Una Herramienta para el Control de la Privacidad contra el Rastreo en la Web

Jagdish Prasad Achara
INRIA Grenoble
France
Email: jagdish.achara@inria.fr

Javier Parra-Arnau
Dept. Computer Science and Mathematics
Universitat Rovira i Virgili
Email: javier.parra@urv.cat

Claude Castelluccia
INRIA Grenoble
France
Email: claude.castelluccia@inria.fr

***Resumen*—El contenido y los servicios gratuitos que nos ofrece la Web son a menudo sostenidos por la publicidad. La proliferación de anuncios cada vez más invasivos y personalizados, sin embargo, ha propiciado la aparición de una gran variedad de herramientas cuya principal función es bloquear estos anuncios. El problema de dichas herramientas es su poca flexibilidad, pues sólo permiten al usuario elegir entre bloquear todos los anuncios o permitirlos. En este artículo, investigamos una tecnología que tiene pretende devolver al usuario el control sobre el rastreo y los anuncios de Internet. A diferencia de las herramientas actuales, nuestra tecnología permite al usuario elegir en qué categorías de páginas Web no quiere ser rastreado, y en cuáles no le importaría serlo con tal de apoyar al creador de contenidos. Hemos implementado esta herramienta en forma de plug-in para Chrome, y hemos evaluado el impacto económico que podría tener en la Web.**

***Palabras clave*—bloqueo de anuncios (*ad blocking*); economía (*economy*); privacidad (*privacy*).**

I. Introducción

En la industria de márketing online, la habilidad de las empresas de anuncios para rastrear y construir perfiles sobre la actividad de navegación de los usuarios es lo que permite servicios de anuncios más efectivos y personalizados. Sin embargo, la intrusión de estas prácticas y la creciente invasión de la publicidad digital ha suscitado en estos últimos años serias preocupaciones con respecto a la privacidad de usuario y la usabilidad Web. Según encuestas recientes, a dos de cada tres usuarios de Internet les preocupa el hecho de que su comportamiento sea examinado sin su conocimiento y consentimiento [1]. Numerosos estudios en esta misma línea reflejan el creciente nivel de ubicuidad y abuso de la publicidad en línea, que es percibida por los usuarios como una degradación significativa de su experiencia de navegación [2], [3], [4].

En respuesta a estas preocupaciones, en los últimos años hemos sido testigos de la aparición de una enorme variedad de herramientas de bloqueo de anuncios y anti-rastreo, cuyo objetivo principal es devolver el control a los usuarios sobre la publicidad y el rastreo. En esencia, los bloqueadores de anuncios y los anti-rastreadores monitorizan todas las conexiones de red que pueden iniciarse cuando el navegador carga una página, y evitan aquellas que se establecen con terceros y pueden corresponder a anuncios. Con este fin, estas herramientas utilizan una serie de listas negras de rastreadores que sus empresas desarrolladoras mantienen de forma manual.

Al margen de la polémica suscitada por el uso de tales listas —especialmente después de la revelación de que Adblock Plus [5] recibía dinero de las compañías de publicidad para eliminarlas de ellas [6]—, el principal problema de estas herramientas es que fueron concebidas sin tener en cuenta dos elementos fundamentales: en primer lugar, el papel crucial de la publicidad en línea como el principal apoyo del contenido y los servicios "gratuitos" de Internet y, en segundo lugar, el beneficio social y económico de la publicidad no intrusiva y racional. Mientras que los bloqueadores de anuncios y anti-rastreadores constituyen un primer intento en esta batalla por recuperar el control sobre la publicidad y el rastreo en Internet, lo cierto es que estas herramientas son extremadamente limitadas y radicales en su enfoque: los usuarios *sólo* pueden elegir entre bloquear o permitir el rastreo, y por lo tanto, entre bloquear o permitir *todos* los anuncios[1].

Con alrededor de 200 millones de personas en todo el mundo que utilizan regularmente estos bloqueadores de anuncios y anti-rastreadores, el modelo económico que sustenta la Web está en serio peligro [4]. Esto ha suscitado un intenso debate sobre la ética de estas tecnologías y la necesidad de una solución que logre un mejor equilibrio entre el modelo de negocio dominante en la Web, la privacidad de usuario y la usabilidad de Internet.

Creemos que la solución pasa necesariamente por dotar a los usuarios de un control real sobre el rastreo, y que esto sólo puede lograrse mediante tecnologías que hagan cumplir sus preferencias reales, y no las opciones radicales y binarias que ofrecen los bloqueadores de anuncios y anti-rastreadores actuales. De hecho, según estudios recientes, dos de cada tres usuarios de estas herramientas no están en contra de los anuncios y aceptarían el compromiso que viene con el contenido "gratis" [7]. La condición para ello es la publicidad sea un proceso transparente y los usuarios tengan el control sobre la información personal que se recaba sobre su actividad en la red [8].

En este artículo, presentamos *MyTrackingChoices*, una tecnología que permite a los usuarios forzar sus preferencias de rastreo, poniendo a su disposición un control exhaustivo sobre los datos de navegación recopilados por las empresas

[1]Debemos destacar que los anuncios entregados por entidades distintas de los editores (i.e., propietarios de las páginas Web) representan la gran mayoría de los anuncios hoy en día en la Web.

de publicidad y rastreadores. Al adoptar nuestra tecnología, los usuarios pueden seleccionar las categorías de las páginas Web donde no quieren ser rastreados por dichas empresas. Además, como los usuarios de nuestra herramienta aprecian el papel fundamental que la publicidad y el rastreo tienen como sostenedor de los servicios "gratuitos" de Internet, los usuarios están dispuestos a ser rastreados en categorías de páginas no sensibles, por ejemplo, aquellas relacionadas con noticias y deportes. El objetivo final de nuestra herramienta es dotar a los usuarios de ciertas garantías en términos a privacidad y experiencia de navegación, manteniendo al mismo tiempo el modelo de negocio actual de publicidad.

II. Soluciones existentes: inconvenientes y perspectivas

Debido a la proliferación de anuncios intrusivos e invasivos[2], el uso de herramientas de bloqueo de anuncios se ha convertido en una norma más que en una excepción. En esencia, dichas herramientas pueden clasificarse en "bloqueadores de anuncios" y "anti-rastreadores". Los primeros tienen como principal objetivo el bloqueo de anuncios, mientras los segundos, aunque responden a motivaciones distintas relacionadas con la privacidad y la transparencia, en última instancia también bloquean anuncios. Algunos ejemplos de bloqueadores de anuncios son AdBlock y AdBlock Plus. Entre los anti-rastreadores más populares encontramos Ghostery, Disconnect y PrivacyBadger. En términos de funcionalidades, los bloqueadores de anuncios del primera grupo eliminan toda la publicidad, mientras que los anti-rastreadores permiten a los usuarios bloquear un rastreador en particular, una categoría en concreto de rastreador (e.g., si su función es el análisis de datos o la entrega de anuncios), o bien evitar el rastreo en dominios específicos.

Inconvenientes. Los actuales bloqueadores de anuncios constituyen una solución radical, pues plantean al usuario dos opciones extremas: bloquear o permitir todos los anuncios. La opción por defecto es, precisamente, la eliminación de toda publicidad, sin que el usuario pueda elegir alguna opción sobre dicha eliminación, ni ajustar el impacto que la utilización de la herramienta pueda tener en el modelo económico de Internet.

Los anti-rastreadores, por otro lado, permiten una cierta configuración, pues dan a los usuarios la opción de decidir qué rastreadores o categorías de rastreadores no quieren que les sigan en la Web. Sin embargo, creemos que la mayoría de los usuarios no están preocupados con los rastreadores en sí, sino con otra dimensión del rastreo, i.e., en qué tipo de páginas están siendo rastreados. En este sentido, Ghostery y herramientas similares permiten a los usuarios bloquear los rastreadores en base al dominio, es decir, pueden especificar dominios concretos donde no quieren ser seguidos. Sin embargo, pensamos que este nivel de granularidad basado en el dominio no es un enfoque apropiado. A continuación, justificamos nuestro argumento:

[2]Se considera que un anuncio es invasivo si esconde contenido o si aparece aleatoriamente en la pantalla haciendo que la experiencia de navegación se degrade.

1. Dado el gran número de dominios, es casi imposible para un usuario determinar y predefinir todos aquellos dominios en los que no desea ser rastreado.
2. Algunos dominios pueden incluir páginas pertenecientes a distintas categorías, por ejemplo, con contenido sensible como "salud" o "religión", y con contenido no sensible, como "deportes" [9]. Los dominios que pertenecen a la categoría de "noticias", e.g., `elpais.com` y `elmundo.es`, son buenos ejemplos, ya que por lo general incluyen páginas de una gran variedad de categorías (deportes, economía, política, salud, viajes, etc.) Por lo tanto, a excepción de algunos dominios en los que todas sus páginas se incluyen en una misma categoría, parece tener más sentido un bloqueo en base a la categoría del contenido de la página, más que en función del dominio.
3. El bloqueo de los rastreadores en base a la categoría de las páginas visitadas (en lugar de hacerlo por cada dominio) hace que sea más fácil la configuración, pues la selección de las preferencias de bloqueo sólo requiere una única interacción por parte del usuario.

Con la excepción de PrivacyBadger, otro de los problemas de las actuales propuestas es que su funcionamiento se basa en la utilización de listas negras de rastreadores y anuncios, que ellos mismos mantienen y que, en algunos casos, confeccionan con la ayuda de comunidades de usuarios. El uso de estas listas por parte de AdBlock Plus, actualmente el bloqueador de anuncios más popular, ha provocado una gran controversia al conocerse que su desarrollador aceptaba dinero de algunas empresas de publicidad para evitar ser incluidas en ellas [6], [10]. Por otra parte, el mantenimiento de estas listas no es práctico, puesto que la industria de publicidad de Internet es muy dinámica y continuamente aparecen nuevos rastreadores y compañías de anuncios.

Iniciativas autorreguladoras. En los últimos años, varias iniciativas de la industria de la publicidad en línea han intentado dar respuesta a las preocupaciones de los usuarios de los bloqueadores de anuncios, como la usabilidad de la Web o los anuncios invasivos [11], [12]. Sin embargo, muchos de esos esfuerzos han ido encaminados a mejorar el proceso de distribución de la publicidad, y ninguno de ellos ha conseguido devolver el control a los usuarios.

Por otra parte, estas iniciativas no tienen en cuenta las motivaciones de privacidad que estos usuarios pueden tener, a pesar de que diversos estudios indican que un porcentaje elevado de ellos bloquea los anuncios por esta razón [4], [13]. El programa LEAN del Interactive Advertising Bureau (IAB), una de las mayores organizaciones de anuncios con cerca de 5 500 empresas y editores, menciona determinados aspectos relacionados con la no invasividad de los anuncios, pero obvia el riesgo de privacidad que supone la publicidad personalizada [11]. De la misma manera, el "Acceptable Ads Manifesto" incluye cinco puntos para mejorar la calidad de los anuncios, pero ninguno de ellos hace referencia a las prácticas de rastreo y creación de perfiles que vulneran ostensiblemente la privacidad de los usuarios en la red [12].

Tabla I
CATEGORÍAS DE CONTENIDO WEB EN LAS QUE UN USUARIO PUEDE BLOQUEAR LA PRESENCIA DE RASTREADORES.

adulto	economía	aficiones & intereses	política
agricultura	educación	hogar	inmuebles
animales	familia	ley	religión
arquitectura	moda	militar	ciencia
arte & entretenimiento	folklore	noticias	sociedad
automoción	comida & bebida	finanzas personales	deporte
negocios	salud & forma física	animales domésticos	tecnología & computación
carrera profesional	historia	filosofía	viajar

Otros ejemplos de iniciativas autorreguladoras son Your Online Choices [14] y Do Not Track (DNT) [15]. La primera es una plataforma mediante la cual los usuarios pueden comunicar a las empresas de anuncios si desean dejar de recibir anuncios personalizados[3]. Mediante la segunda iniciativa, los usuarios pueden notificar si quieren ser rastreados por terceros a través de *cookies* HTTP. El inconveniente de estas propuestas, sin embargo, es que los usuarios no tienen control sobre si sus preferencias de rastreo o anuncios son respetadas.

III. CONTROL REAL SOBRE EL RASTREO WEB

En este artículo, presentamos MyTrackingChoices[4], una herramienta que pretende dar solución a los problemas descritos en la Sección II. La solución que proponemos está dirigida a usuarios que, sin estar en contra de los anuncios per se, desean bloquearlos para evitar el rastreo, la construcción de perfiles y las inferencias que puedan extraerse de su navegación Web.

Nuestra solución parte de la suposición de que la mayoría de los usuarios no quieren ser rastreados en páginas de contenido sensible. Sin embargo, para apoyar a los proveedores de contenidos, aceptan ser rastreados —y por tanto recibir anuncios— en aquellas páginas cuyo contenido no consideran sensible.

El propósito de esta solución es sostener el actual modelo económico de la Web basado en publicidad, a diferencia de las soluciones actuales. La idea es que los usuarios puedan elegir las categorías de páginas que son sensibles a la privacidad y bloquear a los rastreadores, y así los anuncios, presentes en dichas páginas. Este nivel de granularidad supone una enorme ventaja con respecto a los actuales bloqueadores de anuncios y anti-rastreadores, de acuerdo con lo argumentado en la Sección II. Al mismo tiempo, dicha flexibilidad permite que los usuarios puedan seguir recibiendo anuncios adaptados a sus intereses, aunque obviamente en aquellas categorías que no hayan sido declaradas como sensibles.

Nuestro enfoque no sigue una política de "todo o nada" en comparación con los bloqueadores de anuncios actuales. Nuestra herramienta, además de dar la opción de escoger las categorías sensibles, también permite al usuario elegir si no quiere ser rastreado en páginas concretas, sin bloquear todas las páginas de esa misma categoría.

[3]Los anuncios de Internet pueden personalizarse en base a la localización del usuario, al contenido de la página visitada, a los intereses de navegación, o pueden no estar adaptados a ninguno de estos aspectos.

[4]https://chrome.google.com/webstore/detail/mytrackingchoices/fmonkjimgifgcgeocdhhgbfoncmjclka

III-A. *Detalles de implementación*

Hemos desarrollado MyTrackingChoices en forma de plug-in de navegador para Google Chrome[5]. En términos de funcionalidades, nuestra herramienta permite a los usuarios bloquear a los rastreadores —y por lo tanto, a los anuncios de terceros— en base a un conjunto de categorías predefinidas de contenido, y por página Web. Si los usuarios no están de acuerdo con la categorización de una página Web, éstos pueden cambiar la categoría asignada por el plug-in. Los usuarios también pueden ver la lista de rastreadores presentes en una página. Actualmente, MyTrackingChoices es compatible con páginas en español, inglés, francés e italiano.

Nuestra herramienta se compone de tres módulos principales, en concreto, un categorizador, un módulo de selección de política, y otro módulo de bloqueo de rastreadores. A continuación, describimos brevemente cada uno de estos módulos. A lo largo de esta descripción utilizaremos indistintamente los términos página Web y URL.

III-A1. Categorizador: Este módulo clasifica las páginas visitadas por el usuario en un conjunto predefinido de tema de interés. El módulo emplea una taxonomía jerárquica de 2 niveles, compuesta por 32 *categorías de nivel superior* y 330 *categorías de nivel inferior o subcategorías*. En aras de la usabilidad, la versión actual del plug-in permite que los usuarios seleccionen únicamente las categorías de nivel superior. La Tabla I muestra estas categorías temáticas.

El algoritmo de categorización integrado en nuestra herramienta está parcialmente inspirado en la metodología presentada en [17] para clasificar anuncios no textuales en categorías de interés. Nuestro algoritmo también se basa en la taxonomía disponible para el plug-in Firefox Interest Dashboard [18] desarrollado por Mozilla.

Nuestro categorizador se nutre de dos fuentes de datos previamente clasificados. En primer lugar, una lista de URLs, o más específicamente, dominios y *hostnames*, que es consultada para determinar la categoría de la página. En segundo lugar, una lista de unigramas y bigramas [19] que se utiliza cuando la búsqueda por URL no devuelve resultado. El primer tipo de datos se justifica por el hecho de que una parte relativamente pequeña de toda la Web representa la mayor parte de las visitas. Además, es evidente que las operaciones de búsqueda por datos pre-clasificados requieren pocos recursos computacionales en el navegador y pueden ser más precisas. El

[5]Actualmente se puede descargar de Chrome Web Store [16].

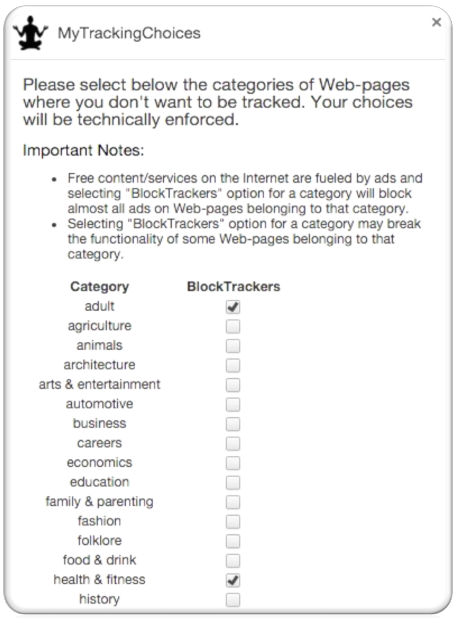

Figura 1. Panel de configuración de políticas de bloqueo por categorías.

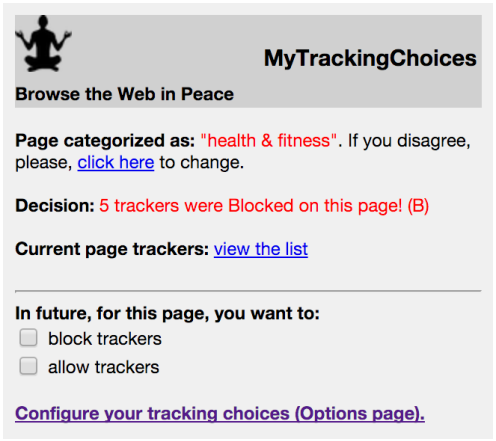

Figura 2. Menú principal de nuestra herramienta.

segundo tipo de información, por el contrario, se justifica como plan alternativo y nos permite aplicar heurísticas de lenguaje natural a las palabras disponibles en la URL, título, palabras clave y contenido.

Para casi cada una de las categorías de nivel superior, la versión actual del plug-in incorpora los 500 sitios más populares de Internet, de acuerdo con `alexa.com`[6]. Además, la lista de direcciones URL incluye las páginas clasificadas por el plug-in de Mozilla (alrededor de siete mil). Por otro lado, el número de unigramas y bigramas en inglés es de aproximadamente 76 000. Disponemos también de tres listas adicionales, aunque de un menor número de entradas, para los idiomas español, francés e italiano. Para compilar todas estas listas de palabras, nos hemos basado en los siguientes datos:

- una versión refinada de los datos de categorización disponibles en el plug-in Firefox Interest Dashboard;
- un subconjunto de los términos de inglés de WordNet 2.0 [20] para los que WordNet Domain Hierarchy [21], [22] propociona una etiqueta de dominio;
- un subconjunto de los términos disponibles en WordNet 3.0 Multilingual Central Repository [23], para permitir la categorización de las páginas escritas en los idiomas comentados anteriormente;
- y una lista de mapeo entre las entradas de las versiones 2.0 y 3.0 de WordNet [24].

El categorizador recurre a estas listas sólo cuando el *hostname* y el dominio no se encuentran en la base de datos de URLs. Cuando esto sucede, el algoritmo intenta clasificar la página utilizando las unigramas y bigramas extraídos de los siguientes campos de datos: URL, título, palabras clave y contenido. Dependiendo del campo de datos en cuestión, el categorizador asigna diferentes pesos a los unigramas y bigramas correspondientes. Al hacerlo, podemos reflejar el hecho de que los términos que aparecen en la URL, el título, y en especial las palabras clave especificadas por el editor (si están disponibles), suelen ser más descriptivas y explicativas que aquellas incluidas en el cuerpo de la página.

Como se hace habitualmente en el campo de búsqueda y recuperación de la información, nuestro clasificador de páginas también utiliza la frecuencia de término y la frecuencia inversa de documento (TF-IDF, *term frequency-inverse document frequency*) [25]. Dicho de otro modo, ponderamos la/s categoría/s resultante/s en función de la frecuencia de ocurrencia de los unigramas y bigramas correspondientes, y en base a una medida de su frecuencia dentro de la Web.

En términos de almacenamiento, la lista entera de unigramas, bigramas y sus correspondientes valores IDF ocupa aproximadamente 1 megabyte en formato comprimido. Por último, una inspección manual de los resultados de la categorización para una gran número de páginas Web indica que el algoritmo es, en casi todos los casos, ciertamente preciso. Sin embargo, sería necesario llevar a cabo una evaluación más rigurosa del categorizador.

III-A2. Módulo de políticas: Este módulo es el responsable de aplicar las políticas de bloqueo definidas por los usuarios. Los usuarios pueden definir dichas políticas mediante el panel de configuración mostrado en la Fig. 1.

Además de estas políticas, los usuarios pueden configurar el bloqueo en base a direcciones URL. Esto permite a los usuarios ser un poco más granulares si, ocasionalmente, no están conformes con las políticas de bloqueo por categoría. Por ejemplo, en un escenario en el que un usuario ha bloqueado una categoría de páginas Web, pero quiere dar soporte a una página específica en esa categoría al permitir sus anuncios. La Fig. 2 muestra el menú principal de MyTrackingChoices, donde es posible definir estas políticas de filtrado por URL.

El funcionamiento de este módulo se describe a continuación. Cuando un usuario visita una página, el módulo espera a que el categorizador le envíe la categoría correspondiente.

[6]Los motores de búsqueda no se incluyen obviamente en esta lista, y la categorización de la página de resultados se realiza precisamente en base a la consulta (aparece en la título de la propia página) y al contenido.

Equipado con la URL y dicha categoría, el módulo decide si las conexiones de red de terceros deben ser bloqueadas o no. En este punto, queremos destacar que si existe un conflicto entre las políticas por página y por categoría, el plug-in da preferencia a las primeras. En caso de que una página sea clasificada en múltiples categorías, si alguna de esas categorías ha sido seleccionada por el usuario como sensible, entonces las conexiones de red en la página son bloqueadas.

III-A3. Módulo de bloqueo: Este módulo es responsable de bloquear las conexiones de red de terceros en aquellas páginas clasificadas como sensibles. Lleva a cabo dos tareas principales. Primero, la búsqueda de dominios de terceros, y después, el bloqueo. Para encontrar los dominios de terceros, sólo tenemos que comprobar si los dominios de las conexiones de red coinciden con el dominio que el usuario introdujo en la barra de direcciones. Si no coinciden, entonces dichos dominios son considerados como dominios de terceros.

El siguiente paso es bloquear las conexiones de red de este tipo de dominios. Sin embargo, el bloqueo de las conexiones de red de todos los dominios de terceros podría dañar la funcionalidad de ciertas páginas Web. Esto se debe a que algunas páginas descargan contenidos de otros dominios (e.g., de un dominio distinto pero que pertenece al mismo editor, o un proveedor de contenidos). Para evitar esto, los bloqueadores de anuncios existentes mantienen una lista de dominios de empresas de anuncios y rastreadores. Típicamente, esta lista también incluye expresiones regulares que se utilizan para identificar sus correspondientes dominios.

MyTrackingChoices adopta un enfoque distinto para averiguar qué conexiones de terceros deben bloquearse, con tal de evitar que la funcionalidad de la página visitada se vea alterada. En lugar de basarse en estas listas, que son de un tamaño considerable y notablemente variables con el tiempo, nuestra herramienta parte de un conjunto reducido de dominios, esenciales para el funcionamiento de una página. En nuestra experiencia, una lista de este tipo es más fácil de mantener y de procesar.

En concreto, MyTrackingChoices bloquea un dominio de terceros que no se incluye en esta lista, y lo clasifica como rastreador si el dominio en cuestión ha sido observado en tres o más dominios distintos visitados por el usuario. Este modo de operación implica por tanto que, para que el plug-in sea completamente funcional, el usuario debe visitar algunas páginas después de la instalación. En esencia, utilizamos esta heurística para evitar mantener una lista de dominios de terceros que proporcionan servicios de contenido a los editores. Por ejemplo, `lemonde.fr` utiliza el dominio `lemde.fr` para descargar parte del contenido de sus páginas. Sin embargo, este dominio de terceros únicamente se encuentra en `lemonde.fr`. Por lo tanto, dominios de terceros como éste no serían clasificados como rastreadores.

Por último, querríamos destacar la imposibilidad de un rastreador de sortear el bloqueo realizado por nuestra herramienta. Como los bloqueos se llevan a cabo en el lado del usuario, una vez identificadas las conexiones de terceros, es ciertamienta imposible que el rastreador responsable de una conexión pueda evitar que se produzca la desconexión. Habiendo dicho esto, nuestra herramienta podría ser detectada por un publicador de la misma manera que se identifican a los usuarios de bloqueadores de anuncios y antirastreadores.

IV. Evaluación

A la espera de un mayor despliegue de nuestra herramienta, en esta sección llevamos a cabo una breve evaluación de la misma. En concreto, estudiamos el impacto económico que su adopción podría tener en la Web, y analizamos su rendimiento en términos de precisión en la categorización de contenido y carga computacional. Dicho análisis se ha llevado a cabo a partir de los datos recopilados de un conjunto de usuarios de nuestra herramienta. En particular, se han capturado los datos de navegación, rastreadores y anuncios de 96 usuarios, durante un periodo de 5 semanas. El conjunto de datos está formado por 86 922 páginas visitadas y 27 861 anuncios.

En primer lugar, analizamos las preferencias de bloqueo seleccionadas por los usuarios. Particularmente interesante fue observar que el 30.02 % de los usuarios optó por bloquear todas las categorías. Este tipo de usuarios responde claramente al perfil típico de usuario de los bloqueadores de anuncios, cuya motivación es evitar cualquier tipo de rastreo y/o los anuncios. Evidentemente, como estos usuarios no desean ejercer un control granular sobre el rastreo, nuestro plug-in no contribuiría a reducir el impacto en la Web. Por contra, el 69.98 % restante de usuarios decidió bloquear, en promedio, 8.11 categorías (de un total de 32).

En términos de páginas bloqueadas, observamos que las 3 categorías más afectadas fueron “adultos”, “religión” y “salud & forma física”. Estas categorías en realidad son consideradas como sensibles por la ley de protección de datos europea [9]. Las 3 categorías menos afectadas fueron “tecnología & computación”, “ciencia” y “noticias”. Respecto a los anuncios, las categorías de páginas con mayor publicidad fueron “arte & entretenimiento”, “tecnología & computación” y “noticias”. Por otro lado, como era de esperar, las páginas con mayor número de anuncios bloqueados fueron las clasificadas como “adultos”, “salud & forma física”, “política” y “finanzas personales”.

Del análisis de páginas y anuncios bloqueados, la conclusión más destacable es que tan sólo fueron bloqueados el 33.19 % de las páginas visitadas y el 23.8 % de los anuncios entregados. Teniendo en cuenta que sólo en 2015 el bloqueo de anuncios supuso unas pérdidas para los editores de cerca de 22 000 millones de euros [4], si todos los bloqueadores de anuncios fueran sustituidos por nuestra herramienta, los creadores de contenido habrían ahorrado 16 720 millones de euros.

Por otro lado, también evaluamos nuestro plug-in en términos de precisión y carga computacional. En los experimentos llevados a cabo, MyTrackinChoices clasificó correctamente casi todo el contenido Web. Esta conclusión se basa en la rectificación que los propios usuarios pueden hacer de la clasificación de la páginas. En tan sólo 18 de las 12 543 páginas distintas de nuestro conjunto de datos, los usuarios

reportaron una categoría temática distinta de la obtenida por nuestro categorizador.

Finalmente, evaluamos el tiempo de carga de página, utilización de memoria y de CPU, y comparamos estos resultados con algunos de los bloqueadores de anuncios y anti-rastreadores más populares, en concreto, con uBlock Origin, SuperBlock Adblocker, Ghostery, AdRemover, Adguard AdBlocker, AdBlock Pro, Adblock Plus y AdBlock.

Los resultados de este análisis fueron obtenidos después de 10 visitas consecutivas a estas 3 páginas Web: `ara.cat`, `lemonde.fr` and `nytimes.com`. La principal conclusión que extraemos de esta comparativa es que nuestro plug-in, a pesar de su mayor complejidad, superó a la mayoría de las herramientas estudiadas. En términos de tiempo de carga de página, MyTrackingChoices mostró unas prestaciones similares a Ghostery and uBlock Origin, las dos herramientas que encabezaron este aspecto. En cuanto a utilización de memoria, nuestra herramienta requirió, en promedio, entre 37 y 39 MB, ocupando la tercera posición, detrás de Ghostery y uBlock Origin en `lemonde.fr` and `nytimes.com`, y la segunda detrás de Ghostery en `ara.cat`. Por último, con tan sólo un 2 % de utilización de procesado, nuestro plug-in superó todas las herramientas en consumo de CPU. En resumen, el rendimiento de MyTrackingChoices estuvo por encima de la media en esas tres páginas, a pesar de la incorporación de funcionalidades de privacidad más sofisticadas.

V. Conclusión y Trabajo Futuro

Las herramientas existentes basadas en el bloqueo de cualquier forma de publicidad no han devuelto al usuario el control sobre el rastreo y la publicidad, y lo que es peor, amenazan seriamente el actual modelo económico de la Web.

En este artículo, proponemos una herramienta que tiene como objetivo dotar al usuario de ese control y, al mismo tiempo, lograr un equilibrio entre la privacidad de usuario y la economía de Internet. MyTrackingChoices es una herramienta que permite a los usuarios ejercer un control más flexible y exhaustivo sobre las entitades que rastrean su navegación en la red. En lugar de plantear un control binario, como ofrecen los bloqueadores de anuncios y anti-rastreadores actuales, nuestro plug-in permite a los usuarios elegir las categorías de las páginas que son sensibles en términos de privacidad para ellos, y donde no desean ser rastreados. Como resultado de esta elección, los usuarios pueden controlar el perfil de navegación recopilado por companías de anuncios, rastreadores y brokers de datos.

Nuestros resultados experimentales muestrán cómo la adopción de nuestra herramienta podría ayudar a reducir significativamente el impacto económico que los actuales bloqueadores de anuncios tiene en Internet. Como líneas de trabajo futuro, querríamos llevar a cabo una evaluación mucho más minuciosa de dicho impacto con una población de usuarios mayor.

Agradecimientos

Este trabajo está financiado en parte por Inria Project Lab CAPPRIS. J. Parra-Arnau es beneficiario de una beca posdoctoral Juan de la Cierva, FJCI-2014-19703, del Ministerio de Economía y Competitividad.

Referencias

[1] K. Purcell, J. Brenner, and L. Rainie, “Search engine use 2012,” Pew Internet, Amer. Life Project,” Res. Rep., Mar. 2012.

[2] “The state of online advertising,” Adobe, Tech. Rep., 2012, accessed on 2015-09-11. [Online]. Available: http://www.adobe.com/aboutadobe/pressroom/pdfs/Adobe_State_of_Online_Advertising_Study.pdf

[3] G. Marvin, “Consumers now notice retargeted ads,” Marketing Land, Tech. Rep., Dec. 2013, accessed on 2015-08-12. [Online]. Available: http://marketingland.com/3-out-4-consumers-notice-retargeted-ads-67813

[4] “The cost of ad blocking,” PageFair,” Res. Rep., Aug. 2015.

[5] “Adblock plus,” accessed on 2015-10-22. [Online]. Available: https://adblockplus.org

[6] R. Cookson, “Google, Microsoft and Amazon pay to get around ad blocking tool,” *Financial Times*, Feb. 2015, accessed on 2014-03-10. [Online]. Available: http://www.ft.com/cms/s/0/80a8ce54-a61d-11e4-9bd3-00144feab7de.html

[7] “Adblock plus user survey results, part 3,” Eyeo, Tech. Rep., Dec. 2011, accessed on 2015-07-11. [Online]. Available: https://adblockplus.org/blog/adblock-plus-user-survey-results-part-3

[8] D. Rogers, “How business can gain consumers’ trust around data,” Nov. 2015, accessed on 2015-11-03. [Online]. Available: http://www.forbes.com/sites/davidrogers/2015/11/02/how-business-can-gain-consumers-trust-around-data/

[9] “Handbook on European data protection law,” http://www.echr.coe.int/Documents/Handbook_data_protection_ENG.pdf, 2014, [Online; accessed 17-February-2016].

[10] “Rip: Adblock plus,” http://www.engadget.com/2016/02/12/rip-adblock-plus/, Blog on Engadget.

[11] “Getting LEAN with Digital Ad UX,” http://www.iab.com/news/lean/, 2015, [Online; accessed 17-February-2016].

[12] “Acceptable Ads Manifesto,” https://acceptableads.org/, 2015, [Online; accessed 17-February-2016].

[13] “Profiling Adblockers,” http://www.globalwebindex.net/blog/profiling-adblockers, 2015, [Online; accessed 17-February-2016].

[14] “YourOnlineChoices,” European Interact. Digit. Advertising Alliance. [Online]. Available: http://www.youronlinechoices.com/

[15] “Tracking preference expression (DNT),” Tech. Rep., Aug. 2015. [Online]. Available: http://www.w3.org/TR/tracking-dnt/

[16] “MyTrackingChoices: Available to download at Chrome Web Store,” https://chrome.google.com/webstore/detail/mytrackingchoices/fmonkjimgifgcgeocdhhgbfoncmjclka?hl=fr, 2016, [Online; accessed 21-February-2016].

[17] A. Kae, K. Kan, V. K. Narayanan, and D. Yankov, “Categorization of display ads using image and landing page features,” in *Proc. ICDM Workshop Large-Scale Data Min.: Theory, Appl.* ACM, 2011, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/2002945.2002946

[18] “Firefox interest dashboard,” Nov. 2014, accessed on 2015-05-02. [Online]. Available: https://www.mozilla.org/en-US/firefox/interest-dashboard/

[19] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[20] G. A. Miller, “WordNet: A lexical database for english,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[21] B. Magnini and G. Cavaglià, “Integrating subject field codes into wordnet,” in *Proc. Lang. Resource, Evaluation (LREC)*, June 2000, pp. 1413–1418.

[22] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta, “Revising wordnet domains hierarchy: Semantics, coverage, and balancing,” in *Proc. Post-COLING Workshop Multiling. Ling. Resources*, Hangzhou, China, Aug. 2004, pp. 101–108.

[23] A. Gonzalez-Agirre, E. Laparra, and G. Rigau, “Multilingual central repository version 3.0: upgrading a very large lexical knowledge base,” in *Proc. Global WordNet Conf.*, 2012.

[24] J. Daudé, , L. Padró, and G. Rigau, “Validation and tuning of wordnet mapping techniques,” *Proc. Int. Conf. Recent Adv. Nat. Lang. Process. (RANLP)*, Sept. 2003.

[25] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.