SOCIEDADE EDUCACIONA DE SANTA CATARINA – SOCIESC

CENTRO UNIVERSITÁRIO SOCIESC – UNISOCIESC

CAMPUS ANITA GARIBALDI

DOUGLAS ZIETZ

RENAN AUGUSTO REDEL

ANÁLISE DO TRÁFEGO PROVENIENTE DE PUBLICIDADES EM UM ISP

JOINVILLE

2021

DOUGLAS ZIETZ

RENAN AUGUSTO REDEL

ANÁLISE DO TRÁFEGO PROVENIENTE DE PUBLICIDADES EM UM ISP

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação do Centro Universitário Sociesc, como requisito parcial para a obtenção do Título de Bacharel em Engenharia de Computação.

Orientador: Dr. Ricardo Jose Pfitscher

JOINVILLE

2021

DOUGLAS ZIETZ

RENAN AUGUSTO REDEL

ANÁLISE DO TRÁFEGO PROVENIENTE DE PUBLICIDADES EM UM ISP

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação do Centro Universitário Sociesc, como requisito parcial para a obtenção do Título de Bacharel em Engenharia de Computação.

Joinville, 28 de junho de 2021.

BANCA EXAMINADORA

Professor Dr. Ricardo Jose Pfitscher

Professor Me. Anderson José de Souza

Professor Me. Claudinei Dias

RESUMO

Com o constante aumento de disponibilidade da internet nos últimos anos, e com a grande oferta de conteúdo, a *priori* de forma gratuita, houve um aumento relevante em relação ao consumo de dados na rede, e com a publicidade online não é diferente. Tal projeção tem promovido um aumento da quantidade de anúncios nos últimos anos. Esta ferramenta de divulgação proporciona que os anunciantes consigam alavancar seu capital por meio de sua plataforma e permite uma divulgação de produtos e serviços dentro da rede. Entretanto, o consumo destes dados derivados de anúncios, que em sua maioria são indesejados, incidem negativamente na franquia e a disponibilidade do plano de pacote de dados contratado pelo usuário. Neste contexto, esse trabalho tem como objetivo determinar o quão impactante é uso de dados provenientes de propagandas, identificar a utilização de bloqueadores de anúncio e sua eficácia, e caracterizar padrões de consumos com relação à quantidade de tráfego durante o período de análise – em um pequeno provedor de internet. Este trabalho utilizou de um segmento da área de atendimento do provedor, na qual foi analisado o consumo de dados de 33 equipamentos. Os resultados foram exibidos por via de gráfico e tabelas e mostraram que 18% do volume total de tráfego analisado é relacionado com conteúdo de publicidade.

Palavras-chave: Publicidade. ISP. Anúncios. Tráfego de dados. Consumo de dados. Ads.

ABSTRACT

With the constant internet access increase in recent years, and with a large offer of content, a *priori* free of charge, there has been a significant increase of data consumption on the network, and the online advertising follow this trend. Such a projection has promoted an increase in the number of advertisements in recent years. This dissemination tool offers advertisers the ability to leverage their capital through its platform and allows the dissemination of products and services within the network. However, consumption of these data derived from advertisements, which for the most part are unwanted, has a negative impact on the deductible and an availability of the data plan contracted by the user. In this context, this work aims to determine how impactful the use of data from advertisements is, to identify the use of ad blockers and their effectiveness, and to characterize consumption patterns - in a small internet provider. This work uses a segment of the service area of the provider, in which the data consumption of 33 internet nodes was analyzed. The results were identified throught of graphs and classifications and it was estimated that approximately 18% of the total volume of traffic relates to advertising content.

Keywords: Advertising. ISP. Adverts. Data traffic. Data consumption. Ads.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CSS	Cascading Style Sheets
DNS	Domain Name System
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
ISP	Internet Service Provider
SPAM	Sending and Posting Advertisement in Mass
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
GB	Gigabyte
MB	Megabyte

SUMARIO

1 INTRODUÇÃO

Com o crescimento exponencial que a internet teve nas últimas décadas, as propagandas na web se transformaram em um negócio altamente lucrativo para as grandes empresas, que disponibilizam suas plataformas de anúncio e para as empresas que utilizam destas para divulgar seus produtos amplamente (STATT, 2020). Porém, esse tráfego originário das propagandas onera o consumidor final, consumindo a sua banda de rede para realizar o *download* de um conteúdo que nem sempre é desejado.

Durante o acesso a páginas na internet, são executados, com o uso de diversos protocolos de comunicação, uma série de troca de mensagens entre o consumidor final e o servidor da internet. Caso a comunicação e acesso sejam efetivados, haverá como resultado uma página da internet, com diversos objetos, tais como links, imagens, textos e propagandas, que serão requisitados pelo navegador e, por consequência consumindo banda do contratante final. Por este motivo, é importante realizar a análise do consumo de dados provenientes de propagandas, a fim de mensurar o quanto de propagandas são consumidos no dia a dia.

Alguns trabalhos anteriores abordaram a análise de consumo de dados provenientes de propagandas. SIQUEIRA (2019) realizou a análise de consumo de propagandas durante o acesso a alguns sites, de maneira controlada. Entretanto, essa foi uma análise momentânea do consumo de dados por propagandas e teve um resultado representativo apenas para o cenário residencial observado. Para se obter um resultado que seja mais significativo, é necessário o uso de dados de acesso provenientes de diversos usuários. Assim, a medição em provedores de acesso à internet surge como uma opção interessante.

Provedores de internet disponibilizam acesso e navegação na internet a grandes grupos de pessoas por meio de planos de acesso (KUROSE e ROSS, 2017; WHATISMYIPADDRESS, 2018). Assim, todos os dados utilizados na navegação web destes usuários, transitam pelos servidores do provedor de internet, possibilitando assim, a análise em detalhe do que é consumido por um espectro maior de usuários.

Tendo em vista o contexto exposto, entende-se que com o uso de dados anônimos provenientes de um provedor de internet de pequeno porte, coletados por

meio do uso de uma ferramenta de rastreamento de pacotes, seja possível mensurar o quanto de banda de rede é consumido exclusivamente por propagandas. Assim é possível estabelecer os objetivos gerais e objetivos específicos deste trabalho de conclusão de curso. Os quais estão descritos na Seção 1.1.

1.1 OBJETIVOS E OBJETIVOS ESPECÍFICOS

O objetivo geral desta pesquisa é determinar o quanto de banda de rede é utilizado exclusivamente por tráfego de propagandas. Deste modo. É possível elencar os seguintes objetivos específicos:

- Realizar uma revisão bibliográfica referente a comunicação na Web;
- Pesquisar e definir diferentes tipos de publicidades online;
- Coletar e analisar os dados de propagandas em um pequeno provedor de acesso à internet;
- Caracterizar possíveis padrões de consumo relacionados a quantidade de tráfego consumido pela população analisada.

Para a realização desta pesquisa, foi realizada uma pesquisa da revisão bibliográfica por meio de livros, artigos, sites, relatórios, entre outros, a fim de elaborar um referencial teórico completo e atualizado do tema em questão.

1.2 ORGANIZAÇÃO TEXTUAL

O presente texto está organizado da seguinte forma: o Capítulo 2 traz a fundamentação teórica dos temas abordados pela pesquisa. O Capítulo 3 discute a metodologia utilizada juntamente com seus processos e suas etapas de elaboração. No Capítulo 4 está disposto os resultados obtidos através das análises. O Capítulo 5 traz a conclusão dos autores, juntamente com propostas de continuidade para alguns trabalhos futuros.

2 FUNDAMENTAÇÃO TEORICA

Neste capítulo são apresentados alguns conceitos que serão importantes para a compreensão deste trabalho, tais como o funcionamento de requisições web, protocolos de rede, definições de propagandas e seus tipos, entre outros.

2.1 NAVEGAÇÃO WEB

Para entender como as propagandas são inseridas na navegação dos usuários, primeiramente é necessário ter conhecimento de como a navegação Web é realizada na internet. A internet é uma rede que interconecta bilhões de dispositivos pelo mundo (KUROSE e ROSS, 2016), atualmente está presente em praticamente todos os dispositivos que temos contato, de computadores a dispositivos vestíveis.

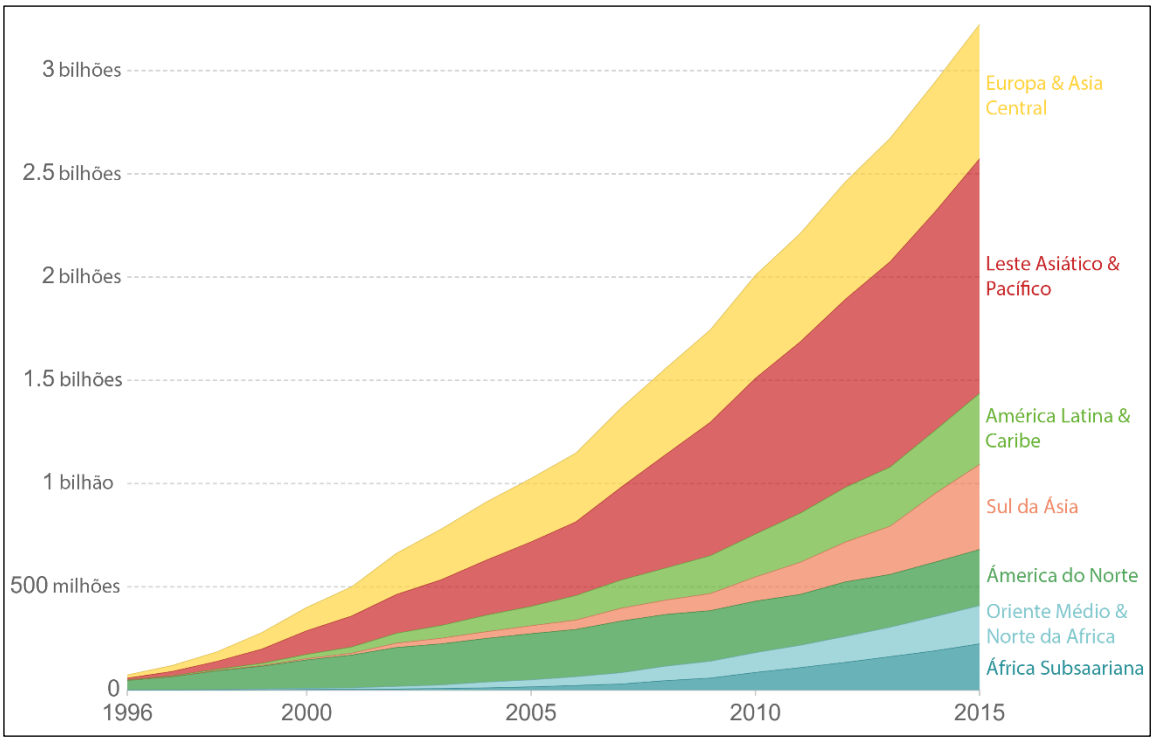
O *FEDERAL NETWORKING COUNCIL* (Definition of "Internet", 1995) definiu a internet como:

> (i) É logicamente ligado entre si por um espaço de endereço exclusivo global baseado no Internet Protocol (IP) ou suas extensões posteriores / follow-ons.
> (ii) É capaz de suportar comunicações usando o Transmission Control Protocol / Internet Protocol (TCP / IP) suíte ou suas extensões posteriores / follow-ons, e / IP ou outros protocolos compatíveis.
> (iii) Fornece, usa ou torna acessível, seja pública ou privada, serviços de alto nível em camadas sobre as comunicações e infraestruturas relacionadas como aqui descrito.

Para realizar a comunicação entre os dispositivos, é necessário fazer uso de protocolos de troca de mensagens implementando através do acesso a uma rede que realizará a interconexão entre os dispositivos de redes distintas. Essa rede de acesso, também conhecida como link de acesso, ou enlace de acesso, é fornecida por um provedor de acesso à internet (ISP – do inglês *Internet Service Provider*).

O uso da internet vem crescendo exponencialmente ano a ano. De acordo com (CISCO, 2017), em 2021 o tráfego anual da internet chegará ao valor histórico de 3,3 zettabytes (2^{70} bytes). Isso se deve a maior alcance da internet, conforme mostra a Figura 1, ainda, vale ressaltar o crescimento do número de pequenos ISPs regionais, oferecendo acesso à internet de forma ampla nas regiões que estão inseridos.

Figura 1 - O crescimento da internet



Fonte: (OUR WORLD IN DATA, 2016).

A Web, ou *World Wide Web*, é responsável por uma das maiores fatias de dados na internet. Conforme (TREVISAN, GIORDANO, *et al.*, 2020), um dos impactos mais relevantes na Web é ocasionado pelos serviços de *stream* de vídeos. Basicamente, a Web é assim denominada pois ela é caracterizada pela interconexão de páginas HTML (*HyperText Markup Language*) que são conectadas entre si, formando assim um comportamento similar a uma teia. Estas páginas são hospedadas em servidores que executam em infraestruturas de rede que compõe a Internet (KUROSE e ROSS, 2016). Ou seja, a Web é um dos serviços que executa sobre a rede de redes, denominada, internet.

2.1.1 A rede de redes

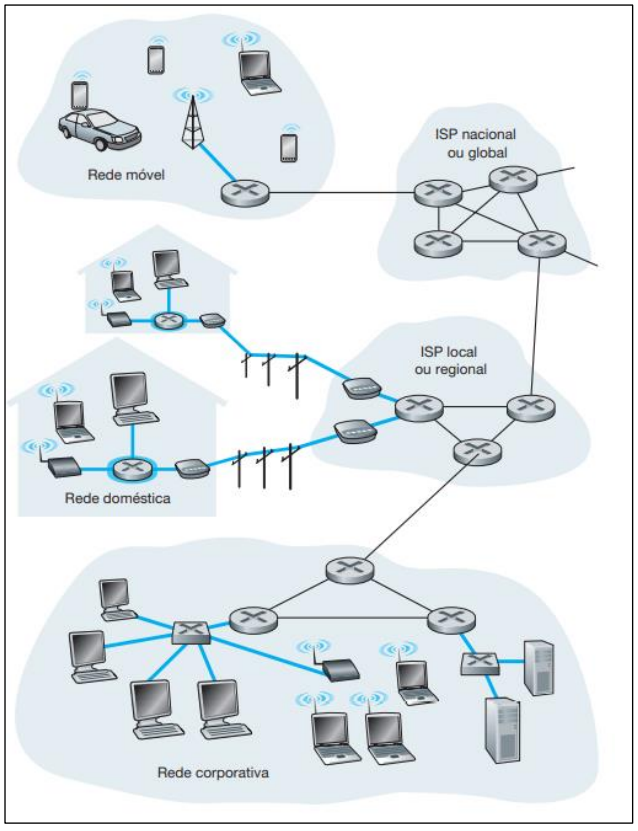
Como mencionado anteriormente, para interligar os diversos computadores e servidores de conteúdo entre si, se faz necessário a existência de ISPs (KUROSE e ROSS, 2016). ISPs também conhecidos como provedores de acesso/conexão, são em sua totalidade pessoas jurídicas que exercem a função de disponibilizar, distribuir e possibilitar o acesso à internet para seus clientes. Como menciona LEONARDI (2012) para a caracterização de provedor de internet, basta que a empresa forneça

acesso suficiente a internet, não sendo obrigatório ou necessário prestarem outros serviços adicionais, tais como correios eletrônicos ou hospedagens de sites.

Pode-se destacar que tais provedores não utilizam apenas cabeamento para a distribuição de seu produto, existem inúmeras operadoras que fornecem seus serviços de forma móvel com o uso de diversas tecnologias (KUROSE e ROSS, 2016).

A Figura 2 representa alguns tipos de redes nas quais os provedores atuam. Os blocos situados a esquerda da figura, indicam a zona na qual os ISP fornecem seus serviços, rede móveis, domésticas e corporativas, onde sua distribuição pode ocorrer tanto pela forma de rede móvel, quanto pela forma de cabeamento.

Figura 2 - Atuação ISPs.



Fonte: (KUROSE e ROSS, 2016).

2.1.2 Requisições de Acesso

O processo inicial de acesso a uma página de um servidor na web se inicia no navegador. Após realizar uma solicitação de acesso a uma página da web o navegador irá examinar o conteúdo da primeira página em busca dos demais objetos a serem requisitados para exibir a página em sua totalidade. Assim, uma sequência

de solicitações é realizada em série e de forma transparente para o usuário final (KUROSE e ROSS, 2016).

Páginas da web são constituídas por objetos, sendo arquivos HTML, imagens, vídeos, *scripts* (KUROSE e ROSS, 2016). Dentre os diversos objetos necessários para exibição do conteúdo final, as propagandas são inseridas para viabilizar a exibição do conteúdo de forma 'gratuita' ao usuário. Portanto, as propagandas fazem parte destas páginas, e podem ser introduzidas de diversas formas, tais como banners, pop-ups, texto etc. A Figura 3 demonstra alguns elementos presentes numa página da web comum que podem passar despercebidos por um usuário comum, tais como imagens, vídeos e publicidade.

Figura 3 - Elementos de uma página da web.

Fonte: r7.com (2020, adaptado).

Para compreender melhor como as propagadas são inseridas e a sua origem, faz-se relevante o estudo dos serviços envolvidos ao longo da comunicação desde a requisição inicial até a exibição da página completa no navegador. Ao digitar o endereço de um site em um navegador, primeiramente o navegador analisa o que está sendo digitado, e verifica se é uma pesquisa ou um endereço da internet (COPES, 2018). Caso seja um endereço, primeiro passo é a resolução do endereço IP deste site (KUROSE e ROSS, 2016). Para tanto, é enviada uma requisição oriunda da camada de aplicação para um servidor DNS (*Domain Name System*) que recebe esta solicitação identifica o endereço IP do servidor de origem e retornar ao usuário o endereço IP de destino (maiores detalhes sobre o DNS na Seção 2.1.3).

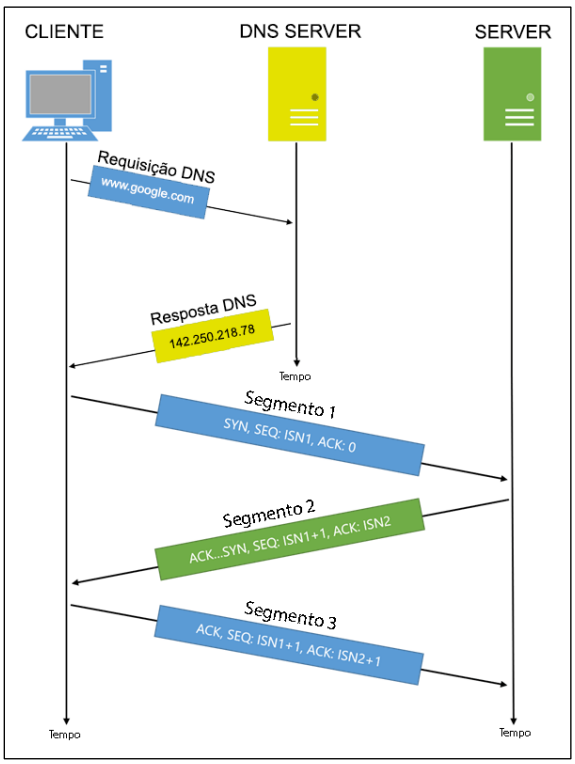
Uma vez identificado o endereço IP de destino, é aberta uma conexão soquete TCP (*Transmission Control Protocol*) entre o cliente e o destinatário, e é enviado através do protocolo TCP um segmento do TCP SYN (primeiro segmento do *three-way handshake*, Figura 4), utilizando a porta 80.

O *three-way handshake* ou *handshake* de três vias, é a forma que uma conexão TCP é estabelecida entre cliente-servidor (FERNANDES, 2018). A Figura 4 demonstra o funcionamento do *three-way* handshake durante o estabelecimento de uma nova conexão. Os segmentos do *three-way handshake* são marcados por três principais *flag bits*:

- SYN: sincronizar;
- ACK: reconhecimento;
- ISN: sequência numérica inicial.

O primeiro segmento, é feito pelo cliente que está solicitando a conexão. Por ele é enviado um segmento especial de TCP, o qual possui a *flag* SYN ativa e um número de sequência inicial (KUROSE e ROSS, 2016). O servidor destinatário, recebe o segmento, e retorna para o cliente um novo segmento especial de TCP, contendo a *flag* SYN em conjunto com a *flag* ACK. A sequência numérica enviada anteriormente, é acrescida em um. Por fim, para que a conexão seja estabelecida, o cliente envia um novo segmento contendo apenas uma *flag* ACK enquanto a sequência numérica permanece a mesma que foi recebida na segunda etapa.

Figura 4 - *Three-way Handshake*



Fonte: (FERNANDES, 2018, adaptado).

Após estabelecida a conexão TCP com o servidor de destino, é enviado pelo cliente um datagrama IP contendo uma requisição HTTP, através do soquete TCP para o servidor de destino. O servidor de destino realiza a leitura da requisição HTTP, e responde para o cliente a página através de uma resposta HTTP. A Tabela 1 mostra os possíveis retornos que podem existir. Por fim, o navegador interpreta as informações e exibe o conteúdo (KUROSE e ROSS, 2016).

Tabela 1 - Respostas HTTP

Intervalo	**Retorno**
100 – 199	Informações
200 – 299	Sucesso
300 – 399	Redirecionamento
400 – 499	Erro de cliente
500 – 599	Erro do servidor

Fonte: os autores (2020).

Todo o processo de requisição é realizado para cada objeto de uma página, tais como imagens, vídeos, *scripts etc.* Páginas com muitos objetos, requerem mais

tempo para serem exibidas e consequentemente acabam consumindo uma maior banda de internet. Ainda, ressalta-se que as requisições de objetos em uma mesma página podem ser feitas a domínios diferentes, o que requer a resolução do IP associado a cada domínio. Este entendimento é importante para compreender como é possível identificar objetos relativos à propaganda nessas páginas. Como mencionado anteriormente, o serviço de DNS é responsável pela tradução entre nome e IP, e será descrito na Seção 2.1.3.

2.1.3 Domain Name System

O *Domain Name System (DNS)* é um serviço de diretório cuja principal função é caracterizada por fazer a tradução dos nomes dos hosts para endereços IPs específicos. Pode-se dizer que o DNS facilita o uso do usuário, possibilitando que ele não precise digitar o endereço IP do site, mas sim o seu domínio.

Como (KUROSE e ROSS, 2016, p. 96) mencionam, o DNS é um banco de dados distribuído executado de forma hierárquica nos servidores de DNS, sendo um protocolo de camada de aplicação que permite que hospedeiros consultem o banco de dados distribuído. O protocolo DNS utiliza o protocolo UDP e faz uso da porta 53.

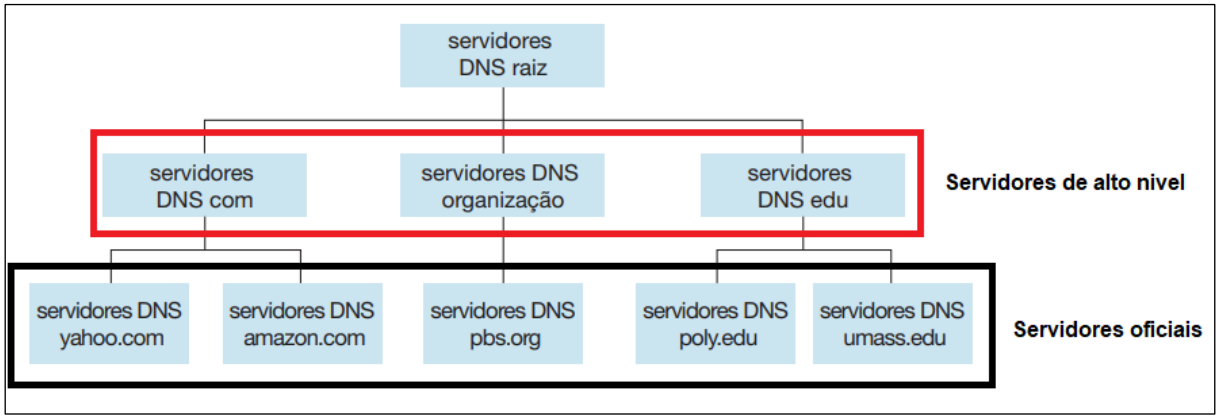
Com relação aos passos de seu funcionamento (KUROSE e ROSS, 2016, p. 98) exemplificam, quando alguma aplicação necessita esse serviço de tradução, ela acionará o lado DNS do cliente, relatando o nome do hospedeiro que precisa ser traduzido. O DNS do hospedeiro realiza a consulta na rede. A partir daí, o DNS do hospedeiro do usuário assume o controle, enviando uma mensagem de consulta para a rede. Em seguida recebe a sua resposta, contendo o mapeamento desejado que é passado para a aplicação no cliente.

Por se tratar de uma rede de tamanho quase imensurável compostas por bilhões de nós, a internet utiliza muitos servidores DNS, que são organizados de forma hierárquica e distribuída por todo o mundo, e podem ser classificados de 3 formas: Servidores raiz, de alto nível e oficiais.

A Figura 5 representa a hierarquia dos servidores. Na parte superior observa-se os servidores raiz, responsáveis por indexar os servidores de maior complexidade. Em vermelho, tem-se os servidores de alto nível, eles são os agregadores da camada inferior. Por fim, na camada inferior pode-se observar os servidores oficiais, formado

pelas organizações que possuem hospedeiros que possam ser acessados via internet.

Figura 5 - Hierarquia dos servidores



Fonte: (KUROSE e ROSS, 2017 adaptado).

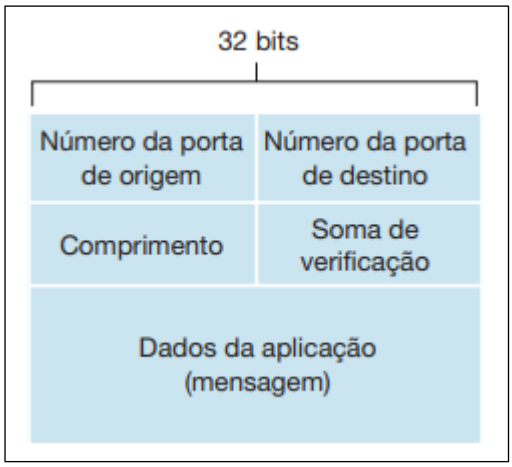
2.1.4 UDP

O UDP (*User Datagrama Control) é* considerado por vários autores, um protocolo irmão do TCP utilizado na camada de transporte, atuando como intermediário entre operações e aplicações, agindo de forma menos confiável e mais simples pois não é baseado em conexões. (KUROSE e ROSS, 2013, p. 69) mencionam que “o UDP provê um serviço não confiável de transferência de dados — isto é, quando um processo envia uma mensagem para dentro de um socket UDP, o protocolo não oferece garantias de que a mensagem chegará ao processo receptor.”

Seu funcionamento baseia-se no transporte de datagramas entre sua origem e seu destino, sem correção de erros e sem entrega confiável, com o intuito de agilizar o processo de comunicação. Quando é solicitado ele envia informações a seu destinatário, sem se preocupar se o remetente recebeu devidamente. Quando há perda de pacote, o mesmo não pode ser recuperado e o restante continua a ser enviado. Devido à sua agilidade, esse protocolo é utilizado por diversos serviços que necessitam de rápida resolução, tais como, o DNS (PRAMATAROV, 2018).

Na Figura 6, pode-se ver a estrutura do segmento UDP. A estrutura possui 5 campos. Número de origem e destino da porta, que indicam as respectivas portas usadas na comunicação. Campo comprimento detalha em bytes o comprimento do segmento, incluindo o cabeçalho. Soma de verificação é usado para verificação de erros e dos dados compartilhados.

Figura 6 - Estrutura do segmento UDP



Fonte: (KUROSE e ROSS, 2016).

2.1.5 TCP

O TCP (*Transmission Control Protocol*) é um protocolo de rede da camada de transporte do modelo TCP/IP. Teve sua primeira menção em 1974 (CERF e KAHN, 1974). Para a comunicação HTTP, é utilizado a porta 80 (KUROSE e ROSS, 2016, p. 236).

Opera segundo o modelo cliente-servidor, necessitando que o início da comunicação seja solicitado pelo cliente. Para se iniciar a comunicação, é necessário realizar uma autenticação inicia, o *three-way handshake*. Após realizado, é iniciada uma comunicação ponto a ponto entre o cliente e o servidor atuando de forma *full-duplex*, ou seja, podendo receber e transmitir dados ao mesmo tempo.

Toda a troca de pacotes realizada possui garantia de entrega dos pacotes além de uma entrega ordenada. Permite recuperação de pacotes perdidos ou a eliminação de pacotes duplicados. A ordenação de entrega é feita com o uso números de sequência, que é uma sequência numérica que se incrementa a cada pacote.

Quando se comparado com o protocolo UDP, o TCP possui uma velocidade de transmissão inferior, devido a necessidade de realização de verificação de pacotes, *three-way handshake*, reenvio de pacotes, presente no protocolo TCP. O protocolo TCP é utilizado em comunicação em que não possa haver perca de comunicações entre o requisitante e o servidor. Assim, serviços que priorizam a qualidade dos dados se aproveitam dos serviços do TCP, tais como, as aplicações via Web.

2.2 PUBLICIDADE ONLINE

Em meados de 1990 com o “*boom*” do uso da internet e a criação de interfaces gráficas para navegação (ALVAREZ e GOMES, 2010), muitas empresas começaram a utilizar esta nova tecnologia como forma de negócio (LEINER, KAHN, *et al.*, 2009, p. 30). Novos meios de se comunicar com os seus possíveis clientes foram sendo criados, a fim de alcançar novos públicos.

RICHARDS e CURRAN definem a publicidade como:

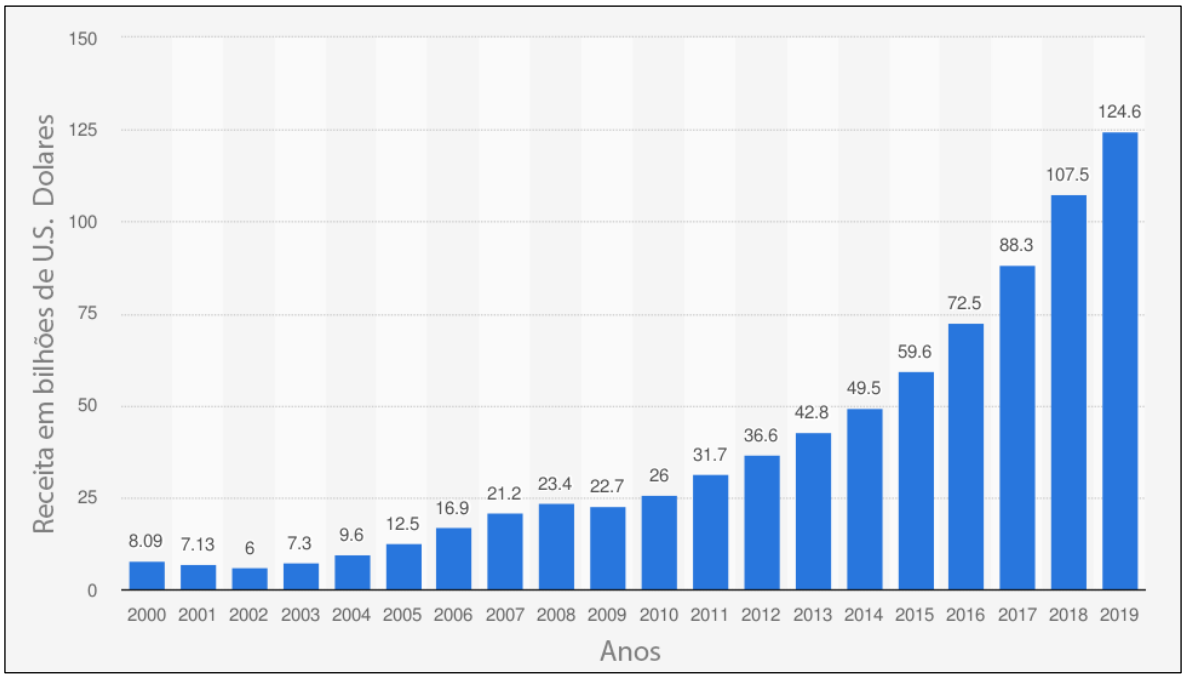
> Publicidade é uma forma de comunicação paga, de uma fonte conhecida, projetada para persuadir o receptor a realizar alguma ação, agora ou no futuro (2002, p. 74, tradução nossa).

Segundo MCMILLAN (2004, p. 3), a publicidade online consiste em uma mensagem indireta, paga por alguém que possa ser identificado – anunciante -, criada por uma pessoa ou empresa específica – intermediário –, comunicada por um veículo de anúncios – disseminador – com o objetivo de influenciar a tomada de decisões do receptor final – consumidor.

Disseminadores são empresas que divulgam os anúncios publicitários para a grande massa. Em 2019, aproximadamente 75% da receita destinada a publicidade online no mundo, foi destinada a 10 grandes empresas (ENBERG, 2019).

Com a rápida popularização da internet (ver Figura 1), a receita destinada a publicidades online cresceu exponencialmente. A Figura 7 mostra a receita em bilhões de dólares dos últimos 19 anos, a qual é possível perceber um crescimento exponencial ano a ano, sendo o último valor registrado de 124,6 bilhões de dólares apenas nos Estados Unidos. Conforme pesquisa *Digital Adspend* realizada anualmente pelo IAB, de 2018 a 2019, houve um crescimento de 23% no valor declarado de investimento em publicidades online no Brasil. Crescimento esse que vem se mostrando constante ao longo dos anos, e não demonstra que irá reduzir no curto prazo, visto a criação de novas formas de anúncios na internet (FREBERG, MCGAUGHEY, *et al.*, 2011, p. 1).

Figura 7 - Receita de publicidade online nos Estados Unidos (em dólares)



Fonte: (IAB, PWC, 2020).

Com a alta demanda dos anunciantes por usuários que venham consumir suas publicidades, o consumo de dados é altamente impactado. Segundo SIQUEIRA (2019, p. 39), em testes realizados em um ambiente controlado, o consumo de banda de rede por parte de conteúdos publicitários, foi de aproximadamente 31% do total trafegado. Assim, entende-se que o uso de publicidade implícita nas páginas web pode ser considerada como uma contrapartida ao oferecimento de serviços "gratuitos" ao usuário final, uma vez que esses conseguem visualizar os mesmos somente tendo um dispositivo com acesso à internet, sem custo adicional direto.

MCMILLAN (2004, p. 7) sugere que existam quatro diferentes tipos de publicidades online:

1. Anúncios com o propósito de desenvolvimento da marca: é a forma mais tradicional de anúncios na internet. Seu propósito é desenvolver a marca junto ao cliente e realizar a comunicação que o anúncio está se dispondo. É comumente apresentado em websites das mais diversas formas, como botões, *hyperlinks*, banners etc.
2. Comunicação corporativa: são propagandas enviadas diretamente para seus clientes, na forma de mensagens. Este tipo de comunicação, é considerado como SPAM quando o consumidor, não opta por receber.

3. Comunicação direta: forma de comunicação difícil de ser classificada como publicidade, devido a sua forma intrusiva de ser realizada e sua personalização. Se caracteriza pelo envio de mensagens diretamente para consumidores específicos, com o objetivo de haver uma ação de resposta.
4. Transações eletrônicas: mensagens com o propósito de conduzir transações diretamente com os seus consumidores.

O proposito e qual seu nível de atuação, serão exemplificados na Tabela 2. Com a evolução da internet, os tipos de anúncios não são limitados aos demonstrados na Tabela 2. Existe uma infinidade de novos tipos de anúncios na internet; programadores estão em constante desenvolvimento de novas formas de tornar angariar cliques em links e obter a ação desejável do anúncio. Por outro lado, existem iniciativas que visam desenvolver ferramentas para limitar o download desses anúncios, denominadas bloqueadores.

Tabela 2 - Tipos de anúncios na Internet

Proposito	Atuação: Externa	Controlada
Comunicação	**Desenvolvimento de marca** Formato pequeno: Botões; Hyperlinks. Banners: Mídia interativa; Expansível; Flutuante. Tela toda: *Pop-ups;* Grandes Mídias interativas; Sobreposição de sites. Espaço pago: Conteúdo patrocinado; Hyperlink pago. Outros: Análises pagas; Vídeos pagos; Artigos pagos.	**Comunicação Corporativa** Envio de mensagens: *Newsletters;* Busca de mensagens: Chat; Blogs; *Webcasts*.
Apelo à uma ação	**Mensagem diretas** Todos os formatos acima que tenham um apelo para uma ação, geralmente com um link.	**Transações eletrônicas** *E-commerce;* Marketing direto.

Fonte: (MCMILLAN, 2004, p. 23 - tradução nossa. Adaptado).

2.3 BLOQUEADORES DE ANÚNCIOS

Bloqueadores de anúncios são softwares que realizam a filtragem dos dados que estão sendo recebidos, bloqueando ou liberando determinados tipos de conteúdo. Esse filtro é realizado utilizando listas de bloqueio, contendo os principais divulgadores de anúncios da internet. O bloqueio efetivamente é feito sobre o código-fonte da página por meio de uma análise do conteúdo que será baixado, podendo este ser um anúncio, rastreadores de navegação, itens de redes sociais, pop-ups etc. (PALANT, 2008).

Como aponta a Tabela 3, existem atualmente, uma infinidade de opções de bloqueadores de anúncios. Podemos classificá-los em três categorias (PAGEFAIR, 2020):

Tabela 3 - Categorias de bloqueadores de anúncios

Categoria	Aplicação
Desktop	Softwares e extensões instalados diretamente no computador/browser. **Ex: Adblock plus, adguard, uBlock Origin.**
Mobile	Aplicativos e extensões instalados em dispositivos moveis. Existem versões que realizam o bloqueio dos anúncios por padrão, e outras versões que precisam ser habilitados. **Ex: Brave, Adblock Plus, Adguard.**
Cross-plataform	Serviços *cross-plataform*, que utilizam em geral, bloqueios de tráfego de mais baixo nível. **Ex: Pi-hole, NordVPN, Adguard DNS.**

Fonte: (PAGEFAIR, 2020, p. 5, adaptado).

O uso de bloqueadores de anúncios vem crescendo de forma constante nos últimos anos. De 2016 a 2019, houve um aumento de 64% no seu uso, totalizando aproximadamente 527 milhões de usuários (PAGEFAIR, 2020).

A motivação do uso de bloqueadores de anúncios, é em sua grande maioria por conta de anúncios irritantes e intrusivos, (AN, 2016) destaca que 64% dos entrevistados utilizam bloqueadores devido a isso, sobretudo, 77% dos entrevistados concordam que nem todos os anúncios devem ser bloqueados. Com a expansão do

uso dos bloqueadores de anúncios, entendeu-se necessário a criação de uma lista de exclusão, em que conste apenas sites com anúncios não intrusivos. Desde 2017 a iniciativa *AcceptableAds*[1] rege uma lista de sites com anúncios aceitáveis que é adicionada à lista branca dos principais bloqueadores de anúncios, além de incentivar o melhor uso da propaganda.

2.3.1 Funcionamento dos bloqueadores de anúncios

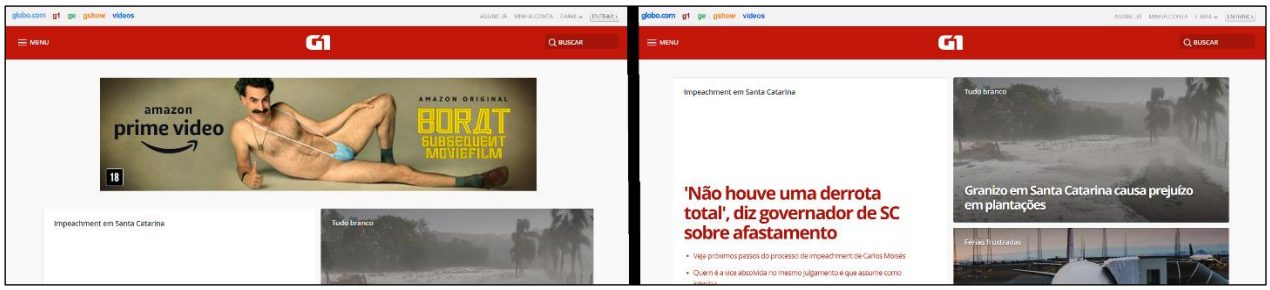
A forma simples de resumir o mecanismo de bloqueio de anúncios é: “[...] um bloqueio que não permite o navegador ou aplicativo realizar o download de conteúdo não desejado.” (ADGUARD, 2017, p. 1, tradução nossa).

Detalhadamente, cada bloqueador, possui uma forma de prevenir a exibição de propagandas. De modo geral, os bloqueadores de anúncios da categoria de desktop funcionam com dois princípios: bloqueio de requisições *HTTP* e ocultação de elementos (ADBLOCK PLUS).

- Bloqueio de requisições: previne que o navegador receba os dados provenientes dos anúncios. Esse bloqueio é feito utilizando como base os filtros pré-existentes;
- Ocultação de elementos: com as publicidades não sendo exibidas, existiram lacunas no conteúdo a ser exibido (Figura 8). Com a utilização de seletores e regras *CSS* no código-fonte da página, é possível reposicionar elementos da página.

[1] Disponível em : https://www.acceptableads.com

Figura 8 - Ocultação de elemento



Fonte: compilação dos autores[2] (2020).

Para dispositivos moveis existem duas formas de realizar o bloqueio de anúncios (ADGUARD, 2017):

- Aplicativos: o bloqueio é realizado a nível de rede, filtrando toda a conexão que passa pelo dispositivo;
- Edição de arquivos *hosts*: se baseia em editar os arquivos *hosts* dos dispositivos. Arquivos *hosts* são responsáveis por associar endereços de rede a endereços de IPs.

Os bloqueadores da categoria *cross-plataform* que utilizam bloqueio de conteúdo através do DNS, não são capazes de realizar a ocultação de elementos, realizando um bloqueio semelhante ao feito pela edição de arquivos *hosts*, porém, são capazes de bloquear todos os anúncios dos dispositivos conectados na rede (ADGUARD, 2017).

Tendo em vista o comportamento das requisições na web, e os métodos que as propagandas são inseridas, é necessário compreender quais são as técnicas e ferramentas existentes para coleta e análise de tráfego. Neste contexto, a Seção 2.4 discute as tecnologias para captura de pacotes.

2.4 CAPTURA DE PACOTES

A captura de pacotes consiste em um processo de monitoramento de pacotes que são transmitidos na rede, por meio de *hardware* ou *software*. Com o uso deste tipo de ferramenta, é possível obter informações como tipo, tamanho, estrutura e dados dos pacotes que estão sendo transmitidos. (BAN, CHO, *et al.*, 2007, p. 1)

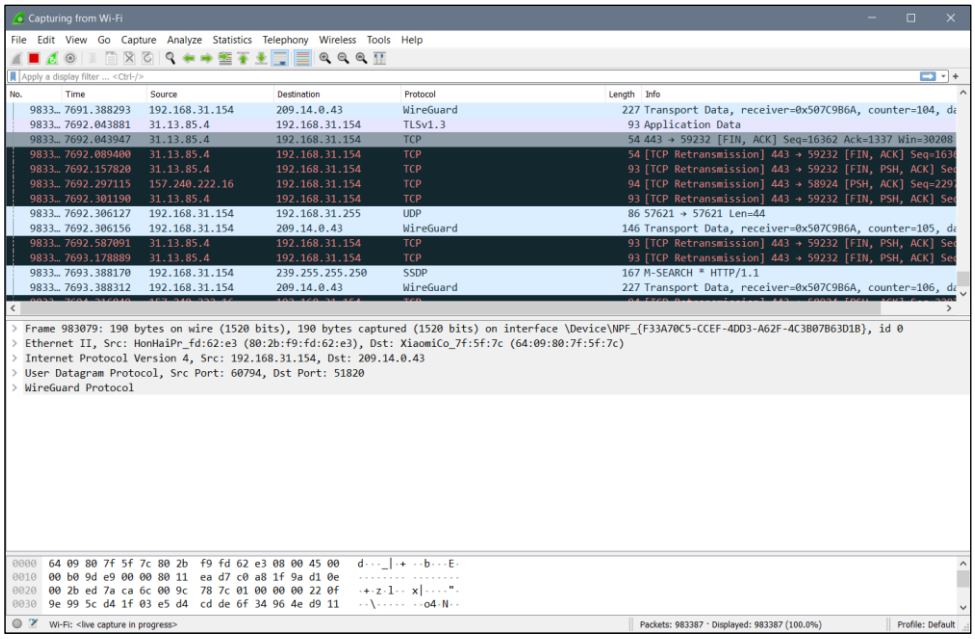
[2] Montagem realizada a partir de imagens coletadas do site g1.globo.com.

2.4.1 Wireshark

O Wireshark é um software *cross-platform* analisador de pacotes, utilizado principalmente para solução de problemas em redes. Tal ferramenta permite que os usuários visualizem todos os pacotes que estão sendo transmitidos na rede, além de poder visualizar detalhes como tamanho, dados que estão sendo transmitidos, protocolo etc. (WIRESHARK, 2020).

Um dos recursos primários do Wireshark, é o seu sistema de captura de pacotes, permitindo que o usuário realize a captura de todos os pacotes transitados em uma interface de rede, em determinado protocolo e horário, além de permitir a posterior análise dos dados capturados. Como exemplificado na Figura 9, a seguir:

Figura 9 - Interface Wireshark



Fonte: os autores (2020).

A Figura 9 mostra a interface principal do Wireshark, podendo ser possível visualizar os pacotes que estão sendo transmitidos em tempo real além de detalhes específicos de cada protocolo/pacote. Incluem-se aqui, o tipo de protocolo, o IP de origem e destino e as portas de origem e destino. Essas informações são primordiais para identificação dos fluxos de dados que sejam provenientes de publicidade no tráfego analisado.

2.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O presente capítulo demonstrou o estudo do referencial teórico para a total compreensão dos tópicos abordados.

Através deste foi possível entender o princípio de funcionamento das requisições web e o funcionamento das partes envolvidas nestas requisições, tais como ISP, e protocolos DNS, TCP e UDP. Além disso, também foi discutido o que é a publicidade online e os seus diversos tipos.

Por fim, o capítulo trouxe uma discussão sobre os bloqueadores de anúncios e os princípios básicos das ferramentas para captura e análise de tráfego na internet.

3 PROCEDIMENTOS METODOLÓGICOS

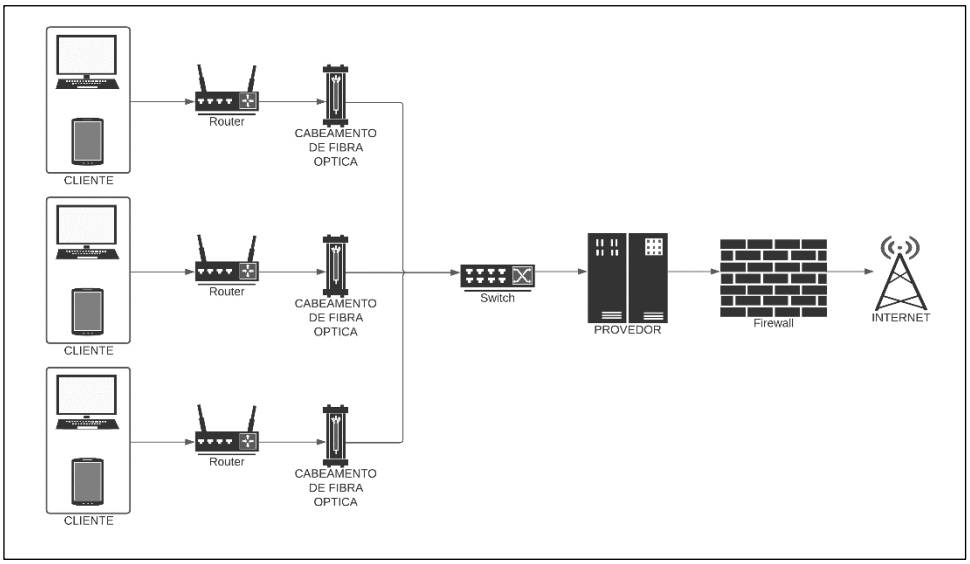
Com o objetivo de analisar o tráfego do ISP, a metodologia foi elaborada em etapas, que estão dispostas da seguinte maneira: ambiente de pesquisa, a qual descreve de onde foram obtidos os dados; procedimento experimental, a qual menciona o passo a passo na análise dos dados obtidos, com o intuito de viabilizar a última etapa; e por fim o algoritmo, o qual automatiza a análise em larga escala e permite modular que tipo de informação buscamos.

3.1 AMBIENTE DA PESQUISA

O estudo de análise foi realizado em conjunto com a empresa Zlink.net, localizada em Garuva - SC, a qual disponibilizou os dados para estudo.

A Figura 10 mostra a topologia de rede do ISP de maneira simplificada. Os modelos dos equipamentos serão omitidos devido a política da empresa.

Figura 10 - Topologia ISP



Fonte: os autores (2021).

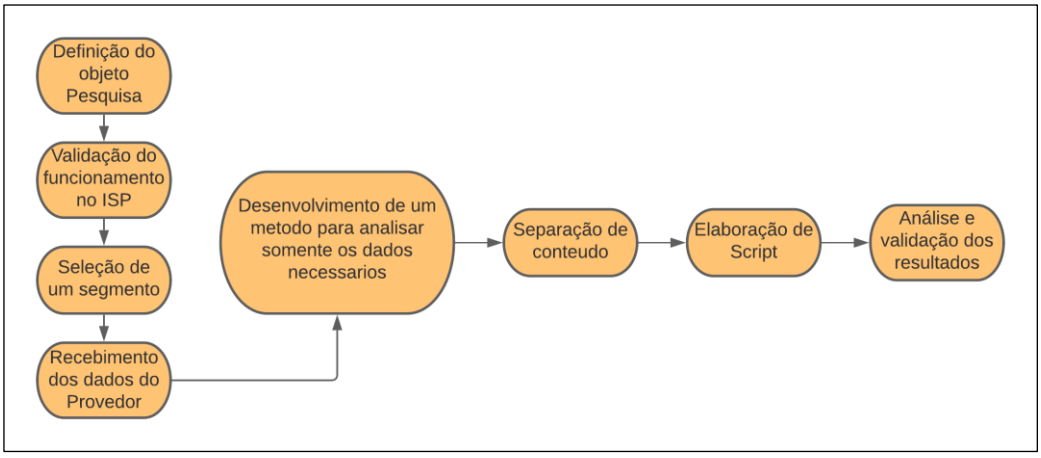
Conforme ilustrado pela Figura 10 a topologia possui em uma de suas extremidades os clientes que realizam os seus acessos e requisições. As requisições destes acessos são direcionadas a um concentrador que consolida as fibras óticas e direciona os dados para o switch local da empresa. Por fim as requisições passam por um *firewall* que filtra as requisições, e possibilita o armazenamento dos dados que trafegaram na rede. As requisições validas, são direcionadas para a Internet.

Os dados coletados correspondem ao tráfego de internet de um segmento da rede que representa um ponto de acesso regional a um bairro da cidade, este ponto se manteve com uma média de 33 equipamentos ativos no período de 13 a 27 de fevereiro de 2021. Este tráfego de rede foi capturado através da ferramenta de captura de pacotes disponível no PAN-OS, e salvo em arquivos do tipo PCAP, o que possibilitou uma análise mais detalhada dos dados, inclusive, através da ferramenta Wireshark.

3.2 ETAPAS DA PESQUISA

O ciclo inicial da pesquisa, iniciou-se com a identificação e definição do escopo da pesquisa, uma etapa onde foi definido qual seria o objetivo do projeto, seguiu-se estudando e identificando o funcionamento do ISP para que fosse necessário encaminhar de maneira mais efetiva a análise. Tendo em vista que não seria possível realizar uma análise geral de todo o ISP, pois ocorreria grande acúmulo de tráfego de dados por se tratar de muitos clientes, foi selecionado apenas um segmento da rede. Neste segmento foi possível executar todos os testes e métodos para extração dos dados necessários para a posterior elaboração de um *script* capaz de estratificar os dados de propaganda e separá-lo do tráfego normal de forma automatizada. A Figura 11 ilustra as etapas mencionadas.

Figura 11 - Etapas da pesquisa



Fonte: os autores (2021).

3.3 PROCEDIMENTO EXPERIMENTAL

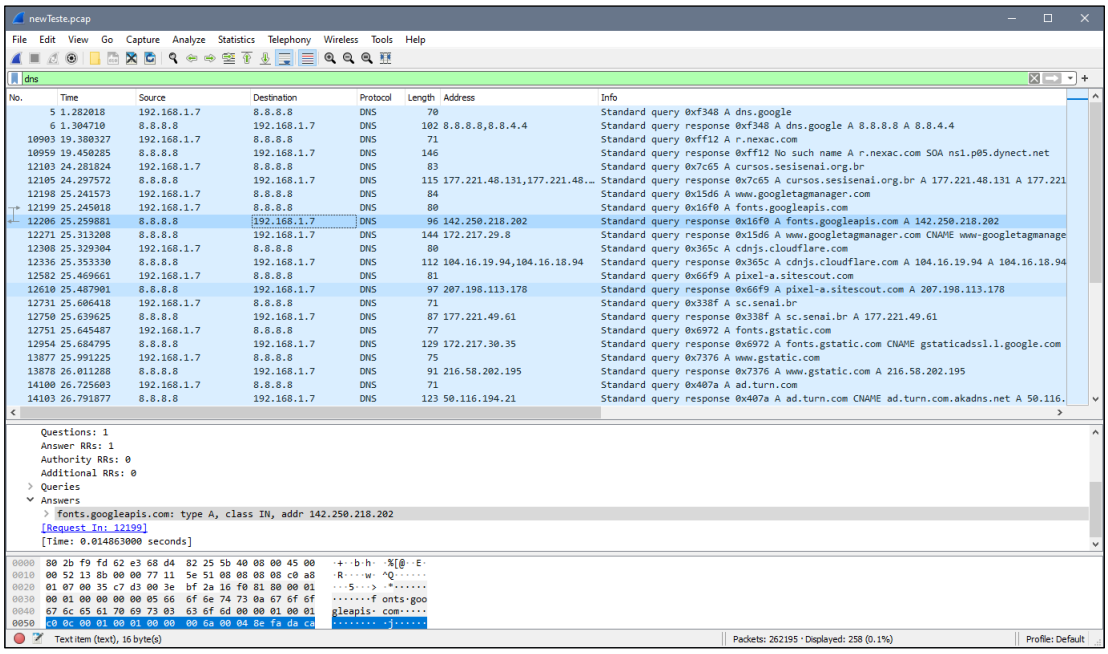
Com o uso do software Wireshark, foi possível realizar alguns procedimentos experimentais com o intuito de definir as etapas futuras para a realização da análise

do tráfego capturado anteriormente no ISP. Nesta etapa, serão descritos os procedimentos realizados para a extração e captura dos dados de rede.

3.3.1 Filtragem de dados:

A partir dos dados capturados pelo ISP, foi realizada a filtragem dos pacotes. A Figura 12 ilustra a tela do software Wireshark e lista todos os pacotes trafegados durante o período informado. É possível também visualizar o número do pacote em sequência, o tempo de chegada do pacote, o endereço IP de origem e destino, além de listar o protocolo utilizado na comunicação, tamanho do pacote e informações extras como por exemplo o endereço IP resolvido pelo DNS.

Figura 12 - Filtragem de dados



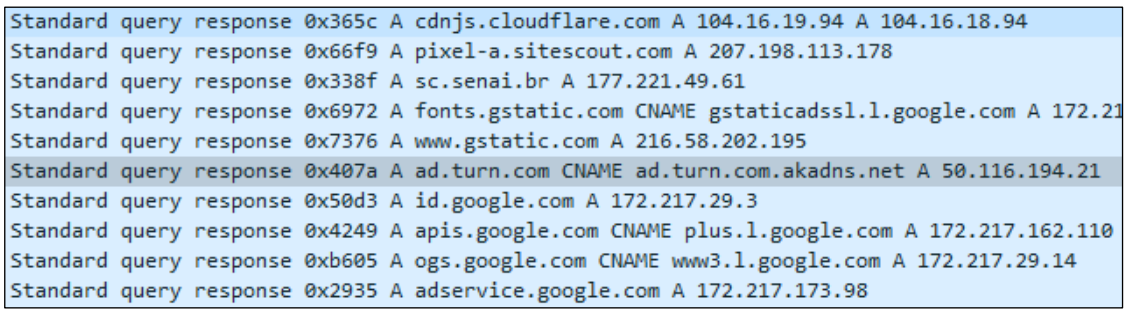
Fonte: os autores (2021).

Com o uso do sistema de filtros disponível no software, foi possível filtrar todos os pacotes transmitidos através do protocolo DNS *over* UDP e posteriormente analisar a forma de resposta obtida do servidor DNS. De acordo com o exposto na Seção 2.1.3, tal etapa permite identificar o domínio de origem e destino das requisições, o que, comparado a uma lista de serviços de propaganda torna possível a contabilização do tráfego originário destas fontes. Para realizar a identificação das propagandas, foi utilizada uma lista de domínios identificados previamente como “propaganda”. Esta lista é proveniente de um dos bloqueadores de anúncios mais populares na atualidade

e possui mais de 1 milhão de domínios conhecidos como propagandas (NICOLAS, 2019).

A Figura 13 mostra um fragmento da tela do Wireshark com algumas respostas obtidas do servidor DNS para o computador local, na qual é possível identificar o domínio resolvido e os endereços IPs que respondem a esse determinado domínio.

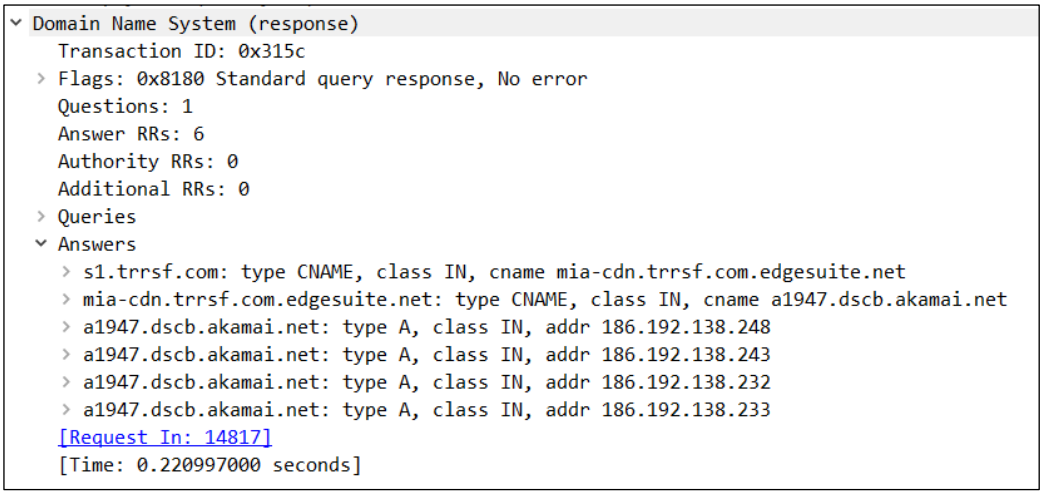
Figura 13 - Respostas DNS



Fonte: os autores (2021).

Com base nas respostas obtidas do DNS, é possível determinar o domínio que está sendo resolvido e todos os endereços IP que o domínio em questão responde. A Figura 14 ilustra um domínio o qual originou uma resposta com mais de um endereço IP. Podemos notar que o domínio “a1947.dscb.akaamai.net” responde para os endereços IPs “186.192.138.248”, “186.192.138.243”, “186.192.138.232” e 186.192.138.233”

Figura 14 - Domínios e IPs



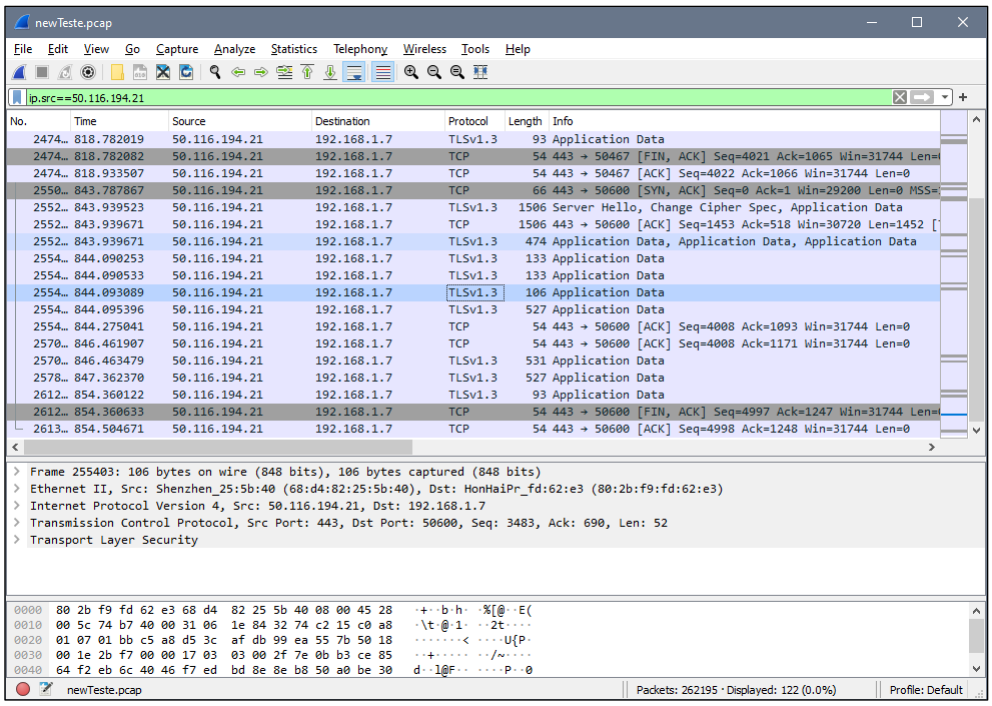
Fonte: os autores (2021).

Este tipo de resolução de endereços amplia consideravelmente as possibilidades de origem de tráfego, permitindo que um único domínio, tenha tráfego oriundo de um ou mais endereços IP.

3.3.2 Classificação dos dados

Uma vez realizada a captura e filtragem dos dados, faz-se necessária a classificação dos mesmos quanto ao tipo de informação trafegada. A Figura 15 mostra a realização da filtragem de pacotes baseado nos endereços IPs cujos domínios são conhecidos fornecedores de propaganda.

Figura 15 - Classificação dos dados



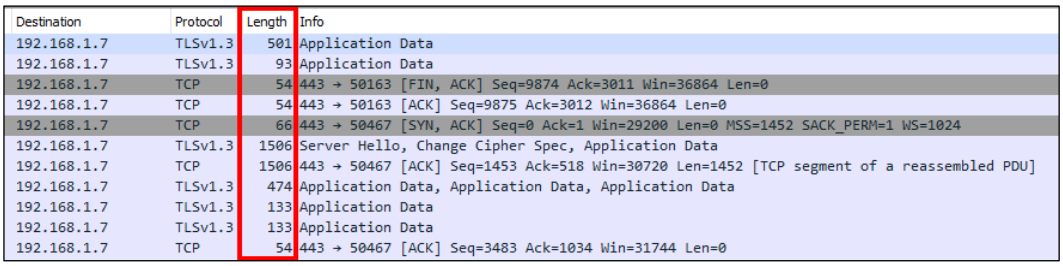
Fonte: os autores (2021).

Com o uso dos filtros disponíveis no software, foi possível a identificação dos pacotes oriundos da rede externa com destino à rede interna dos clientes (*download*). Permitindo assim a classificação dos dados entre tráfego normal e propagandas, para a posterior consolidação dos dados.

3.3.3 Consolidação dos dados

Após a aplicação dos filtros necessários, obtém-se o resultado mostrado na Figura 16, no qual é possível identificar nas abas do software, o destinatário dos pacotes (equipamento local), o protocolo de comunicação, o tamanho do pacote e informações extras.

Figura 16 - Tamanho dos pacotes



Destination	Protocol	Length	Info
192.168.1.7	TLSv1.3	501	Application Data
192.168.1.7	TLSv1.3	93	Application Data
192.168.1.7	TCP	54	443 → 50163 [FIN, ACK] Seq=9874 Ack=3011 Win=36864 Len=0
192.168.1.7	TCP	54	443 → 50163 [ACK] Seq=9875 Ack=3012 Win=36864 Len=0
192.168.1.7	TCP	66	443 → 50467 [SYN, ACK] Seq=0 Ack=1 Win=29200 Len=0 MSS=1452 SACK_PERM=1 WS=1024
192.168.1.7	TLSv1.3	1506	Server Hello, Change Cipher Spec, Application Data
192.168.1.7	TCP	1506	443 → 50467 [ACK] Seq=1453 Ack=518 Win=30720 Len=1452 [TCP segment of a reassembled PDU]
192.168.1.7	TLSv1.3	474	Application Data, Application Data, Application Data
192.168.1.7	TLSv1.3	133	Application Data
192.168.1.7	TLSv1.3	133	Application Data
192.168.1.7	TCP	54	443 → 50467 [ACK] Seq=3483 Ack=1034 Win=31744 Len=0

Fonte: os autores (2021).

A parte em destaque na Figura 16, refere-se aos tamanhos dos pacotes obtidos, que são representados em bytes. Com a soma dos dados desses pacotes, é possível calcular o tráfego oriundo dos filtros aplicados. O intuito da aplicação dos filtros é poder retirar uma margem percentual dos dados referentes ao consumo de propagandas. Isto pode ser feito através da extração dos dados totais de propagandas dividido pelos dados totais obtidos.

Visto a grande quantidade de pacotes obtido provenientes do ISP, foi identificada a necessidade de criação de um *script* que auxiliasse e automatizasse a contabilização destes dados.

3.4 *SCRIPT* DE CONSOLIDAÇÃO

O *script* foi elaborado utilizando a linguagem de programação Python, em conjunto com a biblioteca DPKT[3], que permite a leitura de arquivos no formato PCAP, ideal para buscas e análises em capturas TCP/IP.

O *script* foi elaborado em três etapas, sendo a primeira a responsável pelo carregamento de todos os dados, a segunda etapa realiza a filtragem dos pacotes e identifica quais domínios com propagadas foram acessados e por fim, é realizado a leitura de todo o tráfego capturado, respectivamente as linhas de código 147, 148 e 149 conforme ilustrado na Figura 17.

[3] Disponível em: https://github.com/kbandla/dpkt

Figura 17 - Etapas do *Script*

Fonte: os autores (2021).

3.4.1 Carregar dados

A primeira etapa do *script* é responsável por realizar a leitura de todos os dados que são utilizados durante a execução. Primeiramente o *script* carrega a lista de domínios com propagandas, e salva todas as ocorrências em uma lista de *strings*. Foi utilizado uma listagem contendo 1.335.182 (um milhão, trezentos e trinta e cinco mil, cento e oitenta e dois) domínios identificados como provedores de propagadas.

A realização da leitura do arquivo PCAP é representada na Figura 18 através das linhas 92 a 95.

Figura 18 - Leitura dos dados

Fonte: os autores (2021).

O arquivo ilustrado na Figura 18 como “dados.pcap” contém todo o tráfego capturado anteriormente. Conforme é realizado a leitura do arquivo PCAP, a segunda etapa do algoritmo é executada em paralelo, poupando tempo de análise.

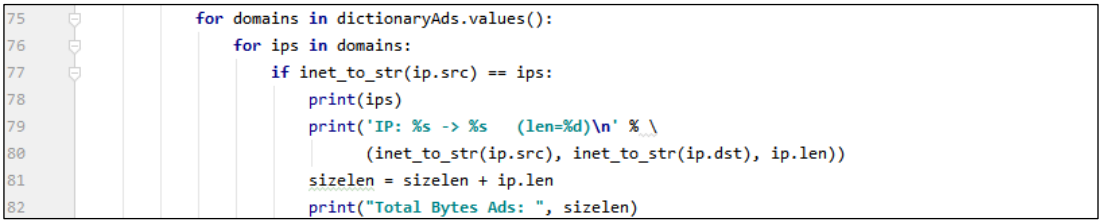
3.4.2 Filtragem dos dados

A segunda etapa do *script* é responsável pela filtragem dos dados carregados. Esta etapa é executada em paralelo com a etapa anterior.

Conforme o *script* realiza a leitura do arquivo PCAP, é realizada a análise das respostas DNS. Estas respostas trafegam sobre o protocolo UDP, permitindo assim o acesso aos domínios e endereços IPs que foram resolvidos pelo servidor DNS. Para cada pacote, é realizada uma verificação se o domínio resolvido está presente na lista de domínios identificados como provedores de propagadas. Caso este domínio seja reconhecido como uma propaganda, os seus endereços IPs e seu endereço de domínio são salvos num dicionário do Python. Por fim, conforme ilustra a Figura 19, é

realizada uma busca em todo o arquivo PCAP a fim de determinar a quantidade de propagandas trafegado.

Figura 19 - Filtragem dos dados



```
75          for domains in dictionaryAds.values():
76              for ips in domains:
77                  if inet_to_str(ip.src) == ips:
78                      print(ips)
79                      print('IP: %s -> %s   (len=%d)\n' % \
80                            (inet_to_str(ip.src), inet_to_str(ip.dst), ip.len))
81                      sizelen = sizelen + ip.len
82                      print("Total Bytes Ads: ", sizelen)
```

Fonte: os autores (2021).

O trecho de código exibido na Figura 19, é um segmento do *script* desenvolvido e utilizado na filtragem dos dados coletados anteriormente. A linha 75 indica o início da estrutura de repetição *for*, que irá percorrer todos os itens do dicionário (*dictionaryAds*), nesta etapa, são os domínios salvos anteriormente. Na linha 76, temos uma nova estrutura de repetição, que irá percorrer todos os possíveis endereços IPs que cada domínio possui e buscar em todo o tráfego proveniente destes IPs. Todo o tráfego é identificado pela linha 77 e os valores trafegados salvos numa variável acumuladora (*sizelen*), disposta na linha 81. Após a execução, é disponibilizado o total em bytes identificado.

3.4.3 Tráfego completo

A última etapa do algoritmo consiste na leitura de todo o arquivo PCAP e identificação da quantidade de dados que trafegou da internet para o cliente (*download*). A Figura 20 apresenta parte do *script* utilizado para a leitura do tráfego completo. Na linha 41, há uma operação condicional que valida se o pacote analisado possui a sua origem um endereço IP externo e salva os valores do tráfego numa variável incrementável, conforme linha 44. Se faz também necessário a substituição da *string* “IPLOCAL” pelo endereço IP local do equipamento que está sendo analisado. Endereço foi obtido com uma simples análise individual dos arquivos disponibilizados pelo provedor de internet.

Figura 20 - Tráfego completo

```
41              if (inet_to_str(ip.src) != 'IPLOCAL'):
42                  print('IP: %s -> %s   (len=%d)\n' % \
43                      (inet_to_str(ip.src), inet_to_str(ip.dst), ip.len))
44                  sizelen = sizelen + ip.len
45          print("Total Bytes Trafegados: ", sizelen)
```

Fonte: os autores (2021).

Ao final da execução é possível realizar uma análise comparativa entre a quantidade de tráfego de propagandas e o tráfego total.

3.4.4 Armazenamento da base de propagandas

Devido a grande quantidade de dados analisada, se fez necessário a divisão de todo o tráfego em diversos arquivos individuais. Durante está a análise o *script* foi executado isoladamente para cada arquivo, sendo uma execução para cada dia analisado.

Com o intuito de preservar a base de propagandas que foram identificadas nos dias anteriores a execução, a base de dados é salva utilizando as linhas de código ilustradas na Figura 21.

Figura 21 - Armazenamento das fontes de propagandas

```
97          with open('dados.p', 'rb') as fp:
98              dictionaryAds = pickle.load(fp)
155     with open('dados.p', 'wb') as fp:
156         pickle.dump(dictionaryAds, fp, protocol=pickle.HIGHEST_PROTOCOL)
```

Fonte: compilação dos autores[4] (2021).

As linhas de código 97 e 98, referem-se a leitura da base de dados já existente. Estas linhas de código, são executadas em conjunto com o carregamento de dados (verificar seção 3.4.1. Carregar dados) e salvam a base de dados existente num dicionário do Python. As linhas 155 e 156 ilustram as linhas de códigos utilizadas para salvar a base de dados num arquivo denominado “dados.p”. Foi utilizado o modulo Pickle[5], disponível no Python 3.5. Este modulo, implementa protocolos binários para serialização e desserialização de uma estrutura Python (PYTHON SOFTWARE FOUNDATION, 2021).

[4] Montagem realizada a partir de imagens coletadas do *script* elaborado pelos autores.

[5] Disponível em : https://docs.python.org/3/library/pickle.html

3.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo, apresentou todo o processo metodológico utilizado nesta pesquisa.

Foram expostas as etapas iniciais para a definição do tema da pesquisa e viabilidade de análise junto ao ISP. Após isso, já com uma base de dados foi possível elaborar um plano de execução da análise desta pesquisa com o uso do software Wireshark, tornando possível a visualização e busca manual do tráfego de propagandas, utilizando os endereços IP das resoluções DNS como fonte de busca aliado a uma base de domínios já reconhecidos como provedores de propagandas.

Identificado o grande volume de dados trafegados no período analisado, se viu necessário a automatização dessa análise com o uso de um *script*. Este foi elaborado utilizando a linguagem Python em conjunto com a biblioteca DPKT, que permitiu a análise completa dos arquivos disponibilizados pelo ISP. O *script* foi dividido em três partes principais, sendo o carregamento de dados, a filtragem dos dados e a leitura de todo o tráfego. Foi também necessária a utilização da biblioteca Pickle, que permitiu a reutilização do tráfego identificado como propaganda, de diferentes dias sem que houvesse a necessidade de uma resposta DNS para um determinado domínio.

4 RESULTADOS OBTIDOS

Este capítulo irá apresentar os resultados que foram obtidos através da análise do tráfego de dados dos clientes da empresa Zlink.net.

Será apresentado o volume de dados trafegado referente a publicidade online durante o período, além de explicar como foi realizado o cálculo dos dados apresentados.

4.1 PREPARAÇÃO DA BASE DE DADOS

A completa execução da análise desta pesquisa, foi possível após a preparação da base de dados para a sua análise através do *script*. Durante a elaboração do *script*, foi constatada a impossibilidade de análise de pacotes do protocolo QUIC devido a cifragem do tráfego presente neste protocolo. Para tal análise, a medição aqui realizada não é eficiente e foi desconsiderada, podendo ser abordada em trabalhos futuros. Foi realizada a remoção de todos os pacotes trafegados no protocolo QUIC dos arquivos analisados. Maiores detalhes sobre as razões da remoção deste protocolo serão discutidos no decorrer desta seção.

A base de dados disponibilizada pelo ISP continha 15 arquivos no formato PCAP de tamanhos variados. Cada um destes arquivos teve em média 103,2 GB por dia analisado. A somatória de todos os arquivos, chegou a aproximadamente 1,548 TB de tráfego.

Devido a quantidade de arquivos com grandes tamanhos, foi necessário executar o *script* de análise em dois dispositivos com o objetivo de otimizar o tempo de análise dos mesmos, o Quadro 1 especifica as configurações de cada equipamento. Ambos os dispositivos possuem configurações equiparáveis, por conta disso, a diferença de tempo de processamento entre os dispositivos será irrelevante para essa análise.

Quadro 1 - Configurações dos dispositivos

Identificação	Dispositivo 01	Dispositivo 02
Processador	Intel Core i5-8250U @1.80GHz	AMD Ryzen 3 1200 3.10 GHz
Memória RAM	8,00 GB	8,00 GB
Sistema operacional	Linux Kernel versão 5.8, GNOME 3.38, POP!_OS 20.10	Windows 10 Pro

Fonte: os autores (2021).

O tempo médio de execução do *script* foi de aproximadamente 23 horas para cada dia analisado, o que representa aproximadamente 7 GB por hora de execução do *script*. O tempo total de execução do *script* foi de aproximadamente 349 horas. O Quadro 2 detalha o tempo de execução do *script* em horas, o tamanho do arquivo para cada dia analisado em GBs referente ao tempo de execução.

Quadro 2 - Dados de execução

Dia de Fevereiro	**Tempo de Execução**	**Tamanho do arquivo**
13	25	101,00
14	29	144,55
15	21	109,50
16	24	103,19
17	30	118,02
18	23	108,33
19	22	99,46
20	26	111,23
21	29	125,97
22	21	96,18
23	19	84,55
24	16	75,45
25	19	87,53
26	21	83,25
27	23	100,70
	Desvio Padrão: 4,01	**Desvio Padrão:** 17,7
	Total: 349	**Total:** 1548

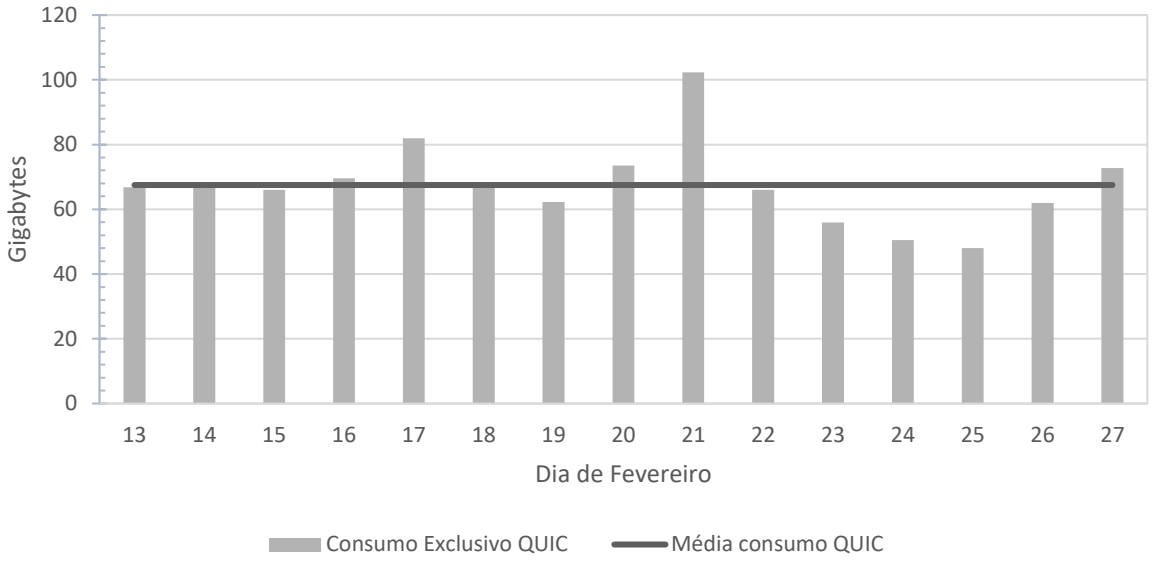
Fonte: os autores (2021).

4.2 VOLUME DE DADOS

Foi realizada a execução do *script* para a análise dos arquivos disponibilizados pelo ISP. Primeiramente foi realizada a medição de todo o tráfego capturado. Verificou-se que a quantidade em GBs trafegado em cada dia do período analisado

chegou à uma média de 178,11 GB de tráfego geral por dia, sendo destes 11,3% referente a propagandas. Considerando que os percentuais se mostraram muito abaixo dos resultados apresentados pelo trabalho de Siqueira (2019), o qual constatou que o consumo de tráfego por propagandas é de aproximadamente 31%, analisou-se qual seria o motivo relacionado à essa diferença. Após análise, verificou-se que boa parte dos pacotes trafegados, são do protocolo QUIC. O Gráfico 1 ilustra o tráfego exclusivo no protocolo QUIC no período analisado. Pode-se notar que a quantidade de GBs trafegados exclusivamente pelo protocolo QUIC é bastante elevado, sendo consumido em média 67,5 GBs por dia.

Gráfico 1 - Consumo com protocolo QUIC

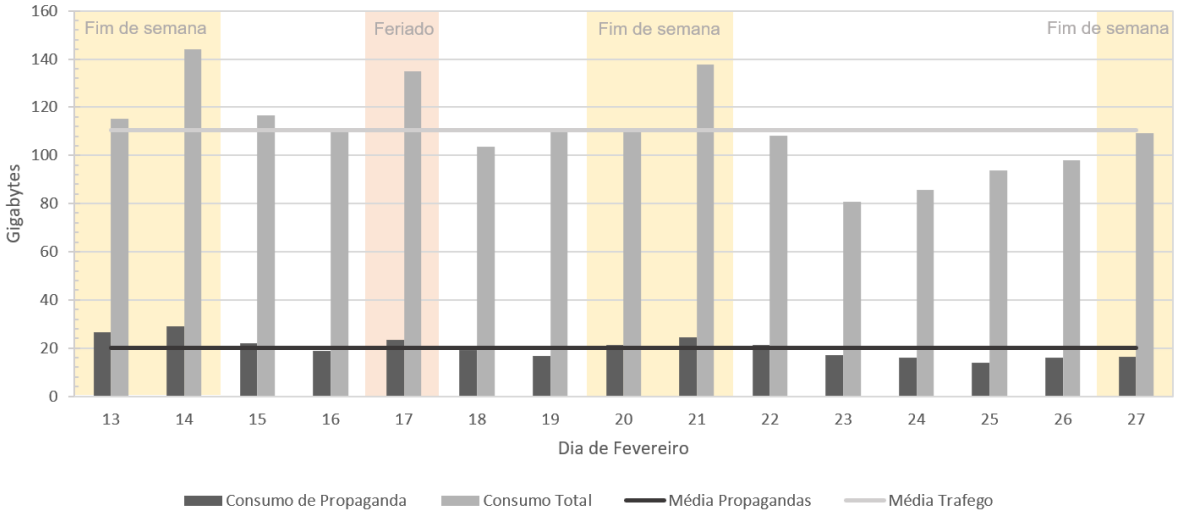


Fonte: os autores (2021).

Por ser tratar de um protocolo cifrado, impossibilitando a identificação das resoluções *DNS over QUIC*, decidiu-se por realizar a medição desconsiderando tal protocolo. Por conta disso, grande parte do consumo relacionado a vídeos, não foi considerado na análise, uma vez que a maioria dos domínios de provedores de vídeos (por exemplo Youtube) são resolvidos através do protocolo QUIC (ALLEVEN, 2017). Outra ressalva que se faz, é que propagandas oriundas do mesmo endereço do conteúdo, não são identificadas como propagandas e sim como tráfego normal.

Foi realizada uma nova análise desconsiderando o protocolo QUIC no tráfego total, e chegou-se aos resultados apresentados no Gráfico 2. Chegou-se a uma média de tráfego de 110 GB por dia analisado. Obteve-se também uma média de 20,2 GB de consumo de propagandas, o que representou 18,2% do tráfego diário analisado.

Gráfico 2 - Consumo de Propagandas sem QUIC



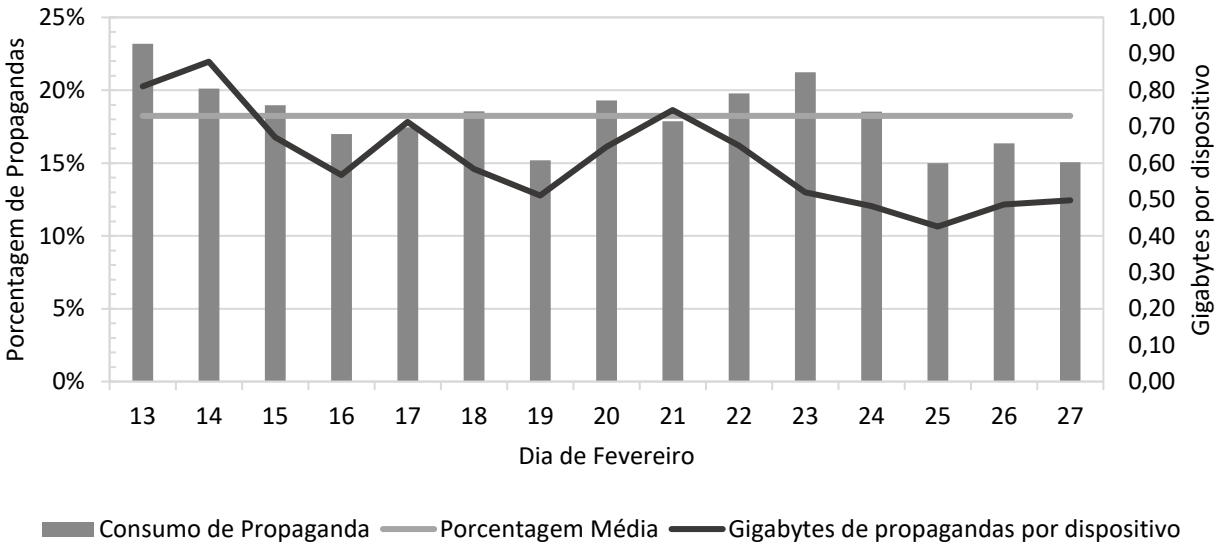
Fonte: os autores (2021).

O consumo total de dados se mostrou-se acima da média geral, em apenas 6 dias de análise (dias 13, 14, 15, 17, 19 e 21), sendo que os dias 13, 14 e 21 são dias de fim de semana. O consumo de propagadas excedeu a média geral em 8 dias de análise (dias 13, 14, 15, 18, 20, 22, 23 e 24).

Notou-se um pequeno aumento no consumo geral de dados trafegados nos domingos (aumento de 27,3%) e feriados (aumento de 21,8%) quando se comparado a média geral. No consumo de propagandas, ocorreu aumento nos domingos (aumento de 42,5%) e feriados (aumento de 16,4%). Não foi possível identificar se estes números se tratam de um padrão de consumo no decorrer do mês devido ao período de análise disponibilizado pelo ISP não abranger períodos maiores ao aqui apresentando. Uma análise estatística do consumo destes dados num período maior será considerada para trabalhos futuros.

O consumo por unidade dispositivo analisado, se mostrou constante durante os testes. A média de consumo ficou em aproximadamente 3,35 GB de dados por dia e o desvio padrão foi de 0,544 GB. A média de propagandas por dispositivo é apresentada no Gráfico 3 em conjunto com a porcentagem de propagandas. A média de propagandas ficou em aproximadamente 610 MB por dia em cada dispositivo analisado, havendo picos nos dias 13, 14, 17 e 21 com desvio padrão de 133 MB.

Gráfico 3 - Consumo por dispositivo

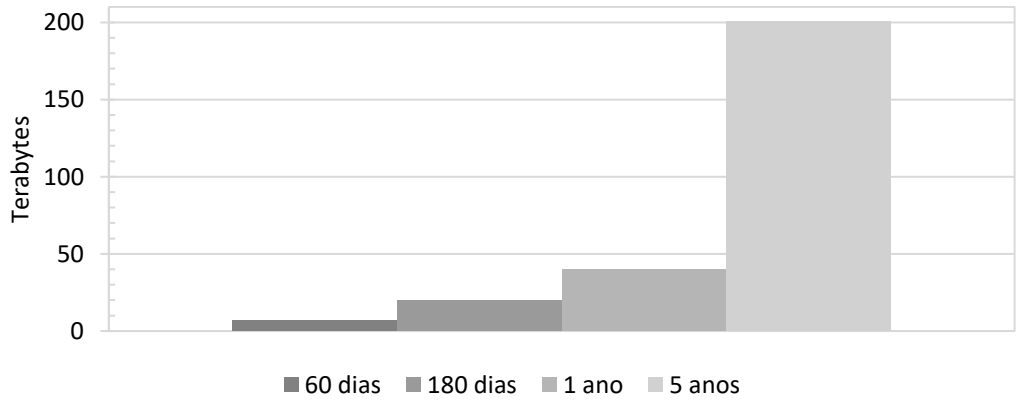


Fonte: os autores (2021).

4.3 ESTIMATIVAS

Este capítulo trata-se de projeções futuras com relação à um longo período de análise, como forma de visualização de resultados em larga escala. São dados estimados com base nos resultados obtidos no trabalho. O Gráfico 4, representa a estimativa de análise de dados desconsiderando o protocolo QUIC no tráfego total e utilizando a mesma população de análise (33 dispositivos). Considerou-se realizar estimativas em períodos de 60 dias (2 meses), 180 dias (6 meses), 365 dias (1 ano) e 1825 dias (5 anos).

Gráfico 4 – Estimativa de consumo



Fonte: Os autores (2021)

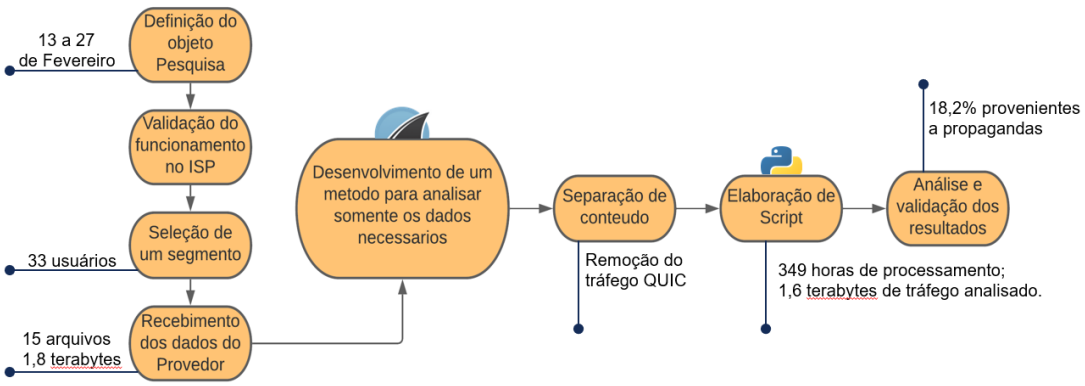
Obteve-se como resultado uma estimativa de consumo no período de 60 dias, um tráfego médio de 6,6 TBs, em 180 dias uma utilização média de 19,8 TBs, em 1 ano um consumo médio de 39,6 TBs e em 5 anos aproximadamente 200 TBs de dados relacionados a propagandas.

4.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo apresentou uma avaliação sobre o consumo de propagandas numa determinada população da cidade de Garuva – SC. Esta avaliação foi realizada para verificar o quanto o consumidor final é impactado por um tráfego nem sempre desejado.

A Figura 22 indica o resultado das etapas da pesquisa. Realizamos primeiramente a definição do objetivo e períodos da pesquisa, o período de análise ficou estipulado em 13 a 27 de fevereiro de 2021. Em seguida, realizamos a validação do funcionamento do ISP e a viabilidade da coleta dos dados. Selecionamos um segmento da rede que seria analisado, definimos como 33 clientes. Recebemos do ISP um total de 15 arquivos, cada um referente a um dia do período da pesquisa, totalizando aproximadamente 1,8 TBs. Realizamos com a ajuda do *software* Wireshark, uma filtragem experimental, com o intuito de identificar as informações que seriam mais relevantes para a nossa pesquisa. Nesta etapa de filtragem experimental, identificamos uma limitação da desta pesquisa, que foi a impossibilidade de análise de tráfegos que utilizem *DNS over QUIC*, devido a cifragem utilizada pelo protocolo. Foi realizada a separação dos dados disponibilizados pelo ISP, removendo todo o tráfego proveniente do protocolo QUIC.

Figura 22 - Resultado das etapas da pesquisa



Fonte: Os autores (2021)

Após a elaboração do *script,* foi realizado a execução do mesmo em dois dispositivos simultaneamente, com o intuito de otimizar o tempo de análise total. O tempo de execução total de toda a análise foi de aproximadamente 359 horas com um total de 1548 TB analisados.

Após realizada a análise de todos os arquivos disponibilizados, chegou-se a um resultado da medição de tráfego de propagandas. Notou-se que o consumo exclusivo de propagandas se mostrou bastante elevado, considerando que este tipo de conteúdo, normalmente não é de interesse do consumidor final. De todo o tráfego analisado, em média 18,2% estão relacionados a publicidade na internet. Em linhas gerais, isto significa que de todo o acesso realizado a internet por um cliente, quase um quinto é referente a propaganda.

5 CONCLUSÃO

Com o crescimento da internet, a propagação de propagandas se tornou um dos mercados mais rentáveis atualmente, sendo a principal forma de negócio de diversas empresas principalmente do ramo da tecnologia. Esta forma de monetização, nem sempre possui um conteúdo que é de interesse do consumidor final, este que acaba sendo onerado com o tráfego indesejado de publicidades em páginas da internet, aplicativos, vídeos etc. Por um lado, os provedores de conteúdo oferecem um serviço “gratuito”, por outro, para oferecer esse serviço, eles são suportados com o oferecimento de propagandas. Este trabalho então, teve por objetivo avaliar o quanto do tráfego acessado por clientes de um ISP, são relativos à propaganda.

O primeiro passo da execução desta pesquisa foi a obtenção dos arquivos referentes de tráfego de um ISP. Após diversos testes, foi elaborado um *script* automatizador para viabilizar a análise do tráfego capturado. Por fim, com a união do *script* e dos arquivos disponibilizados pelo ISP, foi possível mensurar a quantidade aproximada de publicidades consumidas por clientes de um determinado provedor da cidade de Garuva – SC.

Após a execução do *script*, concluiu-se que, em média, 18,2% de todo o conteúdo acessado na internet é relacionado a publicidade. Durante o período analisado, a média de consumo de publicidade diária é de aproximadamente 610 MB por equipamento/residência. Ainda, percebeu-se um aumento de consumo em dias específicos da semana, tais como domingos (aumento de 42,5% no tráfego total) e feriados (aumento de 16,4% no tráfego total), contudo o valor proporcional do tráfego propagandas se manteve dentro dos desvios da média.

Com base no tema aqui analisado, são propostos alguns trabalhos futuros para continuidade desta pesquisa. A primeira proposta seria consultar o tráfego de acesso de uma base de consumidores maior com o objetivo de busca de possíveis padrões de consumo. A segunda proposta, sugere analisar o consumo total de propagandas com a inclusão do protocolo QUIC ou outros protocolos de DNS cifrado. Por fim, uma terceira proposta, seria realizar a análise de consumo de dados relacionados a propagadas em um ambiente corporativo.

REFERÊNCIAS

ADBLOCK PLUS. Adblock Plus filters explained. **Adblock Plus**. Disponivel em: <https://adblockplus.org/filter-cheatsheet>. Acesso em: 24 Outubro 2020.

ADGUARD. How ad blocking works: the might behind the magic. **Adguard**, 2017. Disponivel em: <https://adguard.com/pt_br/blog/how_adblocking_works.html>. Acesso em: 24 Outubro 2020.

ALLEVEN, M. YouTube driving more QUIC-based traffic on mobile: Vasona. **Fierce Wireless**, 2017. Disponivel em: <https://www.fiercewireless.com/wireless/youtube-driving-more-quic-based-traffic-mobile-vasona>. Acesso em: 22 maio 2021.

ALVAREZ, M.; GOMES, V. Publicidade na Web. **Papo Nada Furado**, 2010. Disponivel em: <https://paponadafurado.webnode.com.br/seminarios/publicidade-na-web/>. Acesso em: 03 Outubro 2020.

AN, M. Why People Block Ads (And What It Means for Marketers and Advertisers). **Hubspot**, 2016. Disponivel em: <https://blog.hubspot.com/marketing/why-people-block-ads-and-what-it-means-for-marketers-and-advertisers>. Acesso em: 17 Outubro 2020.

BAN, S. J. et al. Implementation of IEEE 802.15.4 Packet Analyzer. **International Journal of Computer, Electrical, Automation, Control and Information Engineering**, 2007.

BARBOSA, K. R. D. S.; PEREIRA, S. J. E. **Análise Passiva do Tráfego DNS da Internet Brasileira**. UFAM. Manaus. 2009.

CERF, V. G.; KAHN, R. E. A Protocol for Packet Network Intercommunication. **IEEE Transactions on Communications Technology**, 22, Maio 1974. 637 - 648.

CISCO. Cisco Visual Networking Index Predicts Global Annual IP Traffic to Exceed Three Zettabytes by 2021. **Cisco**, 2017. Disponivel em: <https://newsroom.cisco.com/press-release-content?articleId=1853168>. Acesso em: 07 novembro 2020.

COPES, F. How HTTP requests work. **Flaviocopes**, 2018. Disponivel em: <https://flaviocopes.com/http-request/#the-http-protocol>. Acesso em: 08 Novembro 2020.

ENBERG, J. What’s Shaping the Digital Ad Market. **eMarketer**, 2019. Disponivel em: <https://www.emarketer.com/content/global-digital-ad-spending-2019>. Acesso em: 04 Outubro 2020.

FEDERAL NETWORKING COUNCIL. Definition of "Internet". **nitrd**, 1995. Disponivel em: <https://www.nitrd.gov/fnc/internet_res.pdf>. Acesso em: 20 Setembro 2020.

FERNANDES, J. P. The TCP 3-Way HandShake. **JohnpFernandes**, 2018. Disponivel em: <https://www.johnpfernandes.com/2018/12/08/the-tcp-3-way-handshake/>. Acesso em: 09 Novembro 2020.

FREBERG, K. et al. Who are the social media influencers? A study of public perceptions of personality. **Fuel Energ Abstr**, Março 2011. Acesso em: 03 Outubro 2020.

IAB, PWC. **Internet advertising revenue report**. Pwc. [S.l.]. 2020.

KUROSE, J.; ROSS, K. **Redes de Computadores e a Internet:** Uma Abordagem Top-Down. 7ª. ed. [S.l.]: Pearson Education - Brazil, 2016.

LEINER, B. M. et al. **Brief History of the Internet**. [S.l.]: Computer Communication Review, v. 39, 2009. Disponivel em: <https://www.internetsociety.org/internet/history-internet/brief-history-internet/>. Acesso em: 20 Setembro 2020.

LEONARDI, M. **Internet:** elementos fundamentais. in Responsabilidade Civil na Internet e nos demais meios de comunicação. 2ª. ed. São Paulo: Saraiva, 2012.

MCMILLAN, S. J. **Internet Advertising:** One Face or Many? [S.l.]: [s.n.], 2004.

NICOLAS. **PiHole**, 2019. Disponivel em: <https://github.com/nicholasb2101/PiHole>. Acesso em: 3 Abril 2021.

OUR WORLD IN DATA. Internet users by world region. **OurWorldInData**, 2016. Disponivel em: <https://ourworldindata.org/grapher/internet-users-by-world-region>. Acesso em: 03 Outubro 2020.

PAGEFAIR. Growth of the Blocked Web. **Blockthrough**, 2020. Disponivel em: <https://blockthrough.com/2020/02/06/2020-adblock-report-3/>. Acesso em: 20 Outubro 2020.

PALANT, W. Filtering HTML code in Adblock Plus. **AdblockPlus**, 2008. Disponivel em: <https://adblockplus.org/blog/filtering-html-code-in-adblock-plus>. Acesso em: 18 Novembro 2020.

PARETA, J. S.; UROZ, I. C.; ROS, P. B. Demystifying content-blockers: A large scale study of actual performance gains, Barcelona, 05 Fevereiro 2020.

PRAMATAROV, M. Why does DNS use UDP? **Cloudns**, 2018. Disponivel em: <https://www.cloudns.net/blog/dns-use-udp/>. Acesso em: 10 Novembro 2020.

PYTHON SOFTWARE FOUNDATION. pickle — Python object serialization. **Python**, 2021. Disponivel em: <https://docs.python.org/3/library/pickle.html>. Acesso em: 30 abril 2021.

RICHARDS, J. I.; CURRAN, C. M. **Oracles on "Advertising":** Searching for a Definition. 2ª. ed. [S.l.]: [s.n.], v. XXXI, 2002.

SIQUEIRA, A. F. M. F. **Quanto eu pago pelos teus anúncios? Uma análise de consumo de bytes por propaganda no acesso a páginas web**. Centro universitário Sociesc. Joinville, p. 50. 2019.

STATT, N. YouTube is a $15 billion-a-year business, Google reveals for the first time. **TheVerge**, 2020. Disponivel em: <https://www.theverge.com/2020/2/3/21121207/youtube-google-alphabet-earnings-revenue-first-time-reveal-q4-2019>. Acesso em: 18 Novembro 2020.

TREVISAN, M. et al. Five Years at the Edge: Watching Internet From the ISP Network. **IEEE/ACM Transactions on Networking**, 30 Janeiro 2020. 561 - 574.

WHATISMYIPADDRESS. What is an Internet Service Provider? **WhatisMyIpAddress**, 2018. Disponivel em: <https://whatismyipaddress.com/isp>. Acesso em: 18 Novembro 2020.

WIRESHARK. About WS. **Wireshark**, 2020. Disponivel em: <https://www.wireshark.org/index.html#aboutWS>. Acesso em: 03 Novembro 2020.

GLOSSÁRIO

Banner:	Curta mensagem publicitária em um site da Internet, com link para a página do anunciante.
Cross-platform:	Palavra de origem inglesa que significa ferramenta disponível em múltiplas plataformas.
Flag:	Item utilizado para indicar um estado particular da conexão ou prover algumas informações adicionais.
Full-duplex:	Capacidade de se comunicar de forma bidirecional.
Host:	Arquivo utilizado pelo sistema destinado a relacionar endereços IP com domínio.
Hyperlink:	Referência a outro documento ou site.
Pop-up:	Tipo de janela que se abre ao visitar uma página *web*.
Script:	Conjunto de instruções para que uma função seja executada em determinado ambiente.
Three-way handshake:	Palavra de origem inglesa que significa cumprimento de três etapas.