Universidade de Aveiro
Ano 2022

Ricardo Emanuel Couto Madureira

Quadro para avaliar a privacidade dos utilizadores de um sítio Web

Universidade de Aveiro
Ano 2022

Ricardo Emanuel Couto Madureira

Quadro para avaliar a privacidade dos utilizadores de um sítio Web

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica do Professor Doutor André Ventura da Cruz Marnôto Zúquete, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro e do Doutor Hélder José Rodrigues Gomes, Professor Adjunto na Escola Superior de Tecnologia e Gestão de Águeda da Universidade de Aveiro.

o júri

presidente

Professor Doutor Paulo Jorge Salvador Serra Ferreira, Professor Associado, Universidade de Aveiro

Vogais

Arguente Principal: Professora Doutora Ana Rita Costa Bonifácio Selores dos Santos, Professora Adjunta, Universidade de Aveiro

Orientador: Professor Doutor André Ventura da Cruz Marnoto Zúquete, Professor Auxiliar, Universidade de Aveiro

Agradecimentos

O primeiro agradecimento terá de ser para a minha mãe, porque sem o seu apoio, a todos os níveis, nunca teria começado esta etapa da minha vida, por isso a ela o meu eterno obrigado. Quero também deixar uma palavra de agradecimento à minha irmã, porque, apesar de ela não ter conhecimento disso, ajudou-me muito a acabar este trabalho. De seguida, quero agradecer ao Professor André Zúquete e ao Professor Hélder Gomes por todo o conhecimento que me transmitiu e a disponibilidade de ambos para esclarecer qualquer dúvida que tive durante o desenvolvimento deste trabalho. A ambos, quero agradecer a paciência, as conversas sobre os mais variados tópicos e todas as experiências que partilhámos que me ajudaram a ter uma visão diferente da qual tinha quando começámos este trabalho. Finalmente, deixo um obrigado a todos os colegas com quem tive a oportunidade de trabalhar e que, também, me ajudaram a formar a pessoa que sou hoje.

palavras-chave *cookie*, rastreamento, privacidade, utilizador, sítio Web

Resumo

Para proteger a privacidade dos utilizadores, foram implementadas diversas políticas de privacidade sobre a utilização de tecnologias de rastreio, para possibilitar a aquisição do consentimento dos utilizadores antes da recolha de dados. Mas será que um sítio Web cumprindo todas estas políticas respeita por completo a privacidade de um utilizador? Não arranja outras formas subliminares para violar a privacidade de um utilizador e desta forma rastreá-lo sem o seu consentimento?

A presente dissertação tem por objetivo fazer uma análise a um conjunto variado de sítios Web, com o intuito de certificar que estes sítios Web estão de acordo com um conjunto de requisitos relacionados com a privacidade a partir de serviços que o utilizador esteja a usufruir. Neste sentido foi criado um quadro de avaliação que proporcionou uma classificação, tendo em conta atributos implementados num sítio Web que são considerados prejudiciais para a privacidade de um utilizador. Desta forma os sítios Web foram separados por categorias de jornais, compras, bancos, governo, instituições universitárias e sítios Web relacionados com transportes e reservas online. A categoria de jornais foi a que demonstrou ter um nível de privacidade mais grave para o utilizador com uma percentagem de 80% e a categoria de universidades foi a que demonstrou ter uma privacidade muito boa para o utilizador com uma percentagem de 100%.

Keywords *cookie*, tracking, privacy, user, website

Abstract

To protect user's privacy, several privacy policies on the use of tracking technologies have been implemented, to make it possible to acquire user's consent before collecting data. But does a Website complying with all these policies fully respect a user's privacy? Doesn't it come up with other subliminal ways to violate a user's privacy and thus track him without his consent?

The purpose of this dissertation is to analyze a variety of Websites, in order to certify that these Websites comply with a set of requirements related to privacy from services that the user is enjoying. In this sense, an evaluation table was created that provided a classification, taking into account attributes implemented in a Website that are considered harmful to the privacy of a user. In this way the Websites were separated by categories of newspapers, shopping, banking, government, university institutions, and transportation-related websites and online booking. The newspaper category was the one that was shown to have the most severe level of privacy for the user with a percentage of 80% and the university institutions category was the one that was shown to have very good privacy for the user with a percentage of 100%.

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

CAPTCHA - Completely Automated Public Turing test to tell Computers and Humans Apart

CDN - content delivery network

CTR - Click-through rate

DNS - Domain Name System

HTML - Hypertext Markup Language

HTTP - Hypertext Transfer Protocol

KPI - Key Performance Indicator

NSA - National Security Agency

PNG - Portable Network Graphic

SSL - Secure Sockets Layer

TCP/IP - Transmission Control Protocol/Internet Protocol

URL - Uniform Resource Locator

1. Introdução

O utilizador tem de estar consciente que quando acede a um sítio na Web, poderá estar sujeito a um conjunto de ameaças que podem, por exemplo, ter como objetivo identificar o utilizador para depois criar um perfil sobre o mesmo. Existem diversas técnicas que permitem concretizar esse objetivo, como por exemplo, a análise dos endereços IP e a utilização abusiva de *JavaScript* e *Flash* (Zviran, 2008). Pretende-se definir um quadro que permita avaliar qual o nível de perigosidade em termos de violação da privacidade que um sítio Web potencialmente representa para os utilizadores, com base nas tecnologias "perigosas" identificadas nos recursos Web fornecidos pelo sítio Web. É dessa forma possível que, por exemplo, esteja a ser obtida do utilizador, eventualmente sem o seu conhecimento, informações que poderão levar à sua identificação e consequentemente distinguir os diferentes utilizadores que acedem a um determinado serviço. Há inúmeras técnicas de interação com o ambiente computacional do cliente que podem ser usadas para melhorar a sua experiência no contacto com o serviço, mas que também podem ser usadas para revelar quem ele é de uma forma sub-reptícia. Isto faz com que a confiança dos utilizadores na Web seja posta em causa.

É, então, relevante ter um quadro de avaliação que permita medir, consoante as técnicas usadas, se um determinado sítio Web oferece um nível de privacidade considerável ao utilizador. Para realizar essa tarefa, pode recorrer-se a um conjunto de plataformas disponíveis na Web que fazem a análise externa a servidores Web. O *WBF Analyzer* (de Matos & Feitosa, 2021) e a *PrivacyScore* (Maass et al., 2017) são dois exemplos deste tipo de métodos e plataformas, respetivamente. Neste método (*WBF Analyzer*) e plataforma (*PrivacyScore*), observa-se a existência de grandes diferenças em relação ao que se pretende demonstrar neste estudo. Por exemplo, a plataforma *PrivacyScore* realiza uma classificação sobre os diversos sítios na Web que analisa, mas foca-se maioritariamente em questões relacionadas com a segurança e não com os mecanismos que irão ser abordados neste estudo.

1.1. Objetivos

A presente dissertação tem como objetivos a deteção e caracterização das várias técnicas sub-reptícias que podem ser utilizadas por diversos sítios Web e a criação de um quadro de avaliação que produza uma classificação para sítios na Web em função das técnicas encontradas nos recursos Web fornecidos pelo site. Essa classificação será realizada através de uma plataforma que irá demonstrar aos utilizadores que os serviços, implementados nos sítios na Web, estão de acordo com um conjunto de requisitos relacionados com a privacidade. É de realçar que o objetivo não é detetar se um determinado sítio Web está a fazer o rastreio de um determinado utilizador, mas sim identificar as tecnologias utilizadas por um sítio Web e, em função do potencial destas para serem utilizadas para rastreio do utilizador, produzir uma classificação de perigosidade.

2. Contextualização

A presente dissertação tem por objetivo criar um quadro de avaliação que produza uma classificação para os sítios Web, para tal é necessário compreender a forma como estes sítios Web funcionam. Este capítulo irá contextualizar o leitor sobre o ambiente em que esta dissertação se insere. É feita uma breve explicação sobre o conceito de sítio na Web, a forma como funciona, o que se entende por rastreamento na Web e uma breve introdução sobre alguns conceitos que são necessários para se entender a forma como o rastreamento na Web pode ser utilizado sem o consentimento de um utilizador.

2.1 Um sítio na Web

Um sítio Web pode ser entendido como um tipo de programa informático, ou seja, utiliza tecnologias online (incluindo navegadores) para realizar uma enorme variedade de diferentes tarefas. Pode servir para todo o tipo de propósitos diferentes, como por exemplo, encomendar comida, reservar férias, entre outras finalidades. Alternativamente, um sítio Web pode ser algo tão simples como um formulário de contacto ou calculadoras *online*.

Os sítios Web recuperam e guardam informação através de *scripts*. Esta informação pode assumir qualquer tipo de forma, os exemplos mais comuns que ocorrem nos sítios Web podem ser através de carrinhos de compras, sistemas de gestão de conteúdos e formulários *online* (RingCentral, 2022).

2.1.1 Como funciona uma aplicação Web

Uma aplicação Web caracteriza-se por dois componentes fundamentais, sendo eles designados por cliente e servidor, tipicamente em máquinas diferentes, como ilustrado na Figura 1 (T. M. W. Docs, 2022). Observando estes componentes, é importante referir que os mesmos comunicam segundo um modelo cliente-servidor, utilizando o protocolo HTTP para comunicar, em que o cliente pede ao servidor recursos Web que são disponibilizados em formato HTML e renderizadas no ecrã pela aplicação Web.

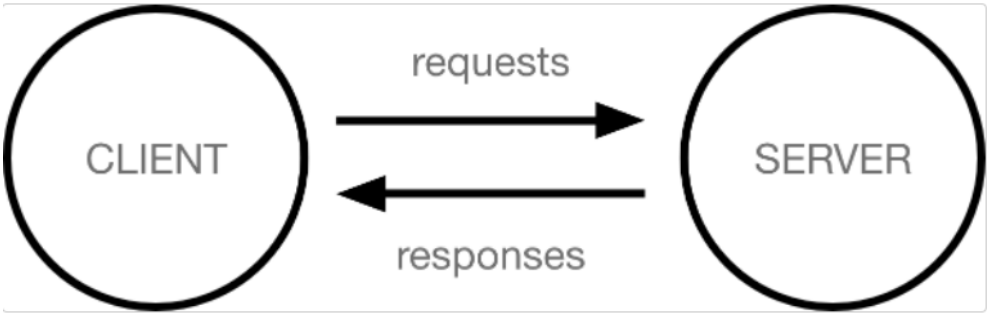


Figura 1 - Introdução ao funcionamento de um sítio Web

Fonte: (T. M. W. Docs, 2022)

A entidade cliente pode ser entendida como uma aplicação que está inserida num dispositivo normal de um utilizador ligado à Internet, como por exemplo um computador ou um telemóvel. No entanto estas entidades também podem ser vistas como um *software* de acesso à Web que se encontra disponível nestes dispositivos mencionados, como por exemplo, um navegador (*Firefox*, *Chrome*, entre outros*)*.

As entidades designadas por servidor também podem ser entendidas como sendo aplicações que correm em servidores e atendem a pedidos de clientes. Quando um dispositivo do lado do cliente pretende obter um recurso Web, é gerada uma cópia do mesmo e descarregado do servidor para a máquina do cliente. Para além destas entidades, cliente e servidor, também é preciso dar a devida atenção a outros aspetos importantes relativamente ao funcionamento de uma aplicação Web.

TCP/IP

O termo TCP/IP, que tem como significado *Transmission Control Protocol*/Internet *Protocol*, pode ser entendido como uma pilha de protocolos que permitem a comunicação entre dispositivos na Internet.

Concretamente, os dispositivos ligados à Internet são identificados por um endereço IP, que nalguns casos é estático e noutros é dinâmico. Os servidores Web podem ser identificados através de nomes de domínio. A conversão entre estes endereços IP é feita pelo serviço de DNS que será explicado com maior detalhe no ponto seguinte.

DNS

Esta funcionalidade designada por *Domain Name System*, pode ser entendida como um livro de endereços para os sítios Web. Quando se digita um endereço Web num navegador, o navegador usa o componente DNS desse endereço para encontrar o endereço IP do sítio Web antes de lhe poder aceder via TPC/IP.

O processo de como esta funcionalidade ocorre é através de uma conversão do nome do sítio Web, como por exemplo "site1.com", num endereço IP do computador do servidor. É dado um endereço IP a cada dispositivo e esse endereço é necessário para encontrar o dispositivo apropriado - tal como um endereço de rua é utilizado para encontrar uma determinada casa. Quando um utilizador acede a um recurso, deve ocorrer uma tradução entre o que um utilizador escreve no navegador Web e o endereço da máquina necessário para localizar esse recurso (DNS, 2022).

Hypertext Transfer Protocol (HTTP)

Trata-se de um protocolo de camada da aplicação para a transmissão de recursos, como por exemplo, documentos HTML. Foi desenvolvido para comunicações entre navegadores e servidores Web. Um fluxo típico deste protocolo envolve uma máquina cliente que faz um pedido a um servidor, que depois envia uma mensagem de resposta.

Este protocolo utiliza métodos de pedido específicos a fim de executar várias tarefas. Todos os servidores HTTP utilizam os métodos *GET* e *HEAD*, mas nem todos suportam os restantes métodos que irei apresentar:

- *GET:* solicita um recurso específico na sua totalidade;
- *HEAD:* solicita um recurso específico;
- *POST:* adiciona conteúdos, a um recurso existente;
- *PUT:* modifica um recurso existente;
- *DELETE:* apaga um recurso específico.

A comunicação usando o protocolo HTTP não oferece qualquer segurança na comunicação entre clientes e servidores Web. O protocolo HTTPS veio colmatar essa deficiência garantindo a confidencialidade e autenticação.

Protocolo HTTPS

Desta forma, um pedido HTTP é a forma de como as plataformas de comunicação da Internet, tais como os navegadores web, pedem as informações que necessitam para carregar um sítio web. O protocolo HTTPS (*Hypertext Transfer Protocol Secure*) é uma versão segura do protocolo HTTP que utiliza o protocolo SSL/TLS para controlo da confidencialidade e autenticação. Este protocolo torna possível, por exemplo, aos utilizadores que acedem a um sítio Web através do navegador transmitir dados sensíveis de forma segura através da Internet. O protocolo HTTPS adiciona controlo da confidencialidade e autenticação ao protocolo HTTP:

Confidencialidade dos servidores - Sendo o protocolo HTTP originalmente concebido como um protocolo de texto, é vulnerável a escutas e ataques. Ao incluir o SSL/TLS, o HTTPS impede que os dados enviados pela Internet sejam compreendidos por entidades terceiras.

Autenticação - Ao contrário do protocolo HTTP, o protocolo HTTPS inclui uma autenticação robusta através do protocolo SSL/TLS. O certificado SSL/TLS de um sítio web inclui uma chave pública, que um navegador web pode utilizar para confirmar que os recursos enviados pelo servidor, como por exemplo páginas HTML, foram assinados digitalmente por alguém na posse da chave privada correspondente (SSL.com, 2021).

Pode-se observar de uma forma ilustrativa, a diferença entre estes dois protocolos na Figura 2 (HARNISH, 2021).

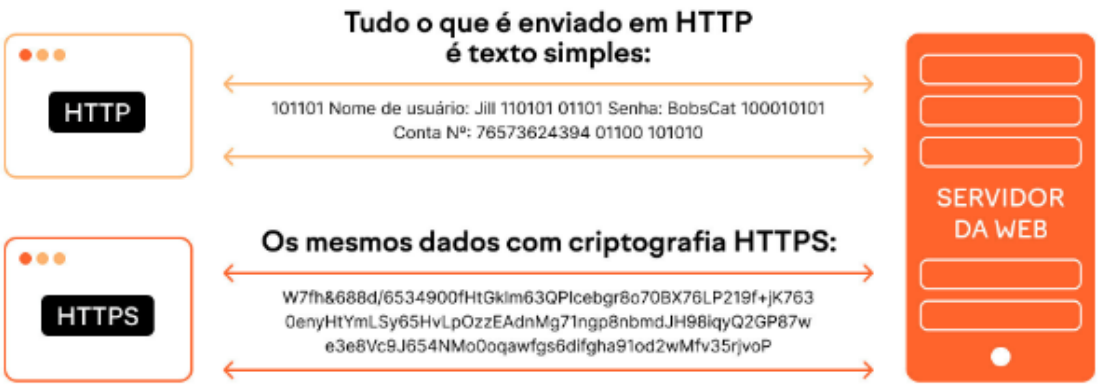


Figura 2 - Diferença entre HTTP e HTTPS

Fonte: (HARNISH, 2021)

Cabeçalhos *Referrer*

Cada pedido HTTP que é feito através da Internet transporta consigo uma série de dados codificados que transportam diferentes tipos de informação. Em cada pedido HTTP, existe os cabeçalhos que contem informações de texto armazenadas na forma de pares chave- valor. Estes cabeçalhos comunicam informações essenciais, como por exemplo, o navegador que o utilizador está a utilizar e quais os dados que estão a ser solicitados (Cloudflare, 2022).

Estes cabeçalhos, podem incluir a classe *Referrer*, que indica a origem ou o URL do recurso Web a partir do qual o pedido foi feito. Um URL pode ser entendido como um endereço de um recurso na Web.

Para demonstrar este funcionamento de uma forma mais clara, na Figura 3, o cabeçalho *Referrer* inclui o URL completo da página do "*site-one*" no contexto da qual a solicitação foi feita.

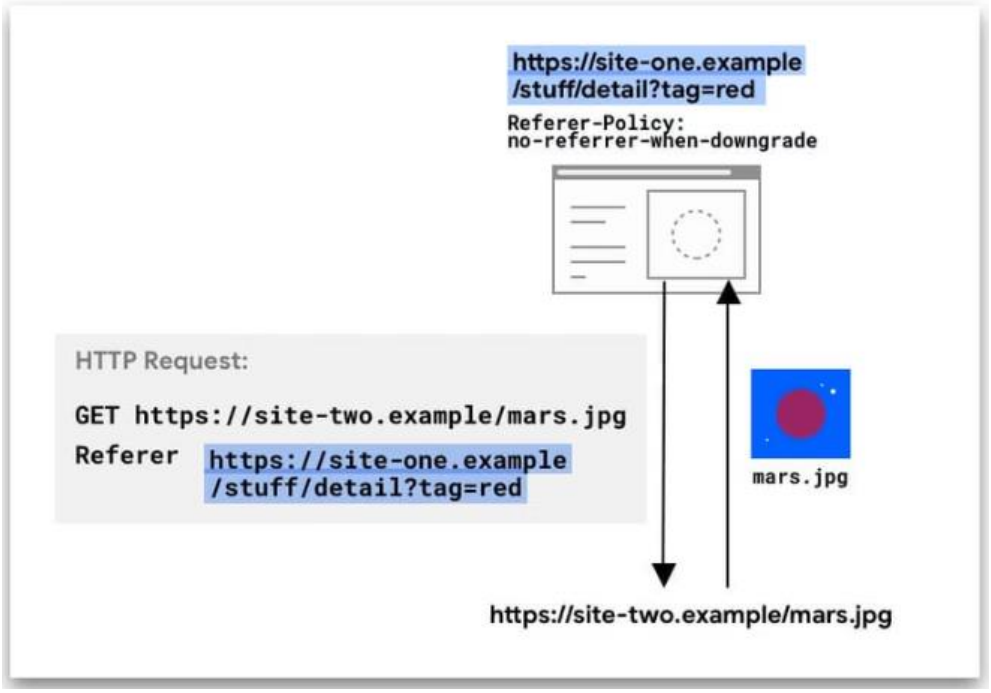


Figura 3 - Exemplo de um cabeçalho Referrer

Fonte: (Cloudflare, 2022)

No valor associado à classe *Referrer* podem ainda estar presentes diferentes tipos de solicitações:

- Solicitações de navegação, como por exemplo, quando um utilizador seleciona uma hiperligação;

- Solicitações de recursos, quando um navegador solicita imagens, *iframes*, *scripts* e outros recursos que uma página precisa para realizar determinados serviços.

Classe *Referrer-Policy* nos cabeçalhos *Referrer*

O cabeçalho *Referrer-Policy* controla a quantidade de informação de referência (enviadas pelo cabeçalho *Referrer*) que deve ser incluída nos pedidos. A *Referrer-Policy* é utilizada para manter a privacidade do sítio Web de origem enquanto procura recursos ou navegação.

As informações estabelecidas pela *Referrer Policy* podem ser entregues para um pedido através de vários métodos. Pode ser feito através da utilização do cabeçalho ou através de elementos no HTML onde este recebe a palavra-chave do *Referrer* como um valor que, por sua vez, permite a definição da *Referrer-Policy* (GeeksforGeeks, 2021). Este cabeçalho tem diversas configurações necessárias para a sua utilização. Dependendo da escolha da configuração, as informações disponíveis no *Referrer* podem ser:

- Não conter nenhuma informação, ou seja, nenhum cabeçalho estará presente;
- Conter apenas a origem;
- Conter o URL completo, mais especificamente, a origem, caminho e a *string* de consulta (*query string*).

Portanto, as configurações possíveis para este atributo são as seguintes:

no-*Referrer* - significa que o servidor deve instruir o navegador para nunca enviar o cabeçalho *Referrer* com pedidos feitos a partir do domínio. Também inclui hiperligações para páginas no mesmo domínio (HELME, 2017).

Exemplo: Se um ficheiro em “https://example.com/page.html” estabelece uma configuração de "**no-*Referrer***", então a navegação para https://example.com/ (ou qualquer outro URL) não enviaria nenhum cabeçalho *Referrer*.

no-*Referrer*-when-downgrade – Neste caso, o navegador não envia o cabeçalho *Referrer* durante uma navegação de HTTPS para HTTP, mas envia o URL completo durante navegação para qualquer destino (HELME, 2017).

Exemplo: Se um ficheiro em https://example.com/page.html estabelece uma configuração de "**no-*Referrer*-when-downgrade**", então as navegações para um URL do tipo https://not.example.com/ enviarão um cabeçalho *Referrer* com um valor de https://example.com/page.html, uma vez que a origem de nenhum dos recursos é um URL

não-potencialmente digno de confiança. As navegações a partir dessa mesma página para **http**://not.example.com/ não enviariam um cabeçalho *Referrer*.

Origin – O navegador enviará sempre o cabeçalho *Referrer* para a origem a partir da qual o pedido foi feito (HELME, 2017).

Exemplo: Se um ficheiro em https://example.com/page.html estabelece uma política de " **origin** ", então a navegação para qualquer origem enviaria um cabeçalho *Referrer* com um valor de https://example.com/, mesmo para URLs que não são potencialmente confiáveis.

origin-when-cross-origin - O navegador enviará o URL completo para pedidos com a mesma origem, mas apenas envia a origem quando os pedidos são de origem cruzada (*cross-origin)* (HELME, 2017).

Exemplo: Se um ficheiro em https://example.com/page.html estabelece uma política de " **origin-when-cross-origin**", as navegações para https://example.com/not-page.html enviariam um cabeçalho *Referrer* com um valor de "https://example.com/page.html".

As navegações a partir dessa mesma página para https://not.example.com/ enviariam um cabeçalho *Referrer* com um valor de https://example.com/, mesmo para URLs que não sejam potencialmente dignos de confiança.

Same origin – Significa que o navegador define um cabeçalho *Referrer* apenas em pedidos com a mesma origem. No caso do destino ser diferente, não será enviado qualquer cabeçalho *Referrer* (HELME, 2017).

Exemplo: Se um ficheiro em https://example.com/page.html estabelece uma política de "**same origin**", então as navegações para https://example.com/not-page.html enviariam um cabeçalho *Referrer* com um valor de https://example.com/page.html.

As navegações a partir dessa mesma página para https://not.example.com/ não enviariam um cabeçalho *Referrer.*

strict-origin - Envia apenas a origem como referência quando o nível do protocolo de segurança permanece o mesmo (HTTPS→HTTPS), mas não o envia para um destinatário menos seguro (HTTPS→HTTP) (W3C, 2017).

Exemplo: Se um documento em https://example.com/page.html estabelece uma política de "**strict-origin**", então as navegações para https://example.com/not-page.html enviariam um cabeçalho *Referrer* com um valor de https://example.com. Da mesma forma, navegações para https://not.example.com enviariam um cabeçalho *Referrer* com um valor de https://example.com/.

As navegações a partir dessa mesma página para **http**://not.example.com não enviam um cabeçalho *Referrer*.

strict-origin-when-cross-origin - Envia a origem, caminho e a *string* de consulta ao efetuar um pedido da mesma origem, mas só envia a origem quando o nível do protocolo de segurança permanece o mesmo (HTTPS→HTTPS) e não envia os cabeçalhos para destinatários menos seguros (HTTPS→HTTP) (W3C, 2017).

Exemplo: Se um ficheiro em https://example.com/page.html estabelece uma política de "**strict-origin-when-cross-origin**", então as navegações para https://example.com/not-page.html enviariam um cabeçalho *Referrer* com um valor de https://example.com/page.html.

As navegações a partir dessa mesma página para https://not.example.com/ enviariam um cabeçalho *Referrer* com um valor de https://example.com/. As navegações a partir dessa mesma página para **http**://not.example.com/ não enviariam nenhum cabeçalho *Referrer*.

unsafe-url - Envia a origem, o caminho e a cadeia de consultas quando efetua qualquer pedido, independentemente da segurança (W3C, 2017).

Exemplo: Se um ficheiro em https://example.com/sekrit.html estabelece uma política de "**unsafe-url**", então a navegação para http://not.example.com/ (e todas as outras origens) enviaria um cabeçalho *Referrer* com um valor de https://example.com/sekrit.html.

Classe cache-control nos cabeçalhos

Por predefinição, os recursos são sempre autorizados a serem armazenados em cache por qualquer tipo de cache. O Cache-control é um cabeçalho utilizado para especificar políticas de cache do navegador tanto em pedidos de clientes como em respostas de servidores. As políticas incluem a forma como um recurso é colocado em cache, onde é colocado em cache e a sua idade máxima antes de expirar.

Para tal, as diversas configurações para este cabeçalho são as seguintes (Imperva, 2022):

Public – Esta configuração indica que um recurso pode ser armazenado em cache por qualquer cache.

Private – Esta configuração indica que um recurso é específico do utilizador, ou seja, pode ser colocado em cache, mas apenas no dispositivo do cliente. Por exemplo, uma resposta de página Web marcada como privada pode ser colocada em cache por um navegador, mas não por uma rede de entrega de conteúdo, mais concretamente, a CDN.

No-cache – Indica que um navegador pode armazenar uma resposta, mas deve primeiro submeter um pedido de validação ao servidor de origem.

No-store - Ao estabelecer esta indicação, significa que os navegadores não estão autorizados a armazenar em cache uma resposta e devem retirá-la do servidor cada vez que é solicitada. Um exemplo desta configuração é para dados sensíveis, tais como dados bancários que dizem respeito aos utilizadores.

No-Transform - Os *proxys* intermédios por vezes alteram o formato das imagens e ficheiros de modo a melhorar o desempenho. A diretiva de não transformação informa os *proxys* intermediários para não alterarem o formato ou os recursos.

Percebendo todas estas características relacionadas com um sítio Web, de que forma é que se pode estabelecer uma ligação com todas estas funcionalidades para se entender o funcionamento de um sítio Web?

Pode ser visto da seguinte forma, quando um utilizador visita um sítio Web:

1. O navegador vai ao servidor DNS e encontra o endereço real do servidor onde o sítio Web se encontra.
2. O navegador envia uma mensagem de pedido para o servidor, pedindo-lhe que este envie uma cópia do sítio Web ao cliente. Esta mensagem e todos os outros dados enviados entre o cliente e o servidor, são enviadas através de uma ligação à Internet utilizando TCP/IP.
3. Se o servidor aprovar o pedido do cliente, o servidor envia ao cliente uma mensagem "200 OK", o que significa "Pode visualizar este sítio Web "e posteriormente começa a enviar os ficheiros do sítio Web para o navegador como uma série de pequenos pacotes de dados.

4. O navegador reúne estes pequenos pacotes numa página Web completa e mostra-as ao utilizador.

2.2 O que é rastreamento na Web?

Este termo chamado de rastreamento na Web pode ser entendido como a recolha e partilha de informação sobre a atividade de um utilizador na Internet - o que ele faz online e como o faz. O rastreamento na Web, permite que várias empresas tenham uma compreensão mais completa das preferências dos utilizadores.

O rastreamento de utilizadores para fins analíticos é bastante comum, onde a ferramenta analítica mais popular é o *Google Analytics*. Estes sítios Web utilizam geralmente um *software* analítico para obter informações sobre os seus utilizadores. Isto pode incluir a demografia dos visitantes e a forma como estes utilizam o sítio Web. Por exemplo, como chegam ao sítio Web e quantas páginas visitam. Um inquérito realizado em 2017 concluiu que 79% dos sítios Web utilizam técnicas de rastreamento para recolherem dados dos utilizadores (Crawford, 2020) .

Muitas funções de um sítio na Web não funcionarão sem alguma forma de rastreamento. Por exemplo, os sítios Web rastreiam os utilizadores para os manter ligados enquanto navegam em páginas diferentes. Quando o rastreamento nos sítios Web é bem realizado, os seguimentos aos utilizadores podem ser benéficos tanto para os utilizadores individuais como para as empresas. Por exemplo, este tipo de rastreamento pode ajudar o utilizador a lembrar os *logins* e preferências, para que não precise de começar do zero cada vez que visita um sítio Web. Da mesma forma, pode ajudar as empresas a melhorar a forma como o respetivo sítio Web é apresentado e utilizado pelos utilizadores, através de métricas melhoradas de CTR (taxa de *click-through*), taxa de conversão, aumento das inscrições nos respetivos sítios Web e valor de vendas.

Portanto, o rastreamento de sítios Web pode ser utilizado de uma forma útil mas também pode ser utilizado para outros fins. Destacam-se dois exemplos:

- Quando um utilizador procura um restaurante no *Google* e o serviço lhe fornece uma lista de restaurantes na sua área local, é porque o motor de pesquisa sabe onde o utilizador está localizado.

- Quando uma loja de comércio eletrónico mostra uma lista de produtos recomendados ao utilizador, esta loja sabe que o utilizador gosta de certas características pois rastreou itens que o utilizador consultou ou comprou anteriormente.

Sem a tecnologia de rastreio de sítios Web, os dois exemplos acima referidos, ou não existiriam ou existiriam de uma forma menos abusiva. Portanto, a prevalência do rastreio de sítios Web, a falta de transparência sobre a recolha de dados, a forma como utilizam esses dados e quem tem acesso aos mesmos, demonstra que existem diversos problemas com esta prática.

É, em parte, devido a estas questões que países e regiões de todo o mundo introduzem leis para regulamentar a forma como os sítios Web podem recolher dados para rastrear os utilizadores.

2.3 Que dados é que podem recolher os sítios Web?

Os sítios Web recolhem uma diversa série de dados para diferentes utilizações. Isto inclui dados fornecidos através de formulários, como por exemplo, o endereço eletrónico e informações sobre cartões de crédito, bem como outros tipos de informação. Alguns exemplos de dados serão demonstrados nos seguintes pontos (*Cookie*Pro, 2020):

- Endereços IP para determinar a localização de um utilizador.
- Informação sobre a forma como o utilizador interage com os sítios Web. Por exemplo, quais são os elementos em que clicam e quanto tempo permanecem numa página.
- Informação sobre navegadores e o dispositivo com que o utilizador acede ao sítio Web.
- Saber a atividade de navegação em diferentes sítios Web. Desta forma tem-se a possibilidade de saber os interesses individuais do utilizador, hábitos de compra, problemas que enfrentam, entre outros.

Nem todos os sítios Web recolhem todos os dados acima referidos. Alguns até nem recolhem qualquer tipo de dados, ou seja, tudo dependerá do serviço que o sítio Web estiver a prestar e a forma como o sítio Web é implementado.

É importante também lembrar que os sítios Web não são a única forma de as empresas recolherem dados sobre os utilizadores. As empresas também recolhem dados de aplicações para *smartphones*, altifalantes inteligentes, e-mails, entre outros.

2.4 Como os sítios Web rastreiam a atividade dos utilizadores

Quando um utilizador visita um sítio Web, os dados são recolhidos a partir do seu dispositivo e navegador, que podem ser utilizados para adaptar a sua experiência ou recolher informações sobre como e por onde está a navegar.

Os dois principais tipos de rastreamento no sítio Web são designados por entidades primeiras (first-party) e entidades terceiras (third-party). O rastreamento por entidades primeiras é efetuado pelo sítio Web que o utilizador está a visitar. Este tipo de rastreamento monitoriza o comportamento do utilizador para o relembrar de certas preferências, tais como, o tipo de conteúdo que normalmente escolhe, as suas definições de idioma, localização, entre outras informações. Normalmente, não existe preocupação com este tipo de rastreamento, pois este é utilizado para melhorar a experiência do utilizador. Porém, a situação torna-se mais critica se os sítios Web que realizam este tipo de rastreamento venderem os seus dados, por exemplo, a entidades publicitárias.

O rastreamento por entidades terceiras acontece quando outras partes, para além do sítio Web que o utilizador está a usufruir, também rastreiam a sua atividade. Por exemplo, pode visitar um sítio Web de notícias, sem percecionar que esse sítio Web também suporta *cookies* de entidades terceiras que rastreiam o seu comportamento. O rastreamento do comportamento de entidades terceiras é frequentemente utilizado para ajudar os anunciantes a adaptarem os seus anúncios às suas preferências.

Existem diversas razões pelas quais um sítio Web pode incluir serviços de entidades terceiras. Os serviços provenientes de entidades terceiras podem ter diversos objetivos, como por exemplo, utilizar o *Google Analytics*, que é considerado uma entidade terceira, para obter informações sobre o tráfego do seu próprio sítio Web. No entanto, a *Google* pode utilizar esses dados para fins de marketing. Portanto, as entidades terceiras têm diversas finalidades, tais como (*Cookie*Yes, 2022):

- Serviços de publicidade que são utilizados para identificar os consumidores, rastrear o seu comportamento de navegação, prever os seus interesses quando

realizam uma compra e mostrar aos utilizadores anúncios que representam os seus interesses.

- Sistemas de medição de audiências que permitem aos proprietários do sítio Web aprenderem sobre o que os utilizadores visitam e os tipos de serviços que utilizam.
- Serviços de comunicação social que têm dois objetivos principais: a incorporação do conteúdo gerado pelo utilizador numa dada página e permitir a partilha destes conteúdos na rede social escolhida.

Portanto, a incorporação de entidades terceiras pode ser utilizado de uma forma positiva, mas também pode ser utilizado para fazer um rastreamento sub-reptício a um utilizador. Isto deve-se ao facto das entidades terceiras não estarem a ser controladas pelo sítio Web de entidade primeira. Desta forma, um dos indicadores de que uma entidade terceira poderá estar a fazer um rastreio sub-reptício a um utilizador é quando estas implementam os seus próprios *cookies*. Consequentemente, através de uma combinação de *cookies* ou de outras tecnologias de rastreamento, as entidades terceiras poderão extrair os históricos de navegação e desta forma conseguirem identificar os utilizadores.

2.4.1 Mecanismos utilizados para rastrear a atividade dos utilizadores

Existem diferentes tipos de mecanismos utilizados para realizar um rastreio aos utilizadores. De seguida serão apresentados os diferentes tipos de mecanismos e será feita uma breve introdução sobre alguns conceitos que são importantes para posteriormente se entender o seu funcionamento.

Mecanismos de Rastreamento da Sessão

O Rastreamento da sessão é utilizado para registar as ações dos utilizadores num sítio Web. Eventos tais como o preenchimento de um formulário, são tratados sequencialmente e resumidos numa sessão para poderem ser registados como clientes individuais (Wiki, 2021).

Existem diversas formas de rastrear as sessões do utilizador e podem ser descritas da seguinte forma (Haq, 2022):

- ***Cookies*** - É uma identificação atribuída aos utilizadores quando estes visitam um sítio Web pela primeira vez. Sempre que, o navegador a partir do qual o sítio Web

foi acedido, se ligar ao servidor receberá o *cookie* que foi dado anteriormente pelo servidor. Desta forma o servidor reconhecerá que é o mesmo utilizador.

- **Campos de formulário ocultos** - São utilizados para inserir informação, como por exemplo, um identificador específico daquele utilizador, quando é enviado um pedido ao servidor. Estes campos não são diretamente visíveis para o utilizador, mas cada vez que um pedido é enviado ao servidor, a identificação específica é também enviada como parte do pedido permitindo que o servidor reconheça o utilizador.
- **Reescrita do URL -** Trata-se da adição de um ID de sessão único no próprio URL quando um pedido é enviado para o servidor.

Mecanismos de rastreamento com base no armazenamento

Estes mecanismos de rastreamento dependem do armazenamento explícito de dados nos computadores dos utilizadores. Estes métodos são geralmente os mais utilizados e são muito mais avançados comparativamente com os métodos baseados em sessões e onde as suas capacidades são também mais elevadas. Tal deve-se ao facto de terem a capacidade de reconhecer a instância particular de um navegador ou de um sistema operativo, dependendo da técnica.

Conceito de cache na Web

Antes de ser abordado com mais detalhe o que são mecanismos de rastreamento baseados em cache (onde será explicado com maior detalhe no capítulo 3), será necessário para o leitor compreender o funcionamento de uma cache na Web.

Os utilizadores quando utilizam um serviço de um sítio Web têm a tendência de descarregar o mesmo conteúdo várias vezes. Sem uma cache na Web adequada, cada vez que um utilizador fizer um pedido, a resposta deve vir do servidor de origem. Quando vários utilizadores solicitam um conteúdo ao mesmo tempo, os tempos de resposta podem aumentar e provocar uma sobrecarga no servidor. Deste forma, com uma cache na Web, o conteúdo

fica mais próximo dos utilizadores finais, melhorando assim os tempos de resposta (Gibb, 2016).

Conceito de *Fingerprinting*

Da mesma forma que o leitor necessita de perceber o funcionamento de uma cache na Web, neste mecanismo será necessário entender o que é *fingerprinting* ou impressão digital e como pode ser fundamental no rastreamento de um utilizador (que será demonstrado com mais detalhe no capítulo 3).

Uma impressão digital nada mais é do que informação recolhida. É um conjunto de atributos, recolhidos, por exemplo, do sistema do utilizador cuja combinação de valores é muito provavelmente única para cada dispositivo, formando um identificador do dispositivo. Estes atributos do sistema podem incluir, por exemplo, o fuso horário, a resolução de ecrã, versão do software instalado, as fontes instaladas, os *plugins* instalados, e os *cookies* ativados. Quanto mais diversos são os valores dos atributos, mais são considerados identificadores, uma vez que os seus valores não são partilhados por muitos dispositivos, o que acontece, por exemplo, com a versão do sistema operativo. As impressões digitais podem ser utilizadas como identificadores a fim de rastrear o dispositivo dos utilizadores nos sítios Web, garantindo a associação entre as sessões de navegação e até criando uma ligação à identidade do utilizador. Sempre que um utilizador acede a uma determinada página Web, que inclui a utilização de *software* de recolha de impressões digitais, a impressão digital do dispositivo é recolhida e comparada com uma base de dados de dispositivos conhecidos. Se o dispositivo for desconhecido, é adicionado à base de dados. Desta forma, cada vez que um utilizador visitar uma página Web monitorizada, a base de dados irá ser aumentada com as informações dos dispositivos. Das primeiras ocasiões onde se ouviu falar deste mecanismo foi no *The Washington Post*, quando *Edward Snowden* disse nas informações que deu aos jornalistas, que a NSA estava a utilizar a resolução do ecrã dos computadores para identificar terroristas. Atualmente, as empresas de publicidade estão a utilizar estas mesmas técnicas (Rowe, 2019).

3. Estado da arte sobre técnicas de rastreamento

As técnicas de rastreio necessitam de armazenar, entre diversos pedidos HTTP realizados à entidade que gere esse estado, algum tipo de dados sobre o dispositivo do utilizador ou sobre o mesmo. Tal pode ser alcançado através de vários mecanismos, sendo o método mais comum os *cookies*. Nesta secção irão ser analisadas algumas técnicas de rastreio (ver Tabela 1) que podem pôr em causa a privacidade do utilizador.

Tabela 1 - Mecanismos e técnicas de Rastreio

Fonte: Elaboração própria

Mecanismo de Rastreamento	Nome	Breve Descrição	Possiveis Tecnologias
Mecanismos de rastreamento com base no armazenamento	Cookies HTTP	Ficheiro de texto contendo informação com dados dos utilizadores	HTTP headres, JavaScript
	Cookies de entidades terceiras	Um cookie de entidades terceiras é um cookie que pertence a um domínio que não o exibido em primeira mão no navegador	HTTP headres, JavaScript
	Cookie Synchronizatioon	Metodo para contornar a política same-origin	Flash
	SuperCookies	Recolha de informação sobre os utilizadores de várias formas	Flash/Java
	Flash Cookies	São mais persistentes e podem conter ate 100KB de informação comparativamente ás HTTP Cookies	Flash/Java
	Evercookies	É uma API JavaScript para produzir cookies extremamente persistentes	Flash/Java
	LocalStorage	Introduzido pelo HTML5, oferece possibilidades de guardar dados pelos sites	HTML5
Mecanismos de rastreamento baseados na cache dos navegadores	Loading performing test	Tempo de carregamento de um objeto para identificação do utilizador	Server-side measurements, JavaScript
Utilização da colaboração incosciente do utilizador	Captchas	O Captchas manipula letras e números e confia na capacidade humana para determinar quais os símbolos que estão a ser representados	HTML5, JavaScript, CSS
	Utilização de Event Tagging para rastreamento através do campo de formulário	Event Tags regista os cliques e impressões para, por exemplo, anunciantes	HTML5, JavaScript
	Clickjacking	Método para apresentar elementos sensiveis fora do contexto no site	HTML5, JavaScript, CSS
Mecanismos de rastreamento baseado em Fingerprinting	Device Fingerprinting	Recolher informações do sistema	IP Address, TCP headers, HTTP headers, JavaScript, Flash
	Operating System Instance Fingerprinting	Recolher informações do sistema operativo do utlizador	JavaScript, Flash, Java, ActiveX
	Browser Related Fingerprinting	Recolher informações sobre o navegador do utilizador	HTML5, JavaScript, CSS

3.1. Mecanismos de rastreamento com base no armazenamento

Realizada uma breve introdução sobre este mecanismo no capítulo 2, irei abordar a técnica atualmente mais conhecida que pertence a este mecanismo designada por *cookies* HTTP onde será dada uma descrição sobre o seu funcionamento, características e a forma de como pode ser usada para fazer um rastreio sub-reptício a um utilizador. Posteriormente, serão abordadas outras técnicas relacionadas com este mecanismo.

3.1.1. Cookies HTTP

Um *cookie* HTTP é um pequeno fragmento de dados do utilizador (cada um está limitado a 4 KB de informação), que o servidor Web envia para o navegador do utilizador. O navegador do utilizador, pode armazenar estes dados e enviá-los de volta para o mesmo servidor, numa próxima requisição (que por exemplo, poderá ser uma criação ou alteração de conta).

Esta técnica oferece diversas possibilidades e utilidades aos sítios na Web, mas de acordo a com *Mozilla* (Network, 2022a), possui três principais objetivos:

- Gestão das sessões dos utilizadores –O servidor ao ter armazenada esse *cookie* tem a capacidade de, numa próxima sessão, relembrar ao utilizador o seu *login*, carrinho de compras, pontuações de jogos, ou qualquer outra informação.
- Personalização – Relembrar as preferências do utilizador, temas e outras definições fornecidas pelo mesmo;
- Rastreamento – Seguir e analisar o comportamento do utilizador.

Os *Cookies* HTTP, que possuem um identificador próprio, são usados para o armazenamento no lado do navegador do utilizador. A implementação do *cookie* no navegador do utilizador pode ser feita da seguinte forma, o servidor, ao receber uma requisição HTTP proveniente do navegador, envia um cabeçalho *Set-Cookie* como resposta a esse requisito. O *Set-Cookie*, que envia *cookies* do servidor para o navegador, pode ser configurado conforme se observa na Figura 4.

```
Set-Cookie: id=a3fWa; Expires=Wed, 21 Oct 2015 07:28:00 GMT; Secure; HttpOnly
```

Figura 4 - Definição de set-cookie

Fonte: (MDN, 2021)

Antigamente esta era a única forma possível de armazenamento, porém, atualmente surgiram novas técnicas como a utilização de APIs de armazenamento. Uma desvantagem dos *cookies* HTTP comparativamente às APIs é que os *cookies* HTTP têm de ser enviados em todas as requisições ao servidor Web, o que leva a uma diminuição na performance de um sítio Web. Exemplos de APIs modernas são o Web storage API (técnica que será descrita detalhadamente no ponto número sete) e IndexedDB.

Existem dois tipos de *cookies* HTTP, *cookie* de sessão e *cookie* persistente. Os *cookies* de sessão expiram quando o utilizador encerra o navegador, ao contrário dos *cookies* persistentes, que ficam guardadas no dispositivo do utilizador e possuem uma data de expiração, que é emitida pelo servidor Web.

Compreendendo o funcionamento de um *Cookie* HTTP, observa-se que esta técnica tem uma funcionalidade legitima e oferece diversas vantagens aos utilizadores. No entanto, estas funcionalidades podem ser manipuladas, por um rastreador, cujo objetivo é pôr em causa a privacidade do utilizador. Surge então algumas diferentes possibilidades para se utilizar esta técnica de uma forma sub-reptícia.

Experiências realizadas no artigo elaborado pelos autores Li, T. C., Hang, H., Faloutsos, M., & Efstathopoulos, P. (2015) diferenciam as *cookies* de rastreamento e não rastreamento com base no seu tempo de expiração e tamanho do campo de valor *(<cookie-value>).* Constataram que mais de 90 % dos *cookies* de rastreamento têm uma vida útil superior a 1 dia, ao contrário dos *cookies* de não rastreamento. Como o valor contido por um *cookie* deve ser suficientemente longo para se conseguir distinguir um utilizador, 80 % dos *cookies* de rastreamento têm os seus valores superiores a 35 caracteres. Para evitar esta deteção, os sítios na Web tentam dividir o identificador do utilizador em vários *cookies*.

Antes mencionar outra possibilidade para utilizar esta técnica de uma forma sub-reptícia, é necessário abordar-se uma funcionalidade, pertencente aos *cookies* HTTP, designada de *domain-specific* (Network, 2022a). Esta funcionalidade pode ser entendida como, por exemplo, um *cookie* definido por “dissertação.com” não pode ser envidado para outro domínio, como por exemplo “tesedoutoramento.com”. Os *cookies* têm um atributo de

domínio que pode ser especificado e caso não seja, o *cookie* só pode ser enviado para esse domínio em específico, excluindo subdomínios. Porém, se o atributo de domínio for definido como "dissertação.com", o *cookie* pode ser enviado para subdomínios do tipo "dissertação2.dissertação.com". Teoricamente, esta funcionalidade ajuda a combater a violação da privacidade de um utilizador pois a política, designada por *same-origin*, impede que os *cookies* sejam carregados através de um contexto *third-party* para um contexto *first-party*. Tal significa que o *third-party tracker* não tem a possibilidade de ler *cookies* definidos por outro domínio de *first-party*. No entanto, os rastreadores encontraram uma forma de contornar esta política, onde por exemplo conseguem reunir informações através dos URLs. A técnica que é utilizado para contornar esta política, designa-se por sincronização de *cookies (Cookie Synchronization)* (Papadopoulos et al., 2019). Em suma, esta técnica poderá também pôr em causa a privacidade do utilizador e será explicada com mais detalhe (mais especificamente, no ponto três da secção número I) no decorrer desta dissertação.

Alguns navegadores tentam aumentar o nível de privacidade dos utilizadores, bloqueando a definição e leitura de *cookies* por entidades terceiras. Contudo, esta funcionalidade do navegador pode ser facilmente contornada através do redirecionamento do utilizador, realizado por um Java*Script*, para uma página Web de entidades terceiras. Esse redirecionamento, tem por objetivo definir ou ler os *cookies*. A Figura 5 apresenta um exemplo para ilustrar como um serviço de uma entidade terceira pode, de certa forma, rastrear e "roubar" dados sobre o comportamento e histórico de um utilizador em vários sítios na Web (Sikkeland, 2020). Devido à política *same-origin* (Network, 2022b), que impede que os *cookies* sejam carregados através de entidades terceiras para entidades primeiras. Se um utilizador visitar um sítio chamado "theonion.com" que concede um *cookie*, uma entidade terceira "tracker.com" não poderá, por omissão, visualizar este *cookie*. A forma para evitar esta funcionalidade passa por a entidade terceira (neste caso, "tracker.com") definir o seu próprio *cookie*. Para tal, quando um utilizador visitar o "theonion.com", que tem em sua posse um anúncio do domínio "tracker.com", este último, pode definir um *cookie* único. Desta mesma forma, quando o utilizador visitar o "cnn.com", que também incorpora um anúncio do "tracker.com", o mesmo *cookie* é enviado de volta para o "tracker.com". Assim, o rastreador sabe quais os sítios que o utilizador visitou (Bujlow et al., 2017). O rastreador pode guardar um perfil de navegação para este utilizador e também armazenar o seu histórico de navegação. São capazes também, de registar quantas

vezes o utilizador visita cada sítio na Web e quanto tempo permaneceram nesses mesmos sítios. Se um rastreador tiver ligação a um número considerável de sítios na Web, pode potencialmente, saber tudo sobre a navegação de um utilizador.

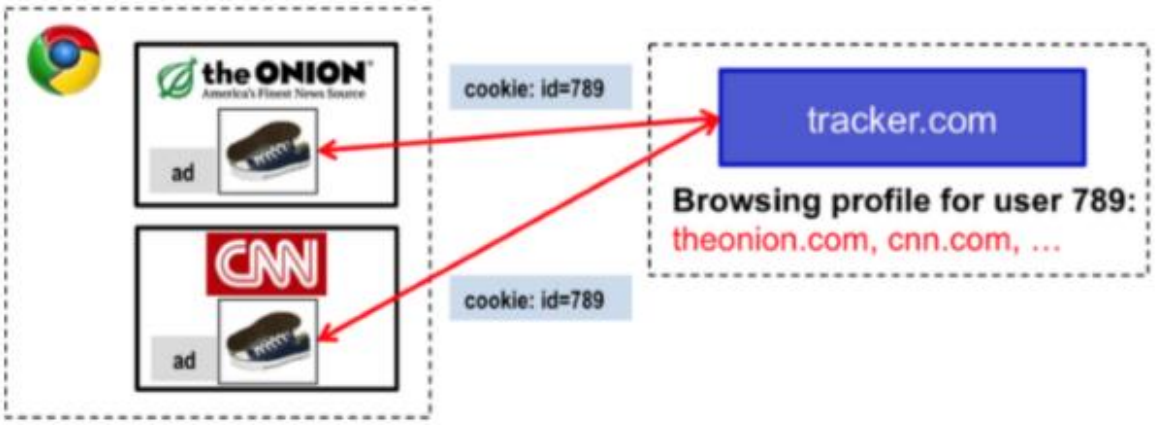


Figura 5 - Third-party rastreio baseado em cookies

Fonte: (Sikkeland, 2020)

3.1.2. Cookies de entidades terceiras

Um *cookie* de entidades terceiras é um *cookie* que pertence a um domínio que não o exibido em primeira mão no navegador. Este tipo de *cookies* de entidades terceiras são frequentemente utilizados para fins de rastreamento. Por exemplo, quando um utilizador visita um sítio Web, o seu navegador envia pedidos sobre tudo o que se encontra nesse sítio Web, incluindo o código incorporado para realização da partilha de redes sociais na página, como por exemplo, o *Facebook*, *Google Analytics* e recursos CDN. Os *cookies* definidos pelos recursos de entidades terceiras no navegador são enviados juntamente com estes pedidos. O nome do sítio Web que envia o pedido encontra-se no cabeçalho *Referrer* do pedido. Desta forma, entidades terceiras utilizam os *cookies* com os dados do cabeçalho *Referrer* e descobrem os hábitos de navegação dos utilizadores. Por norma, pode-se impedir este rastreamento desativando os *cookies* de entidades terceiras no *Firefox* e no *Chrome*. No entanto, isto pode ter um impacto negativo na experiência de navegação do utilizador.

Portanto, para se contornar esta opção existe um atributo nas *cookies* designado por samesite, que permite aos administradores restringir quais os pedidos que adicionam *cookies*.

Este atributo permite remover certos *cookies* de pedidos que não tenham sido emitidos pelo próprio sítio Web, em vez de desativá-los a todos. Pode observar-se um exemplo da definição deste atributo na Figura 6.

```
Set-Cookie: CookieName=CookieValue; SameSite=Lax;

Set-Cookie: CookieName=CookieValue; SameSite=Strict;
```

Figura 6 - Definição do atributo samesite

Fonte: (Albeniz, 2022)

Alguns dos valores possíveis para esta configuração são:

- strict – Se utilizar o valor *Strict*, os *cookies* não serão enviados quando uma entidade terceira emite os pedidos. Porém, esta opção pode ter um impacto negativo na experiência de navegação do utilizador. Por exemplo, se um sítio Web que o utilizador está a visitar contiver esta configuração, poderá ser solicitado ao utilizador que volte a iniciar a sessão de *login*, uma vez que o *cookie* não será enviado juntamente com o pedido.

- lax – proporciona um equilíbrio razoável entre a segurança e usabilidade para os sítios Web que tencionem, por exemplo, manter a sessão de *login* do utilizador após a chegada do utilizador a partir de uma ligação externa. Com esta configuração, o envio de *cookies* por parte de entidades terceiras só será realizado se estas entidades terceiras emitirem um pedido GET que afete o nível superior de navegação do utilizador, ou seja, que afete o URL na barra de navegação. Por exemplo, quando é realizado o carregamento de uma imagem, esta operação não altera o nível superior de navegação, o que significa que o *cookie* não vai ser envidado juntamente com o pedido.
- none – Esta configuração não dará qualquer tipo de proteção. O navegador anexa os *cookies* em todos os contextos de navegação em todos os sítios Web.

Portanto, o valor lax é uma boa escolha para *cookies* que afetem a exibição do sítio Web. O valor strict é útil para *cookies* relacionados com as ações que o utilizador está a realizar

(Albeniz, 2022). De seguida mencionam-se algumas técnicas de rastreio mais comuns que podem ocorrer com entidades terceiras.

Publicidade de entidades terceiras

A publicidade direcionada é talvez a maior utilização por parte destas entidades para fazer rastreio a um utilizador. Existem várias empresas de publicidade, mas a rede *Doubleclick* da *Google* é a mais conhecida. Os sítios Web podem alojar anúncios fornecidos, por exemplo, pelo doubleclick.net e incorporá-los no seu sítio Web como uma imagem ou *iframe*. Quando um utilizador visita um sítio Web, que incorpora anúncios do doubleclick.net, poderá ser criado um *cookie* para esse utilizador. Isto permite ao doubleclick.net rastrear os utilizadores entre sítios Web, uma vez que o mesmo *cookie* é automaticamente incluído em todos os pedidos feitos ao doubleclick.net. Desta forma quando um utilizador visitar outro sitio Web, desde que este inclua anúncios do doubleclick.net, o utilizador esta na mesma sujeito a um rastreamento sub-reptício (Roesner et al., 2012).

Na Figura 7 é ilustrado num sítio Web designado por "site1.com" que estabelece um anúncio da entidade terceira doubleclick.net onde é incorporado através de um *iframe*. Quando um utilizador visita o "site1.com", é feito um pedido do doubleclick.net para carregar o anúncio, onde este tem a possibilidade de definir um *cookie* único para esse utilizador. O *cookie* é propriedade da entidade doubleclick.net.

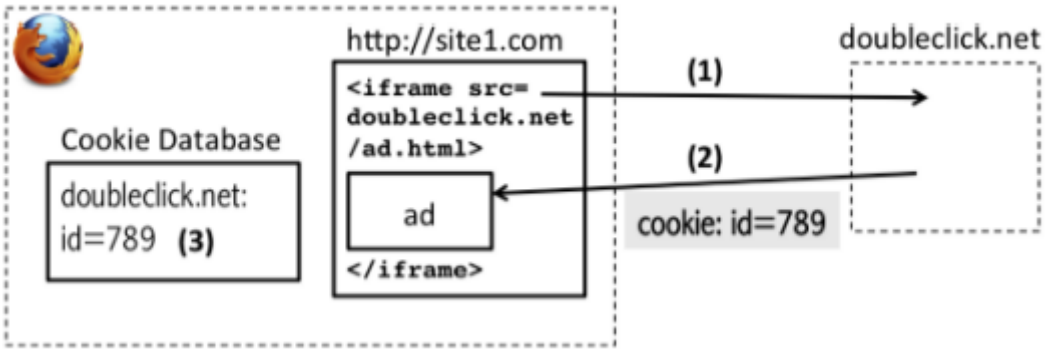


Figura 7 - Rastreio de entidades terceiras do doubleclick.net estabelecido no site "site1.com"

Fonte: (Roesner et al., 2012)

Publicidade de entidades terceiras através de *Popups*

Este tipo de rastreamento destina-se a utilizadores que bloqueiam todos os *cookies* de entidades terceiras nas definições do seu navegador ou que estão a utilizar ferramentas de privacidade que limitam a utilização de *cookies* de entidades terceiras. Devido à possibilidade de a maioria dos anunciantes definirem *cookies*, este método contorna estes bloqueios de forma que os *cookies* de entidades terceiras passem a ser de entidade primeira. Como as janelas de *popups* têm benefícios para a publicidade (por exemplo, melhorar a captura da atenção do utilizador), colocam a técnica de rastreamento numa posição de entidade primeira sem o consentimento do utilizador. A partir daí, a técnica lê os *cookie*s de entidade primeira, sem sofrer alterações pelo bloqueio de *cookie*s de entidades terceiras (Roesner et al., 2012).

Publicidade de entidades terceiras através de *Widgets* sociais

Os *widgets* sociais referem-se geralmente aos botões das redes sociais, como por exemplo o botão *like* do *Facebook* e o botão do *Twitter Share*. Estes *widgets* são geralmente incorporados em vários sítios Web como elementos Web de entidades terceiras e funcionam através da utilização de *cookies*. Por exemplo, quando um utilizador visita um sítio Web que incorpora um destes *widgets*, é feito um pedido a esse sítio Web, que estabelece um *cookie* para o utilizador. Desta forma, este método pode ser utilizado como uma técnica de rastreamento de entidades terceiras sem que seja necessário o utilizador interagir com o *widgets*. Se o utilizador receber um *cookie* do *Facebook* através deste método e mais tarde visitar o sítio Web "*Facebook*.com" diretamente, um *cookie* de entidade primeira será definido. Desta forma, o *cookie* da entidade primeira pode ser incorporado ao *cookie* de entidade terceira, permitindo ao *Facebook* rastrear o utilizador através dos sítios Web.

Um exemplo deste processo pode ser ilustrado na Figura 8. O sítio Web "site1.com" incorpora um *widgets* do tipo *Facebook* na forma de uma *script*, através da utilização de um elemento *iframe*. Quando o utilizador visitar o "site1.com", é feito um pedido ao *Facebook*, que responde com um *cookie* de entidades terceiras (Roesner et al., 2012).

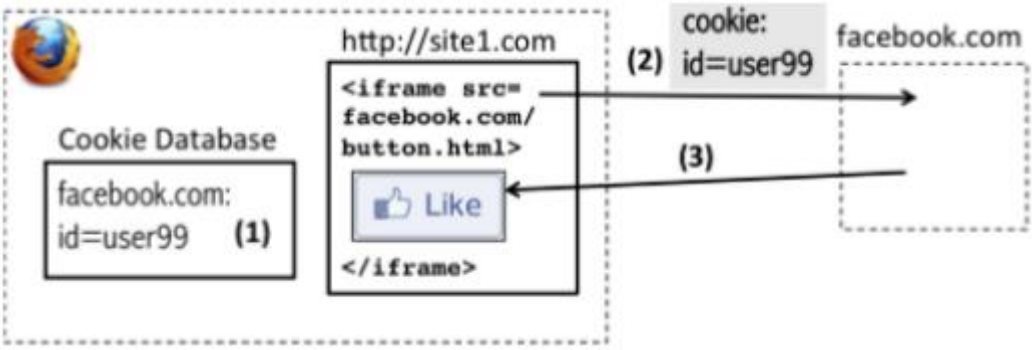


Figura 8 - O Widget social incorporado no site1.com que é utilizado para a entidade terceira Facebook.com estabelecer um cookie

Fonte: (Roesner et al., 2012)

3.1.3. Cookie Synchronization

A sincronização de *cookies* não é um mecanismo de rastreamento, ou seja, pode ser vista como um método para contornar a política *same-origin* e ajudar os rastreadores a partilhar a informação. A forma como os rastreadores utilizam esta funcionalidade pode ser vista da seguinte forma:

- Diferentes rastreadores têm a possibilidade de partilhar os seus identificadores únicos e sincronizá-los, de forma a ter os mesmos identificadores do utilizador em questão, o que permite fazer rastreio dos utilizadores que utilizam um determinado serviço num sítio Web. Para tal irei demostrar um exemplo, mencionado num artigo que foi elaborado pelos autores Papadopoulos, P., Kourtellis, N., & Markatos, E. P. (2019). O utilizador visita o "sítio1.com" que integra um *script* de rastreio do "tracker1.com". Para carregar este *script*, é feito um pedido pelo navegador para o "tracker1.com"; a respetiva resposta contem um *cookie* com um determinado valor, como podemos observar na Figura 9.

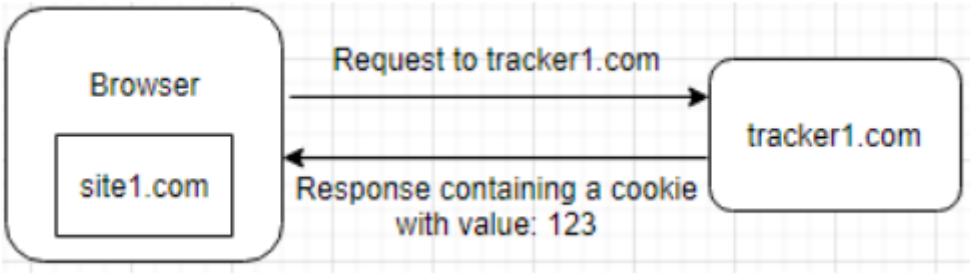


Figura 9 - Exemplo de como funciona uma sincronização de cookies

Fonte: (Papadopoulos et al., 2019)

- O utilizador visita outro sítio *Web*, neste caso "sítio2.com", que incorpora uma *script* de rastreio de outro rastreador, o "tracker2.com". O utilizador recebe um *cookie* com outo valor, que neste caso tem um valor 456 (ver Figura 10).

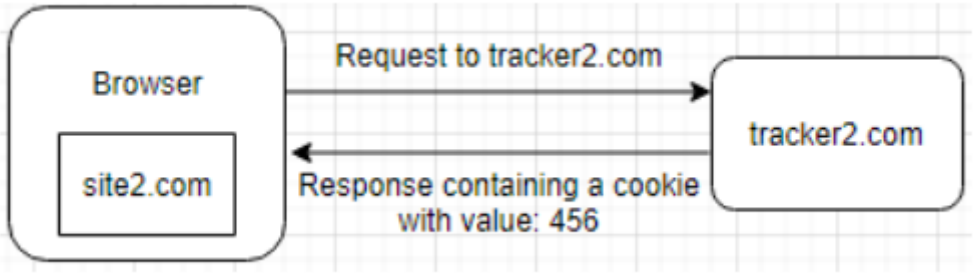


Figura 10 - Exemplo de como uma cookie synchronization funciona

Fonte: (Papadopoulos et al., 2019)

- Por último, é onde acontece a sincronização entre o "tracker1.com" e o "tracker2.com". O utilizador visita outro sítio, "sítio3.com", que incorpora um *script* de rastreamento do "tracker1.com", mas não incorpora nenhum *script* do "tracker2.com". Por consequência, o "tracker2.com" não sabe se o utilizador visitou o "sítio3.com" ou não. No entanto, quando o navegador do utilizador faz um pedido ao "tracker1.com", este responde com um pedido de redirecionamento para o "tracker2.com", o que obriga o navegador a fazer um pedido a este último, a também a facultar o seu *cookie*. Assim, este pedido pode ser construído com um URL personalizado contendo vários parâmetros, que podem incluir tanto o *cookie* único do utilizador para o "tracker1.com", como informações sobre o "sítio3.com".
  O "tracker2.com" sabe agora que o utilizador que é conhecido pela *cookie* com o valor "456", visitou o "sítio3.com". Por fim estes dois *trackers* podem fazer uma junção de informação que foram reunindo ao longo do processo, onde se pode observar com mais simplicidade na Figura 11.

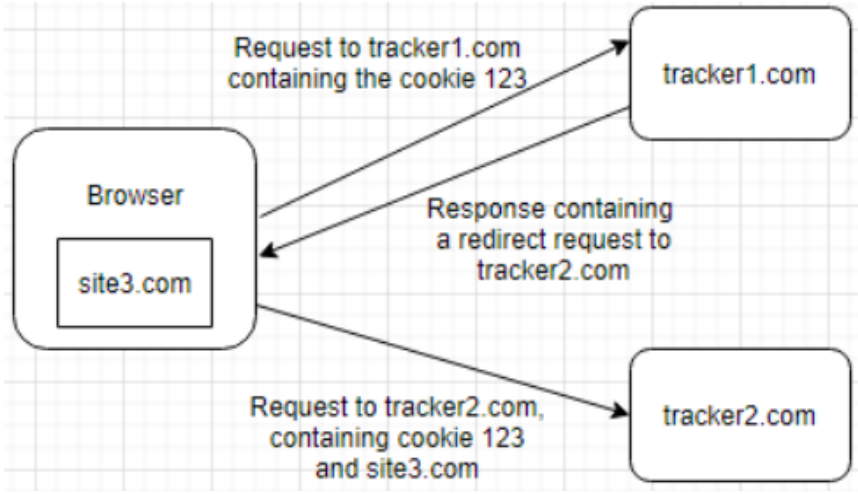


Figura 11 - Exemplo de como uma cookie synchronization funciona

Fonte: (Papadopoulos et al., 2019)

3.1.4. Supercookies

Um *supercookie* é um tipo de *cookie* de rastreamento inserido num cabeçalho HTTP por um fornecedor de serviços de Internet (empresa que fornece serviços de Internet) com o objetivo de recolher dados sobre o histórico e hábitos de navegação na Internet de um utilizador. Um *supercookie* não é tecnicamente um *cookie* HTTP, mas sim uma informação inserida em pacotes de Internet enviados a partir de um dispositivo do utilizador e do serviço a que está ligado (Anon, 2018).

Este tipo de mecanismo, levanta várias preocupações relacionadas com a violação da privacidade, pois ao limpar os dados do navegador não irá anular este mecanismo porque este *supercookie* é guardado no dispositivo do utilizador ao invés de ser guardado no armazenamento do navegador.

Nos dias atuais, os *supercookies* podem ser descritos como técnicas ou procedimentos que englobam três tipos de recolha diferentes (ENISA EU, 2012):

- Recolha de informação sobre os utilizadores de várias formas.
- Utilizam diferentes mecanismos de armazenamento. Dependendo do utilizador, este armazenamento pode ser efetuado através do navegador, através de *cookies Flash* e através de muitos mecanismos de armazenamento introduzidos pelo HTML5.

- Pode levar ao ressurgimento de outros *cookies*. Os *supercookies* são usados principalmente para ressuscitar *cookies* HTTP eliminados pelo utente (voluntariamente ou através da configuração do seu navegador).

Os *supercookies* mais conhecidos nos dias atuais são os *Flash cookies* e os *evercookies*.

3.1.5. Flash Cookies

Os *Flash cookies* foram identificados pela primeira vez em 2009 e foram a técnica mais forte de rastreamento que sucedeu aos *cookies* HTTP [4]. Eles têm certas semelhanças, nomeadamente, têm a possibilidade de guardar informações sobre os utilizadores quando estes estão a utilizar um qualquer serviço num sítio Web.

Os *Flash cookies* foram originalmente criados para servir as necessidades do utilizador. Para tal, têm a funcionalidade de melhorar a experiência do utilizador guardando as suas preferências, dados dos jogos, entre outras coisas. O armazenamento destes fica localizado no computador do utilizador como objetos locais partilhados.

Todo o processo de como um *Flash cookie* deve ser explorado para atingir um determinado objetivo pode ser entendido da seguinte forma: quando um sítio envia um elemento *Flash*, como por exemplo, um elemento publicitário, um pedido é realizado pelo navegador do cliente e um *cookie* é enviado para o utilizador. Como foi mencionado anteriormente, os sítios Web podem incorporar elementos *Flash* de entidades terceiras, o que dá a possibilidade do *Cookie Flash* realizar o rastreamento dos utilizadores.

O que torna o *Cookie Flash* particularmente perigoso e o que pode levar a violações de privacidade, são os mecanismos de *supercookies*, ou seja, são mais persistentes comparativamente com os *cookies* HTTP e podem conter até 100 KB de informação em relação aos *cookies* HTTP, que têm um limite de 4 KB. Uma outra característica de um *Cookie Flash* é que não tem uma data de validade. Independentemente de o utilizador realizar a ativação de definições de privacidade, tais como a limpeza automática da *cache* ou a limpeza do histórico no navegador, isso não tem qualquer efeito nos *Flash cookies.* Pior ainda, devido ao armazenamento destes *cookies* ser realizado no sistema de ficheiros local do utilizador, tal permite que seja feito o rastreamento através de diferentes navegadores.

Uma outra exploração dos *Flash cookies* é como repositório de cópias de segurança para *cookies* HTTP. Por exemplo, se um sítio na Web definir um *cookie* HTTP, o *cookie Flash* pode ser definido com o mesmo nome e valor. Caso o utilizador realize a limpeza do

histórico no seu navegador, o *cookie Flash* pode ressuscitar os *cookies* HTTP, o que se torna uma preocupação de privacidade significativa para o utilizador (Soltani et al., 2009).

Uma outra característica que pode pôr em causa a privacidade do utilizador pode ser o armazenamento de informação por parte de plug-ins, por exemplo, o *plugin Flash* tem a possibilidade de guardar dados de uma forma persistente sem que o próprio navegador tenha controlo sobre esses dados, o que poderá levar a um uso indevido e abusivo da informação do utilizador.

Todos estes mecanismos combinados fazem com que um *cookie Flash* seja um tipo de *supercookie* mais eficaz no rastreamento comparativamente aos *cookies* HTTP.

3.1.6. Ever*cookie*

O projeto *Evercookie* é uma API Java*Script* para produzir *cookies* extremamente persistentes, ou seja, pode-se pensar neste mecanismo como uma rotina de Java*Script* que coloca *cookies* persistentes no computador do utilizador tendo a possibilidade de os guardar em múltiplos locais de uma forma obscura, o que pode ser uma violação à privacidade do utilizador. Um exemplo pode ser esconder texto dentro de um ficheiro de imagem PNG. Se o ficheiro de *cookies HTTP* for apagado, o *Evercookie* tem a possibilidade de o restaurar. O *Evercookie* trabalha de forma a armazenar informações de *cookies* em vários tipos de mecanismos de armazenamento, onde pode incluir (Kamkar, 2010):

- *Cookies* HTTP
- *Flash cookies*
- HTML5 Session Storage
- HTML5 Local Storage
- HTML5 Database Storage via SQLite

A principal característica do *Evercookie* é que tem a possibilidade de recriar os *cookies* HTTP apagados dos repositórios normais dos navegadores. Portanto, os *evercookies* só precisam de estar presentes num repositório para recriar o *cookie* em todos os outros repositórios.

3.1.7. LocalStorage

Esta técnica designada por LocalStorage, é um objeto de armazenamento na Web, em HTML5, para guardar dados do cliente, localmente, no computador de um utilizador. Os

dados armazenados não têm data de validade e existirão até serem eliminados. Em contraste, o armazenamento por sessão, que é outra API de armazenamento na Web em HTML5, apaga os dados armazenados quando o navegador é encerrado.

Portanto quando é utilizado com ponderação, o localStorage pode ser uma poderosa solução de armazenamento de dados, onde de seguida, irá ser mencionado alguns exemplos que fazem uso desta técnica.

HTML5 Local and Session Storage

O HTML5 Web Storage permite armazenar algumas informações no computador do utilizador, semelhante aos *cookies* HTTP, mas de uma forma mais rápida e eficaz. A informação armazenada no Web Storage não é enviada para o servidor Web, ao contrário dos *cookies* HTTP, onde os dados são enviados para o servidor em cada pedido. Além disso, os *cookies* HTTP permitem armazenar uma pequena quantidade de dados (4KB), ao contrário deste mecanismo que permite armazenar até 5MB de dados. Existe dois tipos de Web Storage que diferem quanto à sua localização e tempo de vida útil:

- HTML5 *Local Storage*, que obedece à política *same-origin*, oferece ainda outra possibilidade de rastreamento aos utilizadores. A colocação dos objetos no armazenamento não requer qualquer *plugin* (*Web Storage API - Web APIs | MDN*, 2022). Os objetos são armazenados permanentemente, têm a possibilidade de persistirem até serem removidos pelo sítio Web ou pelo utilizador. Para além desta característica, qualquer objeto pode ter um tamanho até 5 MB, o que oferece uma enorme vantagem comparativamente com os *cookies* HTTP mas, pelo contrário, o utilizador pode apagá-los via navegador, o que não acontece com os *Cookies Flash* (Lawson & Sharp, 2011).

  Um exemplo da utilização do *Local Storage* está identificado num exemplo denominado por *Web Storage Demo*. Esta página Web oferece opções em que o utilizador tem a possibilidade de modificar a cor, fonte e imagem da página Web. Quando procede à escolha de opções diferentes, a página é automaticamente atualizada. Estas opções são armazenadas no *localStorage*, de modo que, quando o utilizador sair da página e voltar a carregar a mesma e as suas escolhas permaneçam inalteradas.
- HTML5 *Session Storage* é muito semelhante ao *Local Storage*, ou seja, obedece à política *same-origin* e a política de armazenamento e os objetos podem ter

tamanhos até 5 MB. No entanto, os objetos só estão disponíveis apenas para a janela do browser atual e são apagados quando a janela é fechada (*Web Storage API - Web APIs | MDN*, 2022).

Um exemplo de uma possível utilização do HTML5 *Session Storage* pode ser através da *Web Storage API* que fornece mecanismos através dos quais os navegadores podem armazenar com segurança, pares chave/valor. Na figura 12 pode-se observar alguns exemplos base de como podemos guardar, obter e remover dados no *SessionStorage*. Uma possível utilização destes comandos poderia guardar, por exemplo, a idade do utilizador daquela sessão e posteriormente apresentar anúncios dedicados a essa faixa etária.

```
myStorage = window.sessionStorage;

//Guardar dados em sessionStorage
sessionStorage.setItem('key', 'value');

//Obter dados guardados de sessionStorage
let data = sessionStorage.getItem('key');

//Remover dados guardados de sessionStorage
sessionStorage.removeItem('key');

//Remover todos os dados guardados de sessionStorage
sessionStorage.clear();
```

Figura 12 - Utilização de alguns comandos básicos da API

*Fonte: (*Web Storage API - Web APIs | MDN, *2022)*

Limitações

- LocalStorage é um mecanismo síncrona ou *synchronous*, termo usado em inglês. Bloqueia a execução do *thread* principal até a operação estar completa, o que tem um efeito negativo no desempenho de um sítio Web, principalmente quando há muitas operações.
- Qualquer código Java*Script* dentro de um sítio Web tem acesso ao LocalStorage, o que significa que está aberto a vários tipos de ataques ou rastreamentos. Um exemplo muito frequente de rastreamento através desta técnica é a possibilidade a um atacante introduzir *script*s do lado do cliente em páginas Web visualizadas

por outros utilizadores. Desta forma, se alguém introduzir o próprio código Java*Script* no seu sítio Web, pode recuperar todos os dados armazenados no LocalStorage e enviá-los para qualquer lugar.

A Tabela 2 faz uma comparação das técnicas *cookies* HTTP, *cookies Flash* e HTML5 Storage relativamente ao local do seu armazenamento, à dimensão máxima permitida, o instante de expiração e a sua acessibilidade. Alguns autores destacam os riscos de privacidade apresentados pelo HTML5, ao contrário de alguns autores, que mencionam que o HTML5 tem um grande potencial para preservar a privacidade (AYENSON* et al., 2009).

Tabela 2 - Características importantes dos cookies HTTP, Flash cookies e HTML5 Storage

Fonte: (AYENSON* et al., 2009)

	HTTP Cookies	Flash Cookies	HTML5 Storage
Armazenamento	4KB	100KB por defeito	5MB por defeito
Expiração	Sessão por defeito	Permanente por defeito	Permanente por defeito
Localização	Num ficheiro SQL (Firefox)	Guardado fora do Browser	Num ficheiro SQL (Firefox)
Acesso	Só através do Browser	Através de vários Browsers na mesma maquina	Só através do Browser

3.2. Mecanismos de rastreamento baseados na cache dos navegadores

O próximo mecanismo de rastreio que irá ser abordado trata-se de um mecanismo baseado na *cache* dos navegadores, que também utiliza métodos baseados no armazenamento. Existem múltiplas camadas de cache. Algumas caches são dedicadas a um único utilizador, outras são dedicadas a vários utilizadores. Algumas são controladas pelo servidor, outras pelo utilizador e outras por intermediários. Na figura 13 pode-se observar as várias camadas de cache entre um navegador e um servidor (Sonzogni, 2022).

Caches de navegadores - Estas caches são dedicadas a um único utilizador e estão implementadas no seu navegador. Melhoram o desempenho, evitando obter a mesma resposta várias vezes.

Local Proxy – Esta cache pode ser instalada pelo utilizador, mas também pode ser gerida por intermediários, como por exemplo, uma empresa, uma organização, ou um fornecedor

de Internet. Os *proxies* locais armazenam frequentemente uma única resposta para múltiplos utilizadores, o que constitui uma cache "pública". Os *proxies* locais têm múltiplas funções.

Cache de servidor de origem / CDN - É controlada pelo servidor. O objetivo da cache do servidor de origem é reduzir a carga no servidor de origem. A forma como realizam esta redução é guardar a mesma resposta para múltiplos utilizadores. Os objetivos de um CDN são semelhantes, mas estão distribuídos por todo o mundo e atribuídos a um conjunto de utilizadores mais próximo para reduzir a latência.

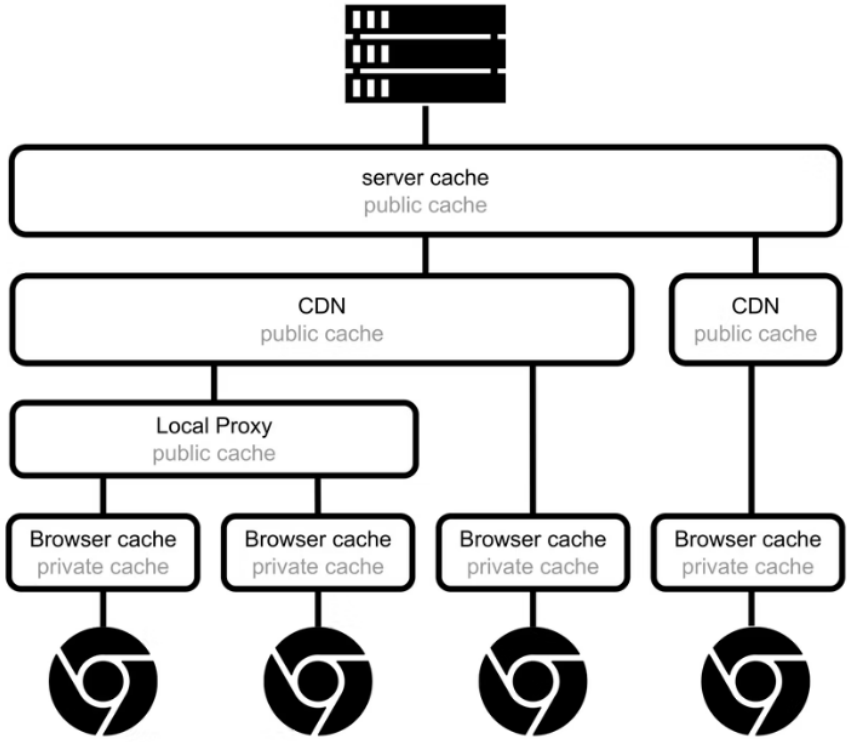


Figura 13 - Camadas de cache entre um navegador e um servidor

Fonte: (Sonzogni, 2022)

Este mecanismo explora as possibilidades de identificar e determinar quais os sítios Web que foram visitados anteriormente através da utilização de várias *caches*. A razão pela qual este mecanismo consegue realizar este tipo de identificações é porque tem guardado na cache, cópias dos ficheiros ou um local de armazenamento temporário que pode ser acedido de uma forma rápida, para assim realizar estas identificações.

3.2.1. *Loading performing test*

Os sítios na Web podem utilizar Java*Script* para detetar o tempo de carregamento de qualquer objeto (como por exemplo, uma imagem) a partir de qualquer URL. O tempo de carregamento pode ser medido, por exemplo por Java*Script*, e comunicado ao serviço Web, que pode avaliar se o objeto está ou não presente na cache do navegador. Desta forma, através de testes num objeto que é sempre acedido quando um utilizador utiliza o sítio Web, o *script* pode avaliar se o sítio *Web* foi previamente acedido (Bujlow et al., 2017). Para se entender melhor esta técnica irei dar um exemplo meramente ilustrativo. Pensemos em três amigos (o Ricardo, a Joana e o Diogo), a Joana visitou o sítio Web pertencente ao Ricardo, que tem como endereço http://www.ricardo.com. Mas o Ricardo quer saber se a Joana já visitou o sítio Web do Diogo (http://www.diogo.com).

Em primeiro lugar, o Ricardo observa o síto Web do Diogo e escolhe um ficheiro que a Joana possa ter interesse em visualizar no sítio Web do Diogo. O Ricardo supõe que ela escolhe o ficheiro que contém o logótipo corporativo do Diogo (ou seja, em http://www.diogo.com/logo.jpg). O Ricardo quer determinar se o ficheiro do logótipo está na cache Web da Joana. Se o ficheiro estiver na sua cache, então ela deve ter visitado o sítio Web do Diogo. Para chegar a esta conclusão, o Ricardo realiza uma implementação, por exemplo um Java*Script*, para proceder ao ataque o Ricardo incorpora esta implementação na sua página inicial. Quando a Joana visitar novamente o site Web do Ricardo, o Java*Script* é automaticamente descarregado e executado no navegador da Joana. A função deste Java*Script* é medir o tempo necessário que leva para aceder ao http://www. diogo.com/logo.jpg no dispositivo da Joana. Se o tempo for inferior a algum limiar (por exemplo, 80 milissegundos), o Ricardo conclui que a Joana esteve no sítio Web do Diogo, caso seja superior ao limiar, conclui-se que ela não esteve no sítio Web.

3.3. Utilização da colaboração inconsciente do utilizador

Neste mecanismo o objetivo é demonstrar como, por vezes, um utilizador pode estar a sofrer um rastreio através de interações inofensivas com o sítio Web. Estas interações, que aos olhos do utilizador parecem ser perfeitamente normais, estão por detrás, por exemplo, a detetar o histórico do navegador do utilizador sem que o mesmo se aperceba, ou seja, de uma forma inconsciente.

3.3.1. Captchas

CAPTCHA significa, teste de Turing Público Completamente Automatizado para fazer uma distinção entre Computadores e Humanos. Por outras palavras, CAPTCHA determina se o utilizador é real ou um robô de spam. O CAPTCHA manipula letras e números e confia na capacidade humana para determinar quais os símbolos que estão a ser representados. A forma como funciona um CAPTCHA é pedir aos utilizadores que identifiquem as letras. Estas letras estão distorcidas de modo que os robôs não sejam capazes de identificá-las. Para passar no teste, os utilizadores têm de interpretar o texto distorcido, escrever as letras corretas, por exemplo, num formulário e enviar esse formulário. Se as letras não corresponderem, os utilizadores são convidados a tentar novamente.

Entendido o conceito de CAPTCHA, este deve ser usado em diversos sítios Web pois tem uma funcionalidade importante. No entanto, podem ser usados de uma forma em que ponha em causa a privacidade do utilizador. Por exemplo, o sítio Web tem inserido na sua página um CAPTCHA onde é solicitado aos utilizadores que escrevam as letras que são mostradas. Porém, estas letras correspondem a hiper*links* para um URL que o atacante deseja investigar, ou seja, cada letra corresponde a um determinado sítio Web que o atacante deseja saber se o utilizador já visitou ou não. Para chegar à conclusão de que o utilizador visitou esse sítio Web, a cor de cada letra que o utilizador insere deve aparecer a preto. Caso contrário se a cor da letra for igual à cor do fundo do CAPTCHA, o sítio Web não foi visitado pelo utilizador.

O atacante deve ter o cuidado de algumas letras que o utilizador escrever, aparecerem sempre a preto, para os casos em que o utilizador não tenha visitado nenhum sítio Web que o atacante está a investigar (Chow et al., 2008).

3.3.2. *Clickjacking*

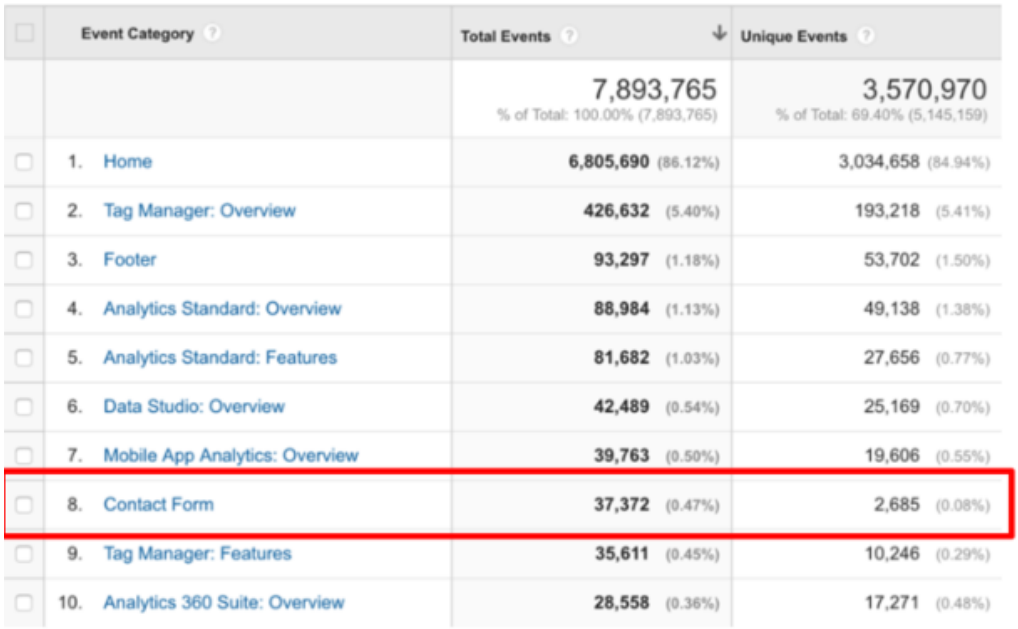
Clickjacking, consiste num atacante que usa múltiplas camadas transparentes ou opacas, para induzir um utilizador a clicar num botão ou link noutra página quando este pretendia clicar na página de nível superior. O utilizador acredita que está a clicar na página visível, mas de facto, está a clicar no elemento invisível. Este elemento invisível pode levar o utilizador para páginas maliciosas ou levar o utilizador a descarregar malware involuntariamente.

Como pode esta técnica ser usada para pôr em causa a privacidade de um utilizador? Um dos exemplos mais notórios foi um ataque contra a página de definições do plugin *Adobe Flash*. Ao carregar esta página para um iframe invisível, possibilita a que um atacante consiga induzir um utilizador, que tem por objetivo alterar as definições de segurança do Flash, a dar permissão para que qualquer *plugin Flash* utilize o microfone e a câmara do dispositivo (Rydstedt, 2022).

3.3.3. Utilização de *Event Tagging* para rastreamento através do campo de formulário

Pode-se encontrar em muitos sítios Web vários tipos de formulários, como por exemplo, formulários de subscrição, formulários de contacto, formulários de inscrição, entre outros. Compreender como os utilizadores, que acedem a um sítio Web, interagem com o formulário e como estão a ter sucesso através do formulário ou a abandonar o formulário, é um indicador chave do desempenho (KPI) importante para o negócio ou para um qualquer serviço no sítio Web (Senol et al., 2022).

Existem muitas ferramentas que têm um único objetivo em específico num campo do formulário, que passa por fazer um rastreamento sub-reptício a um utilizador num sítio Web. Uma possível ferramenta para fazer este tipo de rastreamento é a utilização do *Google Analytics Event Tracking*, onde esta é uma ótima forma de rastrear a interação dos utilizadores num formulário. A grande vantagem de utilizar esta ferramenta é que esta possui a paridade de dados e a capacidade de utilizar esses dados transversalmente com todos os outros dados do *Google Analytics*. Para se entender melhor este tipo de rastreamento, destaca-se um exemplo que é possível observar-se na Figura 14. Na coluna designada por "*Event Category*", pode-se observar a medição da ocorrência de um evento. Quando se seleciona a linha "*Contact Form*", passa-se ao nível seguinte da hierarquia.

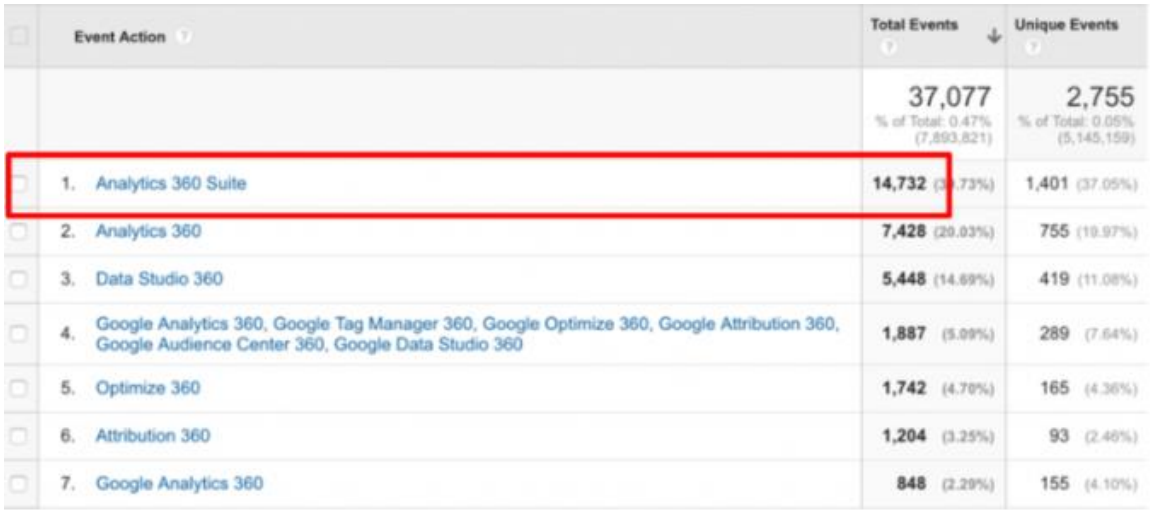


	Event Category	Total Events	Unique Events
		7,893,765 % of Total: 100.00% (7,893,765)	3,570,970 % of Total: 69.40% (5,145,159)
1.	Home	6,805,690 (86.12%)	3,034,658 (84.94%)
2.	Tag Manager: Overview	426,632 (5.40%)	193,218 (5.41%)
3.	Footer	93,297 (1.18%)	53,702 (1.50%)
4.	Analytics Standard: Overview	88,984 (1.13%)	49,138 (1.38%)
5.	Analytics Standard: Features	81,682 (1.03%)	27,656 (0.77%)
6.	Data Studio: Overview	42,489 (0.54%)	25,169 (0.70%)
7.	Mobile App Analytics: Overview	39,763 (0.50%)	19,606 (0.55%)
8.	Contact Form	37,372 (0.47%)	2,685 (0.08%)
9.	Tag Manager: Features	35,611 (0.45%)	10,246 (0.29%)
10.	Analytics 360 Suite: Overview	28,558 (0.36%)	17,271 (0.48%)

Figura 14 - Medição da ocorrência de um evento

Fonte: (Seiden, 2018)

Na Figura 15, observa-se a utilização da secção "Event Action" para capturar o produto que alguém indicou estar interessado. Mais uma vez, se selecionarmos uma fila, por exemplo, 'Analytics 360 Suite', passa-se para o nível seguinte e final da hierarquia, a "Event Label". Neste nível, utilizou-se o campo da etiqueta para capturar o campo do formulário com o qual um utilizador interagiu.

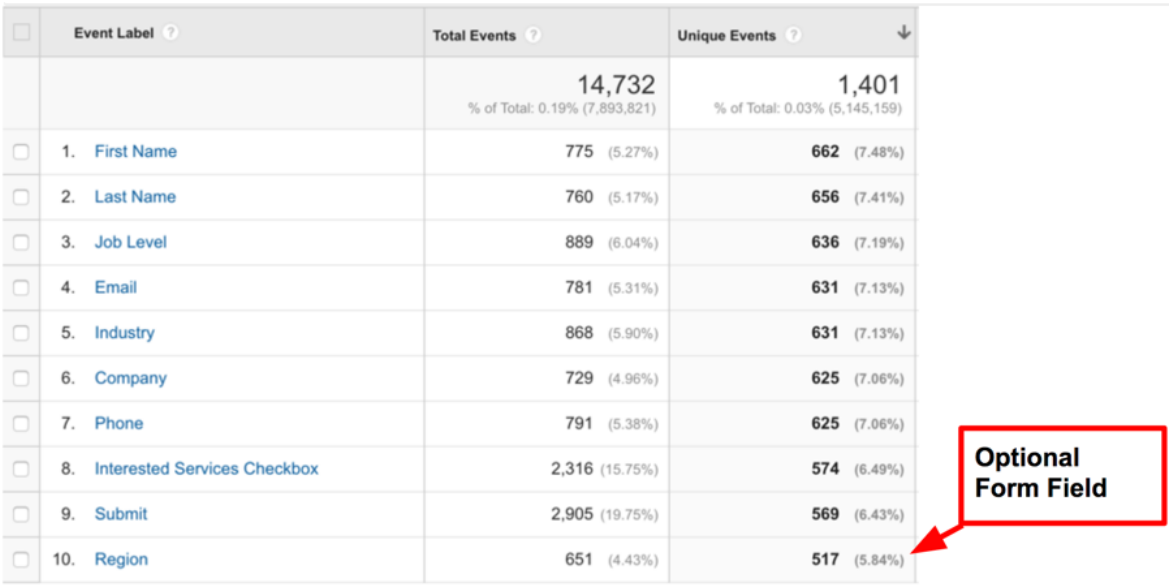


	Event Action	Total Events	Unique Events
		37,077 % of Total: 0.47% (7,893,821)	2,755 % of Total: 0.05% (5,145,159)
1.	Analytics 360 Suite	14,732 (39.73%)	1,401 (37.05%)
2.	Analytics 360	7,428 (20.03%)	755 (19.97%)
3.	Data Studio 360	5,448 (14.69%)	419 (11.08%)
4.	Google Analytics 360, Google Tag Manager 360, Google Optimize 360, Google Attribution 360, Google Audience Center 360, Google Data Studio 360	1,887 (5.09%)	289 (7.64%)
5.	Optimize 360	1,742 (4.70%)	165 (4.36%)
6.	Attribution 360	1,204 (3.25%)	93 (2.46%)
7.	Google Analytics 360	848 (2.29%)	155 (4.10%)

Figura 15 - Secção "Event Action" para capturar o produto interessado

Fonte: (Seiden, 2018)

Como se pode observar na Figura 16, foi recolhido o Nome, Apelido, Email, entre outras informações que são pertencentes ao utilizador (Seiden, 2018).



	Event Label	Total Events	Unique Events
		14,732 % of Total: 0.19% (7,893,821)	1,401 % of Total: 0.03% (5,145,159)
1.	First Name	775 (5.27%)	662 (7.48%)
2.	Last Name	760 (5.17%)	656 (7.41%)
3.	Job Level	889 (6.04%)	636 (7.19%)
4.	Email	781 (5.31%)	631 (7.13%)
5.	Industry	868 (5.90%)	631 (7.13%)
6.	Company	729 (4.96%)	625 (7.06%)
7.	Phone	791 (5.38%)	625 (7.06%)
8.	Interested Services Checkbox	2,316 (15.75%)	574 (6.49%)
9.	Submit	2,905 (19.75%)	569 (6.43%)
10.	Region	651 (4.43%)	517 (5.84%)

Figura 16 - Informação recolhida

Fonte: (Seiden, 2018)

3.4. Mecanismos de rastreamento baseado em *Fingerprinting*

O mecanismo de rastreamento baseado em *fingerprinting*, é um processo em que um sítio Web reúne pequenas informações sobre o dispositivo do utilizador e reúne toda essa informação para formar uma imagem única, ou impressão digital, do dispositivo do utilizador. Desta forma, um utilizador pode estar a ser alvo de rastreamento em diversos sítios na Web pertencentes a diferentes entidades, o que não é possível com a utilização de *cookies*. Este mecanismo pode deixar o utilizador vulnerável pois este não tem como saber se está a ser alvo de rastreamento e não sabe como evitar isso apesar de que, é possível evitar este mecanismo desativando o suporte do Java*Script*, Java e Flash, no entanto, muitos utilizadores não sabem da existência desta solução.

De acordo com o artigo elaborado pelos autores Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., & Vigna, G. (2013), 40 dos 10 000 sítios Alexa, usam *script*s *fingerprinting* de *BlueCava* (DiGioia, 2014), *Iovation* e *ThreatMetrix*. As categorias mais populares que utilizam *fingerprinting* são as de porno, com uma percentagem de 15% e para encontros, com uma percentagem de 12.5%.

De seguida, destaco três subsecções relacionadas com este mecanismo de *fingerprinting*, que estão relacionadas na forma de como um sítio pode recolher informações acerca do utilizador e, desta forma, pôr em causa a sua privacidade.

3.4.1. *Device Fingerprinting*

Device fingerprinting envolve a recolha de informações do sistema. No artigo elaborado pelos autores Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., & Preneel, B. (2013) identifica vários atributos do sistema que podem ser utilizados que podem fazer parte deste conceito de *device fingerprint*, onde estes podem incluir o tamanho do ecrã do dispositivo, versões do sistema operativo e a lista de fontes instaladas. Os sítios na Web podem utilizar Java*Script* e a propriedade *navigator* (M. W. Docs, 2020) para recolher algumas destas informações. O objeto *navigator* pode ser recuperado usando uma propriedade chamada window.navigator (Network, 2020). Este objeto contém informações da lista de plugins instalados no navegador, plataforma do navegador, línguas conhecidas pelo utilizador e muito mais (Mowery & Shacham, 2012).

3.4.2. *Operating System Instance Fingerprinting*

As técnicas de *fingerprinting* nesta categoria recolhem informações sobre o sistema operativo dos utilizadores, onde inclui frequentemente a arquitetura e versão do sistema operativo, em que ambos podem ser reunidos com Flash e Java*Script*. Em Java*Script*, os programadores podem usar a propriedade *navigator.oscpu* que devolve uma *string* com informações sobre o sistema operativo dos utilizadores (M. W. Docs, 2022).

Uma lista das fontes instaladas pelos utilizadores pode também ser reunida usando Java*Script*. A largura, altura, cor e profundidade de pixel do ecrã do utilizador pode ser facilmente recolhido usando Java*script* e a propriedade *window.screen* (W3schools, 2022). O *Flash* pode também detetar as capacidades de áudio dos utilizadores, o que oferece também a possibilidade de saber se os navegadores têm acesso permitido à câmara e ao microfone, entre outras coisas.

3.4.3. *Navegador Related Fingerprinting*

Esta técnica recolhe informações sobre o navegador do utilizador. Num artigo elaborado pelos autores Unger, T., Mulazzani, M., Frühwirt, D., Huber, M., Schrittwieser, S., &

Weippl, E. (2013), os autores identificaram características CSS e HTML5 que podem ser usadas para recolher informação sobre o navegador que pode ser usada como parte da *fingerprint*. Encontraram três métodos relacionados com CSS que podem ser utilizados, mais especificamente propriedades CSS, CSS *selectors* e filtros CSS e onde os navegadores implementam estes de forma diferente. Quando se trata de *fingerprinting* baseadas em HTML5, os autores descobriram que os detalhes de implementação do HTML5 dentro de diferentes navegadores varia e ao olhar para estes detalhes, um navegador pode ser identificado.

3.5. Estado da arte sobre plataformas e ferramentas de avaliação

Feita uma análise sobre as várias técnicas que um sítio na Web pode utilizar para pôr em causa a privacidade de um utilizador, foi realizada uma pesquisa sobre as diferentes plataformas e ferramentas de avaliação que podem automaticamente analisar, sob qualquer forma, as técnicas de violação de privacidade usadas por um determinado sítio Web. Temos de ter em conta que ao longo do tempo tem-se desenvolvido estratégias de reforço da privacidade que podem efetivamente bloquear mecanismos de rastreio, mas, como foi explicado anteriormente, o nosso objetivo passa por analisar se o serviço oferecido através da plataforma Web vai ao encontro com um conjunto de requisitos que garantam ao utilizador que a sua privacidade é respeitada durante a interação com o sítio na Web.

Tendo em conta este pensamento, irão ser apresentados estudos onde realizaram uma avaliação sobre a segurança e privacidade nos sítios na Web.

PrivacyScore

O trabalho realizado no artigo elaborado pelos autores Maass, M., Wichmann, P., Pridöhl, H., & Herrmann, D. (2017) teve por objetivo a introdução do *PrivacyScore*, ou seja, um portal automatizado para *scanning* de sítios Web que permite a que qualquer pessoa possa fazer *benchmark* sobre a segurança e a observação de características relacionadas com a privacidade de vários sítios na Web. Para alem disto, os utilizadores podem controlar a forma como a metodologia de classificação e análise é feita.

Posteriormente, foi feita uma descrição sobre as principais características da plataforma *PrivacyScore*. Como em todos os *scanning* de serviços Web, o *PrivacyScore* permite que o utilizador submeta URL´S dos sítios na Web que pretende analisar em termos de segurança

e privacidade. Os utilizadores podem navegar na base de dados das listas de sítios Web existentes ou criar novas listas de sítios Web. Após a realização desta etapa e feita a seleção dos dados mais relevantes para o utilizador, estes dados são submetidos para o *scanning engine* que irá recolher uma serie de factos utilizando múltiplos *scan modules.*

Os factos recolhidos serão avaliados por propriedades que foram selecionadas pelos respetivos autores e que estarão de acordo com a segurança e privacidade de um sítio na Web. De forma a tornar os resultados mais acessíveis ao utilizador final, os *checks*, que são as técnicas que os autores consideraram ser a melhor forma para classificar um sítio na Web consoante os seus objetivos, são organizados em grupos. O esquema de classificação define como os *checks* devem ser organizados em grupo, como cada resultado possível, consoante o *check*, deve ser classificado e que a importância é associada a cada *check* e respetivo grupo. Os *checks* foram definidos da seguinte forma: *NoTrack, Attacks, EncWeb, and EncMail,* onde cada abreviatura se refere:

- *NoTrack* - Problemas relacionados com a privacidade mais especificamente, rastreamento, análise e serviços de publicidade.
- *Attacks* – Verificam se um sítio na Web define os cabeçalhos HTTP de uma forma segura como *ContentSecurity-Policy* e *X-XSS-Protection,* onde estes protegem os utilizadores de certos ataques.
- *EncWeb* – Verifica se um sítio na Web segue as melhores praticas de implementação TLS, verifica também se estes sítios na Web estão vulneráveis a ataques conhecidos e se contêm conteúdos não encriptados sobre uma página encriptada.
- *EncMail* – Similar ao *EncWeb* mas realiza *checks* que são executados para o servidor de correio primário que está listado no registo de domínio MX.

O esquema de classificação atualmente implementado baseia-se na ordem definida pelo utilizador consoante os seus grupos *check*. Os utilizadores podem manipular a ordem de acordo com as suas preferências. Sítios Web com uma classificação a "bom" (que irá ter uma cor verde) na coluna de verificação da prioridade (coluna mais à esquerda) são elevados para o topo da tabela, seguidos por sítios Web com uma classificação amarela e vermelha no primeiro grupo *check*, respetivamente. Todos os sítios Web com a mesma classificação, ou seja, irão ter a mesma cor no primeiro grupo *check* são ainda classificados de acordo com a sua classificação no grupo *check* seguinte e assim sucessivamente até que todos os grupos

tenham sido considerados e avaliados. Podemos ver um exemplo de um quadro de classificação na Figura 17.

#	URL	Name	Type	NoTrack	Attacks	EncWeb	EncMail	Rating
1	http://www.ibb.de/ (1 failure) / 2017-06-24 @ 18:29:42	IBB Berlin	public					
2	http://www.helaba.de/ / 2017-06-24 @ 18:26:12	Hessische Landesbank	public					
3	http://www.berlinhyp.de/ / 2017-06-24 @ 18:24:42	Berlin Hyp AG	public					
4	http://www.bayernlb.com/ / 2017-06-24 @ 18:24:26	Bayerische Landesbank	public					
5	http://www.bhw.de/ / 2017-06-24 @ 18:23:35	BHW Bausparkasse AG	private					
6	http://www.pfandbriefbank.com/ / 2017-06-24 @ 18:22:38	Deutsche Pfandbriefbank AG	private					
7	http://www.hypovereinsbank.de/ / 2017-06-24 @ 18:22:57	Unicredit Bank AG	private					

Figura 17 - Ranking de uma lista de sítios na Web das home page dos principais bancos alemães

Fonte: (Maass et al., 2017)

Para termos uma melhor noção dos potenciais desta ferramenta, foi realizada uma análise sobre alguns sítios Web portugueses, mais especificamente, sítios de alguns bancos portugueses, seguradoras e o sítio da Assembleia da República. Pode-se observar a sua análise na Figura 18, de realçar que alguns sítios não foram analisados devido à falta de permissão de alguns sítios para a realização deste tipo de analises.

#	URL		Localiza��o	NoTrack	EncWeb	Attacks	EncMail	Rating
1	https://ind.millenniumbcp.pt/	/ 2022-01-05 @ 15:38:45	Portugal					
2	https://www.tranquilidade.pt/	(1 failure) / 2022-01-05 @ 15:39:32	Portugal					
3	https://www.bancobpi.pt/	(1 failure) / 2022-01-05 @ 15:17:30	Portugal					
4	https://www.allianz.pt/	-	Portugal	n/a				
4	https://www.libertyseguros.pt/	-	Portugal	n/a				
4	https://www.parlamento.pt/	-	Portugal	n/a				
4	https://www.creditoagricola.pt/	-	Portugal	n/a				

Figura 18 - Ranking de uma lista de sítios na Web das home page de alguns sítios portugueses

Fonte: Elaboração própria

WBF Analyzer

O trabalho realizado no artigo elaborado pelos autores de Matos, G. F., & Feitosa, E. L. (2021), teve por objetivo a deteção de chamadas *fingerprinting* aos objetos Java*Script* em páginas Web e medir o nível de severidade. Analisando mais em concreto as funcionalidades da *WBF Analyzer*, esta precisa de códigos Java*script* dos URLs que irão ser analisados e onde este processo pode ser realizado através de um *Web crawler*. De seguida irão ser referidas, resumidamente, as fases que serão desempenhadas por este método:

- Pré-Processamento – Nesta etapa realiza-se o pré-processamento do código Java*script*, que de uma forma resumida, contempla uma fase de extração que verifica e limpa o código para que este se torne legível.
- Identificação - o bloco Java*script* é recebido e um *parser* transforma o respetivo código numa AST, que é então repassada para o identificador. Uma AST pode ser entendida como uma representação intermediária do código fonte que favorece análises insensíveis ao fluxo e possibilita a exploração de diversos caminhos a serem verificados. Após o identificador receber a AST como entrada, ele aplicará três regras necessárias para deteção das chamadas de *navegador fingerprinting*. A primeira regra, mais concretamente a regra de extração, percorrerá a AST de entrada para identificar e realizar a extração dos termos. A segunda regra, Normalização, tem por objetivo normalizar a relação entre objetos e propriedades para produzir as chamadas que dão origem ao escopo normalizado. Por fim, a terceira regra, Classificação, aplica a contagem de frequência, a classificação do risco de chamada e a classificação do risco na página Web (Matos, 2021).

De forma a visualizar o funcionamento deste método, os autores testaram a *WBF Analyzer* num conjunto de 50 sítios que o top 50 Alexa é composto no ano de 2021, onde foram extraídos somente os códigos Java*script* das páginas principais de cada sítio.

Como resultado desta análise, a *WBF Analyzer* conseguiu classificar, em relação ao risco, que 81.6% (40) dos sítios com um risco alto, 8.2% (4) dos sítios com um risco médio e 10.2% (5 sítios) com um risco baixo, como se pode observar na Tabela 3.

Tabela 3 - Conjunto de sítios Web e a respetiva classificação quanto ao risco de deteção de fingerprinting

Fonte: Elaboração própria

#	Site	Risco	#	Site	Risco	#	Site	Risco
1	pandas.Tv	Nada encontrado	2	bongacams.com	Alto	3	intl.alipay.com	Médio
4	login.microsoftonline.com	Baixo	5	facebook.com	Alto	6	outlook.live.com	Alto
7	stackoverflow.com	Alto	8	weibo.com	Alto	9	vk.com	Alto
10	www.17ok.com	Baixo	11	twitter.com	Alto	12	www.360.cn	Alto
13	www.aliexpress.com	Alto	14	www.amazon.co.jp	Alto	15	www.adobe.com	Alto
16	www.amazon.com	Alto	17	www.amazon.in	Alto	18	www.baidu.com	Alto
19	www.aparat.com	Alto	20	www.bing.com	Baixo	21	www.csdn.net	Alto
22	www.google.com.hk	Alto	23	www.ebay.com	Alto	24	www.huanqiu.com	Alto
25	www.instagram.com	Baixo	26	www.google.com	Alto	27	www.jd.com	Alto
28	www.linkedin.com	Alto	29	www.naver.com	Alto	30	www.microsoft.com	Alto
31	www.netflix.com	Médio	32	www.office.com	Alto	33	www.okezone.com	Alto
34	www.qq.com	Alto	35	www.reddit.com	Baixo	36	www.shopify.com	Alto
37	www.sina.com.cn	Alto	38	www.sohu.com	Alto	39	www.tianya.cn	Alto
40	www.taobao.com	Alto	41	www.twitch.tv	Alto	42	www.wikipedia.org	Médio
43	www.tmall.com	Alto	44	www.yahoo.co.jp	Alto	45	www.yahoo.com	Alto
46	www.yy.com	Alto	47	www.zhanqi.tv	Médio	48	xinhuanet.com	Alto
49	zoom.us	Alto	50	www.youtube.com	Alto			

WebXray

A ferramenta *WebXray* é utilizada para analisar o tráfego e o conteúdo de páginas Web, extrair políticas legais e identificar as empresas que recolhem dados dos utilizadores. Com uma interface de fácil compreensão para o utilizador na linha de comando, torna o *WebXray* relativamente fácil de utilizar por não programadores e utilizadores com necessidades avançadas poderem analisar milhares de milhões de pedidos tendo um total aproveitamento da arquitetura distribuída do *WebXray*. Portanto, esta ferramenta tem sido utilizada para executar centenas de sessões de navegação simultâneas distribuídas por vários continentes.

Esta ferramenta é desenvolvida principalmente em *Python* e trabalha em várias etapas. Primeiro, uma lista de endereços de sítios Web é passada ao programa. Esta lista é processada para assegurar que todos os endereços estão devidamente formatados. De seguida, cada sítio Web é passado através do módulo de subprocesso *Python* para uma instanciação na linha de comando do navegador *PhantomJS*. Neste contexto, "headless" refere-se ao facto de o navegador correr numa linha de comando e não requerer uma interface gráfica para o utilizador. O *PhantomJS* recebe como argumentos um endereço Web e um programa *Javascript* que é responsável por carregar o sítio Web, processar o título da página e metadados, recolher *cookies* e detetar tanto pedidos HTTP como eventos HTTP recebidos. Dados sobre a página, são armazenados numa base de dados e depois os pedidos HTTP são examinados.

No artigo elaborado pelo autor Libert, T. (2015), os resultados indicaram que nove em cada dez sítios Web divulgam dados de utilizadores a entidades às quais o utilizador provavelmente desconhece, mais de seis em cada dez sítios Web geram *cookies* de entidades terceiras e mais de oito em cada dez sítios Web carregam código *Javascript* de partes externas, onde posteriormente ficam armazenados nos computadores dos utilizadores. Os sítios Web que divulgam dados dos utilizadores contactam uma média de nove domínios externos, indicando que os utilizadores podem sofrer um rastreamento por diversas entidades terceiras em conjunto. Ao rastrear a divulgação, não intencional, de históricos pessoais de navegação na Web, foi revelado que um conjunto de empresas americanas recebem a grande maioria dos dados dos utilizadores.

Outras plataformas de recolha de dados para medição da privacidade

Feita uma análise sobre os vários estudos que foram implementados para chegar à realização de uma avaliação tendo em conta a preservação da privacidade na informação do utilizador, foi feito de seguida, o estudo das diferentes plataformas disponíveis que podem medir, de alguma forma, a privacidade de um sítio na Web automaticamente e recolher esses dados:

- *FourthParty* (Mayer & Mitchell, 2012)– Trata-se de um plugin do *Firefox* para instrumentar o navegador para observar o trafego de HTTP, DOM *Windows*, *cookies*, carregamento de recursos e também tem a possibilidade de instrumentar

a API Java*script*. Suporta um subconjunto da plataforma OpenWPM, que será explicada com mais detalhe no decorrer da dissertação e por fim, este *plugin* pode ser utilizado para detetar *fingerprinting*.

- *Chameloen crawler* (Ghostwords, 2016) **–** É um *tracker* baseado em Chromium onde tem a possibilidade de utilizar uma extensão no navegador *Chameleon* para detetar o *fingerprinting* do mesmo. É automatizado, mas apenas deteta uma parte da superfície da *fingerprinting*.
- *TrackingObserver* (Roesner, 2014) – É uma extensão que deteta, realiza a medição e bloqueia *third-party Web trackers*. Não utiliza *blacklists*, mas tem a possibilidade de detetar o comportamento nos navegadores.
- *FPDetective* [20] - Trata-se de uma plataforma que realiza a deteção e análise de *fingerprinting* do navegador. O seu foco principal centra-se na deteção de *fingerprinting* ao invés de confiar em *fingerprinters* conhecidas, mas a realização da sua construção teve por objetivo a analise destas mesmo e grande parte da sua funcionalidade tem por objetivo suportar isso.
- O *OpenWPM* (Steven Englehardt, 2013) – É uma ferramenta *open-source* que tem por objetivo medir a privacidade em sítios na Web, utiliza uma versão automatizada do *navegador Firefox10* e suporta também a recuperação automática de falhas neste navegador. Esta ferramenta usa, em conjunto, o *Selenium11* (Selenium, 2022) que se trata de um *Web driver*, que pode ser utilizado nos navegadores *Firefox*, *Chrome*, *Internet Explorer* e *PhantomJS*, que foi implementada para detetar e caracterizar comportamentos de rastreio *online*. Esta ferramenta dá a possibilidade de usá-la em diferentes navegadores, o que se torna numa grande vantagem.

4. Seleção do tipo de análise e plataforma

O presente capítulo irá focar-se na análise das entidades terceiras que poderão estar presentes num sítio Web. Serão apresentadas justificações para a seleção deste foco, pois considera-se ter uma maior violação da privacidade para o utilizador quando este usufrui de um serviço num sítio Web.

Posteriormente, será feita uma seleção da plataforma que irá permitir analisar os sítios Web considerando o foco que foi selecionado.

4.1. Problemas de privacidade relacionada com entidades terceiras

Como foi observado, existem diversos mecanismos que têm por objetivo rastrear um utilizador, sem o seu consentimento, quando este está a interagir com um sítio Web. Portanto, ao decidir-se qual o foco, pensou-se em todas as formas que poderão estar presentes num sítio Web que tivessem algum impacto na privacidade quando o utilizador interage com um sítio Web. Desta forma, chega-se à conclusão do impacto que os serviços de entidades terceiras e a forma como estas têm, por exemplo, a facilidade de acompanhar as atividades de navegação de um utilizador. Uma dessas formas de acompanhar a atividade do utilizador é, por exemplo, a definição de *cookies* por parte das entidades terceiras pois estas não têm a necessidade de definir este tipo de tecnologia. Consequentemente, esta definição de *cookies* poderá ser um indicador de que uma entidade terceira estará a fazer um rastreamento ao utilizador. Desta forma, serão analisadas as implicações relacionadas com a privacidade com a possibilidade de rastreamento por parte de entidades terceiras nos sítios Web. Será feita uma análise sobre a informação do histórico de navegação que está disponível para entidades terceiras e como essa informação é identificável.

I. Informações disponíveis

O histórico de navegação na Web está obrigatoriamente associado à informação pessoal de um utilizador. Os recursos que um utilizador usufrui podem revelar diversas informações como a sua localização, interesses, compras, entre outras.

Quando um recurso de entidade primeira incorpora conteúdos de entidades terceiras, esta entidade é normalmente informada sobre o URL do recurso da entidade primeira através de um cabeçalho. Se um recurso incorporar uma *script* de uma entidade terceira, esta também

descobrirá, por exemplo, o título da página do sítio Web e em alguns casos, as entidades primeiras transmitem ainda mais informação de forma voluntária às entidades terceiras.

II. Identificação

Um histórico de navegação na Web é frequentemente identificado. Os autores Mayer and Mitchell (2012), analisaram uma taxonomia de cinco formas em que um histórico de navegação pode ser identificado.

1. **Uma entidade terceira também pode ser uma entidade primeira** – Uma entidade terceira pode ser uma entidade primeira num outro contexto, ou seja, num contexto em que o utilizador forneceu voluntariamente a sua identidade. O *Facebook*, por exemplo, tem mais de 800 milhões de utilizadores e impõe uma exigência em que os utilizadores forneçam o seu verdadeiro nome ao serviço. Quando um recurso inclui um *widget* social do *Facebook* proveniente de entidades terceiras, o *Facebook* identifica o utilizador para personalizar o *widget* (Mayer & Mitchell, 2012).
2. **Uma entidade primeira vende a identidade do utilizador** - Uma entidade primeira do sítio Web fornece voluntariamente a identidade de um utilizador para entidades terceiras, a troco de pagamento. Em certos casos, fizeram um negócio modelo do mesmo, em forma de sorteio ou *quiz* gratuito. Vários fornecedores de dados publicitários (por exemplo, *Datalogix* (Datalogix, 2022)) compram informação de identificação, recuperam o dossiê do utilizador a partir de uma base de dados de consumidores *offline* e utilizam-na para direcionar a publicidade.
3. **Uma entidade primeira fornece, involuntariamente, a identidade** - Se um sítio na Web coloca informações de identificação num URL ou no título da página, pode estar, involuntariamente, a divulgar informações a entidades terceiras. Num artigo elaborado pelos autores Krishnamurthy, B., Naryshkin, K., & Wills, C. E. (2011), os autores examinaram a interação de 120 sítios Web para descobrirem a fuga de informação para entidades terceiras. Concluíram que 48% sítios Web revelaram o identificador do utilizador num *cabeçalho Referrer*.

4. **Uma entidade terceira utiliza uma exploração de segurança** – Uma entidade terceira pode explorar uma vulnerabilidade de segurança num sítio Web da entidade primeira para descobrir a identidade do utilizador. Num artigo analisado (Narayanan, 2010), o autor mostrou como a captura de *frames* inadequada pode facilitar a identificação de um utilizador.
5. **Re-identificação** - A entidade terceira poderia corresponder históricos de navegação a um conjuntos de dados ou *datasets* identificados para reidentificar OS utilizadores, tal como os autores *Narayanan* e *Shmatikov* fizeram com o *dataset* do Prémio Netflix (Narayanan & Shmatikov, 2008).

4.2. Plataforma selecionada

Definido o aspeto que se considera ser o mais prejudicial e que pode ter um impacto significativo na privacidade do utilizador, o próximo passo foi arranjar uma forma de observar estes aspetos nos sítios Web e retirar as diversas análises. A observação de sítios Web e serviços para detetar, caracterizar e quantificar comportamentos que tenham um impacto na privacidade de um utilizador, provou ter uma enorme influência. Enquanto ferramentas que bloqueiam o rastreamento nos sítios Web são utilizadas apenas por uma pequena minoria e abordam apenas uma parte do problema, a medição da privacidade num sítio Web tem forçado, constantemente, as empresas a melhorarem as suas práticas de privacidade.

Portanto, através de uma pesquisa sobre as diversas plataformas e ferramentas (identificadas e analisadas no capítulo 3) que poderiam identificar todo o tipo de funcionalidades e serviços que uma entidade terceira poderia realizar num sítio Web. Foi encontrada uma plataforma que atinge a maior parte dos requisitos necessários para realizar uma análise que vá ao encontro do nosso objetivo. Desta forma, a plataforma encontrada foi o *WebXray*.

4.2.1. *WebXray*

O *WebXray* é utilizado para detetar pedidos e *cookies* de entidades terceiras. A forma como esta ferramenta realiza este tipo de operações pode ser entendida da seguinte forma: ao receber uma lista de URLs, o *WebXray* carrega cada página no navegador mostrando de uma forma mais realista o comportamento real de um utilizador. Durante o carregamento da

página não é necessária qualquer interação, o que significa que as notificações de aceitação de *cookies* não são executadas e todos os *cookies* são definidos sem o consentimento do utilizador (Libert, 2022).

Devido a todas estas funcionalidades, os utilizadores têm a possibilidade de aceder a relatórios com diversas informações, como por exemplo, números médios de entidades terceiras e *cookies*, domínios e elementos de entidades terceiras mais comuns, volumes de dados transferidos, obter informação sobre a utilização de encriptação SSL. Para complementar, o *WebXray* utiliza uma biblioteca personalizada de propriedade de domínios para identificar a origem do fluxo de dados de um determinado domínio de entidades terceiras para um proprietário empresarial e, se aplicável, para as empresas de origem. O esquema de dados utilizado para a biblioteca personalizada de propriedade do domínio é flexível, permitindo a geração de relatórios personalizados, bem como extensões de autoria para adicionar fontes de dados adicionais (Dussutour, 2020). Para além da monitorização de conteúdos e *cookies*, esta ferramenta pesquisa e extrai *links* para políticas de privacidade numa dada página. O texto de todas as ligações é avaliado para encontrar correspondências numa lista de termos associados com políticas de privacidade. Uma vez descobertas as ligações às políticas, realiza-se a recolha e análise de políticas de privacidade.

Uma grande vantagem desta ferramenta é que permite analisar as páginas à escolha do utilizador. A forma para realizar esta operação será através da colocação de todos os endereços das páginas que se deseja analisar num ficheiro de texto e colocar este ficheiro no diretório “page_lists”. Uma chamada de atenção que os autores mencionam é que os endereços têm de começar por "http://" ou "https://", caso contrário, o *WebXray* não os reconhecerá como endereços válidos. Uma vez colocada a lista de páginas que se deseja analisar, pode-se prosseguir para a execução da ferramenta. Tem-se também a possibilidade de analisar uma única página através do comando “python3 run_*WebXray* .py -s” e de seguida o endereço da página que queremos analisar.

Limitações

Neste ponto, destaca-se algumas limitações relacionadas com esta plataforma que foram identificadas na utilização da mesma:

- O conjunto de páginas pode não ser totalmente abrangido e consequentemente não apresentar o conteúdo que desejamos analisar.

- Pode potencialmente falhar em alguns mecanismos de rastreio ou ser sinalizado como "bot", resultando assim numa subcontagem de atributos.
- As contagens de certos atributos podem ficar aquém do número real de pedidos de entidades terceiras que são feitos. Feita experiências em certas análises, deparou-se que os resultados, em relação aos pedidos, têm um limite inferior, comparativamente ao montante real que poderá ser superior. Esta limitação foi também mencionada no artigo (Libert, 2015). No entanto, dada a extensa prevalência de pedidos de entidades terceiras, este constrangimento serve apenas para destacar a magnitude e o alcance dos resultados.

5. Criação do quadro de avaliação

Definida a ferramenta que irá permitir analisar os diversos sítios Web, é então necessário definir os atributos que poderão violar a privacidade de um utilizador de uma forma sub-reptícia. Desta forma, estes atributos serão selecionados para constituírem o quadro de avaliação que permitirá classificar os sítios Web que serão posteriormente selecionados.

5.1. Definição de *cookies* por parte de entidades terceiras

Como mencionado no decorrer desta dissertação, a definição de *cookies* por parte das entidades terceiras pode ser um indicador de que uma entidade terceira poderá ter intenções de realizar um rastreamento sub-reptício ao utilizador. Portanto, será abordado uma classe, relacionada com os *cookies*, designado por samesite. Cuja entidade terceira pode utilizar com a intenção de rastrear um utilizador.

Problemas de privacidade

Como mencionado no capítulo 2, a definição de *cookies* por parte de entidades terceiras pode causar diversas preocupações ao utilizador. Portanto, foi criado uma classe designada por samesite com o objetivo de impedir um qualquer rastreamento por parte das entidades terceiras. Porém, esta classe pode possuir uma configuração que qualquer atacante pode usufruir para contornar estas funcionalidades. Essa configuração designa-se por none (Krawczyk, 2022). O que significa que um atacante pode utilizar esta configuração para comunicar claramente o seu desejo intencional de enviar um *cookie* através de entidades terceiras para um uso ilegítimo.

5.2. Definição de cabeçalhos por parte de entidades terceiras

Esta funcionalidade poderá ser usada para diversos fins, como por exemplo análise de conteúdo, exploração de *login* ou otimização de *caching*. Por outro lado, existem utilizações mais problemáticas, tais como rastreamento, roubar informação ou, de uma forma sub-reptícia, criar fugas de informação inadvertidas sobre informação sensível.

Classe *Referrer-Policy*

Serão mencionadas algumas configurações possíveis para esta classe que podem ser atribuídas aos pedidos feitos pelas entidades terceiras com o objetivo de infringir a privacidade de um utilizador.

Problema de privacidade

O *Referrer* e toda a sua informação disponível pode ser muito benéfico para determinados serviços. Por exemplo, um serviço analítico pode usar toda a informação disponível para determinar que 50% dos visitantes do sítio Web "site-two.example" vieram do sítio Web "social-network.example". Por outro lado, quando o URL está completo, ou seja, quando a informação disponível no URL contém o caminho e a *string* de consulta, que é enviada pelo cabeçalho *Referrer* entre origens, pode causar impactos significativos na privacidade de um utilizador. No exemplo da Figura 19, os URL's contêm informações privadas e por vezes até com informações de identificação ou confidenciais relacionadas com o utilizador. Portanto, deixar passar estas informações entre pedidos de entidades terceiras pode comprometer a privacidade dos utilizadores num sítio Web.

Figura 19 - Exemplo de URL´s com informações privadas

Fonte: (Nalpas, 2020)

Portanto, das configurações mencionadas no capítulo 2 referente a esta classe, existem algumas que poderão não oferecer uma privacidade total ao utilizador. Para uma solicitação de origem cruzada (cross-origin), ou seja, para solicitações em os sítios Web acedem a conteúdos que estão hospedados noutros domínios, a configuração **no-*Referrer*-when-downgrade** partilha toda a informação contida no URL. Portanto, esta configuração nunca será uma boa opção para aumentar a privacidade de um utilizador. A configuração **strict-**

origin-when-cross-origin também partilha informação para outros domínios, mas neste caso só irá partilhar a origem do sítio Web. Se um sítio Web utiliza HTTPS, este não vai desejar que os URL´S contenham informações em solicitações não HTTPS, pois desta forma qualquer pessoa na rede poderá visualizá-las. Para tal, a configuração **origin-when-cross-origin,** não seria uma boa opção para aumentar a privacidade de um utilizador. Para finalizar, a única configuração que qualquer entidade terceira, localizada num sítio Web, deve evitar por completo é a utilização da configuração **unsafe-url,** visto que envia toda a informação possível no URL ao efetuar um qualquer pedido, independentemente da segurança da camada de transporte (Nalpas, 2020).

Realça-se que algumas configurações não foram mencionadas, nomeadamente as configurações **Origin, Same origin, strict-origin,** pois o objetivo é analisar a quantidade de informação que é passada para uma entidade terceira. Logo, estas configurações são as mais pertinentes de utilizar quando se quer analisar informação que ocorre dentro do mesmo domínio.

Classe cache-control

Da mesma forma que algumas configurações na classe *Referrer-Policy* podem ser usadas para roubar informação aos utilizadores, algumas configurações na classe cache-control podem ser estabelecidas, de igual forma, a entidades terceiras e ter também um impacto negativo na privacidade de um utilizador.

Problemas de privacidade

A utilização desta classe nos cabeçalhos pode ter diversas vantagens, da mesma forma que pode ser usado para violar a privacidade de um utilizador. Por exemplo, os cabeçalhos podem permitir que um conteúdo autenticado seja colocado em cache, o que leva a que as sessões possam ser partilhadas entre utilizadores que utilizam o mesmo servidor *proxy*. Um dos objetivos dos cabeçalhos de cache que estão devidamente configurados, é evitar ter informação personalizada guardada em *proxies*. O servidor necessita de incluir cabeçalhos apropriados para indicar se a resposta pode ser guardada em cache ou não.

Desta forma, uma página pode ser marcada com a configuração "**private**" ou "**public**". Para impedir que um intermediário guarde um recurso em cache, utiliza-se a configuração "**private** ". Por outro lado, existem configurações que são por vezes utilizadas de uma forma

inadequada. Por exemplo, a opção "**no-cache**" implica apenas que o *proxy* deve verificar cada vez que um recurso é solicitado se este ainda é válido, mas pode ainda assim armazenar o recurso. Uma melhor opção é a "**no-store**", que impedirá que o pedido e a resposta sejam armazenados pela cache. A opção "**no-transform**" evitará que *proxies* intermediários alterem o formato dos recursos, para assim não haver qualquer tentativa de incluir *script*s de rastreio ao utilizador nesses ficheiros alterados. A única configuração que qualquer entidade terceira, localizada num sítio Web, deve evitar é a "**public**", pois esta configuração permite que uma resposta possa ser colocada em cache e ser mostrada a um utilizador diferente. Isto torna-se realmente um problema quando um sítio Web, com serviços de autenticação, tem em sua posse dados privados dos utilizadores. Desta forma, esses dados correm o risco de serem exibidos. De uma forma geral, só se deve utilizar esta configuração para páginas estáticas ou páginas que devolvam os mesmos dados.

5.3. Localstorage nos sítios Web

Como já foi mencionado no estado da arte, o uso de mecanismos de localstorage pode levar ao rastreio de um utilizador. Portanto qualquer uso, por parte de um sítio Web, de um qualquer mecanismo deste tipo é um indicador que poderá estar a ser feito um rastreamento sem o consentimento do utilizador. Portanto irão ser destacados alguns problemas que, qualquer entidade terceira localizada num sítio Web, possa usufruir ao usar mecanismos de localstorage e colocar a privacidade de um utilizador em risco.

Problemas de privacidade

O localstorage partilha muitas das características de um *cookie*, incluindo os mesmos riscos de privacidade. O armazenamento de algo sensível, como uma palavra-passe num ficheiro de localstorage, simplifica o processo para um rastreador pois não precisa de carregar este ficheiro no seu próprio navegador, ao contrário da utilização de *cookies*.

Um outro fator que torna este mecanismo muito problemático é que com o localstorage, não há armazenamento por parte do servidor, ou seja, não existe uma base de dados sobre a qual um programador tenha controlo. Desta forma, os programadores não têm forma de atualizar o código ou a informação quando esta é armazenada. Teria de ser o utilizador a apagar o ficheiro com a informação manualmente, o que exigiria encontrá-la, o que por vezes é complicado (Tal, 2020).

5.4. Protocolos de segurança e privacidade de um sítio Web

Para tornar a navegação de um utilizador num sítio Web mais segura e com uma maior privacidade, foi criado a Transport Layer Security (TLS), o sucessor da Secure Sockets Layer (SSL). TLS é um protocolo de segurança concebido para melhorar a segurança dos dados para as comunicações através da Internet. Quando um servidor e um cliente comunicam, um TLS bem configurado assegura que nenhuma entidade terceira possa alterar qualquer mensagem (Centre, 2021).

Serviços de entidades terceiras

O que se pretende analisar neste atributo, em relação às entidades terceiras pertencentes a um determinado sítio Web, é a forma como o transporte de conteúdos ocorre entre estas entidades terceiras e o próprio sítio Web. O objetivo não é avaliar se a configuração do TLS num sítio Web está implementada da forma mais correta, mas sim observar se a segurança na camada de transporte da informação se mantém o mesmo. Por exemplo, quando um navegador de um sítio Web invoca recursos, em que a camada de transporte sofre uma alteração, ou seja, passar de HTTPS para HTTP. Tal alteração é um indicador de que essa entidade terceira não pode ser confiável pois poderá ter intenções de violar a privacidade de um utilizador. Portanto, o objetivo é avaliar se existe a invocação de recursos por HTTP.

6. Classificação do quadro de avaliação

Um indicador de qualidade global, de alto nível, proporciona uma forma fácil de classificar a qualidade na forma de como as entidades terceiras de um sítio Web estão implementadas. Para o cálculo deste indicador, foi então criado um grau de violação da privacidade que os atributos anteriormente mencionados podem ter na navegação de um utilizador num sítio Web (Tabela 4).

Legenda	Muito Grave	Grave	Bom	Muito Bom

Tabela 4 - Legenda para a classificação

Fonte: Elaboração própria

De seguida e tendo em conta este grau de violação da privacidade, foi então atribuída uma cor a cada configuração no respetivo atributo de forma a entender-se como as configurações podem prejudicar ou não um certo utilizador. É possível observar-se esta atribuição nas seguintes tabelas.

Samesite	Nivel de privacidade
None	Muito Grave
Lax	Grave
Strict	Bom
Não utiliza	Muito Bom

Tabela 5 - Atributo samesite

Fonte: Elaboração própria

Referrer-Policy	Nivel de privacidade
Unsafe-url	Muito Grave
No-referrer-when-downgrade	Grave
Origin-when-cross-origin	Grave
Strict-origin-when-cross-origin	Bom
No-referrer	Muito Bom

Tabela 6 - Atributo Referrer-Policy

Fonte: Elaboração própria

Cache-Control	Nivel de privacidade
Public	Muito Grave
No-cache	Grave
No-Transform	Bom
No-Store	Muito Bom
Private	Muito Bom

Tabela 7 - Atributo cache-control

Fonte: Elaboração própria

Por último, foram criados os seguintes casos, tendo em conta o grau de violação da privacidade, que um atributo pode ter tendo em conta a forma de como está implementado num sítio Web:

- **Muito grave:** Um sítio Web recorre a entidades terceiras que utilizam a classe samesite = **none** ou a classe *Referrer-Policy* = **unsafe-url** ou a classe cache-control = **public** ou que explore mecanismos de localstorage ou que haja invocação de recursos por HTTP.
- **Grave:** Um sítio Web recorre a entidades terceiras que utilizam a classe samesite = **lax** ou a classe *Referrer-Policy* = **no-*Referrer*-when-downgrade/ origin-when-cross-origin** ou a classe cache-control = **no-cache** ou que explore mecanismos de localstorage ou que haja invocação de recursos por HTTP.
- **Bom:** Um sítio Web recorre a entidades terceiras que utilizam a classe samesite = **strict** ou a classe *Referrer-Policy* = **strict-origin-when-cross-origin** ou a classe

cache-control = **no-transform** ou que não explore mecanismos de localstorage ou que não haja invocação de recursos por HTTP.

- **Muito Bom:** Um sítio Web recorre a entidades terceiras que não estabeleçam *cookies*, ou que utilizam a classe *Referrer-Policy* = **no-*Referrer*** ou a classe cache-control = **private/no-store** ou que não explore mecanismos de localstorage ou que não haja invocação de recursos por HTTP.

Encontram-se apresentados, na Tabela 8, os parâmetros anteriormente mencionados para realizar a classificação. Tendo em conta esta métrica, a forma utilizada para chegar a uma classificação final de um sítio Web foi fazer a contagem da frequência com que aparece cada configuração e correspondê-la ao nível de privacidade a que pertence. No final fazer a contagem da cor predominante e atribuir o nível de privacidade ao sítio Web. Por exemplo, um sítio Web que apresente três configurações que são classificadas como "Muito Grave", o sítio Web será classificado como "Muito Grave".

	Nivel de Privacidade			
Atributos	**Muito Grave**	**Grave**	**Bom**	**Muito Bom**
SameSite	none	lax	strict	Não Utiliza
Referrer-Policy	unsafe-url	no-referrer-when-downgrade / origin-when-cross-origin	strict-origin-when-cross-origin	no-referrer
Cache-control	public	no-cache	no-transform	private/no-store
Local Storage	Utiliza	Utiliza	Não Utiliza	Não Utiliza
TLS	Invoca recursos	Invoca recursos	Não invoca recursos	Não invoca recursos

Tabela 8 - Parâmetros para realizar a classificação.

Fonte: Elaboração própria

6.1. Resultados obtidos

Tendo em conta a classificação anteriormente realizada, foram definidos alguns sítios Web da realidade portuguesa que possam ser analisados e desta forma, testar a classificação. Como foi mencionado na secção 4, a ferramenta utilizada para a análise dos diversos sítios Web foi a *WebXray*, onde foi passado o URL do sítio Web que se pretende analisar e posteriormente foi recolhida toda esta informação que ficará guardada numa base de dados. Desta forma, da extensa informação que a ferramenta oferece, foi observada toda a informação relacionada com os *cookies* e a forma de como são estabelecidas e exploradas pelas entidades terceiras. De seguida, observou-se o tipo de *Referrer-Policy* que era estabelecido entre os pedidos por parte das entidades terceiras e do respetivo sítio Web e também o tipo de cache-control estabelecido entre os pedidos realizados pelas entidades terceiras e o sítio Web. Por último, foi também observado se as entidades terceiras exploravam mecanismos de localstorage e se nos pedidos entre estas entidades terceiras e o sítio Web havia invocação de recursos por HTTP.

Portanto, os sítios Web foram separados por categorias para assim ser mais fácil a compreensão para o leitor dos diversos resultados que se irão obter. As categorias dos sítios Web selecionadas foram de jornais, bancos, compras, governo, instituições universitárias e sítios Web relacionados com transportes e reservas *online*.

Categoria de jornais

Os jornais que serviram de base a esta classificação foram: Jornal de notícias, Record, Abola, Averdade, Sapo, Cmjornal, diário de notícias, observador, Jornal de Negócios e Diário de Aveiro.

URL	Samesite	Referrer-Policy	Cache-control	Local Storage	TLS	Nivel de Privacidade
https://www.jn.pt/	none	No-referrer-when-downgrade	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.record.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.abola.pt/	none	Strict-origin-when-cross-origin	public	Utiliza	Não invoca recursos	**Muito Grave**
https://averdade.com/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.sapo.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.cmjornal.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.dn.pt/	none	No-referrer-when-downgrade	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://observador.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.jornaldenegocios.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.diarioaveiro.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**

Tabela 9 - Classificação para categoria de jornais

Fonte: Elaboração própria

Categoria de compras

Os sitos Web, onde se pode realizar diversos tipos de compras, que serviram de base a esta classificação foram: Olx, Continente, dott, Wook, Prozis, Farfetch, goldpet, Pcdiga, StandVirtual e KuantoKusta.

URL	Samesite	Referrer-Policy	Cache-control	Local Storage	TLS	Nivel de Privacidade
https://www.olx.pt/	none	Unsafe-url	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.continente.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://dott.pt/pt	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.wook.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.prozis.com/pt/pt	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.farfetch.com/pt	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://goldpet.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.pcdiga.com/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.standvirtual.com/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.kuantokusta.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**

Tabela 10 - Classificação para categoria de compras

Fonte: Elaboração própria

Categoria de bancos

Os sitos Web de instituições bancárias que serviram de base a esta classificação foram: Crédito Agrícola, Banco BPI, Novobanco, CaixaGeralDepositos, MillenniumBCP, Santander, Banco Montepio e Eurobic.

URL	Samesite	Referrer-Policy	Cache-control	Local Storage	TLS	Nivel de Privacidade
https://www.creditoagricola.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.bancobpi.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.novobanco.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.cgd.pt/Particulares/Pages/Particulares_v2.aspx	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://ind.millenniumbcp.pt/pt/particulares/Pages/Welcome.aspx	Não Utiliza	Não Utiliza	Não Utiliza	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.santander.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.bancomontepio.pt/particulares	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.eurobic.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.bpg.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.bancoctt.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**

Tabela 11 - Classificação para categoria de bancos

Fonte: Elaboração própria

Categoria de governo

Os sitos Web de instituições relacionadas com o governo que serviram de base a esta classificação foram: Serviços públicos (eportugal), Polícia de segurança publica (psp), Cartão jovem, Direção geral da educação (dge), Parlamento, Câmaras Municipais de Lisboa, Porto e Coimbra, Exército Português e Presidência.

URL	Samesite	Referrer-Policy	Cache-control	Local Storage	TLS	Nivel de Privacidade
https://eportugal.gov.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Bom**
https://www.psp.pt/Pages/homePage.aspx	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.cartaojovem.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Bom**
https://www.dge.mec.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.parlamento.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.lisboa.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.cm-porto.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.exercito.pt/pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.presidencia.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.cm-coimbra.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**

Tabela 12 - Classificação para categoria de governo

Fonte: Elaboração própria

Categoria de Universidades

Os sitos Web de instituições universitárias que serviram de base a esta classificação foram: ipb (Instituto Politécnico de Bragança), ipv (Instituto Politécnico de Viseu), upt (Universidade Portucalense), uc (Universidade de Coimbra.), uminho (Universidade do Minho), ulisboa (Universidade de Lisboa), feup (Faculdade de Engenharia da Universidade do Porto), isep (Instituto Superior de Engenharia do Porto), ua (Universidade de Aveiro), ulusiada (Universidade Lusíada).

URL	Samesite	Referrer-Policy	Cache-control	Local Storage	TLS	Nivel de Privacidade
https://portal3.ipb.pt/index.php/pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.ipv.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.upt.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.uc.pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.uminho.pt/PT	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.ulisboa.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Bom**
https://sigarra.up.pt/feup/pt/web_page.inicial	Não Utiliza	Não existe pedidos	Não existe pedidos	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.isep.ipp.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.ua.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.por.ulusiada.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**

Tabela 13 - Classificação para categoria de universidades

Fonte: Elaboração própria

Categoria de Transportes e sitios Web de reservas

Os sitos Web de instituições relacionadas com transporte e de reservas online que serviram de base a esta classificação foram: cp, rede-expresso, flytap, easyJet, metroporto, britishairways, swiss, airbnb, edreams e trivago

URL	Samesite	Referrer-Policy	Cache-control	Local Storage	TLS	Nivel de Privacidade
cp.pt/passageiros/pt	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Bom**
https://rede-expressos.pt/pt	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Bom**
https://www.flytap.com/pt-pt/	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.easyjet.com/pt	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.metrodoporto.pt/	none	Origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Grave**
https://www.britishairways.com/travel/home/public/pt_pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.swiss.com/pt/pt/homepage	none	Strict-origin-when-cross-origin	no-cache	Não Utiliza	Não invoca recursos	**Muito Bom**
https://www.airbnb.pt/	Não Utiliza	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Bom**
https://www.edreams.pt/	none	Strict-origin-when-cross-origin	no-cache	Utiliza	Não invoca recursos	**Muito Grave**
https://www.trivago.pt/	Utiliza	No-referrer-when-downgrade	no-cache	Não Utiliza	Não invoca recursos	**Grave**

Tabela 14 - Classificação para categoria de transportes e reservas

Fonte: Elaboração própria

Para se observar de uma forma geral o nível de privacidade dos sítios Web, apresenta-se na Figura 20 as percentagens dos respetivos níveis de privacidade consoante a categoria a que foram atribuídos.

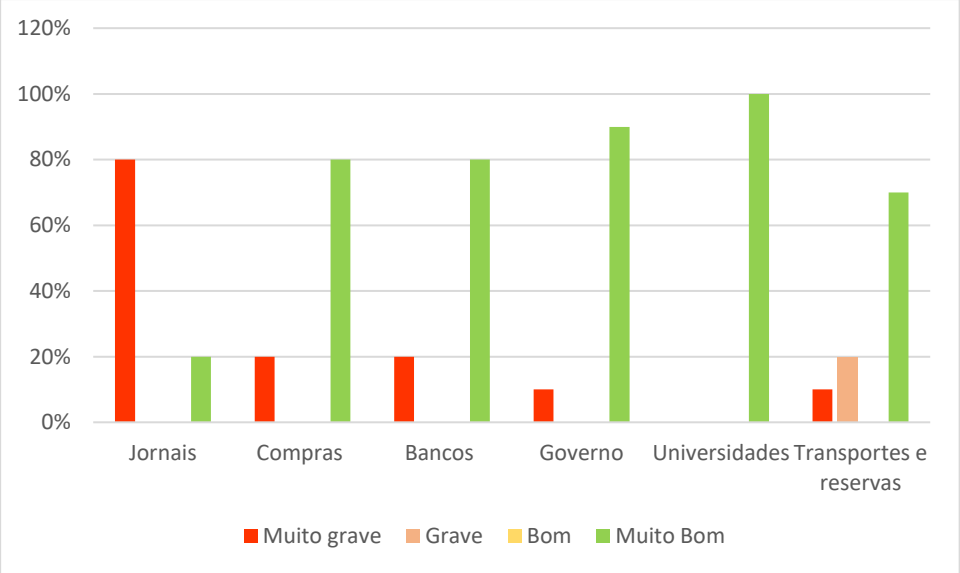


Figura 20 - Níveis de privacidade consoante a categoria

Fonte: Elaboração própria

6.2. Discussão dos resultados

Realizada então a recolha dos diversos dados que permitiu classificar os sítios Web, pode-se observar algumas considerações. Antes de se abordar os diversos resultados obtidos, realça-se que ao longo da classificação houve casos, apesar de poucos, em que a classificação dava um empate relacionado com a atribuição no nível de privacidade dos atributos. Por exemplo, havia sítios Web onde dois atributos eram classificados como "Muito Grave" e dois atributos que eram classificados como "Grave". A forma para resolver estes empates foi considerar sempre o pior caso, ou seja, observando o exemplo mencionado, o sítio Web foi classificado como "Muito grave" pois era o pior caso em termos de violação da privacidade para um utilizador.

Desta forma, ao visualizar a Figura 20, comprovou-se que a categoria de jornal viola de uma forma muito grave a privacidade de um utilizador. Nesta categoria, observa-se que na maior parte dos casos as entidades terceiras podem estabelecer *cookies* de uma forma livre, ou seja, através da configuração "none". Desta forma, têm a possibilidade de utilizar toda informação contida nos *cookies* para os seus próprios benefícios, seja de uma forma positiva ou negativa. Observou-se também que se utiliza excessivamente a configuração "strict-origin-when-cross-origin" que, apesar de ser uma configuração utilizada pela maior parte dos sítios Web aqui apresentados, não oferece uma total privacidade ao utilizador, uma vez

que as entidades terceiras ficam desta forma a saber a origem do sítio Web que o utilizador visitou. A razão pela qual a grande maioria dos sítios Web utiliza esta configuração nos cabeçalhos *Referrer-Policy*, é devido à segurança que esta oferece pois não pode haver, entre os pedidos de um sítio Web e uma entidade terceira, troca de URL's em que o nível de segurança não seja o mesmo. Mas como o objetivo desta dissertação é analisar a privacidade que um atributo oferece, não se teve em conta o grau de segurança. Realço também a quantidade de entidades terceiras que têm a possibilidade de explorar mecanismo de localstorage nos sítios Web, onde desta forma têm a possibilidade de rastrear um utilizador sem o seu consentimento.

Comprovou-se também que a categoria relacionada com instituições universitárias oferece uma grande privacidade aos utilizadores. O que gerou particularmente uma certa surpresa foi na categoria de compras onde os sítios Web, em grande escala, respeitam a privacidade de um utilizador. Na minha opinião, estava convicto que estes sítios Web iriam ter uma classificação muito pior pois estão sujeitos a uma quantidade elevada de entidades terceiras e desta forma haver uma maior tentação em violar e roubar qualquer tipo de informação aos utilizadores.

7. Conclusão

Após ter sido executado este trabalho e, em jeito de conclusão, é importante retirar algumas conclusões.

Realça-se que avaliar a privacidade de algo, independentemente do conteúdo de avaliação, é complicado pois não existe um meio termo, ou seja, ou aquele conteúdo ou serviço oferece privacidade a um utilizador ou não oferece. Porque ou se revela informação acerca de um utilizador ou não se revela informação. Desta forma, este linear de raciocino vai ao encontro dos resultados que se obteve no capítulo anterior, pois como se pode observar todos os sítios Web obtiveram classificações na maior parte entre dois valores, mais especificamente "Muito Grave"" ou "Muito Bom". Houve poucas classificações que se situassem entre o "Grave" ou "Bom", onde estas classificações podem ter sido influenciadas na forma como foi atribuído o nível de privacidade aos atributos.

Observando a escolha dos atributos que fizeram parte desta classificação, pode-se fazer uma comparação com a literatura ou estado da arte aqui realizado. De realçar que na maior parte dos sítios Web, como foi também realçado na discussão dos resultados, as entidades terceiras exploram de uma forma abusiva os *cookies* para os seus próprios objetivos e vai ao encontro do que foi aqui estudado. Sendo a técnica *cookies* ou *cookies* de entidades terceiras, a principal forma que um sítio Web ou entidades que estejam ligados ao mesmo (como as entidades terceiras), pode implementar para retirar certas informações sobre os utilizadores que estejam a usufruir dos seus serviços ou fazer um rastreamento subliminar a estes. Destaca-se também a exploração, por parte das entidades terceiras em mecanismos de *localStorage*, concluindo que esta técnica é utilizada por diversos motivos, sejam esses motivos de uma forma não abusiva para o utilizador ou com a intenção de retirar algo mais ao utilizador, o que também foi mencionado no estado da arte. Uma outra técnica também mencionada no estado da arte e que foi avaliada por esta classificação foi a utilização de cache na Web e a forma como pode ser manipulada para retirar qualquer tipo de informação. A configuração que foi utilizada pela maior parte dos sítios Web aqui analisados, obriga a que seja o servidor a decidir se a informação proveniente de uma entidade terceira pode ser guardada na cache, o que oferece alguma preservação da informação. Porém, consoante a decisão do servidor, pode ainda assim guardar qualquer tipo de informação e desta forma ser utilizada para qualquer fim.

Concluindo desta forma a ligação feita pelo estado da arte e a análise feita na escolha dos atributos utilizados para esta classificação, destaca-se uma limitação observada tendo em conta também a literatura analisada nesta dissertação. Constatou-se que nenhuma plataforma ou ferramenta se foca em todos os mecanismos que foram analisados na revisão de literatura. Por exemplo, a plataforma WBF Analyzer foca-se única e exclusivamente na deteção de chamadas *fingerprinting* nas páginas Web, ou seja, não é o que se pretende apenas demonstrar nesta avaliação. Sendo assim, a limitação que se encontra é que a escolha dos atributos para esta classificação não abrangeu a maior parte das técnicas mencionadas. Um exemplo disso é a análise de chamadas *fingerprinting* e assim saber quais é que podem pôr em causa a privacidade de um utilizador. Portanto uma observação que pode ser feita para um trabalho futuro, será alterar a plataforma de forma a esta também possa detetar chamadas *fingerprinting* nos recursos. Um outro trabalho futuro que pode ser desenvolvido é também ter a possibilidade de alterar a plataforma para detetar quais os sítios Web que utilizam a técnica *Event Tagging* nos campos de formulário. Desta forma, ficava-se a saber quais os sítios Web que utilizam esta técnica, que pode ter um impacto significativo na privacidade de um utilizador e assim enriquecer a classificação que foi desenvolvida nesta dissertação.

REFERÊNCIAS BIBLIOGRÁFICAS

Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., & Preneel, B. (2013). FPDetective: Dusting the Web for Fingerprinters. *Electronic Commerce — Security*, 13. https://doi.org/10.1145/2508859.2516674

Albeniz, Z. (2022). *Security Cookies Whitepaper | Invicti*. https://www.invicti.com/security-cookies-whitepaper/

Anon, D. (2018). *How cookies track you around the web & how to stop them | Privacy.net*. Privacy.Net. https://privacy.net/stop-cookies-tracking

AYENSON*, M. D., WAMBACH†, D. J., SOLTANI‡, A., & HOOFNAGLE**†, N. G. & C. J. (2009). FLASH COOKIES AND PRIVACY II: NOW WITH HTML5 AND ETAG RESPAWNING. *World Wide Web Internet And Web Information Systems*.

Bujlow, T., Carela-Espanol, V., Lee, B. R., & Barlet-Ros, P. (2017). A Survey on Web Tracking: Mechanisms, Implications, and Defenses. *Proceedings of the IEEE*, *105*(8), 1476–1510. https://doi.org/10.1109/JPROC.2016.2637878

Centre, N. C. S. (2021). *Using Transport Layer Security to protect data - NCSC.GOV.UK*. https://www.ncsc.gov.uk/guidance/using-tls-to-protect-data

Chow, R., Golle, P., Jakobsson, M., Wang, L., & Wang, X. (2008). Making CAPTCHAs Clickable. *Proceedings of the 9th Workshop on Mobile Computing Systems and Applications - HotMobile '08*. https://doi.org/10.1145/1411759

Cloudflare. (2022). *What is HTTP? | Cloudflare*. Medium. https://www.cloudflare.com/learning/ddos/glossary/hypertext-transfer-protocol-http/

CookiePro. (2020). *Website Tracking: How Websites Track You - Blog - CookiePro*. https://www.cookiepro.com/blog/website-tracking/

CookieYes. (2022, May 2). *Website Tracking: How and Why Websites Track Users - CookieYes*. https://www.cookieyes.com/blog/website-tracking/

Crawford, E. (2020). *Website Tracking: How Websites Track You - Blog - CookiePro*. https://www.cookiepro.com/blog/website-tracking/

Datalogix. (2022). *Privacy Policy – Datalogix | IT Solution Specialists*. Medium. https://datalogix.ie/privacy-*Policy*/

de Matos, G. F., & Feitosa, E. L. (2021). *WBF Analyzer: Um Método para Detecção de Browser Fingerprinting em Páginas Web*. *i*, 351–364. https://doi.org/10.5753/sbseg.2021.17327

DiGioia, R. (2014, March 25). *BlueCava launches Cross-Screen Audience Association*. https://www.prweb.com/releases/2014/03/prweb11701300.htm

DNS, C. (2022). *What is DNS? | How DNS works | Cloudflare*. Medium. https://www.cloudflare.com/learning/dns/what-is-dns/

Docs, M. W. (2020, December 8). *Navigator - APIs da Web | MDN*. https://developer.mozilla.org/pt-BR/docs/Web/API/Navigator

Docs, M. W. (2022). *Navigator - Web APIs | MDN*. https://developer.mozilla.org/en-US/docs/Web/API/Navigator

Docs, T. M. W. (2022). *How the web works - Learn web development | MDN*. https://developer.mozilla.org/en-US/docs/Learn/Getting_started_with_the_web/How_the_Web_works

Dussutour, C. (2020). *Public websites analysis | Joinup*. https://joinup.ec.europa.eu/collection/open-source-observatory-osor/news/public-websites-analysis

ENISA EU. (2012). Privacy considerations of online behavioural tracking. *European Network and Information Security Agency (ENISA)*, 1–33. www.enisa.europa.eu

GeeksforGeeks. (2021, October 12). *HTTP headers | Referrer-Policy - GeeksforGeeks*. Medium. https://www.geeksforgeeks.org/http-headers-*Referrer-Policy*/

Ghostwords, A. (2016, September 27). *GitHub - ghostwords/chameleon-crawler: Browser automation for Chameleon. - Pesquisa Google*. https://www.google.com/search?q=GitHub+-+ghostwords%2Fchameleon-crawler%3A+Browser+automation+for+Chameleon.&rlz=1C1FCXM_pt-PTPT970PT970&oq=GitHub+-+ghostwords%2Fchameleon-crawler%3A+Browser+automation+for+Chameleon.&aqs=chrome..69i57j69i64.301j0j7&source

Gibb, R. (2016). *What is a Web Cache?* https://blog.stackpath.com/web-cache/

Haq, M. R. ul. (2022). *What is Session Tracking?* https://www.educative.io/edpresso/what-is-session-tracking

HARNISH, B. (2021). *O que é HTTPS: guia definitivo de como funciona o HTTPS*. Medium. https://pt.semrush.com/blog/o-que-e-https/

HELME, S. (2017). *A new security header: Referrer Policy*. https://scotthelme.co.uk/a-new-security-header-*Referrer-Policy*/

Imperva. (2022). *Cache Control*. Medium. https://www.imperva.com/learn/performance/cache-control

Kamkar, S. (2010, October 11). *Evercookie*. https://samy.pl/evercookie/

Krawczyk, P. (2022). *SameSite | OWASP Foundation*. https://owasp.org/www-community/SameSite

Krishnamurthy, B., Naryshkin, K., & Wills, C. E. (2011). *Privacy leakage vs. Protection measures: the growing disconnect*. www.alexa.com

Lawson, B., & Sharp, R. (2011). *HTML INTRODUCING 5 BRUCE LAWSON REMY SHARP*. www.robertnyman.com

Li, T. C., Hang, H., Faloutsos, M., & Efstathopoulos, P. (2015). TrackAdvisor: Taking Back Browsing Privacy from Third-Party Trackers. *Undefined*, *8995*, 277–289. https://doi.org/10.1007/978-3-319-15509-8_21

Libert, T. (2015). *(PDF) Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites*. https://www.researchgate.net/publication/283471153_Exposing_the_Hidden_Web_An_Analysis_of_Third-Party_HTTP_Requests_on_1_Million_Websites

Libert, T. (2022). *webXray*. https://webxray.org/

Maass, M., Wichmann, P., Pridöhl, H., & Herrmann, D. (2017). PrivacyScore: Improving privacy and security via crowd-sourced benchmarks of websites. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10518 LNCS*(September 2016), 178–191. https://doi.org/10.1007/978-3-319-67280-9_10

Matos, G. F. de. (2021). *Dissertacao_GeandroMatos_PPGI.pdf* [Universidade Federal do Amazonas]. https://doi.org/https://tede.ufam.edu.br/bitstream/tede/8442/6/Dissertacao_GeandroMatos_PPGI.pdf

Mayer, J. R., & Mitchell, J. C. (2012). Third-party web tracking: *Policy* and technology. *Proceedings - IEEE Symposium on Security and Privacy*, 413–427. https://doi.org/10.1109/SP.2012.47

MDN. (2021, June 4). *Set-Cookie - HTTP | MDN*. https://developer.mozilla.org/pt-BR/docs/Web/HTTP/Headers/Set-Cookie

Mowery, K., & Shacham, H. (2012). *Pixel Perfect: Fingerprinting Canvas in HTML5*.

http://www.joelonsoftware.com/items/

Nalpas, M. (2020). *Melhores práticas no uso dos cabeçalhos Referrer e Referrer-Policy*. https://web.dev/i18n/pt/*Referrer*-best-practices/

Narayanan, A. (2010). *How Google Docs Leaks Your Identity | 33 Bits of Entropy*. https://33bits.wordpress.com/2010/02/22/google-docs-leaks-identity/

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings - IEEE Symposium on Security and Privacy*, 111–125. https://doi.org/10.1109/SP.2008.33

Network, M. (2020, December 8). *Window.navigator - APIs da Web | MDN*. https://developer.mozilla.org/pt-BR/docs/Web/API/Window/navigator

Network, M. (2022a, March 28). *Cookies HTTP - HTTP | MDN*. MDN Web Docs. https://developer.mozilla.org/pt-BR/docs/Web/HTTP/Cookies

Network, M. (2022b, May 24). *Same-origin Policy - Web security | MDN*. MDN Web Docs. https://developer.mozilla.org/en-US/docs/Web/Security/Same-origin_*Policy*

Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., & Vigna, G. (2013). Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. *Proceedings - IEEE Symposium on Security and Privacy*, 541–555. https://doi.org/10.1109/SP.2013.43

Papadopoulos, P., Kourtellis, N., & Markatos, E. P. (2019). Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. *International World Wide Web Conference Committee*, 11. https://doi.org/10.1145/3308558.3313542

RingCentral. (2022). *What is a Web Application? Definition, Benefits and How it Works | RingCentral UK Blog*. Medium. https://www.ringcentral.co.uk/gb/en/blog/definitions/web-application/

Roesner, F. (2014). *Security and Privacy from Untrusted Applications in Modern and Emerging Client Platforms - Pesquisa Google* [University of Washington]. https://www.google.com/search?q=Security+and+Privacy+from+Untrusted+Applications+in+Modern+and+Emerging+Client+Platforms&rlz=1C1FCXM_pt-PTPT970PT970&oq=Security+and+Privacy+from+Untrusted+Applications+in+Modern+and+Emerging+Client+Platforms&aqs=chrome..69

Roesner, F., Kohno, T., & Wetherall, D. (2012). *Detecting and Defending Against Third-Party Tracking on the Web*. http://tracker2.com/track?cookie_

Rowe, W. (2019). *Fingerprinting Explained: How It Works & How To Block It – BMC Software | Blogs*. https://www.bmc.com/blogs/how-to-block-fingerprinting/

Rydstedt, G. (2022). *Clickjacking | OWASP Foundation*. https://owasp.org/www-community/attacks/Clickjacking

Seiden, K. (2018). *Digital Debrief – Using Event Tagging for Form Field Tracking*. https://www.kristaseiden.com/using-event-tagging-for-form-field-tracking/

Selenium. (2022). *Selenium*. https://www.selenium.dev/

Senol, A., Acar, G., Humbert, M., & Borgesius, F. Z. (2022). *Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission*.

Sikkeland, Ø. (2020). *Protecting User Privacy from Web Tracking Threats* [UNIVERSITY OF OSLO]. https://www.duo.uio.no/handle/10852/79549

Soltani, A., Canty, S., Mayo, Q., Thomas, L., & Hoofnagle, C. J. (2009). Flash Cookies and Privacy. *AAAI Spring Symposium - Technical Report*, *SS-10-05*, 158–163. https://doi.org/10.2139/SSRN.1446862

Sonzogni, A. (2022). *Improve security and privacy by updating HTTP Cache*. https://web.dev/http-cache-security/

SSL.com. (2021). *What is HTTPS? - SSL.com*. Medium. https://www.ssl.com/faqs/what-is-https/

Steven Englehardt, A. N. (2013). Online Tracking: A 1-million-site Measurement and Analysis. *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, 20. https://doi.org/https://doi.org/10.1145/2508859.2516674

Tal, L. (2020). *Is LocalStorage safe to use? | Snyk*. https://snyk.io/blog/is-localstorage-safe-to-use/

Unger, T., Mulazzani, M., Frühwirt, D., Huber, M., Schrittwieser, S., & Weippl, E. (2013). *SHPF: Enhancing HTTP(S) Session Security with Browser Fingerprinting (extended preprint)*. https://www.caniuse.com

W3C. (2017). *Referrer Policy*. https://www.w3.org/TR/*Referrer-Policy*/

W3schools. (2022). *JavaScript Window Screen*. https://www.w3schools.com/js/js_window_screen.asp

Web Storage API - Web APIs | MDN. (2022, June 21). https://developer.mozilla.org/en-US/docs/Web/API/Web_Storage_API

Wiki, R. (2021). *Session Tracking - Ryte Wiki - The Digital Marketing Wiki*. Medium.

https://en.ryte.com/wiki/Session_Tracking

Zviran, M. (2008). User's perspectives on privacy in web-based applications. *Journal of Computer Information Systems*, *48*(4), 97–105. https://doi.org/10.1080/08874417.2008.11646039