UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

SISTEMAS DE INFORMAÇÃO

Matheus Gurjão Heliodoro

Mensurando Cookies e Privacidade Web em um mundo pós-LGPD

Recife

2022

Matheus Gurjão Heliodoro

Mensurando Cookies e Privacidade Web em um mundo pós-LGPD

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Vinicius Cardoso Garcia

Recife

2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Heliodoro, Matheus Gurjão.
Mensurando cookies e privacidade web em um mundo pós-LGPD / Matheus Gurjão Heliodoro. - Recife, 2022.
35 : il.

Orientador(a): Vinicius Cardoso Garcia
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2022.

1. Privacidade. 2. Web. 3. LGPD. 4. Cookies. 5. Rastreamento. I. Garcia, Vinicius Cardoso. (Orientação). II. Título.

000 CDD (22.ed.)

MATHEUS GURJÃO HELIODORO

Mensurando Cookies e Privacidade Web em um mundo pós-LGPD

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Recife, 17 de Outubro de 2022

BANCA EXAMINADORA

_____

Prof. Vinicius Cardoso Garcia(Orientador)

UNIVERSIDADE FEDERAL DE PERNAMBUCO

_____

Profa. Jéssyka Vilela (2º membro da banca)

UNIVERSIDADE FEDERAL DE PERNAMBUCO

Resumo

Quando trafegamos na internet, muitos sites realizam o registro da nossa visita com *cookies* e relacionam as nossas atividades com ele. Com a chegada de leis de proteção de dados, como LGPD e GDPR, duas características importantes são esclarecer o propósito para coleta de dados pessoais e obter o consentimento do usuário. A proposta principal deste trabalho é percorrer os sites mais visitados do ranking da Alexa e verificar quais configuram um *cookie* com o propósito de registro de atividade sem o consentimento do usuário. A principal característica de cookies de registro de atividade é sua persistência entre visitas, sem nenhuma mediação do usuário, o que infringe leis de proteção de dados. Realizamos uma comparação entre a frequência que os sites realizam em regiões com diferentes legislações para descobrir o motivo dos sites se comportarem de diferentes formas com o uso de *cookies*. Os resultados foram que sites com maior tráfego tiveram um aumento no número de cookies persistentes e sites com menor tráfego tiveram uma diminuição. Devido a razões como a recente vigência da LGPD, falta de conhecimento da lei por parte de domínios nacionais, interpretações abertas sobre o conceito de dados pessoais e alternativas ao uso de cookies para rastreamento.

Palavras-chave: Privacidade, Registro de atividade, Cookies, LGPD, GDPR, Rastreamento.

Abstract

When we use the web, a lot of websites register our visit with cookies and relate our activity with it. With the advent of data protection laws, such as LGPD and GDPR, two important characteristics are purpose and consent for collecting the data from the user. The main purpose of this work is to navigate the most visited websites from the Alexa ranking and verify which of them set a cookie with the purpose of registering user activity without their consent. The main property of activity register cookies is their persistente between visits without any user mediation, which violates data protection laws. And a check for the percentage of sites that configure a cookie for user activity changed from regions with different legislations to discover why sites behave so differently when it comes to cookie usage. And the results were sites with bigger traffic raised persistent cookie usage while sites with less traffic reduced persistent cookie usage. Because of reasons like the recent validity of LGPD, lack of knowledge about the new Brazilian law from national domains, open interpretation as to what is personal data and alternatives to tracking.

Keywords: Privacy, Activity record, Cookies, LGPD, GDPR, Tracking.

Lista de Ilustrações

Lista de Abreviaturas e Siglas

LGPD	Lei geral de proteção de dados
GDPR	General Data Protection Regulation
HTTP	HyperText Transfer Protocol
CCPA	California Consumer Privacy Act

Sumário

1 INTRODUÇÃO

Ao acessarmos um site nos deparamos com a mensagem “Este site utiliza *cookies”*, após ler os termos de serviço descobrimos que o site coleta *cookies,* que são registros de dados criados enquanto o usuário utiliza um website, para melhorar nossa experiência e que as informações são compartilhadas com terceiros. Embora poucos sites nos permitam escolher exatamente quais *cookies* habilitar(como cookies essencial ou de terceiros), a maioria deles dá opções como “aceitar” ou “rejeitar” e outros simplesmente subtendem que você aceita todos os *cookies* por utilizar o site (Imagem 1).

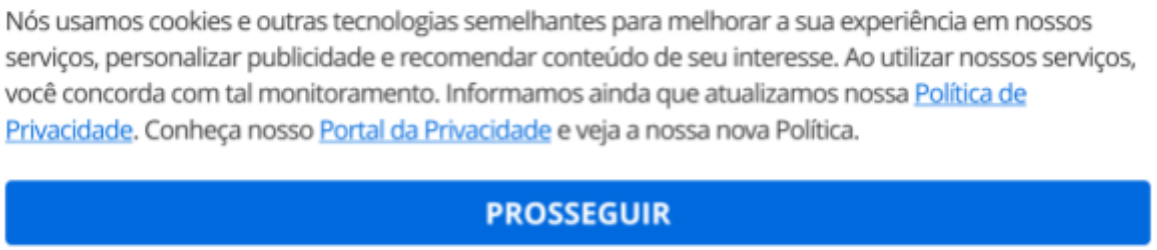

Imagem 1 - popup de cookies no site globo.com
(fonte: autor)

Com a chegada da LGPD (Lei Geral de Proteção de Dados Pessoais) no Brasil, algo como os exemplos acima estariam infringindo princípios de finalidade e transparência. E é perceptível que a indústria multibilionária do marketing tenha predisposição a infringir de alguma forma na privacidade do usuário para obter melhores resultados em sua publicidade (Imagem 2). Pela LGPD dados coletados sem identificação não precisam passar pelo tratamento da lei, o que dá uma margem para sites coletarem dados assim que o site é acessado.

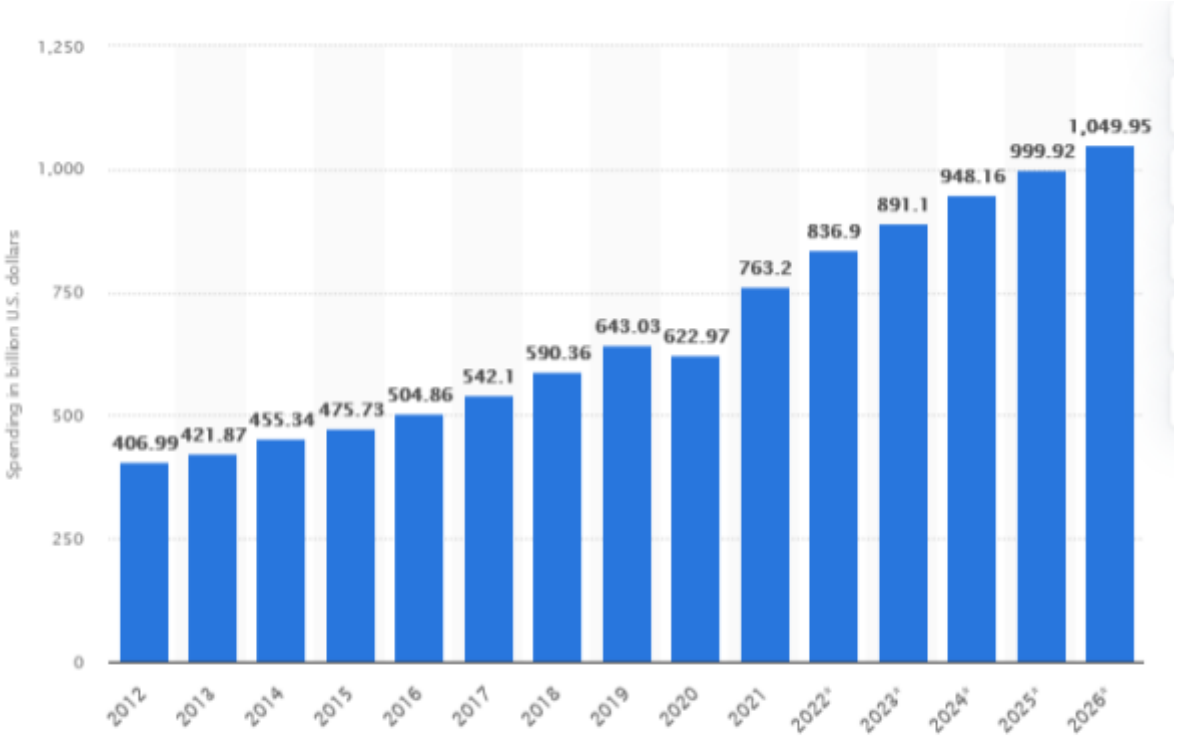

Imagem 2 - Lucro de mídias de publicidade em 2012 a 2026[1] (em bilhões de dólares)
(A. GUTTMANN, 2021)

[1]

Com o intuito de mensurar o impacto que a LGPD trouxe para o ecossistema web, navegamos pelos sites mais visitados no ranking da Alexa de 2018 e verificamos quais deles coletam dados a partir do momento que o site é visitado através de *cookies*. Para efeitos comparativos iremos utilizar estudos anteriores que também buscaram o uso de cookies persistentes na GDPR (General Data Protection Regulation) e o ranking da Alexa de 2018 como base (escolhido devido a quantidade de pesquisas já feitas, pela sua reputação e para efeitos comparativos com o estudo realizado sobre GDPR), isso pois as leis são análogas entre si.

Em 2019, a quantidade de pesquisas e estudos realizados envolvendo tanto a LGPD quanto a sua aplicabilidade em sistemas de informação era baixa. Por conta disso, é possível deduzir que a partir disso existe uma alta demanda por estudos abrangentes dessa temática, especialmente relacionados aos impactos trazidos na privacidade da web com cookies. (TEPEDINO, 2020). Desde 2022 houveram mais estudos com a temática, entretanto, relacionados a Cookies ainda são poucos.

[1] De 2022 em diante, foram realizadas previsões.

Como objetivo principal deste trabalho temos:

- Comparar uso de cookies persistentes no ambiente web em regiões com e sem vigência da LGPD.

Já para objetivos secundários temos:

- Verificar se organizações buscaram conformidade com a LGPD da mesma forma que com a GDPR;
- Verificar se a mesma conformidade também aconteceu em domínios nacionais;
- Analisar se a LGPD trouxe algum impacto para privacidade na web.

A organização do trabalho se dá pela seguinte forma:

1. Introdução
2. Referencial teórico
3. Metodologia
4. Resultados
5. Conclusão
6. Trabalhos Futuros
7. Referências bibliográficas

Neste capítulo abordamos o contexto, motivação, apresentação do problema, justificativa e a proposta do trabalho, além dos objetivos e organização do documento. No próximo capítulo iremos apresentar o referencial teórico e explicar como algumas tecnologias como *cookies* funcionam e o que as leis de proteção de dados dizem sobre essa forma de rastreio de usuários.

2 REFERENCIAL TEÓRICO

Neste capítulo iremos esclarecer alguns conceitos introduzidos no trabalho como *cookies* e abordar um pouco sobre as leis de proteção de dados brasileira e europeia. A CCPA (California Consumer Privacy Act) ficou fora do escopo pois embora também seja uma lei de proteção de dados, o intuito principal era a comparação entre a LGPD e a GDPR, mas também iremos comentar sobre ela.

2.1 WEB COOKIES

Web cookies são pequenos registros de dados criados pelo servidor enquanto o usuário utiliza um website e podem servir múltiplos propósitos. Os mais comuns são gerenciamento de sessão (logins, carrinho de compras, qualquer coisa que o servidor precise lembrar), personalização (configurações de usuário) ou tracking (gravar e analisar comportamento de usuários) (MOZILLA, 2021).

Após sua criação, cookies são enviados de volta para o servidor manter o registro e rastreio de atividades por parte do usuário, por conta disso, é aconselhável que cookies limitem a quantidade de dados, uma vez que eles são enviados em todas as requisições HTTP. Depois de receber a requisição inicial do cliente, cookies podem ser configurados pelo servidor a partir do parâmetro “set-cookies”. No lado do cliente o cookie mantém diversos parâmetros, mas os principais são o seu nome e valor (MOZILLA, 2021).

Cookies com o intuito de coletar dados sobre o usuário precisam persistir entre visitas. Cookies persistentes são emitidos automaticamente após a primeira visita ao website, e tem um tempo de expiração estendido, sendo usados para conectar visitas do mesmo usuário em um longo período. Cookies de sessão por outro lado duram um período muito mais curto e são perdidos ao fechar o browser (DABROWSKI; MERZDOVNIK; ULLRICH; SENDERA; WEIPPL, 2021).

O foco deste trabalho é web cookies, uma vez que ela é uma das principais maneiras de obter atividade de usuários. Outra forma de obter dados de usuários na Web é através de *fingerprinting*, ou seja, obtendo metadados como geolocalização, sistema operacional,

resolução de tela, etc. mas não há uma documentação concisa disso uma vez que é muito difícil encontrar evidências de *fingerprinting* com consistência (KRETSCHMER; PENNEKAMP; WEHRLE, 2021).

Outro método de obter atividade de usuários mas que ainda não teve adoção foi o Topics, desenvolvido pela Google e que utiliza o mecanismo de autenticação do browser como forma de manter o registro de atividade através de uma API. A tecnologia será implementada nos sites que desejarem e os anúncios podem ser feitos a partir de proximidade com 150 tópicos diferentes. O plano é implementar a API no navegador principal da Google, o Chrome (GOEL, 2022).

2.2 GDPR

A GDPR (*General Data Protection Regulation*)[2] é uma lei de segurança e privacidade aplicada a qualquer organização que precise coletar ou armazenar dados de indivíduos residentes na União Europeia. A regulamentação entrou em vigor em 25 de maio de 2018 e impõe multas que podem chegar a dezenas de milhões em euros para qualquer empresa que viole suas normas.

A GDPR foi o passo inicial da Europa para regulamentar o tratamento de dados sensíveis de indivíduos por parte das organizações. Os princípios da GDPR são:

1. **Lawfulness, fairness and transparency**: processamento de dados deve ser dentro da lei, justo e transparente;
2. **Purpose limitation**: os dados devem ter um propósito claro e legítimo;
3. **Data minimization**: os dados coletados devem ser os mínimos possíveis para fins de cumprir com os propósitos;
4. **Accuracy**: os dados devem ser atualizados e precisos;
5. **Storage limitation**: os dados devem ser armazenados apenas pelo período necessário;
6. **Integrity and confidentiality:** o processamento deve ser feito de forma a garantir a confidencialidade, integridade e segurança;

[2] Disponível em https://gdpr-info.eu/

7. **Accountability**: o data controller (entidade responsável por decidir como e porque os dados serão processados, nesse caso a organização) é responsável por seguir com todos esses princípios (WOLFORD, 2021).

Os direitos do titular na GDPR são:

1. **Direito ao acesso**: o direito de o titular de obter do controlador a confirmação do processamento de seus dados e, além disso, solicitar o acesso a suas informações pessoais
2. **Direito à retificação** : o direito do titular de garantir que seus dados pessoais sejam precisos e atualizados conforme necessário.
3. **Direito à exclusão ou esquecimento**: o direito do titular de solicitar ao controlador a exclusão de seus dados pessoais.
4. **Direito à oposição e restrição de processamento**: o direito do titular de se opor ao processamento de seus dados e, até mesmo, restringi-los caso assim deseje.
5. **Direito à portabilidade de dados:** o direito do titular de obter suas informações em um formato estruturado e legível por máquina ou ter seus dados transferidos para outra organização se possível.
6. **Direito à informação:** o direito do titular de ser informado sobre como e por que seus dados pessoais estão sendo processados. Além disso, eles têm o direito de saber se os dados estão sendo compartilhados com terceiros. Isso pode ser resolvido por meio da identificação das bases legais adequadas para processar dados.
7. **Direito à notificação:** em caso de violação de dados, os titulares deverão ser informados no prazo de 72 horas após o conhecimento da violação.

(WOLFORD, 2021)

A GDPR é muito menos aberta a interpretações que a LGPD e tem um trecho onde explicitamente comenta sobre o uso de *cookies* em sites, enquanto que a LGPD não tem trechos explícitos sobre *cookies* o tratamento de dados pessoais um pouco mais aberto a interpretações.

2.3 LGPD

A LGPD (Lei Geral de Proteção de Dados Pessoais)[3] é uma lei de segurança e privacidade aplicada a qualquer organização que precise coletar ou armazenar dados de indivíduos residentes em território nacional brasileiro. A regulamentação foi aprovada em agosto de 2018, entrando em vigor em agosto de 2020 e assim como a GDPR ela impõe sanções jurídicas para organizações que violarem suas normas.

A LGPD é fortemente inspirada na GDPR, seus princípios são:

1. **Princípio da finalidade:** Assim como o *purpose limitation*, os dados devem ter uma finalidade clara e legítima, informada ao titular.
2. **Princípio da adequação:** O tratamento de dados deve ser adequado a sua finalidade.
3. **Princípio da necessidade:** O tratamento de dados deve ser limitado ao mínimo necessário para realização de suas atividades.
4. **Princípio do livre acesso:** Os dados devem ser livres para consulta de seus titulares.
5. **Princípio da qualidade dos dados:** Os dados devem ser atualizados e precisos.
6. **Princípio da transparência:** Os titulares devem ter explicação do tratamento que será realizado de seus dados.
7. **Princípio da segurança:** Medidas de segurança devem ser tomadas para preservar e proteger os dados dos titulares.
8. **Princípio da não discriminação:** Os dados não podem ser usados para fins discriminatórios ou ilícitos.
9. **Princípio da responsabilização e da prestação de contas:** O responsável por utilizar os dados deve comprovar que é capaz de cumprir os princípios acima (PESTANA, 2020).

Os direitos do titular na LGPD são:

1. Confirmação da existência de tratamento;
2. Acesso aos dados;
3. Correção de dados incompletos, inexatos ou desatualizados;
4. Anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com o disposto na LGPD;

[3] Disponível em http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

5. Portabilidade dos dados a outro fornecedor de serviço ou produto, mediante requisição expressa, de acordo com a regulamentação da Autoridade Nacional, observados os segredos comercial e industrial;
6. Eliminação dos dados pessoais tratados com o consentimento do(a) titular, exceto nas hipóteses previstas no art. 16 da Lei;
7. Informação das entidades públicas e privadas com as quais o Controlador realizou uso compartilhado de dados;
8. Informação sobre a possibilidade de não fornecer consentimento e sobre consequências da negativa;
9. Revogação do consentimento, nos termos do § 5.º do art. 8.º da Lei.

(PESTANA, 2020)

Um dos principais questionamentos que se tem com relação a LGPD é o que constitui um dado pessoal, já que o escopo da lei se dá a partir de dados pessoais, ou seja, dados que permitem identificar o indivíduo. Uma vez sendo mais aberta a interpretação, é comum que dê margem para sites não considerarem que dados coletados sejam pessoais.

2.3.1 Legislação antes da LGPD

- **Constituição federal brasileira de 1988**: os pontos principais de privacidade na constituição de 88 não consideravam meios eletrônicos, sendo invioláveis "a intimidade, a vida privada, a honra e a imagem das pessoas" e "o sigilo da correspondência e das comunicações telegráficas, de dados e das comunicações telefônicas"
- **Código de defesa do consumidor de 1993**: o código de defesa do consumidor já considera meios eletrônicos em bancos de dados, sendo eles "preservados, mantidos em sigilo e utilizados exclusivamente para os fins do atendimento"
- **Marco Civil da Internet de 2013**: a lei abrange tópicos como a neutralidade de rede e liberdade de expressão na internet, além de "autodeterminação, privacidade, confidencialidade e segurança das informações e dados pessoais prestados ou coletados" (ASSIS E MENDES, 2022).

Tanto a LGPD quanto o Marco Civil da Internet foram grandes avanços no que se trata de privacidade de dados pessoais de cidadãos do país, especialmente quando as únicas outras leis que tratavam de proteção de dados de indivíduos eram o código de defesa do consumidor e a constituição federal. Isso sem contar com outros tópicos abordados como a neutralidade de rede, algo que teve repercussões em 2013 com projetos de lei como SOPA e PIPA nos Estados Unidos, que tiravam a neutralidade de rede e que se tornaram extremamente polêmicos mas nunca avançaram e foram arquivados (ARAÚJO, 2014).

2.4 CCPA

A CCPA (California Consumer Privacy Act)[4] é uma lei de segurança e privacidade aplicada a qualquer organização que precise coletar ou armazenar dados de indivíduos residentes na Califórnia, estado dos EUA. Assim como as leis citadas anteriormente ela dá ao usuário direitos como:

- Direito de saber quais informações pessoais são coletadas e seu uso;
- Direito de requisitar deleção de informações pessoais;
- Direito de optar por não ter sua informação pessoal vendida;
- Direito de não ter discriminação por utilizar a CCPA (Califórnia, 2018).

É uma lei de proteção de dados um pouco mais rasa e provavelmente a lei menos rígida dentre as três. Como dados pessoais são uma parte importante deste trabalho, iremos ver qual é a definição da CCPA para dados pessoais, ou seja, aquilo que está no escopo de proteção pela lei. Segundo a CCPA:

> "dados pessoais são qualquer informação que identifica, se relaciona e pode estar conectada com você, incluindo nome, cpf, email, produtos comprados, histórico de browser, geolocalização, digitais e inferências e outras informações que possam criar um perfil sobre suas preferências e características" (Califórnia, 2018).

Ou seja, algo como informações obtidas através de *cookies* persistentes estariam no escopo da CCPA.

[4] Disponível em https://oag.ca.gov/privacy/ccpa

2.5 TRABALHOS SIMILARES

A principal inspiração para este trabalho foi o artigo "*Measuring Cookies and Web Privacy in a Post-GDPR World*" dos autores Dabrowski, Merzdovnik, Ullrich, Sendera e Weippl (2021), que também realiza um estudo sobre *cookies* persistentes só que no contexto da GDPR. As soluções propostas na conclusão do estudo para não violar a GDPR eram: "não utilizar cookies persistentes, usar cookies persistentes apenas ao obter o consentimento do usuário ou apenas não oferecer o serviço para a região europeia". Para realizar a comparação, o estudo realizou as visitas aos sites em uma região na europa e outra nos estados unidos, iremos utilizar os resultados das métricas das duas regiões para comparação neste estudo.

O trabalho utilizou filtros para remover sites de terceiros, também criamos um filtro a partir de uma revisão manual dos resultados para recriar ao máximo o trabalho original, os filtros estão no repositório do github e também são comentados na seção de metodologia. O artigo original comenta sobre o uso de filtros para *cookies* de terceiros mas não entra em detalhes dos filtros utilizados.

Outro trabalho relacionado foi o artigo "*An Empirical Study of Web Cookies*" de 2016, que busca analisar mudanças em cookies usados nos sites mais visitados da internet. Esse trabalho foi importante para nos permitir identificar os principais métodos usados por sites para registro de dados e atividade do usuário. O estudo também comenta o impacto de cookies de terceiros, que permitem identificar a atividade de usuários através de múltiplos sites (CAHN; ALFELD; BARFORD; MUTHUKRISHNAN, 2016).

Por fim, um estudo geral sobre o estado da arte de trabalhos que buscaram medir o impacto da legislação europeia na web após 3 anos de vigência. O estudo intitulado "*Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web*" coletou trabalhos com diversas propostas, desde análise de cookies, banners, consentimento, fator psicológico e termos de serviço (KRETSCHMER; PENNEKAMP; WEHRLE, 2021).

Neste capítulo foi realizada uma explanação dos conceitos do estudo, incluindo *cookies*, leis de proteção de dados e métodos de rastreamento de usuário na *web*. No próximo capítulo

iremos apresentar o método de coleta de dados e entrar em mais detalhes de como foi obtido os *cookies* em duas visitas a cada site no ranking da Alexa.

3 METODOLOGIA

A metodologia utilizada neste trabalho envolveu o uso de um script de automação Selenium que realiza visitas aos sites, verificando quais cookies estão presentes. É realizado o fechamento do navegador (aplicação) e sua reabertura utilizando o mesmo perfil, que verifica novamente cookies em uso e retorna quais cookies apareceram nas duas visitas. O navegador utilizado foi o *chromium*[5] em modo *headless* [6] e a geolocalização foi em Recife, Brasil, para melhor reproduzir os estudos no artigo da GDPR.

Executamos uma visita ao site e registramos valores diferentes dependendo se o site tem cookies entre visitas, se não tem cookies entre visitas ou se o site não respondeu. Descartamos os valores de sites que não deram resposta e comparamos as porcentagens de sites com cookies e sem cookies entre visitas. Essas visitas foram realizadas entre julho de 2021 e janeiro de 2022, foram obtidos dados dos primeiros 50 mil sites do ranking.

Para este estudo e para melhor comparação com os estudos realizados a partir da amostra de 2016 e 2018 (WOLFORD, 2021), foi utilizado o top 1.000.000 (um milhão) websites de 2018 (AMAZON, 2021)[7] uma vez que essa amostra foi a mesma do estudo anterior. Também foi usado um filtro adicional de sites mais visitados em território nacional, o que nos permite descobrir se entidades nacionais procuraram manter conformidade com a nova lei de privacidade. Vale salientar que o estudo de 2020 na GDPR foi realizado após dois anos de vigência da lei, enquanto que no Brasil houve um período de aviso prévio até a vigência, realizada em setembro de 2020.

Alguns filtros foram usados para remover cookies de terceiros como Google Analytics, e adicionar cookies de sessão específicos de visitantes, uma vez que não foi usado um netlog[8] para obter cookies diretamente da página, filtros de cookies específicos de plataforma também foram adicionados ao whitelist, além de palavras chave. O repositório com o script usado para

---

[5] Disponível em https://www.chromium.org
[6] Navegador sem interface gráfica, executado em linha de comando com o argumento “--headless”
[7] Disponível em http://s3.amazonaws.com/alexa-static/top-1m.csv.zip
[8] Ferramenta de monitoramento de rede, disponível em https://www.chromium.org/developers/design-documents/network-stack/netlog/

visitação dos sites está disponível em https://github.com/MatheusHeliodoro/CookiePrivacyLGPD.

O arquivo *python_test* contém o script selenium que realiza a visita nos sites, para o *chromium*, importamos o *selenium*, juntamente com a extensão do *webdriver* de options, que nos permite configurar o perfil do navegador e o *driverexception*, que vai nos auxiliar caso encontremos sites que não responderam a tempo ou que estejam fora do ar (Imagem 3).

```
1  from selenium import webdriver
2  from selenium.webdriver.chrome.options import Options
3  from selenium.common.exceptions import WebDriverException
```

Imagem 3 - importando selenium e extensões

Criamos um perfil "test" que irá armazenar os *cookies* nas visitas realizadas e configuramos ele com os argumentos de automação, headless, sem extensões de navegador, sem consultas antecipadas no DNS e sem uso de GPU uma vez que o navegador está no modo *headless* (Imagem 4).

```
41  #configurar perfil de usuario em uma pasta selenium
42  options = Options()
43  options.add_argument("user-data-dir=test")
44  options.add_argument("enable-automation")
45  options.add_argument("--headless")
46  options.add_argument("--no-sandbox")
47  options.add_argument("--disable-extensions")
48  options.add_argument("--dns-prefetch-disable")
49  options.add_argument("--disable-gpu")
```

Imagem 4 - Configurações adicionais ao perfil usado nos testes

Criamos um arquivo results.txt que vamos usar para guardar os resultados. Começamos pulando uma linha e inserindo o nome do site visitado, invocamos o driver com as opções configuradas e tentamos acessar o domínio, caso não seja possível (o driver retorne uma *exception*, geralmente pela demora no tempo de resposta no site) inserimos que a página está fora do ar e prosseguimos para a próxima (Imagem 5).

```
5   def get_cookies(domain):
6           results = open("results.txt", "a")
7           results.write("\n")
8           results.write(domain)
9           print(domain)
10          driver = webdriver.Chrome(options=options)
11          try:
12                  driver.get(domain)
13          except WebDriverException:
14                  results.write("page down")
15                  results.write("\n")
16                  driver.quit()
```

Imagem 5 - visita realizada ao site

Obtemos os cookies e inserimos eles em um dicionário com o nome do cookie como chave e o valor dele como valor no dicionário, fizemos isso duas vezes e comparamos os valores entre as duas visitas (imagem 6).

```
18          cookies_list1 = driver.get_cookies()
19          cookies_visita1 = {}
20          for cookie in cookies_list1:
21                  cookies_visita1[cookie['name']] = cookie['value']
22          driver.quit()
```

Imagem 6 - cookies obtidos

Os resultados em comum são colocados no results.txt e o driver é encerrado para ser reaberto na visita ao próximo site da lista. O fechamento do driver é realizado para simular alguém fechando a janela do navegador (Imagem 7).

```
36          cookies_persistentes = set( cookies_visita1.items() ) & set( cookies_visita2.items() )
37          results.write(str(cookies_persistentes))
38          results.write("\n")
39          driver.quit()
```

Imagem 7 - valor inserido no arquivo de resultados e driver encerrado

Já o "*data classifier*" obtém o resultado do script selenium e classifica se o site tem ou não cookies sendo usados, "*national percentage calculator*" calcula a porcentagem para domínios

nacionais (com final .br) e "*third party classifier*" tira o filtro de cookies específicos, também aceitando cookies de terceiros.

Para os filtros no script de coleta sem cookies de terceiros selecionamos alguns de sites específicos como Amazon, Ebay, Yahoo, Tumblr, Reddit, Ali, Discord, Whatsapp, Yandex, Alibaba, dentre outros. Além disso, também foram usados trechos chave no filtro para identificar potenciais cookies de identificação como "visitor" e "id".

Foi realizado uma verificação nos cookies que apareciam com mais frequência, resultando em maioria cookies de terceiros como Google (_ga, _gid, _gcl_au, _gat, _gads, 1P_JAR), Facebook (_fbp), Amazon (_auc, _asc), Bing (_uetvid, _uetsid), Adobe (s_cc, s_ecid) (Imagem 8).

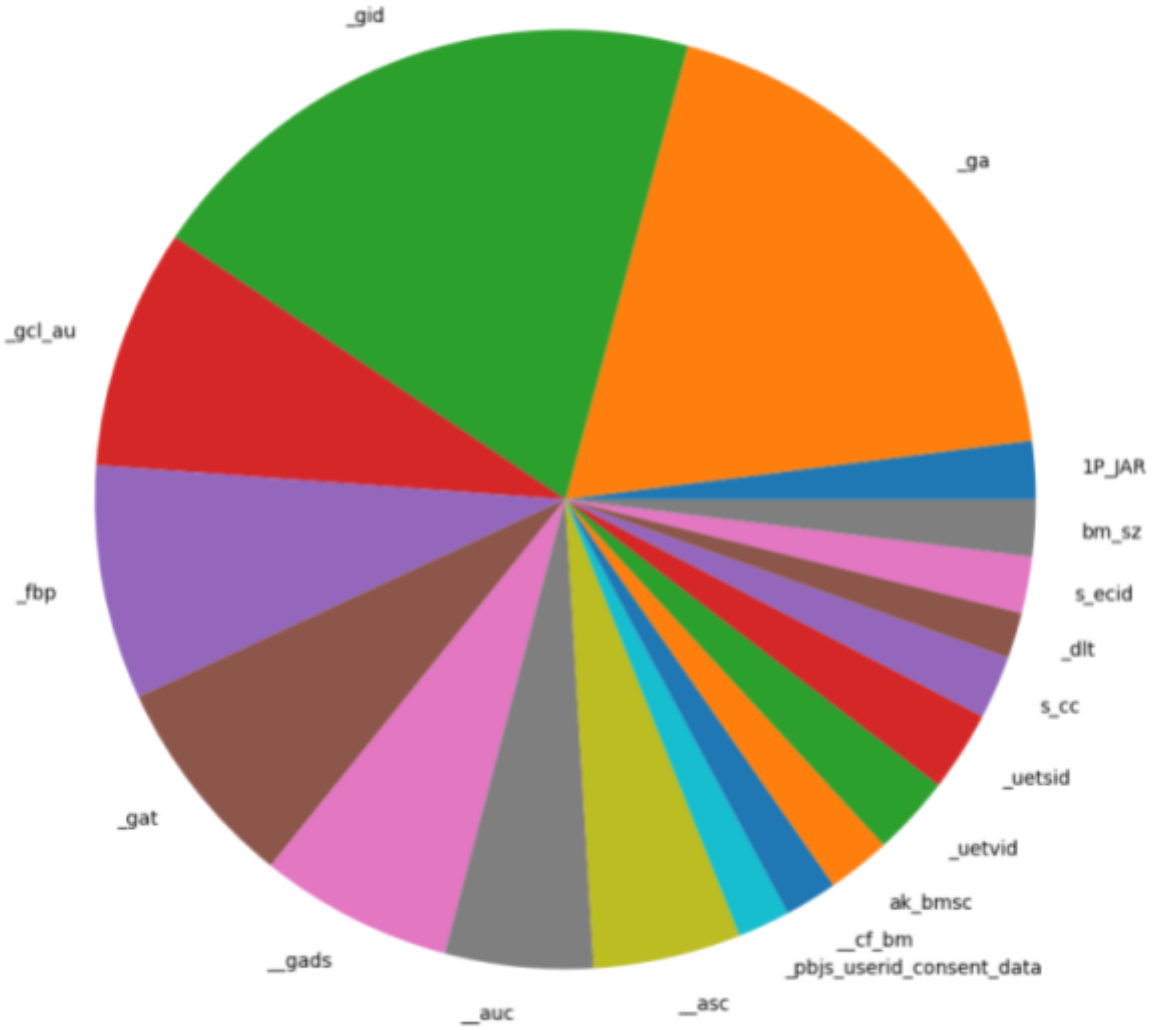

Imagem 8 - Cookies mais encontrados nos sites visitados durante a pesquisa

No repositório também se encontram os quatro gráficos resultantes do estudo que virão a seguir no capítulo resultados. Além do arquivo de resultados que contém cada site visitado e cookies encontrados nas duas visitas. Temos os gráficos para os 50.000 sites visitados, incluindo e sem incluir cookies de terceiros, e gráficos para os 250 sites nacionais visitados, incluindo e sem incluir cookies de terceiros.

Neste capítulo abordamos o método de coleta de dados usado no estudo, detalhamos o funcionamento do script em *selenium* e como foram realizadas as duas visitas através do navegador e a obtenção dos *cookies*. No próximo capítulo iremos apresentar os resultados obtidos e comparar com os estudos anteriores da GDPR.

4 RESULTADOS

Do top 1.000.000 do ranking da Alexa foram coletados 50.000 sites, que revelaram que o uso de cookies persistentes em comparação com outros estudos é menor para sites menos visitados, mas é mais elevado quando comparado com os sites com mais visitas. O estudo de 2020 revelou que sites no top 10(google, youtube, tmall, qq, sohu, baidu, facebook, taobao, 360cn, jd.com e amazon) tem uma porcentagem de 60% no uso de cookies persistentes, enquanto este estudo encontrou uma porcentagem de 100% no uso para os mesmos sites.

Na imagem abaixo no eixo X é possível observar o ranking dos sites coletados, que vai desde 0 até 50000. No eixo Y é possível encontrar a porcentagem do uso de cookies, que se inicia em 90% e vai diminuindo até em volta de 30% ao ultrapassar a marca de 10000 sites registrados.

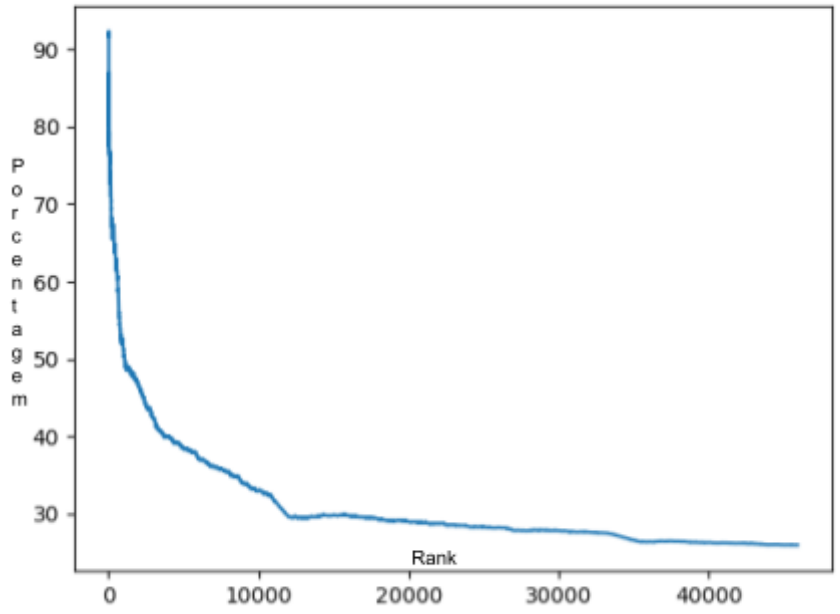

Imagem 9- top 50.0000 sites ordenados por ranking no eixo X e porcentagem no uso de cookies persistentes no eixo Y

Para a imagem abaixo do estudo "*Measuring Cookies and Web Privacy in a Post-GDPR World*" de Dabrowski, Merzdovnik, Ullrich, Sendera e Weippl (2021) temos duas comparações, entre visitas realizadas em território europeu e americano. No eixo X temos o rank total dos sites visitados e no eixo Y vemos a porcentagem de cookies persistentes utilizados. Em azul temos resultados obtidos em 2018 no início da GDPR na região europeia,

enquanto que em vermelho temos os resultados obtidos em território americano no mesmo período.

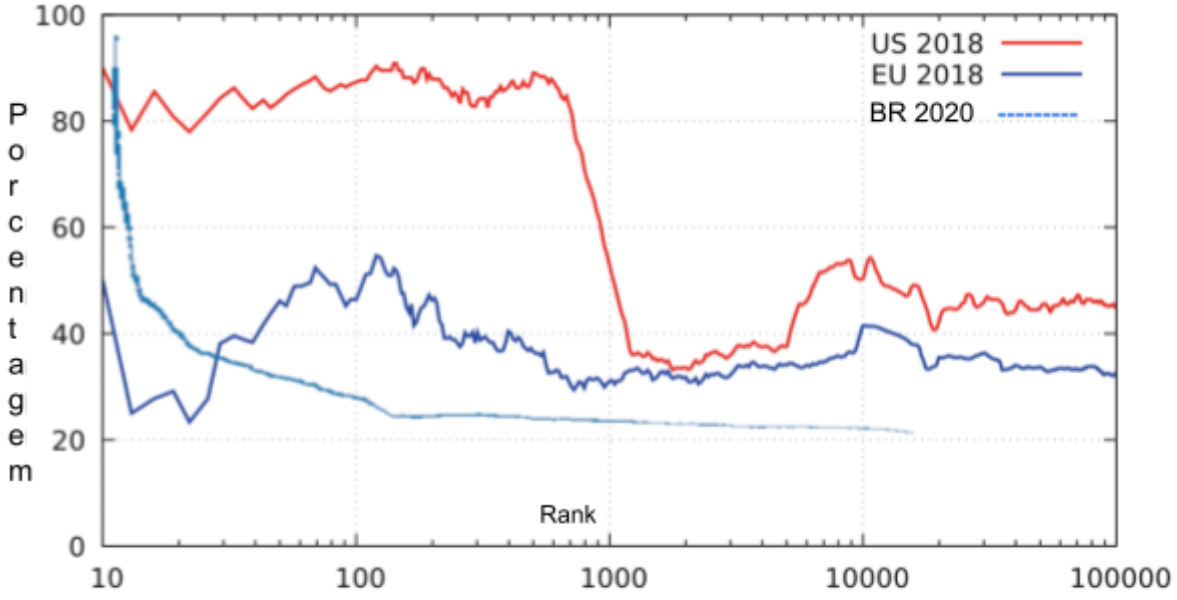

Imagem 10 - porcentagem de sites com uso de cookies persistentes no estudo de 2018 (DABROWSKI; MERZDOVNIK; ULLRICH; SENDERA; WEIPPL, 2021)

Comparando com os resultados do estudo realizado para a lei europeia em 2020 acima, há uma significativa redução no uso de cookies, embora domínios com maior quantidade de visitas ainda tenham uma utilização tão elevada quanto os domínios em ambiente americano, seguindo um padrão de redução exponencial (DABROWSKI; MERZDOVNIK; ULLRICH; SENDERA; WEIPPL, 2021).

Dos 50000 sites visitados, obtemos por volta de 250 sites com domínio nacional, assim como na imagem anterior, temos ranking dos sites no eixo X e porcentagem no uso de cookies no eixo Y.

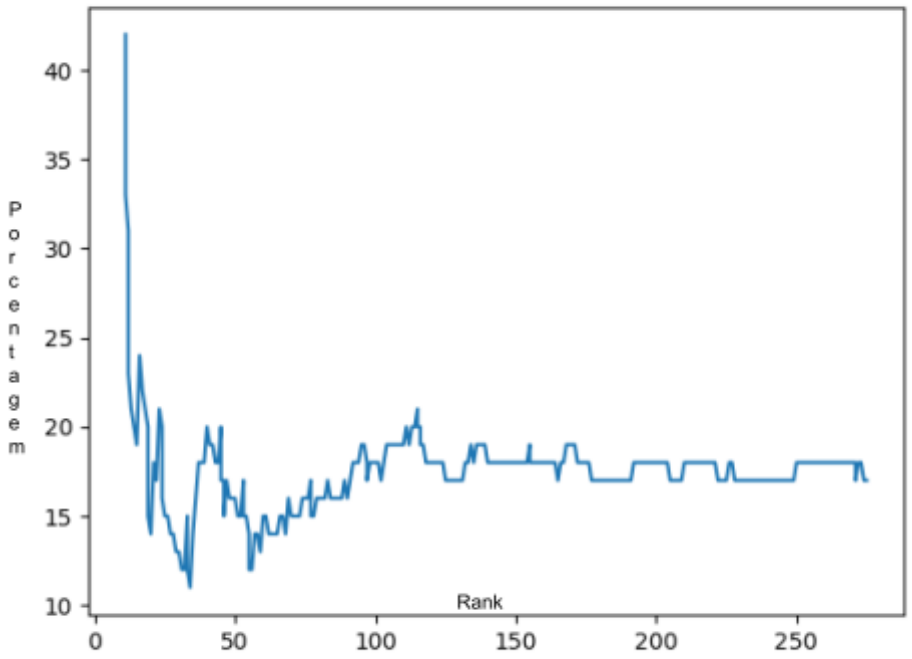

Imagem 11 - porcentagem de uso de cookies em domínios nacionais

Os resultados obtidos em domínios nacionais demonstram que há uma redução significativa no uso de cookies de identificação. Os sites com final em .br apresentaram maior uso de cookies em sites com maior tráfego, com uma porcentagem de 42% de uso de cookies persistentes, e logo entram em queda até chegar em volta de 26% no uso de cookies persistentes. Os sites analisados também foram por ordem de visitação e assim como os sites internacionais, os primeiros têm um número maior de cookies persistentes implementados.

Ranks no eixo X e porcentagem no uso de cookies persistentes no eixo Y:

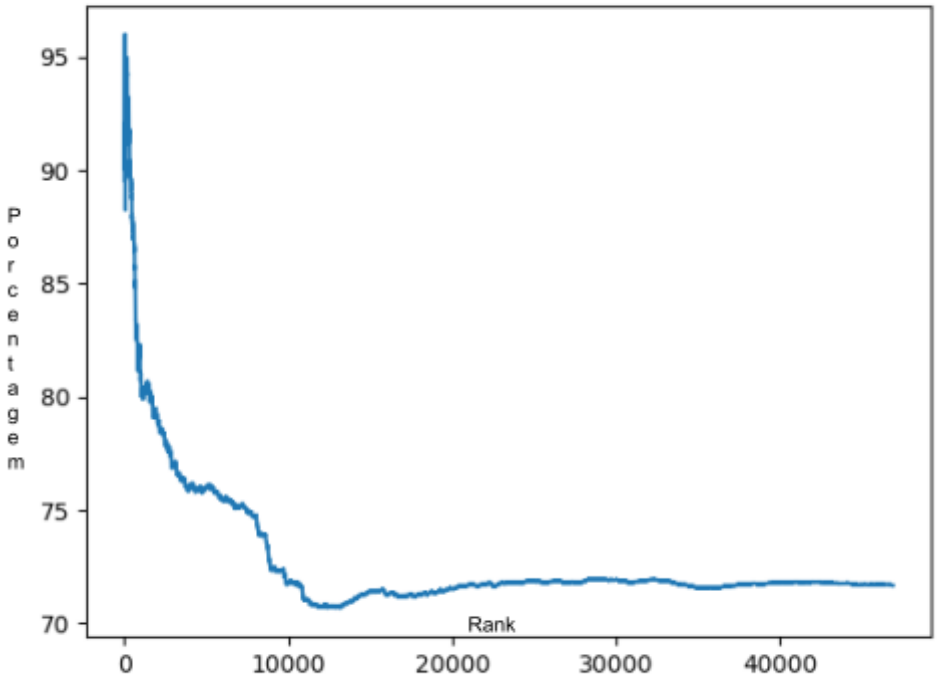

Imagem 12- top 50.0000 sites e porcentagem no uso de cookies persistentes (incluindo terceiros)

Já incluindo cookies de terceiros no ranking de sites internacionais, o número é um pouco maior, uma vez que a implementação de cookies do Google Analytics e da AWS, a porcentagem de uso de cookies persistentes nunca atinge abaixo dos 70%, sendo esse o limiar.

Ranks no eixo X e porcentagem no uso de cookies persistentes no eixo Y:

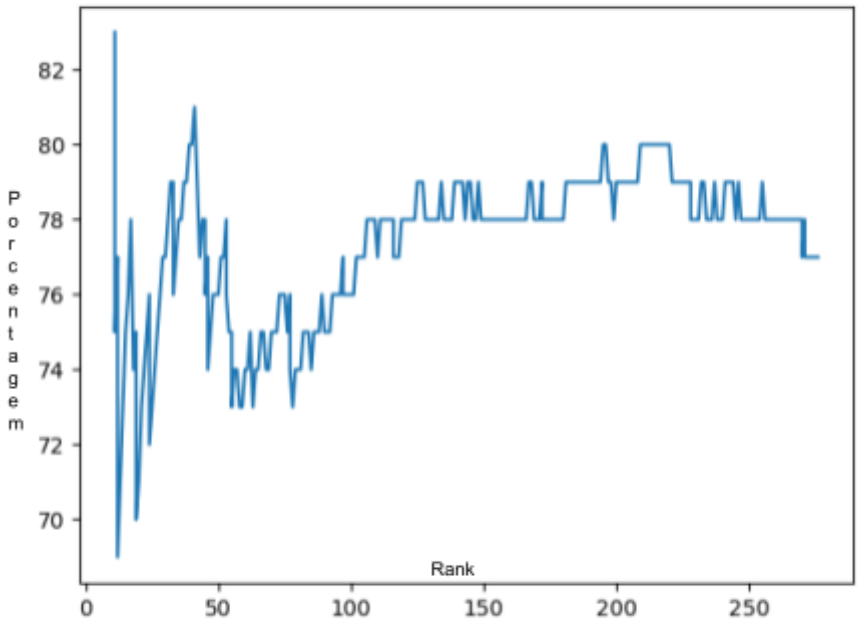

Imagem 13 - porcentagem de uso de cookies em dominios nacionais (incluindo terceiros)

Por final, os resultados obtidos em domínios nacionais incluindo cookies de terceiros é bem mais elevada em comparação aos resultados obtidos em domínios nacionais, mas sofre uma redução ao se comparar com o rank internacional, variando de 80% a 70% o que indica que todos os domínios nacionais tem uma frequência similar no uso de cookies.

Embora a utilização de cookies não seja equivalente a de alguns anos atrás, o impacto na redução do uso de cookies por padrão ao visitar um site não parece ter mudado significativamente em relação ao impacto do uso de cookies em regiões onde a GDPR vigora. A quantidade de cookies sendo usada se assemelha mais a regiões americanas que europeias, principalmente em questão de sites com maior rank.

Neste capítulo obtemos os resultados e realizamos comparações entre o efeito da LGPD e GDPR após as suas vigências. Analisamos também *cookies* de terceiros e como sites com mais visitas não parecem ter uma preocupação com as leis de proteção de dados. No próximo capítulo iremos tirar as conclusões de possíveis motivos que levaram à mudança no uso de *cookies.*

5 CONCLUSÃO

Houve uma queda no uso de *cookies* em geral desde o estudo de 2020, mas continua elevado em sites com maior quantidade de visitas. Em ambos os casos a porcentagem de uso de *cookies* persistentes é maior em domínios internacionais que nacionais. A diferença de *cookies* nacionais com terceiros é bem maior que os cookies internacionais com terceiros. Há múltiplas razões para a redução no uso de cookies, além das leis de proteção de dados, algumas outras como o Topics ou *fingerprinting* que servem como meios alternativos de obter dados de usuários.

Existem formas de amenizar o tracking realizado na web, extensões como ghostery foram comprovadas para ajudar a prevenir boa parte de cookies se estabelecerem, embora outros métodos de tracking como fingerprinting sejam mais difíceis de se detectar (KRETSCHMER; PENNEKAMP; WEHRLE, 2021).

Algumas considerações a serem feitas com relação a adoção da lei de proteção de dados nacional por parte das empresas indica que ainda há um certo desconhecimento e até mesmo negligência com relação a mudança no tratamento de dados pessoais.

Embora pouco tempo tenha se passado desde que a LGPD entrou em vigor, a lei ainda não realizou as sanções por violação de cumprimento, uma vez que entrou em vigência em setembro de 2020. Esse alto número no uso de cookies e desconhecimento da lei pode ser por conta do período curto de transição, uma vez que o estudo da GDPR foi realizado dois anos depois, mas o da LGPD foi realizado alguns meses após a vigência da lei (CORACCINI, 2021).

Por fim, o impacto que a LGPD trouxe no ambiente web ainda não é notável em comparação com a GDPR no ambiente europeu, o uso de cookies vem diminuindo mas isso é percebido mundialmente, não apenas no Brasil, embora a GDPR trate especificamente do uso de cookies, eles podem ser interpretados como dados que permitem identificar o usuário a partir do contexto, e por isso devem ser tratados com cautela.

6 TRABALHOS FUTUROS

O estudo “Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web” reuniu todos os artigos e pesquisas envolvendo impactos da GDPR dentro da Web. Eles não são limitados a tópicos como cookies persistentes, também trazendo outros assuntos como Termos de Serviço, Banners, Consentimento dentro da legislação europeia.

Trabalhos envolvendo o impacto que a LGPD trouxe no Brasil parecem ter sido extremamente escassos quando o assunto engloba Web, tanto por ser um assunto específico quanto pela curta duração desde que entrou em vigor, por conta disso, todos esses estudos reunidos pelo artigo citado acima podem ser reproduzidos no cenário da LGPD. Assim como o que esse trabalho propõe, é possível reproduzir dentro do ambiente da LGPD para verificar se houve algum impacto (KRETSCHMER; PENNEKAMP; WEHRLE, 2021).

7 REFERÊNCIAS

LARA, Letícia; RIBEIRO, Rafael; DE CARVALHO, Rebeca; SUASSUNA, Sílvia. Mudanças e desafios da LGPD. *In*: **Mudanças e desafios da LGPD** . [*S. l.*], 6 out. 2020. Disponível em: https://www.sigalei.com.br/blog/mudancas-e-desafios-da-lgpd. Acesso em: 14 jun. 2021.

TEPEDINO, Gustavo. Desafios da Lei Geral de Proteção de Dados (LGPD). **Revista Brasileira de Direito Civil-RBDCivil**, v. 26, n. 04, p. 11, 2020.

MOZILLA (org.). **Using HTTP cookies**. Disponível em: https://developer.mozilla.org/en-US/docs/Web/HTTP/Cookies. Acesso em: 29 ago. 2021.

DABROWSKI, Adrian; MERZDOVNIK, Georg; ULLRICH, Johanna; SENDERA, Gerald; WEIPPL, Edgar. **Measuring Cookies and Web Privacy in a Post-GDPR World**. Disponível em: https://link.springer.com/chapter/10.1007%2F978-3-030-15986-3_17. Acesso em: 29 ago. 2021.

WOLFORD, Ben. **What is GDPR, the EU's new data protection law?** Disponível em: https://gdpr.eu/what-is-gdpr. Acesso em: 29 ago. 2021.

PESTANA, Marcio. **Os princípios no tratamento de dados na Lei Geral da Proteção de Dados Pessoais**. 2020. Disponível em: https://www.conjur.com.br/2020-mai-25/marcio-pestana-principios-tratamento-dados-lgpd. Acesso em: 1 set. 2021

AMAZON. **Alexa Top 1 Million**. Disponível em: http://s3.amazonaws.com/alexa-static/top-1m.csv.zip. Acesso em: 10 out. 2021.

CAHN, Aaron; ALFELD, Scott; BARFORD, Paul; MUTHUKRISHNAN, S.. An Empirical Study of Web Cookies. In: WWW '16: PROCEEDINGS OF THE 25TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 16., 2016, Wisconsin-Madison. **An Empirical Study of Web Cookies.**

Wisconsin-Madison: Acm, 2016. p. 1-11. Disponível em: https://dl.acm.org/doi/10.1145/2872427.2882991. Acesso em: 12 dez. 2021.

FLEMING, Maria Cristina. **Cookies e LGPD**: o que está errado em 99% dos sites?. O que está errado em 99% dos sites?. 2021. Disponível em: https://www.lgpdazul.com.br/cookies-e-lgpd-o-guia-definitivo/. Acesso em: 12 dez. 2021.

KRETSCHMER, Michael; PENNEKAMP, Jan; WEHRLE, Klaus. Cookie Banners and Privacy Policies: measuring the impact of the gdpr on the web. **Acm Transactions On The Web**, [S.L.], v. 15, n. 4, p. 1-42, 7 jul. 2021. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/3466722.

ASSIS E MENDES. **Histórico das leis de proteção de dados e da privacidade na internet**. Disponível em: https://assisemendes.com.br/historico-protecao-de-dados/. Acesso em: 24 abr. 2022.

CORACCINI, Raphael. **Empresas não conseguem se adaptar à lei de proteção de dados, aponta pesquisa**. 2021. Disponível em: https://www.cnnbrasil.com.br/business/empresas-nao-conseguem-se-adaptar-a-lei-de-protecao-de-dados-diz-pesquisa/. Acesso em: 26 abr. 2022.

ARAÚJO, Cátia Cristina Souza Cruz. **Estudo comparado entre o Stop Online Piracy (SOPA) e o Protect Internet protocol ACT (PIPA) e as leis de pirataria brasileira à luz do Marco civil da internet**. 2014. Disponível em: https://jus.com.br/artigos/33161/estudo-comparado-entre-o-stop-online-piracy-sopa-e-o-protect-internet-protocol-act-pipa-e-as-leis-de-pirataria-brasileira-a-luz-do-marco-civil-da-internet. Acesso em: 30 abr. 2022.

State of California Department of Justice. **California Consumer Privacy Act (CCPA)**. 2018. Disponível em: https://oag.ca.gov/privacy/ccpa. Acesso em: 30 abr. 2022.

GOEL, Vinay. **Get to know the new Topics API for Privacy Sandbox**. 2022. Disponível em: https://blog.google/products/chrome/get-know-new-topics-api-privacy-sandbox/. Acesso em: 30 abr. 2022.

A. GUTTMANN . **Advertising media owners revenue worldwide from 2012 to 2026**. 2021. Disponível em: https://www.statista.com/statistics/236943/global-advertising-spending/. Acesso em: 5 maio 2022.

KRETSCHMER, Michael; PENNEKAMP, Jan; WEHRLE, Klaus. Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web. **Acm Transactions On The Web.** Aachen, p. 1-42. jun. 2021.