# Web Tracking: Cómo nos identifican y monitorizan en Internet

**ESPECIAL VIII - 2016**

www.pabloyglesias.com
contacto@pabloyglesias.com

*El negocio de las grandes compañías tecnológicas se basa en el desarrollo de algoritmos lo suficientemente eficaces como para sacar el máximo valor posible a los datos.*

# El negocio de la hipersegmentación

**En este octavo Especial, quería recopilar todas las técnicas que habitualmente se están usando en el tercer entorno para identificar al usuario. Técnicas a veces basadas en la picaresca, en los diseños no contemplados, y en la tergiversación de su cometido, así como otras específicamente diseñadas para monitorizar nuestros movimientos en la red.**

**En esta carrera encontramos desde negocios digitales que implementan sistemas de Web Tracking con el objetivo de mejorar sus beneficios, pasando por medios que hacen uso de estas tecnologías para conocer mejor a sus usuarios y/o para mostrarles una publicidad más segmentada, y organizaciones políticas y militares que se aprovechan de estas herramientas para mantener un sistema de control, o desarrollar estrategias de espionaje a otras organizaciones o países.**

**Un escenario en el que el usuario tiene lamentablemente las de perder, habida cuenta de que como veremos, muchas de ellas son sencilla y llanamente incontrolables por nuestra parte.**

**Comencemos.**

No le voy a descubrir nada que no haya quedado patente en estos últimos años. **El negocio de buena parte de las grandes compañías tecnológicas se basa precisamente en el desarrollo de algoritmos lo suficientemente eficaces como para sacar el máximo valor posible a los datos que circulan a su alrededor.**

Si algo ofrece el tercer entorno es precisamente la capacidad de cuantificar la complejidad de acciones que un cliente/usuario realiza frente al mismo, con el fin de mejorar cíclicamente el producto, realimentándolo con el feedback que ese cliente/usuario entrega consciente o inconscientemente.

Dejando de lado esa pata analítica del Big Data, en este Especial nos centraremos en esa **primera fase de recolección de datos**, con las herramientas utilizadas por las compañías para tracear y segmentar al usuario. Pero antes, conviene esbocetar **las razones que lleva a un negocio a implantar sistemas de este tipo**, y que son tan variadas como cabría esperar.

Encontramos así compañías que hacen uso de esta información para mejorar sus productos y servicios. Otras, como estrategia de expansión y herramienta de marketing que les permite dirigirse a diferentes públicos objetivo. Algunas hacen negocio directo con los datos *(por ejemplo, revendiéndolos a terceros)*, y otras, indirecto, ofreciendo esa segmentación a terceros para por ejemplo mostrar publicidad o acercar potenciales clientes a otras compañías. También los hay que simplemente necesitan esta información con fines puramente estadísticos *(bien sea para uso interno, bien sea como muestra de la salud del propio negocio de cara a accionistas/inversores)*. Y por supuesto, también encontraremos casos en los que el objetivo pasa por la propia identificación del usuario *(como podría ocurrir en entornos militares o de inteligencia)*.

De cara al usuario, tenemos que entender que esta tendencia tiene sus pros y sus contras:

- Por un lado, el que el negocio que hay detrás sea capaz de segmentar usuarios debería servir para **optimizar nuestra experiencia**, ofreciéndonos aquello que podríamos estar interesados en consumir, y obviando aquello otro que en principio nunca vamos a necesitar. Aquí, el tracking debería jugar a nuestro favor, manteniendo la capacidad competitiva de ese negocio, para beneficio suyo y nuestro.
- Por otro, es de todos sabido que estas estrategias podrían ser consideradas altamente invasivas, atentando contra la privacidad del usuario. En una tienda física, rara vez le van a pedir que se identifique, ¿por qué debería tener que hacerlo en una tienda virtual? Y esa segmentación no siempre va a jugar a nuestro favor, como encontramos en escenarios donde **el precio varía según la información que el sistema ha sido capaz de obtener de nosotros** (1) *(si es la primera vez que visitamos la página, si nuestro dispositivo es considerado gama alta, si ya hemos buscado en otras ocasiones productos semejantes,…)*. En última instancia, esa monitorización podría acabar afectándonos en nuestro día a día, conforme toda esta industria se vaya paulatinamente centralizando en varias grandes corporaciones, capaces de aglutinar y servir esa inteligencia a terceros. Un escenario en el que la monitorización no ocurre de manera dispersa en cada servicio digital, con una tecnología capaz de perseguirnos allá donde vayamos, reconociendo hábitos y costumbres que podrían, sea ahora, sea el día de mañana, pasarnos factura. *Desde la negación de un préstamo hipotecario por un perfil de Facebook que denota una vida de desenfreno, pasando por el aumento del precio de nuestro seguro al considerar que nuestra conducción frente al volante no es la adecuada, hasta escenarios más*

*terroríficos, como el que llevó a la Alemania de Hitler a acabar con prácticamente toda la población judía de Holanda gracias al descubrimiento de un censo nacional convenientemente segmentado por creencias religiosas.*

# Tipologías de tracking de usuarios vía web

**De todos los entornos digitales, la web es la que a priori ofrece un mayor surtido de herramientas para tracear a los usuarios. Y lo hace precisamente por ser un entorno abierto, en comparación con modelos más cerrados como los que encontraremos en el mundo app o en el de los programas de escritorio.**

**Un ecosistema rico en posibilidades, y por ende, cuyo límite solo atiende al límite al que la creatividad humana sea capaz de llegar.**

**Dentro de este escenario, encontramos tres grandes categorías de herramientas de tracking:**

1. De parte del cliente: **Aglutinan todas aquellas técnicas que se aprovechan de elementos que tiene almacenados el propio cliente en su dispositivo.**
2. De inherencia: **Basados en configuraciones de hardware/software del cliente que en su completitud, permiten identificarlo frente a otras configuraciones distintas.**
3. De factores exógenos: **Hablaremos en este apartado de aquellas técnicas que hacen uso de análisis de comportamiento, o que dependen exclusivamente de funcionalidades que únicamente un organismo con gran control en la cadena de suministro de la información podría llevar a cabo.**

**Detallaremos en cada una de ellas las técnicas más representativas *(así como algunas que por su particularidad, me parece oportuno señalar)*, atendiendo a su funcionamiento y a la persistencia que presentan *(dificultad para protegerse de ella)*.**

## Tracking de cliente

**Las cuatro técnicas englobadas dentro de este apartado hacen uso de ficheros e información que el cliente tiene alojados directamente en su propio sistema. Dicha información se saca mediante diferentes funcionalidades del navegador y de las tecnologías web, y se analizan, sea en el propio cliente, sea en el servidor, para que este último ofrezca la respuesta adecuada según esté contemplado en el sistema de monitorización.**

## Tracking de sesión

**Persistencia:** *Baja, mientras dure la sesión activa.*

**En este caso, el negocio hace uso de un identificador que aloja temporalmente en la propia sesión del usuario, normalmente en campos ocultos, propiedades del árbol DOM o formularios que cuentan con un sistema de validación de usuario.**

**Fue históricamente una de las primeras técnicas que permitían ofrecer una suerte de dinamismo a las páginas, aunque ahora, con la evolución de estas tecnologías, se ha quedado obsoleta.**

**Basta con que salgamos de la web y volvamos a entrar para que este tracking se pierda.**

## Almacenamiento en caché

**Persistencia:** *Media, según la configuración del navegador y mientras no expire la caché.*

**El caching web se ha popularizado conforme las páginas hacían acopio de más elementos comunes** *(header, footer, sidebar,…)* **y más recursos gráficos, permitiendo que para cada visita a una página de una web,** no tengamos que volver a descargar absolutamente todos sus elementos**, sino únicamente aquellos que de verdad han cambiado respecto a aquello que ya tenemos cacheado.**

**Es una funcionalidad necesaria en el tercer entorno, que nuevamente se puede aprovechar para alojar en la caché de cada cliente piezas de código o identificadores que luego recuperemos.**

**La persistencia de la caché depende principalmente de dos factores: el cómo tengamos definida la privacidad del navegador, y el tiempo máximo de vida que el administrador le haya otorgado a esa pieza de código o identificador antes de que el cliente tenga que nuevamente descargarlo.**

**Y este es precisamente el elemento que hace que esta técnica no sea tan interesante como la siguiente, habida cuenta de que su éxito depende, hasta cierto punto, de la configuración del cliente.**

**Si este borra el caché del navegador** *(una opción que normalmente encontraremos por Ajustes/Configuración > privacidad)***, o lo tiene configurado para que este paso se haga automáticamente en cada cierre de sesión, se perdería el identificador, y al visitar nuevamente la web, es probable** *(depende de cómo esté desarrollado el sistema)* **que el script de monitorización le asocie al usuario un nuevo ID.**

## Tracking de datos y ficheros locales

**Persistencia:** *Entre media y alta, según las tecnologías utilizadas.*

**En este grupo encontramos técnicas que permiten explotar la información almacenada previamente en ficheros locales del cliente.**

**La más habitual es la utilización de Cookies HTTP, archivos que pueden contener hasta 4KB de información, y cuya eliminación dependerá nuevamente de cómo esté configurado el navegador, o de la acción del propio usuario** *(semejante, en todo caso, al borrado de caché)***.**

**Los llamados** *Local Shared Objects* **(EN) de Adobe Flash permitirían almacenar hasta 100KBs de información, y en esencia, podrían estar funcionando como una técnica separada** *(hasta que Flash sea erradicado del mundo, los navegadores seguirán ofreciendo acceso a estos ficheros)***, o como en su día se demostró con EverCookies, utilizar éstos para reconstruir cookies borradas, haciendo que la persistencia ya no solo dependa de estas últimas, sino también de los** *LSO***.**

**Para terminar, tanto HTML5 con** *WebStorage* **(EN) o** *IndexedDB* **(EN), como Microsoft Silverlight con** *Isolated Storage* **(EN), o Java con** *PersistenceService* **(EN), ofrecen tecnologías que permiten almacenar ficheros en local para su explotación en remoto, y podrían ser perfectamente usados para el tracking de usuarios, con el añadido de que la mayoría de estas técnicas no suelen estar aún contempladas en los parámetros de configuración del navegador, y por ende, suelen requerir un borrado manual, lo que les dota de una persistencia mayor.**

## Tracking mediante tergiversación del HSTS

**Persistencia:** *Alta.*

***Es necesario que todo dispositivo que salga al mercado cuente con unas medidas de seguridad básicas, y entre ellas, la capacidad de recibir actualizaciones críticas de seguridad.***

**El HSTS es un protocolo que permite al navegador asegurarse de que la conexión a una página específica se realiza mediante HTTPs. Es por tanto una funcionalidad dirigida a mejorar la seguridad de parte del cliente.**

**Para ello, hace uso de un listado que el navegador almacena localmente, de manera que cada vez que visitamos una página, consulta ese listado, y si ese dominio está contemplado allí, fuerza a que esa conexión se realice vía HTTPs. En caso contrario, lo lista si la conexión ofrece HTTPs.**

**Y aquí está el quid de la cuestión. Como en su día se demostró,** *SuperCookies (vaya nombrecito le pusieron…)* **permitía que el servidor generase un identificador único a cada cliente analizando las particularidades de esta lista. Los chicos de RadicalResearch mantienen todavía una demo online (2), e incluso en su día se presentó un PoC llamado** *Sniffly* **(3) que analiza el tiempo de respuesta de una web para deducir si tiene o no listada una página específica en su HSTS, y por ende, conocer si el cliente la había o no visitado.**

**La persistencia de esta técnica es por tanto alta, ya que no existe una manera sencilla e inmediata de eliminar esta lista. Se tiene que hacer desde las opciones avanzadas, y en algunos casos, mediante comandos específicos.**

## Tracking basado en la inherencia

**Las técnicas contempladas dentro de esta categoría se aprovechan de elementos que son innatos en el cliente, como es su configuración de hardware, red y software. Y una de ellas, como veremos, está siendo bastante utilizada para fines comerciales.**

### Tracking mediante datos de red

**Persistencia:** *Alta.*

**Esta técnica** *(o conjunto de técnicas, mejor dicho)* **se basan en obtener toda la información posible analizando el tráfico, las cabeceras HTTP o el uso de tecnologías web auxiliares** *(Java, Flash, HTML5,..)* **para conocer de dónde viene la visita.**

**Para ello, un uso tergiversado de WebRTC, una funcionalidad de JavaScript dirigida a mantener videollamadas desde el propio navegador, permitiría a un tercero obtener la IP privada del cliente.**

**La geoposición, aunque no sea todo lo fiable que se pudiera desear, también cuenta con su propia API de HTML5 (EN), y si empezamos a cruzar estos datos con la información del análisis del tráfico se puede obtener con bastante exactitud dónde vive** *(o al menos, desde donde se conecta habitualmente)* **una persona.**

**Eso sí, el navegador suele pedir permiso antes de permitir a una página conocer nuestra posición. El tema de la IP, por ahora, es incontrolable por parte del usuario, a no ser que éste se conecte desde una VPN** *(al final del artículo haré mención a algunos servicios que recomiendo)***.**

## Tracking mediante huella digital

**Persistencia:** *Muy Alta.*

**Este conjunto de técnicas se basan en aprovechar las diferentes configuraciones, tecnologías, plugins y addons que utilizamos en nuestro navegador y sistema operativo, y que son públicas (o accesibles) por parte del servidor, para generar un profiling que pueda ser identificativo del usuario.**

**Entre toda la amalgama de datos recopilados, estarían la versión del navegador y de todas sus extensiones, la resolución de pantalla, las fuentes, los ficheros locales** *(cookies, LSOs,…)***, la zona horaria, el idioma, así como las versiones de tecnologías como Flash o Java que permitirían identificarnos con un error inferior al 5%, según demostró la EFF con** *Panopticlick* **(4).**

**En un paso más allá, se ha llegado a utilizar tecnologías como WebRTC de HTML5 para pintar un canvas que depende única y exclusivamente de las particularidades de cada hardware. De esta manera, un análisis posterior del mismo permite identificar a este usuario respecto a otros, puesto que la imagen resultante es única para cada uno (5).**

**Este tipo de técnicas son prácticamente imposibles de contrarrestar, por lo que considero su persistencia muy alta. Afortunadamente, tienen una debilidad, y es**

**que son incapaces de discernir cuándo ese mismo usuario se conecta desde otro dispositivo.**

**Incluimos al final del artículo un listado de extensiones que podrían ayudar a minimizar el efecto de este tipo de técnicas, a sabiendas de que también puede que dificulten la experiencia de usuario.**

## Tracking mediante factores exógenos

**Quedan por definir aquellas técnicas que dependen de variables que o bien son innatas en el usuario (que no en la tecnología), o bien dependen de elementos de la propia cadena de suministro.**

### Tracking basado en el comportamiento

**Persistencia:** *Muy Alta.*

**Esta técnica aplica inteligencia al comportamiento que tiene el usuario dentro de la página, en una primera fase comprobando su configuración más genérica** *(zona horaria, idioma, histórico,…)***, y en una segunda, mediante algoritmos de** reconocimiento de patrones *(como por ejemplo, dónde suele dejar el ratón o el dedo mientras navega, a qué suele clickar,…)* **y** machine learning**.**

**Tiene como principal ventaja el que permitiría identificar a un usuario indistintamente del dispositivo que esté utilizando, y como principal desventaja, el hecho de que implementar un sistema de este tipo es muy, muy costoso. Tanto como para únicamente estar al alcance de los gigantes de internet, y quizás, de alguna agencia de inteligencia.**

### Tracking mediante el control de la cadena de suministro

**Persistencia:** *Muy Alta.*

**Otra estrategia para mantener monitorizado el tráfico pasaría por incluir identificadores en las cabeceras HTTP que se envían en cada comunicación. Una vez esa petición sale de nuestro terminal, queda expuesta a una posible manipulación por parte de algún elemento de la cadena de suministro, cosa que parece que han estado aprovechando**

***Hay que ser consciente de que siempre, queramos o no, vamos a estar expuestos a algún tipo de tracking.***

**no pocas operadoras de red para identificar con mayor exactitud las actividades de sus clientes (6).**

**Contra este tipo de profiling hay poco que el usuario pueda hacer. Atacar a algo que es innato en nuestra forma de comunicarnos con la máquina, o generar un escenario de man in the middle, queda fuera del control del usuario, aunque sea duro reconocerlo.**

## ¿Qué podemos hacer para protegernos?

**Habida cuenta de la persistencia que demuestran algunas de estas técnicas, hay que ser consciente de que siempre, queramos o no, vamos a estar expuestos a algún tipo de tracking, y que además, mientras más sistemas implementemos para evitarlo, peor experiencia tendremos en la navegación.**

**Sabedores de esto, los primeros pasos serían hacer uso de bloqueadores de publicidad como uBlock (ES) o AdBlock (ES)** *(AdBlock Plus si prefiere permitir que se haga tracking en publicidad no invasiva)*.

**Bloquear JavaScript con NoScript (EN) o ScriptSafe (EN) ayudará también a bloquear todas esas piezas de código que no están convenientemente señaladas como scripts publicitarios** *(y que aún así se usan para esos menesteres, o al menos para tracearnos).* **Privacy Badger (EN), para Chrome, desarrollada por eff.org (EN), permite también bloquear scripts de tracking** *(añadida por petición de un miembro de la Comunidad).*

**Si queremos evitar que las tiendas que visitamos nos cambien el precio por ser recurrentes, tenemos $heriff (EN).**

**Para protegernos de técnicas como la del Canvas Fingerprinting, hay bloqueadores específicos (EN).**

**Si no queremos que una página conozca nuestro paradero, podemos hasta cierto punto engañarla utilizando una VPN geolocalizada en otro punto del globo.**

***AdNauseam* (EN) es un plugin, por ahora, solo disponible para Firefox, y que únicamente funciona junto a AdBlock Edge (EN/** *un fork de AdBlock que no cuenta con whitelist de sites publicitarios*), **que se encarga de hacer click automático en nombre del usuario a cuanta**

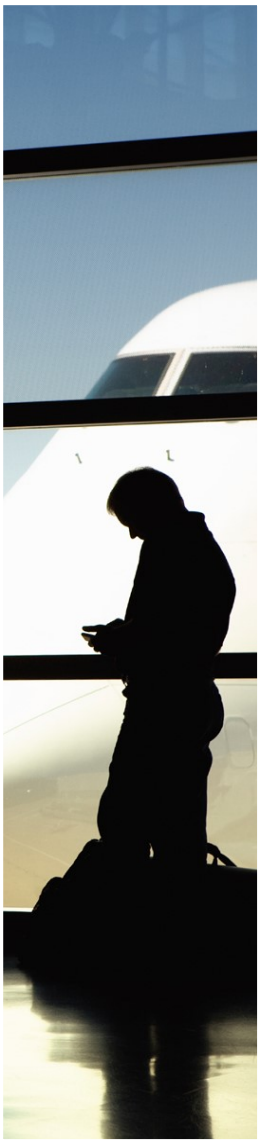
**publicidad el bloqueador es capaz de localizar, sin que éste tenga que hacerlo (y sufra las consecuencias).**

**Por otro lado, *TrackMeNot*, disponible tanto para** Firefox **(EN) como para** Chrome *(EN)* **se encarga de periódicamente** *(por defecto, cada 10 minutos)***, realizar búsquedas en los principales motores de internet con preguntas sacadas aleatoriamente de una serie de listas de palabras, algunas de ellas** consideradas por el Departamento de Seguridad Nacional como potencialmente peligrosas *(EN)***, y por ello, altamente monitorizadas.**

**Estas dos herramientas ofrecen una suerte de** oversharing tergiversado **que minimiza el impacto del tracking web protegiéndonos de sus potenciales peligros** *(segmentación, burbuja de filtros)***.**

**Y si ya lo que queremos es evitar en la medida de lo posible cualquier tracking, aún a sabiendas del sacrificio en usabilidad que vamos a realizar, olvidarnos de Chrome** *(en sí es un tracker de Google)***, Firefox o IE, y apostar por un proyecto que antepone la privacidad por encima de todo como Tor Browser.**

**Eso sí, armándonos de paciencia, y conscientes de que muchas webs ni siquiera van a funcionar. Porque eso mismo que permite a una compañía o una agencia de inteligencia tracearnos, también permite que disfrutemos de una Internet tan rica en formatos y tecnologías.**



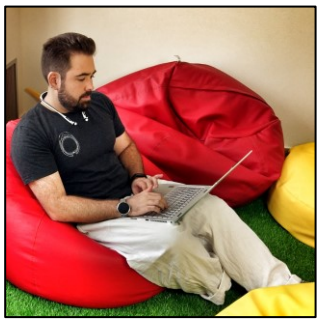
Puede acceder a los últimos informes en la sección archivo de la página

## Referencias

1. Detecting price and search discrimination in internet (Universitat Politecnica de Catalunya/PDF).
2. HSTS Super Cookies Demonstration (RadicalResearch).
3. Sniffly (ToorCon).
4. Panopticlick (EFF).
5. Pixel Perfect: Fingerprinting Canvas in HMTL5 (University of California/PDF/caído).
6. The Rise of Mobile TrackingHeaders: How Telcos Around the World Are Threatening Your Privacy (Access/PDF).

Para realizar comentarios sobre este estudio, por favor, diríjase a la página habilitada para tal fin.

## Información de contacto



Pablo F. Iglesias
**Analista de Información**
**@PYDotCom**
**contacto@pabloyglesias.com**