registrarse para obtener una cuenta de Google, el solicitante debe proporcionar información personal. Es posible que se combinen los datos que proporciona el usuario a través de su cuenta con la información procedente de otros servicios de Google o de terceros para ofrecerle una óptima experiencia y mejorar la calidad de los servicios proporcionados. Para determinados servicios, se le podía ofrecer al usuario la oportunidad de decidir si desea o no la realización de dicha combinación de datos.

- ***Cookies***: al acceder a Google, se envían una o varias *cookies* a su equipo o a otro dispositivo. Las *cookies* se utilizan para mejorar la calidad del servicio, incluidos el almacenamiento de las preferencias del usuario, la mejora de los resultados de búsqueda y de la selección de anuncios y el seguimiento de las tendencias del usuario, como por ejemplo, el tipo de búsquedas que realiza. Google también utiliza *cookies* en los servicios publicitarios para que anunciantes y editores puedan ofrecer y administrar anuncios en Internet y en los servicios de Google.

- **Información de registro**: cuando el usuario accede a los servicios de Google a través de un navegador, una aplicación u otro cliente, los servidores de la empresa registran automáticamente cierta información. Esta información puede contener la solicitud web, la interacción con un servicio, la dirección IP, el tipo y el idioma del navegador, la fecha y la hora de la solicitud y una o varias *cookies* que permiten una identificación exclusiva del navegador o de la cuenta.

- **Comunicaciones de usuarios**: cuando el usuario envía mensajes de correo electrónico u otras comunicaciones a Google, la empresa puede conservar esa información para procesar sus consultas, responder a sus peticiones y mejorar nuestros servicios. Si el usuario envía o recibe mensajes SMS de alguno de los servicios que disponen de esta función, Google puede recopilar y conservar la información asociada a esos mensajes, como el número de teléfono o el contenido del mensaje.

- **Sitios de Google afiliados en otros sitios**: Google ofrece algunos de sus servicios en otros sitios web o a través de ellos. Es posible que la información personal que el usuario facilita a través de estos sitios web se envíe a Google para poder prestar el servicio.

- **Aplicaciones externas**: Google puede poner a disposición de los usuarios aplicaciones externas a través de sus servicios. La información recopilada por Google al habilitar una aplicación externa se procesará de acuerdo con lo estipulado en esta política de privacidad. La información recopilada por

el proveedor de la aplicación externa se regirá por sus propias políticas de privacidad.

- **Datos de ubicación**: Google ofrece servicios que tienen registrada la ubicación del usuario, como Google Maps y Latitude. Si se utilizan estos servicios, Google puede recibir información sobre la ubicación real del usuario o información que se podría utilizar para determinar su ubicación aproximada.

- **Número de aplicación exclusivo**: algunos servicios, como la barra Google, incluyen un número de aplicación exclusivo que no está asociado al usuario ni a su cuenta. Este número y la información sobre la instalación (por ejemplo, el tipo de sistema operativo o el número de versión) se pueden enviar a Google al instalar o desinstalar ese servicio, o cuando dicho servicio establece contacto con los servidores de Google de forma periódica (por ejemplo, para solicitar actualizaciones automáticas del software).

- **Otros sitios**: esta política de privacidad se aplica únicamente a los servicios de Google. Google no ejerce ningún control sobre los sitios que aparecen en los resultados de búsqueda, los sitios que incluyen aplicaciones, productos o servicios de Google ni los enlaces incluidos en sus servicios. Es posible que estos otros sitios envíen sus propias *cookies* u otros archivos al equipo del usuario, recopilen datos o soliciten el envío de información personal.

Respecto al uso que se le da a la información recopilada, Google la utiliza para los siguientes fines:

- Proporcionar, mantener, proteger y mejorar sus servicios (incluidos los servicios publicitarios) y desarrollar nuevos servicios.
- Proteger los derechos o la propiedad de Google o de los usuarios.

Si esta información se va a utilizar con fines distintos al objetivo para el que se ha recopilado, se solicitará el consentimiento previo del usuario.

Localización de los datos

Google procesa la información personal en los servidores de los Estados Unidos de América y de otros países. En algunos casos, la información personal se procesa fuera del país del usuario.

4. Técnicas para proporcionar privacidad

Los mecanismos que se utilizan para controlar el contenido de la huella digital varían sensiblemente en función del proceso de creación sobre el cual se aplican. En este sentido distinguiremos las medidas diseñadas para controlar la creación activa de la huella digital (este proceso se da principalmente con el uso de aplicaciones sociales) de las medidas que evitan la creación pasiva de la huella (este proceso se fundamenta principalmente en la identificación del *browser* del usuario y el análisis de su comportamiento).

4.1. Control de la creación activa de la huella digital

En general, los mecanismos que se utilizan en la actualidad para preservar la privacidad de los usuarios, que de forma activa publican recursos y datos personales en Internet, se fundamentan en tres modelos básicos:

- aplicación del sentido común,
- aplicación de medidas basadas en el control de acceso a recursos y, finalmente,
- utilización de técnicas de perturbación de datos.

4.1.1. Sentido común

Este modelo se basa en no generar contenido susceptible de ser explotado por terceros. Siguiendo esta idea, *Consumer Reports* presenta siete consejos básicos para evitar a los usuarios problemas futuros de privacidad y seguridad. Estos consejos se enumeran a continuación:

Lectura recomendada

(2010). "7 things to stop doing now on Facebook". *Consumer Reports* (vol. 75, núm. 6).

1) Los usuarios no deben utilizar *passwords* sencillos (débiles).

2) Los perfiles de usuario no deben contener fechas de nacimiento completas.

3) Los usuarios deben utilizarse los mecanismos de privacidad proporcionados por las aplicaciones basadas en la Web 2.0 (por ejemplo, redes sociales).

4) Los usuarios no deben incluir nombres o datos de familiares menores de edad en el contenido multimedia publicado en la Web 2.0.

5) Los usuarios no deben indicar la ausencia o presencia en el hogar.

6) Los usuarios deben impedir que los motores de búsqueda (por ejemplo, Google) los encuentren.

7) Los usuarios menores de edad no deben utilizar las redes sociales sin ser supervisados por un adulto.

4.1.2. Control de acceso a los datos personales

Este modelo de medidas de privacidad permite a los usuarios seleccionar las personas que pueden acceder a un dato personal en particular. De esta manera, los recursos publicados por el usuario no se modifican y, por lo tanto, un atacante con suficientes derechos de acceso sería capaz de obtener un perfil completo y real del usuario.

Las medidas de control de acceso son básicas en el proceso de creación activa de la huella digital. Más concretamente, son utilizadas generalmente en aplicaciones Web 2.0 como las redes sociales. Cabe destacar que al ser medidas aplicadas conscientemente por el usuario, no pueden ser utilizadas en el proceso de creación pasiva de la huella digital.

Las medidas de control de acceso se fundamentan en tres tecnologías distintas:

- **Configuración individual de privacidad**. Este mecanismo es el más sencillo de implementar y, de hecho, es el utilizado por defecto en las redes sociales estándar y aplicaciones relacionadas. Se basa en seleccionar entre múltiples opciones de privacidad proporcionadas por la propia aplicación la configuración idónea para cada persona. Asumiendo que el *host* de la aplicación web sea honesto, el sistema es efectivo. No obstante, un estudio elaborado por Bilton (2010) pone de relevancia la complejidad de los controles de privacidad proporcionados por Facebook. Bilton indica que un usuario que desee evitar la diseminación de la mayoría de su información personal deberá utilizar más de 50 botones y seleccionar entre más de 170 opciones distintas. Esta complejidad facilita el error y aumenta la posibilidad de asignar derechos erróneos a entidades deshonestas. Existen también redes sociales basadas en una arquitectura distribuida (por ejemplo, Diáspora) donde cada usuario gestiona su propio servidor web que contiene todos sus recursos y a los cuales puede aplicar las medidas de acceso que considere necesarias. En este caso, la inexistencia de un servidor central proporcionado por una empresa en particular facilita un mejor control por parte del usuario.

- **Utilización de criptografía**. En este caso, los contenidos a publicar se cifran utilizando una cierta clave secreta que solo poseerán aquellas personas autorizadas a acceder a dicho recurso (Graffi y otros, 2009). La criptografía en uso puede ser tanto simétrica como asimétrica. Esta medida de privacidad podría ser aplicada a los contenidos independientemente de la aplicación web, por lo tanto, la seguridad proporcionada es mayor que en el caso anterior. Adicionalmente, la complejidad del sistema se basa en elegir quién tiene acceso a cierto recurso, por lo tanto, la complejidad en la configuración no es especialmente relevante. Como grave inconveniente

de este tipo de medidas podemos indicar que no todas las aplicaciones Web 2.0 las soportan y, por lo tanto, su implantación está fuertemente limitada.

- **Calidad de las relaciones entre el usuario y el resto de entidades**. En este caso, el acceso a los recursos se consigue mediante el análisis de la calidad de la relación entre el usuario que publica el recurso y la entidad que intenta acceder a dicho elemento (Banks y Wu, 2009). Dependiendo de la calidad de dicha relación, el acceso se denegaría o aceptaría. Esta solución se basa en la utilización de un sistema de análisis y contabilización de confianzas. Su objetivo es automatizar el proceso de configuración de la privacidad.

4.1.3. Perturbación de los datos personales

Las medidas basadas en la perturbación se centran en modificar los datos personales para aumentar su ambigüedad y, por lo tanto, la privacidad de los usuarios que los han generado. Este proceso se conoce como generalización.

La generalización de atributos personales es comúnmente utilizada por los métodos de perturbación de bases de datos (Domingo-Ferrer y otros, 2008). Aunque dichas técnicas no se utilizan generalmente para el control de la huella digital en Internet, existen algunos trabajos en la literatura que utilizan este tipo de medidas hasta un cierto nivel.

En este sentido, Hay y otros (2008) presenta un trabajo basado en la anonimización de redes sociales, centrándose específicamente en la perturbación de la red de amigos de cada usuario. Adicionalmente, este trabajo propone la generalización, eliminación y randomización de los atributos personales del usuario.

También es común aplicar técnicas de perturbación a los datos referentes a la localización de los usuarios. Dichos datos se asocian a los recursos publicados por los usuarios. Las medidas propuestas en este ámbito se centran en reducir el nivel de detalle de la localización: hablamos de proximidad en vez de localización precisa (Glassman, 2010).

4.2. Control de la creación pasiva de la huella digital

La generación pasiva de la huella digital de un usuario depende principalmente de la capacidad de identificar correctamente sus distintas interacciones mediante el uso de *cookies* u otros métodos. Por lo tanto, las medidas para evitar o controlar este proceso se basan en impedir la identificación correcta por parte de la entidad que realiza el seguimiento del usuario.

Los sistemas de identificación demuestran que medidas sencillas, como el uso de seudónimos por parte de los usuarios, son totalmente ineficaces debido a que la identificación y análisis del rastro del usuario generalmente ocurre sin que la víctima se percate de la situación. Es la propia máquina del usuario la que delata la identidad real a pesar de que la persona se esconda detrás de un seudónimo. Por ello, las herramientas para preservar la privacidad deben ser más sofisticadas y focalizarse en que la entidad que interactúa con la máquina del usuario no pueda identificar a dicha máquina. Este proceso por extensión proporcionará anonimidad al propietario.

Ved también

Los sistemas de identificación se estudian en el subapartado 1.2 de este módulo.

En la literatura, las herramientas que proporcionan estas características se enmarcan dentro de los esquemas basados en canales anónimos.

En sus inicios, los canales anónimos se centraron en el correo electrónico, debido a la gran importancia de este sistema de comunicación. La primera aplicación práctica surgió gracias a las investigaciones de Paul Baran (1964), que creó un sistema con el que dos personas se podían comunicar mediante la participación de una tercera parte de confianza. Esta entidad se encargaba de recibir las comunicaciones y reenviarlas a su destinatario sin que este (ni nadie que pudiera estar escuchando el canal) supiera que había enviado inicialmente el mensaje. Esto se conseguía mediante la eliminación de la cabecera de identificación e introduciendo los datos de la entidad de confianza. Esta tecnología evolucionó hasta llegar a la actual, conocida como *remailer*.

En 1981, David Chaum presentó en su trabajo "Untraceable electronic mail, return addresses, and digital pseudonymous", una solución para enviar información de forma segura y anónima, iniciando así la investigación sobre los canales anónimos. Esta solución se basaba en la utilización de criptografía asimétrica o criptografía de clave pública para crear un sistema de "mix". Desde entonces se ha avanzado mucho en esta área de la privacidad en Internet y han sido muchas las soluciones planteadas tanto desde un punto de vista teórico como práctico.

Posteriormente, gracias a la aparición de nuevas herramientas y aplicaciones de Internet, surgió la necesidad de garantizar la privacidad durante la utilización de las mismas. En la actualidad, los canales anónimos permiten realizar

de forma anónima todo tipo de actividades en Internet, esto incluye el acceso a páginas web, utilización de motores de búsqueda, envío de correos electrónicos, etc.

De forma general, podemos clasificar los distintos tipos de canales anónimos en tres grupos dependiendo de la tecnología en la que se basan. Estos son:

- nodo central de confianza;
- mix Networks, y
- *onion routing*.

4.2.1. Nodo central de confianza

Dentro de esta categoría encontramos la utilización de *proxies* para suplantar al usuario real delante de un proveedor de servicios web.

Un *proxy* consiste en un servidor que se encarga de aceptar conexiones de un grupo de usuarios y reenviarlas al destino solicitado. El *proxy* esconderá las direcciones IP de los usuarios sustituyéndolas por la suya propia. De esta manera todas las conexiones de los usuarios estarán asociadas a una dirección que no es la suya, manteniendo su privacidad. Los *proxies* se utilizan generalmente para conseguir una navegación web anónima (por ejemplo, acceso a páginas web o utilización de motores de búsqueda).

Recientes innovaciones en este campo han permitido a los *proxies* ofrecer servicios adicionales como son: autorizar filtraciones de contenidos (bloqueo de *cookies*; bloqueo de contenido para adultos, etc.), control de acceso (es necesario iniciar sesión para utilizar el *proxy*), etc.

Cabe destacar que uno de los problemas asociados a los métodos que dependen de un servidor o servidores de confianza es que la amenaza contra la privacidad del usuario se traslada del proveedor de servicios a dichas entidades que conocen todas las interacciones de los usuarios y pueden generar sus perfiles.

En el campo de los *remailers* encontramos varias propuestas. Anon.penet.fi fue el primer *remailer* "honesto". El sistema se basaba en un servidor que contenía una tabla de correspondencia entre direcciones de correo y seudónimos para que los usuarios pudieran enviar y recibir mensajes sin necesidad de identificarse. Anon.penet.fi garantizaba el anonimato eliminando cualquier información que pudiera identificar al usuario emisor (cabeceras, datos dentro del mensaje, etc.).

Remailer

Un *remailer* es un servidor que recibe mensajes, los procesa eliminando las cabeceras, y los dirige hasta el destinatario proporcionando anonimato a los participantes.

Generalmente se aplican técnicas criptográficas y se combinan varios *remailers* para conseguir un mayor grado de anonimato.

Este sistema presentaba importantes vulnerabilidades siendo la más destacada la necesidad de guardar esa tabla que claramente comprometía la privacidad de los usuarios si algún sujeto tenía acceso a ella. Debido a ello, se diseño un nuevo *remailer* llamado Cypherpunk (Parekh, 1996) que introducía importantes cambios respecto a Anon.penet.fi. Las características principales son las siguientes:

- Elimina la tabla de correspondencia.
- El proceso de retransmisión acepta el uso de uno o varios nodos *remailer* encadenados.
- Los elementos identificativos se eliminan y se utiliza criptografía asimétrica para hacer llegar el mensaje al nodo *remailer* deseado.
- Se introduce la posibilidad de introducir órdenes para cada nodo *remailer* de la cadena en su capa respectiva de cifrado. De esta manera, al ciframos el mensaje con la clave pública de un *remailer* concreto se puede indicar alguna función dentro del mensaje (esperar un tiempo para reenviar el mensaje, añadir nuevas cabeceras, etc.).

Cypherpunk es un método más robusto que el anterior, no obstante, también sufre de algunas vulnerabilidades. Por ejemplo, es débil contra un atacante pasivo que escuche la entrada y salida de los *remailers*. Si un nodo envía los mensajes a medida que los recibe es sencillo enlazar la entrada con la salida y descubrir el contenido de los mensajes. Otro grave problema de este esquema es que no tiene en cuenta el tamaño de los mensajes y es posible asociar mensajes de entrada y de salida por el número de bits que ocupan.

Los servidores Nym (Kaashoek, Mazieres, 1998) son un esquema con ciertas semejanzas al *remailer* Cypherpunk. La característica principal de estos servidores es que no mantienen un registro de las personas que utilizan su servicio, y por tanto, en ningún momento se auditan las comunicaciones, manteniendo el anonimato de los usuarios. Su funcionamiento es el siguiente:

- El primer paso es crear una dirección "nym" y configurarla creando un par de claves PGP.
- Este par de claves serán enviadas al servidor nym junto con instrucciones (*reply block*) para informarle de dónde se puede encontrar al usuario para reenviar los mensajes de respuesta. El servidor responde que ha recibido la información.
- A partir de ese momento, el servidor guarda el *reply block* relacionado con la dirección de correo anónima correspondiente.

- Cuando llega un mensaje destinado a una dirección en concreto, este no se guarda sino que se envía directamente al usuario correspondiente utilizando las instrucciones almacenadas. Este comportamiento provoca una debilidad similar a la del Cypherpunk.

Crowds (Reiter y Rubin, 1998) es un sistema colaborativo que pretende proporcionar privacidad a los usuarios que acceden a la web. En este esquema, cada nodo contacta con un servidor central y recibe una lista de participantes. El usuario entonces envía su petición de web a través de otro usuario elegido al azar. Cada nodo que recibe una petición decide al azar si la reenvía a otro nodo de la red o envía la petición directamente al sitio web. La respuesta del sitio web se envía al usuario iniciador siguiendo el mismo camino pero en sentido inverso. Todas las peticiones que atraviesan la red se envían cifradas utilizando criptografía simétrica entre pares de nodos. Los problemas asociados a esta propuesta son:

- El servidor central actúa como cuello de botella.

- La utilización de parejas de claves entre usuarios y que el tipo de grafo que forman los nodos sea completo implica que cada nodo debe guardar un gran número de claves criptográficas.

- Es vulnerable al **ataque del predecesor**: un grupo de atacantes pueden entrar en la red y esperar a que se formen las cadenas de nodos para el envío de mensajes. Si un cierto usuario envía muchos mensajes, aparecerá en la mayor parte de las cadenas de envío y las posibilidades de que los atacantes lo detecten aumentarán. Este ataque depende principalmente del número de atacantes que cooperan y del número de usuarios totales del sistema.

Dentro de esta línea, cabe destacar la existencia de un esquema también basado en el uso de un servidor central de confianza, pero que se centra específicamente en el envío de consultas a motores de búsqueda (por ejemplo, Google). El protocolo *useless user profile* (*UUP*) (Castellà-Roca y otros, 2009) sigue una idea similar a Crowds, ya que un cierto usuario interesado en enviar una consulta no la enviará por sí mismo sino que otro usuario lo hará por él. No obstante, en este caso, la red es completamente dinámica: cuando varios usuarios quieren consultar algo al motor de búsqueda, se ponen en contacto mediante el servidor central que los ayuda a formar un grupo. Una vez formado, los miembros del grupo utilizan un protocolo criptográfico para intercambiarse las consultas entre sí sin saber qué consulta corresponde a cada individuo. Finalmente, cada usuario envía al motor de búsqueda la consulta que le ha correspondido y hace *broadcast* de la respuesta al resto de miembros del grupo. Cada usuario recibe solo la respuesta que corresponde a su pregunta y descarta el resto. Al finalizar este proceso, el grupo desaparece. Si uno de los usuarios quiere realizar otra consulta debe solicitar al servidor central un nuevo grupo.

Este sistema propuesto es muy interesante debido a que elimina los problemas de Crowds respecto al ataque del predecesor y la gran cantidad de claves necesarias. No obstante, este esquema no considera atacantes internos y requiere investigación adicional para cubrir este tipo de atacantes manteniendo un tiempo de ejecución razonable.

4.2.2. Mix networks

Las redes mix se fundamentan en el uso de un grupo de nodos o mixes interconectados entre sí y que forman una red en la cual cada nodo o mix oculta la correspondencia de entrada y salida de sus mensajes mediante criptografía. El objetivo es que los mensajes de salida de la red no puedan relacionarse con los de entrada y por lo tanto, el proveedor de servicios no sea capaz de identificar el emisor original del mensaje. El comportamiento general de las *mix networks* con la presencia de nodos mix deshonestos se basa en que mientras exista una cierta cantidad de nodos honestos en la red, la anonimidad de los usuarios quedará garantizada.

Chaum (1981) fue el primero en introducir este concepto que acabó siendo adoptado y mejorado por otros investigadores.

El primer esquema a considerar fue propuesto por Lance Cottrell y otros (2003) con el sistema llamado Mixmaster. Esta propuesta planea solucionar las vulnerabilidades de Cypherpunk:

- Mixmaster no reenvía automáticamente los mensajes que recibe, sino que los encola hasta que tiene un número determinado. Una vez alcanza la cantidad adecuada, los reenvía al siguiente nodo de forma aleatoria.

- Otro cambio importante es que se especifica un tamaño uniforme por los mensajes que se envían. Si el mensaje es más pequeño, añade bits de relleno y si es demasiado grande, divide el mensaje en bloques del tamaño correspondiente. De esta manera se evita que se puedan asociar mensajes de entrada con los de salida por su tamaño.

- Cada mix sólo conoce el siguiente mix del camino. De esta manera en caso de que uno de los nodos sea malicioso, éste no podrá encontrar la ruta hasta el origen ni el destino del mensaje.

Este sistema cubre los ataques de repetición de mensajes asignando identificadores a los mensajes, y obligando a cada nodo mix a comprobar en una tabla interna si ese identificador ha sido recibido anteriormente. En caso de estar repetido, el mensaje se descarta.

La vulnerabilidad principal del sistema es que solo proporciona anonimidad al que envía el mensaje, se podría decir que sólo cubre el camino de ida del mensaje pero no una posible respuesta.

El sistema Mixminion (Danezis, 2003) se diseñó para solucionar la inexistencia de anonimidad en el camino de respuesta de Mixmaster. Las características principales del sistema son las siguientes:

- Mixminion introduce un sistema de respuesta a mensajes que garantiza el anonimato del emisor y del receptor estableciendo un canal bidireccional de comunicación.

- Las transmisiones en Mixminion se basan en el protocolo TCP y la seguridad en los enlaces se consigue mediante el protocolo TLS.

- En Mixmaster, se evitaba el ataque de repetición guardando los identificadores de los mensajes recibidos. Dado que los identificadores no se guardan indefinidamente, la posibilidad de este tipo de ataques aun persiste. Mixminion soluciona esta situación mediante la asociación de claves a mensajes y un proceso de actualización de dichas claves. Los mensajes que llegan con una clave antigua se descartan automáticamente.

Las redes mix funcionan como una capa de ofuscación entre el usuario que envía el mensaje y el que lo recibe. Uno de los problemas que se presentan en este tipo de aplicaciones es proporcionar garantías a los usuarios de que sus mensajes han sido procesados correctamente por la red. En un escenario donde todos los nodos fuesen honestos esto no sería necesario, no obstante, este requisito no es muy realista, por lo tanto, presentar pruebas de que el proceso se ha seguido correctamente es importante. En este sentido, la red mix propuesta por Chaum incluía un sistema de recibos firmados por los nodos mix diseñado para demostrar su comportamiento honesto. Este campo de trabajo se denomina: "Robust and verifiable mix constructions".

Lectura recomendada

G. Danezis; C. Diaz; P. Syverson (2010). "Systems for Anonymous Communication". *CRC Handbook of Financial Cryptography and Security* (págs. 341-390).

4.2.3. Onion routing

Las redes mix generalmente incorporan un coste en tiempo elevado debido a que cada nodo mix espera un cierto tiempo para obtener varios mensajes, y una vez tiene el número deseado los procesa y reenvía en el orden decidido. Los sistemas basados en *onion routing* eliminan este retraso para poder aumentar la velocidad de transmisión: los nodos de estos sistemas no ofuscan el orden de entrada y salida de los mensajes.

En resumen, el *onion routing* se basa en el uso de circuitos bidireccionales y de baja latencia para proporcionar anonimidad a los usuarios. Lo que estos esquemas ofuscan es la ruta de nodos seguida por el mensaje. Se utiliza criptografía de clave pública para establecer el circuito entre todos los nodos existentes, mientras que los datos se transmiten utilizando criptografía simétrica.

La primera propuesta basada en *onion routing* apareció en la obra de Goldschlag, Reed y Syverson (1996). En este diseño inicial, un primer mensaje abre el circuito a través de la red. Para generar el circuito a cada nodo se le asigna un identificador. Este primer mensaje se cifra utilizando diversas capas de criptografía pública. Cada capa puede ser descifrada solo por el nodo correspondiente utilizando su clave privada. Este primer mensaje lleva material criptográfico compartido entre el emisor y cada nodo del circuito, que será utilizado para enviar los datos en la siguiente fase del protocolo. Los mensajes enviados siempre van cifrados por capas, en el primer mensaje las capas se construyen con claves públicas y en los siguientes mensajes de envío de datos las capas se generan mediante las claves simétricas obtenidas anteriormente. Este tipo de transmisiones cifradas por capas es lo que le da el nombre de *onion* (cebolla en inglés) al protocolo.

El objetivo del *onion routing* es dificultar el análisis de tráfico realizado por un adversario, protegiendo la anonimidad del emisor y receptor del mensaje. No obstante, esta técnica se demostró débil en situaciones de poco tráfico donde un atacante pasivo podía descubrir la coincidencia entre el mensaje de entrada a la red y su salida. A este tipo de ataques se les ha denominado *timing attacks* y *end-to-end correlation attacks* (Raymond, 2000).

Tor posiblemente sea el diseño más relevante basado en *onion routing* (Dingledine y otros, 2004). Tor fue publicado en el 2004 y utiliza una red de *routers* para reenviar trafico TCP. Esta aplicación está específicamente diseñada para tratar tráfico web y se usa generalmente junto con la herramienta Privoxy encargada de eliminar cualquier componente activo existente en las páginas web transmitidas, gestionar las *cookies*, etc.

FoxTor y TorButton

FoxTor y TorButton son dos *plug-ins* para el navegador Firefox que combinan la red TOR con la herramienta Privoxy.

La red Tor está formada por una lista de servidores voluntarios que actúan como los nodos *router* del sistema. Los usuarios que se conectan a la red crean circuitos de tres *routers* eligiéndolos al azar para realizar las transmisiones anónimas. Tor no utiliza el primer mensaje para repartir el material criptográfico simétrico entre los nodos del circuito, en vez de eso sigue un sistema interactivo en el cual el usuario se conecta al primer nodo y le requiere que se conecte al siguiente nodo. Se establece así un canal bidireccional utilizando el protocolo *Diffie-Hellman autenticado* (Diffie y otros, 1992) para compartir claves simétricas. Una vez el circuito ha sido establecido y las claves repartidas, el protocolo

envía los datos siguiendo el procedimiento habitual del *onion routing*. Al igual que la propuesta inicial, Tor tiene dificultades contra un adversario pasivo con una visión global de la red y situaciones de poco tráfico.

Aunque el objetivo de las técnicas basadas en *onion* es reducir el coste en tiempo respecto a las medidas basadas en mix nets, en la práctica, estos mecanismos ralentizan considerablemente la navegación web. Tomando como ejemplo el envío de una consulta a un motor de búsqueda, Boneh y otros (2007) demostraron que realizar dicha consulta con un circuito Tor de dos nodos (por defecto se utilizan tres nodos) requería 25 veces más tiempo que realizar una consulta directa sin anonimato.

Resumen

En este módulo nos hemos centrado en explicar la situación de la privacidad en la era de las tecnologías de la información. Más concretamente, hemos mostrado las implicaciones de la existencia de una huella digital y las medidas que deben tomarse para controlarla.

En el primer apartado hemos tratado el concepto de huella digital y se ha explicado qué contiene y cómo se crea. Hemos hecho especial énfasis en los mecanismos utilizados para identificar a los usuarios cada vez que utilizan servicios o aplicaciones en Internet. La identificación es el paso inicial para realizar el seguimiento y estudio de los usuarios.

En el segundo apartado nos hemos centrado en los perfiles de usuarios. Se ha descrito qué atributos forman un perfil de usuario y se ha explicado con detalle los diferentes escenarios donde dichos datos se obtienen de los propios usuarios. Finalmente, hemos presentado varios ejemplos donde se muestra cómo las empresas explotan los perfiles de usuario para mejorar sus resultados.

En el tercer apartado hemos definido el concepto de privacidad y hemos explicado el funcionamiento de las políticas de privacidad que las empresas de Internet aplican a los datos adquiridos. En este apartado hemos estudiado quién es el propietario de los datos personales una vez que estos han sido obtenidos por las empresas.

Finalmente, en el cuarto y último apartado hemos presentado distintos métodos existentes en la literatura, que han sido diseñados para preservar la privacidad de los usuarios que utilizan los variados servicios presentes en Internet. En este apartado también hemos introducido las ventajas e inconvenientes de las distintas propuestas así como el ámbito de aplicación de cada una.

Actividades

1. En esta actividad estudiaremos qué información referente a nosotros puede encontrarse en Internet. El objetivo es evaluar el nivel de privacidad que hemos perdido y cuál es la tendencia. Para realizar el estudio se elegirán dos compañeros de clase y se presentará una memoria con los apartados siguientes:

- Información encontrada. En el caso de encontrar información sensible se notificará al interesado y en el informe se comentará el tipo de información encontrada sin detallar cuál es dicha información.
- Fuentes de información consultadas.
- Medidas o recomendaciones para disminuir la cantidad de información personal accesible en Internet.

A continuación se presenta un ejemplo de la información que se podría buscar. Este ejemplo solo es una guía, se puede modificar de la forma que se crea conveniente.

- Datos básicos (nombre, alias, edad, teléfonos, fotografía, fecha y lugar de nacimiento, altura, peso, estado civil, familia, etc.).
- Datos médicos (grupo sanguíneo, operaciones, fumador, bebedor, ejercicio físico, etc.).
- Datos legales y financieros (delitos graves, multas de tráfico, tarjetas de crédito, deudas, etc.).
- Aficiones (libros, música, televisión, películas, deportes, etc.).
- Comportamiento en línea (noticias, blog, web, correo electrónico, redes sociales, etc.).
- Personalidad (inteligencia, *myers-briggs*, orientación política, etc.).
- Amenazas potenciales (juegos en línea, interés por el terrorismo, etc.).

2. En esta actividad estudiaremos qué información se puede encontrar en una red social, concretamente la red social Facebook. El estudio se dividirá en una serie de fases que se detallan a continuación, estas fases serán los puntos de la memoria a entregar:

a) Entrada a Facebook. Si disponéis de una cuenta en Facebook se puede utilizar en el estudio. En el caso de no tener una cuenta en esta red social, se puede crear una y añadir algunos amigos o compañeros de clase. En esta primera fase se deben describir los puntos siguientes:

- Cuenta escogida para el estudio.
- Tipo de información que se puede introducir en Facebook. Podéis seguir la guía orientativa utilizada en la primera actividad.
- Herramientas disponibles en Facebook para proteger la privacidad.
- Aplicaciones que nos proporciona Facebook o que están disponibles en Facebook y que se pueden utilizar para obtener información de otros usuarios.

b) Topología de amistades de vuestra red social. Como mínimo se pide que se muestren en el grafo todos los nodos que están a distancia 2 de la cuenta del usuario escogido. La sección debe contener la información siguiente:

- Distancia escogida.
- Representación del grafo.
- En el caso de desarrollar una herramienta para obtener esta información, se debe describir el funcionamiento de esa herramienta. El desarrollo de esta herramienta es opcional.

c) Privacidad de los nodos. Para cada nodo del grafo estudiaremos su nivel de privacidad. La memoria debe incluir la información siguiente:

- Para cada tipo de información que se ha definido en el primer punto se debe indicar cuántos nodos la muestran de forma pública. Esta información se puede presentar de forma agregada (tanto por ciento).
- Información obtenida de forma implícita. Quizás un usuario no aporta información de un cierto tipo (por ejemplo, opción política) pero se puede deducir por comentarios o fotos que sí son mostrados de forma pública.

3. En esta actividad realizaremos un análisis de las políticas de privacidad de Microsoft similar al propuesto en el subapartado 3.2.2 en donde se trataban las políticas de Google.

El estudio se dividirá en una serie de fases que se detallan a continuación, estas fases serán los puntos de la memoria a entregar:

- Analizar qué medidas utiliza Microsoft para identificar a los usuarios y obtener sus datos.
- Averiguar qué tipo de información personal recolecta Microsoft.
- Estudiar de qué manera Microsoft utiliza la información recolectada.
- Averiguar qué políticas de seguridad se aplican a los datos almacenados.

Glosario

broadcast *m* Modalidad de transmisión que prevé el envío de un mensaje a todos los ordenadores conectados a una misma red.

cookie *f* Es un fichero que se envía a un navegador por medio de un servidor web para registrar las actividades de un usuario en un sitio web.

CGI *f* Abreviación de *common gateway interface*, el CGI es un programa de interfaz que permite al servidor de Internet utilizar programas externos para realizar una función específica. También denominado pasarelas o CGI *scripts*, ejecutan un programa y formatean los resultados en HTML de manera que puedan ser presentados en el navegador. Los *scripts* CGI también se usan para introducir una variedad de sistemas de análisis y de tráfico de medida de audiencia de un sitio.

criptografía asimétrica o de clave pública *f* Sistema de encriptación que consiste en utilizar un sistema de doble clave: clave pública y clave privada. La clave pública es conocida por todos y se utiliza para convertir el texto en claro que queremos cifrar en un criptograma, que tan solo podrá volverse a convertir en texto en claro mediante la clave privada, conocida solamente por la persona a la que va remitida la información cifrada mediante la clave pública.

criptografía simétrica *f* Este tipo de criptografía utiliza una única clave para cifrar y descifrar la información. Dado que solo existe una clave para convertir el texto en claro en criptograma y viceversa, esta tiene que ser conocida por las dos partes que quieren intercambiar la información.

dirección IP *f* Es un código 7érico que identifica a un ordenador específico en Internet. Las direcciones de Internet son asignadas por un organismo llamado InterNIC

GPS *m* Sistema global de navegación por satélite (GNSS) que nos permite fijar a escala mundial la posición de un objeto, una persona, un vehículo o una nave.

Html *m* Siglas de *hypertext markup language*. El HTML es el lenguaje informático utilizado para crear documentos hipertexto. El HTML utiliza una lista finita de etiquetas que describe la estructura general de varios tipos de documentos enlazados entre sí en el World Wide Web.

Http *m* Son las siglas de *hypertext transfer protocol*, el método utilizado para transferir ficheros hipertexto por Internet. En el World Wide Web, las páginas escritas en HTML utilizan el hipertexto para enlazar con otros documentos.

malware *m* Cualquier programa, documento o mensaje, susceptible de causar perjuicios a los usuarios de sistemas informáticos.

navegador *m* Un navegador es un programa software que permite ver e interactuar con varios tipos de recursos de Internet disponibles en el World Wide Web.

RFID *f* Acrónimo de *radio frequency identification*, Identificación por radiofrecuencia. Es una tecnología que permite identificar un objeto por radio, mediante una etiqueta (*RFID tag*) que ese objeto lleva adherida o insertada.

sniffer *m* Aplicación de monitorización y de análisis para el tráfico de una red para detectar problemas. Lo hace buscando cadenas 14éricas o de caracteres en los paquetes. Puede usarse ilegalmente para recibir datos privados en una red, además son difíciles de detectar.

spyware *m* El *spyware* es un programa que recopila información de un ordenador y después transmite esta información a una entidad externa sin el conocimiento o el consentimiento del propietario del ordenador.

TCP/IP *m* Son las siglas de *transmission control protocol/Internet protocol*, el lenguaje que rige todas las comunicaciones entre todos los ordenadores en Internet. TCP/IP es un conjunto de instrucciones que dictan cómo se han de enviar paquetes de información por distintas redes. También tiene una función de verificación de errores para asegurarse de que los paquetes llegan a su destino final en el orden apropiado.

TLS *m* Siglas de *transport layer security*. Es el sucesor del protocolo SSL (*secure socket layer*). Ambos son protocolos criptográficos que proporcionan comunicaciones seguras a través de una red, frecuentemente Internet.

URL *m* Siglas de *uniform resource locator*. Es la dirección de un sitio o de una fuente, normalmente un directorio o un fichero, en el World Wide Web y la convención que utilizan los navegadores para encontrar ficheros y otros servicios distantes.

virus *m* Programa creado especialmente para invadir ordenadores y redes y crear el caos. El daño puede ser mínimo, como que aparezca una imagen o un mensaje en la pantalla, o puede hacer mucho daño alterando o incluso destruyendo ficheros.

World Wide Web *f* Literalmente "tela de araña mundial", más conocida como web. Se puede considerar la web como una serie de ficheros de texto y multimedia y otros servicios conectados entre sí por medio de un sistema de documentos hipertexto.

Bibliografía

(2010). "Annual State of the Net Survey". *Consumer Reports* (vol. 75, núm. 6). Consumer Reports National Research Center.

(2010). "7 things to stop doing now on Facebook". *Consumer Reports* (vol. 75, núm. 6).

(Mayo, 2010). "Social savy recruiters utilising social media in their recruitment strategy". *Personnel Today*.

(2011). "Condiciones de servicio de Google". Google: http://www.google.com/accounts/TOS?hl=ES

(2011). "PhoCusWright's Social Media in Travel: Traffic & Activity". PhoCusWright.

(2011). "Políticas de privacidad de Google". Centro de privacidad de Google: http://www.google.es/intl/es/privacy/

(2011). "Your Digital Footprint: how much information about your life gets recorded by big business and Big Brother". *Koppel on Discovery Channel*: http://dsc.discovery.com/convergence/koppel/interactive/interactive.html

Asur, S.; Huberman, B. A. (2010). "Predicting the Future with Social Media". *Proceedings of the 2010 International Conference on Web Intelligence and Intelligent Agent Technology* (págs. 492-499).

Banks, L.; Wu, S. F. (2009). "All Friends are NOT Created Equal: An Interaction Intensity based Approach to Privacy in Online Social Networks". *Proceedings of the 2009 International Conference on Computational Science and Engineering* (págs.970-974).

Baran, P. (1964). "On distributed communications: IX security secrecy and tamper-free considerations".

Berners-Lee, T.; Hendler, J.; Lassila, O. (2001). "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". *Scientific American* (vol. 284, núm. 5, págs. 34-43).

Castellà-Roca, J.; Viejo, A.; Herrera-Joancomartí, J. (2009). "Preserving user's privacy in web search engines". *Computer Communications* (vol. 32, núm. 13-14, págs. 1541-1551).

Chaum, D. (1981). "Untraceable electronic mail, return addresses, and digital pseudonyms". *Communications of the ACM* (vol. 4, núm. 2, págs. 84-88).

Coutu, D.; Palfrey, jr. J. G.; Joerres, J. A.; Boyd, D. M.; Fertik, M. (junio, 2007). "We Googled You". By:, *Harvard Business Review* (vol. 85, núm. 6) 00178012.

Danezis, G.; Dingledine, R.; Mathewso, N. (2003). "Mixminion: Design of a type III anonymous remailer protocol". *Proceedings, IEE Symposium on Security and Privacy* (págs. 2-15).

Diffie, W.; Oorschot, P. C. van; Wiener, M. J. (1992). "Authentication and authenticated key exchanges". *Designs, Codes and Cryptography* (vol. 2, págs. 107-125).

Dingledine, R.; Mathewson, N.; Syverson, P. (2004). "Tor: The second generation onion router". *Proceedings of the 13th USENIX SecuritySymposium* (págs. 303-319).

Domingo-Ferrer, J.; Sebé, F.; Solanas, A. (2008). "An Anonymity Model Achievable Via Microaggregation". *Lectures Notes in Computer Science* (vol. 5159, págs. 209-218).

Eckersley, P. (2010). "How Unique Is Your *Browser*?". Proceedings of the Privacy Enhancing Technologies Symposium.

George Danezis, Claudia Diaz, and Paul Syverson (2010). "Systems for Anonymous Communication". *CRC Handbook of Financial Cryptography and Security* (págs. 341-390).

Glassman, N. (agosto, 2010). "3 Questions About Location-Based Social Networks with face2face". *SocialTimes*.

Goldschlag, D. M.; Reed, M. G.; Syverson, P. F. (1996). "Hiding routing information". *Lecture Notes in Computer Science* (vol. 1174, págs. 137-150).

Graffi, K.; Mukherjee, P.; Menges, B.; Hartung, D.; Kovacevic, A.; Steinmetz, R. (2009). "Practical security in p2p-based social networks". *Proceedings of the IEEE 34th Conference on Local Computer Networks* (págs. 269-272).

Greenberg, A. (2009). "Privacy Groups Target Android, Mobile Marketers". *Forbes,* January, 13th, 2009, http://www.forbes.com/2009/01/12/mobile-marketing-privacy-tech-security-cx_ag_0113mobilemarket.html

Hansell, S. (febr., 2006). "Increasingly, Internet's Data Trail Leads to Court". *The New York Times*.

Hay, M.; Miklau, G.; Jensen, D.; Towsley, D.; Weis, P. (2008). "Resisting Structural Identification in Anonymized Social Networks". *Proceedings of the 2008 Conference on Very Large Databases* (*VLDB*).

Krishnamurthy, B.; Wills, C. E. (2010). "On the leakage of personally identifiable information via online social networks". *ACM SIGCOMM Computer Communication Review* (vol. 40, núm. 1, págs. 112–117).

Laurant, C. (2003). "Privacy & Human Rights 2003-An international survey of privacy laws and developments". Electronic Privacy Information Center.

Leskovec, J.; Adamic, L. A.; Huberman, B. A. (2006). "The dynamics of viral marketing". *Proceedings of the 7th ACM Conference on Electronic Commerce*.

Li, S. (enero, 2011). "Insurers are scouring social media for evidence of fraud". *Los Angeles Times*.

Madden, M.; Fox, S.; Smith, A.; Vital, J. (2007). "Digital Footprints: Online identity management and search in the age of transparency". http://pewresearch.org/pubs/663/digital-footprints

Mazieres, D.; Frans Kaashoek, M. (1998). "The Design, Implementation and Operation of an Email Pseudonym Server". *5th ACM Conference on Computer and Communications Security* (págs. 27-36).

Melenhorst, M.; Grootveld, M.; van Setten, M. (2008). "Tag-based information retrieval of video content", Veenstra, *Proceeding of the 1st international conference on Designing interactive user experiences for TV and video*,.

Milton, N. (mayo, 2010). "Price of Facebook Privacy? Start Clicking". *The New York Times*.

Möller, U.; Cottrell, L.; Palfrader, P.; Sassaman, L. (2003). "Mixmaster protocol - version 3". IETF Internet Draft.

Parekh, S. (agosto, 5, 1996). "Prospects for remailers: where is anonymity heading on the internet?". *On-line journal* (vol. 1, núm. 2): http://freehaven.net/anonbib/cache/remailer-history.html

Pattison, K. (Oct., 2010). "Geolocation Services: Find a Smartphone, Find a Customer". *The New York Times*.

Raymond, J. F. (2000). "Traffic analysis: Protocols, attacks, design issues, and open problems". *Lecture Notes in Computer Science* (vol. 2009, págs. 10-29).

Reiter, M.; Rubin, A. (1998). "Crowds: Anonymity for web transactions". *ACM Transactions on Information and System Security* (*TISSEC*) (vol.1, núm. 1, págs. 66-92).

Rotenberg, M. (2000). "Protecting Human Dignity in the Digital Age". Unesco.

Saint-Jean, F.; Johnson, A.; Boneh, D.; Feigenbaum, J. (2007). "Private web search". *Proceedings of the 2007 ACM workshop on Privacy in electronic society* (págs. 84-90).

Sánchez, D.; Isern, D.; Millán, M. (2010). "Content Annotation for the Semantic Web: an AutomaticWeb-based Approach". *Knowledge and Information Systems* (en prensa).

Singel, R. (agosto, 2009). "You deleted your *cookies*?". *Wired*.

Soderberg, J.; Kakogianni, E. (2010). "Automatic tag generation for photos using contextual information and description logics". *2010 International workshop on content-based multimedia indexing* (págs. 1-7).

Stagg-Macey, C.; Beattie, C. (abril, 2010). "Leveraging social networks: an in-depth view for insurers". *Document*.

Terry, J. (7, febr., 2008). "Leaving a digital footprint: Online activities follow students to job interviews, professional world". *The State News*.

"The Cross-Site Scripting (XSS) FAQ": http://www.cgisecurity.com/xss-faq.html

Wortham, J. (mayo, 2010). "Facebook Glitch Brings New Privacy Worries". *The New York Times*.