Propuesta de Tesina

October 29, 2025

Postulante: Bolzan Francisco
Director: Matic Srdjan
Codirector: Cristia Maximiliano

1 Situación del postulante

Al día 05/10/2025 tengo aprobadas 32 materias de la carrera, restando únicamente la carga en SIU Guaraní de la nota de la materia optativa *Computación Paralela*, la cual tengo aprobada, y realizar la tesina para finalizar la carrera de *Licenciatura en Ciencias de la Computación*. Me encuentro trabajando como ayudante de investigación en el *Instituto Madrileño de Estudios Avanzados - IMDEA*, en el área de Software, por 40 horas semanales.

2 Título

Desarrollo de un *framework* de vanguardia para descubrir técnicas de rastreo en línea en navegadores modernos.

3 Motivación y Objetivo General

El rastreo en línea consiste en la detección e identificación de usuarios y sus comportamientos al navegar la web. Dicha práctica depende profundamente de la generación de identificadores persistentes sobre el protocolo HTTP, el cual es efímero; para esto se abusan diversos mecanismos de persistencia en navegadores, siendo entre ellos el más importante las denominadas *cookies*.

Múltiples investigaciones han estudiado soluciones al problema del rastreo en línea mediante *cookies*, desde sistemas de detección y estudios comparativos hasta sistemas de bloqueo automático de *cookies* utilizadas para rastreo. Estas investigaciones son de gran importancia debido a la constante carrera tecnológica que tiene lugar entre los avances en detección y legislación contra el rastreo en línea, y las nuevas técnicas utilizadas por los rastreadores, muchas veces opacas y evasivas.

Inicialmente el mecanismo mas abusado con propósitos de rastreo eran las *third-party cookies*, las cuales provienen de un sitio distinto a aquel que el usuario esta navegando. Es alrededor de 2016 que surgen legislaciones y marcos regulatorios que buscan restringir dichas capacidades de rastreo, pidiendo consentimiento explicito a los usuarios sobre

que *cookies* aceptan y detallando el propósito de las mismas. Ante esto nuevas técnicas aparecen, entre estás la más replicada siendo el rastreo mediante *first-party cookies* como reemplazo de las *third-party cookies*, aprovechando el rol crítico que estas tienen para la funcionalidad de los sitios web lo cual las hace menos propensas a técnicas agresivas de bloqueo.

Múltiples de las soluciones planteadas en el ámbito académico están basadas en sistemas de *Machine Learning*, los cuales buscan recopilar comportamientos y características de la navegación (como podrían ser datos de la pila de llamadas a funciones, dependencias entre orígenes y exfiltraciones de datos, características de los nombres y valores de las *cookies*, etc) en estructuras ordenadas y catalogadas sobre las cuales luego entrenar un modelo de *ML* orientado a distinguir aquellos comportamientos y características que denotan instancias de rastreo en línea. Estos modelos catalogan las *cookies* de forma probabilística, dicha estrategia falla en dar evidencia clara y replicable la cual pueda ser utilizada tanto por usuarios para juzgar la privacidad de su navegación en la web como por legisladores o investigadores que busquen preservar la integridad de la información de los usuarios. Siguiendo esta línea de investigación, buscamos desarrollar un sistema automatizado y determinista basado en pruebas diferenciales y modelos de inferencia destinado a evaluar el uso de *cookies* en el navegador y descubrir casos de filtraciones de información a entidades externas.

Para lograr este objetivo distinguimos dos etapas del proyecto. Inicialmente nos interesa evaluar sistemáticamente el estado actual del uso de *cookies* en los navegadores modernos, incluyendo sus tipos, orígenes y su duración en el navegador. Por otro lado buscamos desarrollar un sistema automatizado capaz de:

1. Detectar cuando nuevas *cookies* son creadas.
2. Identificar el mecanismo de creación de las mismas.
3. Seguir el uso de dicha *cookie* para descubrir sus posibles exfiltraciones a terceros.
4. Permitir a usuarios evaluar los puntos anteriores sobre sitios de interés y generar reportes sobre los resultados.

4 Fundamentos y estado de conocimiento sobre el tema

Uno de los estudios más influyentes sobre el uso de *cookies* para el rastreo en línea es el realizado por S. Englehardt [13] [12], quien, a través de una herramienta desarrollada en el marco de su investigación, creó uno de los primeros motores de privacidad específicamente diseñados para navegadores modernos, en este caso, para Mozilla Firefox. Su aporte, permitió visibilizar el uso generalizado de *third-party cookies* con propósitos de rastreo en algunos de los sitios web más visitados. Además, el estudio reveló diversas técnicas utilizadas para eludir protocolos de seguridad y normativas de regulación tales como *cookie syncing*, técnica que busca ignorar la *Same-Origin Policy*, y el uso temprano de técnicas de *browser fingerprinting* para la generación de información identificable sobre los usuarios.

Posteriormente, con la adopción de las *first-party cookies* como mecanismo de rastreo,

I. Sanchez-Rola [2] se centró en un análisis más detallado de las acciones y actores involucrados en el rastreo a través de este tipo de *cookies*. Su contribución principal fue la clasificación de los distintos roles y relaciones de cada actor dentro de estos esquemas de rastreo. Además, destacó las técnicas específicas empleadas por cada uno de ellos, lo que permitió evidenciar la compleja red de dependencias, tanto colaborativas como unilaterales, que caracteriza el uso de las *first-party cookies* como medio de rastreo. Este análisis subrayó la amplia adopción de esta práctica y su integración en la web.

Ambas técnicas de rastreo mencionadas (mediante *third-party cookies* y *first-party cookies*) se engloban dentro de lo que se conoce como "rastreo con estados", donde el mecanismo que mantiene el estado de la información son precisamente las *cookies*. En años posteriores, diversos estudios se han centrado en la detección y/o bloqueo de *cookies* en sitios web que emplean este tipo de técnicas. Entre ellos, destacan las previamente mencionadas investigaciones orientadas al entrenamiento de sistemas de *Machine Learning* para clasificar las *cookies* según diversos criterios [1] [10]. Estos sistemas han probado ser altamente efectivos en la identificación de técnicas de rastreo y ademas mantienen una alta granularidad lo que permite a su vez una alta precisión a la hora de bloquear recursos de rastreo que pueden a su vez ser necesarios para la funcionalidad del sitio web.

Otros estudios optaron por rastrear las *cookies* sospechosas hasta sus puntos de exfiltración mediante técnicas de "tintes" [8]. El objetivo en este caso es propagar algún tipo de tinte o identificador desde el origen/creación de una *cookie* hasta sus diversos puntos de exfiltración y luego comparar los dominios de cada componente (*cookie*, script, URL final, etc) a modo de clasificar el comportamiento observado. Dichos sistemas requieren mayor intervención manual que los previamente mencionados pero a cambio son mas deterministas en sus resultados, brindando ademas evidencia clara para la toma de decisiones.

Finalmente, también se encuentran estudios basados en el uso de heurísticas y características específicas de las *cookies* o de los scripts que las generan para la catalogación de las mismas [5] [7] [6] [3]. Este es sin duda el método de detección de rastreo mas básico y que mas mantenimiento manual requiere. A su vez, múltiples de los bloqueadores que encontramos como extensiones de navegadores modernos se basan o implementan este tipo de técnicas junto con listas curadas de dominios previamente identificados como maliciosos.

Por otro lado, también existen técnicas de "rastreo sin estados" que permiten identificar a los usuarios a través de algún tipo de firma digital comúnmente obtenida mediante la combinación de múltiples características identificables y persistentes del dispositivo de navegación utilizado, como pueden ser datos del hardware y configuración del navegador entre otras [4] [9] [11]. Un ejemplo de estas técnicas es el previamente mencionado *browser fingerprinting*, que se ha ido popularizando en los últimos años. Aunque estas técnicas son actualmente menos comunes que las basadas en *cookies*, se están utilizando cada vez más, especialmente en combinación con las anteriores. Sin embargo, en el contexto de este estudio, nos centramos principalmente en las técnicas basadas en estados debido a su mayor granularidad e invasividad. Las técnicas de rastreo con *cookies* no solo tienen la capacidad de identificar a los usuarios, incluso asociándolos con información sensible, como direcciones de correo electrónico, sino que también permiten el seguimiento de comportamientos específicos, como enlaces visitados o productos adquiridos.

5 Metodología y Plan de Trabajo

1. Estudio de técnicas de rastreo en línea mediante el uso de *cookies* y herramientas existentes para la detección de dicha actividad (1 mes).
2. Desarrollo de un estudio diferencial a gran escala para discernir los mecanismos preferidos para la creación de *cookies* (1 mes).
3. Desarrollo del *framework* para la detección automática, mediante pruebas diferenciales e inferencias, de la filtración de información mediante *cookies* (3 meses).
4. Evaluación de uso típico del *framework* desarrollado, usando múltiples sitios y navegadores populares (1 mes).

Referencias

[1] A. Haddi Amjad et al. "Blocking Tracking JavaScript at the Function Granularity". *ACM SIGSAC Conference on Computer and Communications Security* (2024).

[2] I. Sanchez-Rola et al. "Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships". *IEEE Symposium on Security and Privacy* (2021).

[3] M. Ahmad Bashir et al. "Tracing Information Flows Between Ad Exchanges Using Retargeted Ads". *USENIX Security Symposium* (2016).

[4] N. Nikiforakis et al. "Cookieless Monster: Exploring the Ecosystem of Web-based Device Fingerprinting". *IEEE Symposium on Security and Privacy* (2013).

[5] P Nikkhah Bahrami et al. "CookieGuard: Characterizing and Isolating the First-Party Cookie Jar". *ACM Internet Measurement Conference* (2025).

[6] P. Bekps et al. "The Hitchhiker's Guide to Facebook Web Tracking with Invisible Pixels and Click IDs". *The ACM Web Conference* (2023).

[7] P. Papadopoulos et al. "Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask". *The ACM Web Conference* (2019).

[8] Q. Chen et al. "Cookie Swap Party: Abusing First-Party Cookies forWeb Tracking". *The ACM Web Conference* (2021).

[9] S. Boussaha et al. "FP-tracer: Fine-grained Browser Fingerprinting Detection via Taint-tracking and Entropy-based Thresholds". *Privacy Enhancing Technologies Symposium* (2024).

[10] S. Munir et al. "CookieGraph: Understanding and Detecting First-Party Tracking Cookies". *ACM SIGSAC Conference on Computer and Communications Security* (2023).

[11] Peter Eckersley. "How Unique Is Your Web Browser?" *Privacy Enhancing Technologies Symposium* (2010).

[12] Steven Tyler Englehardt. "Automated Discovery of Privacy Violations on the Web". *Princeton University* (2018).

[13] A. Narayanan S. Englehardt. "Online Tracking: A 1-million-site Measurement and Analysis". *ACM SIGSAC Conference on Computer and Communications Security* (2016).