UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas Justino Resende Teixeira

Investigação da *darknet* como fonte de dados para ciberataques

Uberlândia, Brasil

2022

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas Justino Resende Teixeira

Investigação da *darknet* como fonte de dados para ciberataques

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Rodrigo Sanches Miani

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2022

Lucas Justino Resende Teixeira

Investigação da *darknet* como fonte de dados para ciberataques

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Trabalho aprovado. Uberlândia, Brasil, 01 de abril de 2022:

Prof. Dr. Rodrigo Sanches Miani
Orientador

Prof. Dr. Paulo Henrique Ribeiro Gabriel
Professor

Prof. Dr. Ivan da Silva Sendin
Professor

Uberlândia, Brasil
2022

Resumo

A *Dark Web* é uma grande fonte de dados, sejam eles legais ou ilegais. Apesar disso, a navegação na *Darknet* não é tão simples assim devido à falta de padrões em seus links, o que dificulta a localização dos serviços fornecidos. Além disso, pessoas utilizam da sua anonimidade para questionar sem ter medo de serem questionadas. Por isso, é válido pensar na análise de comportamento das pessoas diante a anonimidade, tentar identificar ameaças em seu estágio inicial, e buscar por dados que possam prejudicar um indivíduo, como por exemplo endereço, RG, CPF, entre outros dados sensíveis expostos. Este trabalho possui como objetivo a coleta de links de navegação e também publicações na *Dark Web*, com o intuito de disponibilizar uma base de dados para análise em outros trabalhos. Neste trabalho, foi realizado a indexação de mais de 10.000 *websites* na *Dark Web*, estejam eles funcionais ou não, armazenado o nome da página acessada e o HTTP *status code* para uma melhor visualização se o serviço estava pronto para ser acessado, ou se demandaria uma autenticação. Um desses sites foi escolhido para a extração de perguntas e respostas, sendo esse, um fórum de perguntas e respostas voltado ao público brasileiro em que existe uma seção dedicada para o assunto de *hacking*.

Palavras-chave: Ciberataque, Segurança da informação, *Darknet*, *DarkWeb*, Monitoramento, *Crawler*, *Scraper*.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	*Application Programming Interface*
CSS	*Cascading Style Sheets*
IoC	*Indicator of Compromise*
IP	*Internet Protocol*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
HTTPS	*Hypertext Transfer Protocol Secure*
STJ	Superior Tribunal de Justiça
TOR	*The Onion Router*
URL	*Uniform Resource Locator*
WWW	*World Wide Web*

Sumário

1 Introdução

Na Internet, existem os motores de busca tradicionais, como *Google*, *Bing* e *DuckDuckGo*, que são ferramentas que estão constantemente buscando por web sites e disponibilizando em pesquisas através de palavras chaves, o que constitui a chamada *surface web*. Porém, nem todos os web sites podem ser encontrados por esses buscadores. Quaisquer páginas que não possam ser encontradas na *surface web* fazem parte do que é conhecido como *deep web* (Computer Hope, 2017).

Na *deep web*, existe uma rede chamada *darknet*, que é, não uma rede física a parte, mas uma rede criptografada construída em cima de redes já existentes (GOOS; HARTMANIS; LEEUWEN, 1999), como por exemplo a Internet. A rede *TOR*, um exemplo de *darknet*, e também sendo o foco do desenvolvimento deste trabalho, é um grupo de servidores operados voluntariamente que permite melhorar a privacidade e segurança na Internet. Os serviços *onion* da rede *TOR* permitem que usuários publiquem web sites sem que seja necessário revelar a localização de origem (TOR, s.d.).

Assim como os serviços publicados na Internet constituem a *World Wide Web*, os serviços publicados na *darknet* constituem a *Dark Web* (TIWARI, 2020). Assim como a rede *TOR*, existem também outras redes como por exemplo a *Freenet*, mas o desenvolvimento deste trabalho é realizado em cima da rede *TOR*.

Trabalhos anteriores como por exemplo (KAUR; RANDHAWA, 2020) e (KUMAR et al., 2019), mostram que os cibercriminosos utilizam da rede *TOR* para manter atividades ilegais na *darknet*, o que permite a divulgação de campanhas de furtos massivos de dados, ataques de negação de serviço, *botnets*, entre outros tipos de ameaças, aproveitando da anonimidade fornecida pela rede.

Uma URL (*Uniform Resource Locator*), é o endereço de um recurso único na *Web*. Tais recursos podem ser uma página HTML, um documento CSS, uma imagem, etc (MDN CONTRIBUTORS, 2020). Um exemplo de URL seria <www.portal.facom.ufu.br/graduacao/ciencia-da-computacao>. Um link malicioso ou também chamado de URL maliciosa, é um elemento criado com o objetivo de promover golpes, ataques e fraudes (GATEFY, 2019). Se introduzido no contexto certo, o atacante pode persuadir a vítima a preencher seus dados, como nome completo, CPF, RG, e-mail, senhas, número de cartões de crédito entre outros dados sensíveis. Esse tipo de fraude também recebe o nome de *phishing*.

A coleta dessas URLs maliciosas e de outros indicadores de comprometimento (IoC), pode ser utilizada para antecipar potenciais ataques, identificar o público alvo, ter uma melhor compreensão do objetivo do ataque, ou até mesmo evitar que a campanha

seja disparada, categorizado essa URL como maliciosa, por exemplo.

Tendo a base desses conhecimentos, é admissível que se pense na *darknet* como fonte de dados para tentar prever campanhas que promovam ataques com o intuito de furtar dados direcionados ao público brasileiro. A coleta desses dados abre portas para uma análise mais detalhada tentando identificar e evitar esses ataques com antecedência, tentando mitigar os possíveis efeitos causados.

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo deste trabalho é investigar a *darknet* e identificar sites que promovem discussões sobre temas relacionados à segurança da informação. De posse dessa lista, são desenvolvidos mecanismos para coletar e armazenar as publicações presentes em tais fóruns de forma estruturada para facilitar a análise dos dados posteriormente.

1.1.2 Objetivos específicos

1. Compreender e explorar a *darknet*.
2. Entender o processo de indexação e de busca dela.
3. Desenvolvimento de uma ferramenta para indexar os links de maneira automática.
4. Identificar URLs relevantes.
5. Desenvolver um *crawler* para realizar a extração das postagens.
6. Estruturar os dados para fácil análise a posteriori.

1.2 Justificativa

De acordo com o matemático Clive Humby (ARTHUR, 2013), o recurso mais valioso do mundo não é mais o petróleo, mas sim, dados. Diante disso, cibercriminosos se dedicam à coleta dos dados de usuários da *World Wide Web* (WWW), onde uma enorme quantia de informação é depositada diariamente (DESJARDINS, 2019), visto que em março de 2021, 65,6% da população mundial teve acesso ao serviço (Internet World Stats, 2020).

A comunidade de cibercriminosos vê isso como uma oportunidade para ataques, como por exemplo, realizar o furto e a venda dos dados desses usuários. Segundo o relatório da *PSafe* (PSAFE, 2018), o número de detecção de links maliciosos em 2018 foi de 120,7

milhões. A população do Brasil na época era de 209,5 milhões (WORLDOMETER, 2019) de habitantes, e 79.3% (IBGE, 2020) desses teriam posse de um aparelho celular, o que resulta em aproximadamente 0,73 links maliciosos por pessoa com acesso ao celular.

Com o crescimento da importância dos dados, é evidente que a análise desses também segue com a mesma relevância, assim como o paralelo feito com o petróleo, que quando não refinado, não possui uma grande utilidade. A captura desses dados também deve ser planejada de modo que as informações capturadas sejam úteis para serem processadas posteriormente.

Conseguir acesso aos fóruns de cibercriminosos atuantes na *darknet* através de um monitoramento passivo seria um mecanismo de defesa com bastante potencial, já que ataques poderiam ser previstos, mitigados ou detectados. Diante disso, é possível tomar as devidas medidas de segurança antes mesmo que campanhas que promovam a divulgação de métodos de *phishing* sejam divulgadas.

O presente trabalho está organizado da seguinte forma. O Capítulo 1 traz uma motivação do por quê deste trabalho ser relevante, o Capítulo 2 apresenta trabalhos que abordam temas semelhantes, o Capítulo 3 trata do desenvolvimento de cada etapa, e o Capítulo 4 finaliza o trabalho com as considerações finais.

2 Referencial Teórico

Para o desenvolvimento deste trabalho, é essencial entender alguns tópicos tais como o funcionamento básico de redes de computadores, o que é a segurança da informação no contexto da Internet, e as ameaças que a população está exposta.

2.1 Redes de computadores

Uma rede de computador é um sistema distribuído composto por computadores e outros dispositivos, em que cada um desses elementos é capaz de se comunicar uns com os outros. Para que haja a comunicação entre cada entidade, é necessário que haja um conjunto de regras ou protocolos que cada elemento da rede deve seguir para realizar a comunicação (KIZZA, 2005).

A Internet é uma arquitetura de sistema que revolucionou a comunicação e métodos de comércio permitindo que diversas redes de computadores ao redor do mundo se interconectassem. Às vezes referenciada por "rede de redes", a Internet surgiu nos Estados Unidos nos anos 70, mas não teve visibilidade ao público geral até o início dos anos 90 (KAHN; DENNIS, 2020).

O conjunto de protocolos TCP/IP permite que computadores de todos os tamanhos, de diversos vendedores, executando sistemas operacionais totalmente diferentes, se comuniquem entre si. O que iniciou no fim dos anos 60 como um projeto de pesquisa governamental de troca de pacotes de rede, nos anos 90 se tornou a forma mais utilizada de rede de computadores, e hoje, a base da Internet é formada pelo conjunto de protocolos TCP/IP (STEVENS; WRIGHT, 1996).

Sabe-se que na Internet, todas as mensagens devem ter uma origem e um destino, desse modo, quem almeja utilizar o serviço tendo o principal foco a privacidade e anonimidade acaba sendo prejudicado, pois a Internet não foi projetada com esse objetivo. Porém algumas pessoas se preocupam bastante com sua privacidade e em meados dos anos 90, foi o Governo Federal dos Estados Unidos que demonstrou um maior interesse e iniciou o desenvolvimento de uma nova tecnologia chamada de *Onion Routing* (KAUR; RANDHAWA, 2020).

O conceito de *Onion Routing* ("Roteamento em Cebola"se traduzido ao sentido literal da frase, do inglês) deve-se ao funcionamento da comunicação entre duas partes, que é feita por uma rota aleatoriamente traçada, em que cada nó da rota consegue descriptografar uma parte da mensagem (ou na analogia, retirar uma camada da casca da cebola), e encaminhar a mensagem para o próximo nó, assim repetindo o processo até

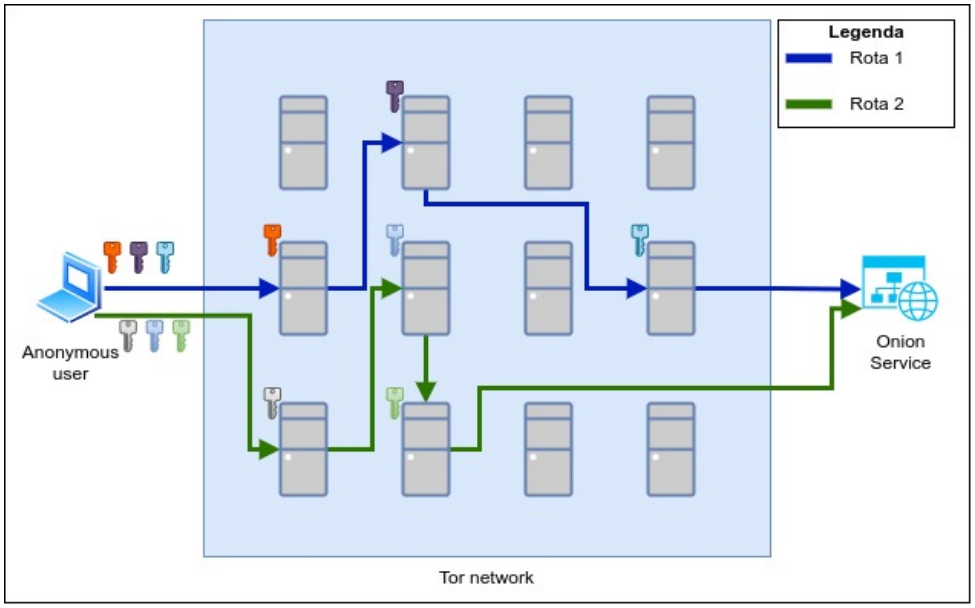

Figura 1 – Diagrama do *Onion Routing*. Fonte: Autoria Própria

o seu destino. Na trajetória da mensagem, apenas o nó adjacente é identificado pelo nó atual, assim garantindo a anonimidade nos dois sentidos de tráfego da mensagem (KAUR; RANDHAWA, 2020). Esse processo pode ser visto no diagrama da Figura 1.

Uma rede que utiliza o *Onion Routing* é classificada como *darknet*. A *darknet* fornece a anonimidade e privacidade que não é fornecida pela Internet, mas para isso é necessário ferramentas mais específicas para o acesso do que quando o objetivo é apenas o acesso à Internet. Alguns exemplos de navegadores que podem ser utilizados para navegar na *darknet* são *The Onion Router* (*TOR*), *FreeNet*, *Riffle*, *Invisible Internet Project* (*I2P*) e *Whonix*(KAUR; RANDHAWA, 2020).

2.2 Segurança da informação

A segurança de computadores pode ser definida como a proteção oferecida a um sistema de informação automatizado para atingir os objetivos apropriados de preservação da integridade, disponibilidade e confidencialidade de ativos de sistemas de informação (incluindo hardware, software, firmware, informações/dados e telecomunicações)(BOSWORTH; KABAY, 2002).

Com essa definição, é possível expandir esse conceito nos seguintes pilares (STALLINGS, 2014):

- **Confidencialidade** - Garante que os dados não fiquem disponíveis e nem sejam revelados a indivíduos não autorizados. Além disso, garante que o indivíduo controle quem tem acesso às suas informações.

- **Integridade** - Garante que um sistema só tenha suas informações atualizadas por quem tenha permissão, e de modo esperado.
- **Disponibilidade** - Garante que os sistemas funcionem prontamente e que não haja negação de serviço a usuários autorizados.

Assim como no meio físico, os dados digitais estão sujeitos a furtos, sequestros e essa é uma grande preocupação para os profissionais de segurança da informação. A área que realiza estudos de falhas e vulnerabilidades em aplicações é essencial para manter a segurança dos dados dos usuários, assim tendo um potencial para ser uma das áreas mais importantes da computação.

Mas nem sempre o problema está localizado nos códigos das aplicações. Apesar de vulnerabilidades nos códigos de aplicações serem praticamente impossíveis de serem evitadas, é possível ver em (MOUTON; LEENEN; VENTER, 2016) que as pessoas ainda estão suscetíveis à manipulação e que o fator humano é um elo fraco.

2.3 Ameaças de segurança

Muitos ataques exploram a psicologia tanto quanto a tecnologia. Quando o atacante deseja dados sensíveis de um usuário, utiliza técnicas de *phishing*. Nesse tipo de ataque, o atacante envia um e-mail para a vítima se passando por uma outra pessoa (KAUR; RANDHAWA, 2020). O conteúdo do e-mail pode ser um link malicioso, um link que aparenta ser genuíno porém ao clicar, a vítima é redirecionada para a página do atacante. Essa página é visualmente idêntica à que a vítima esperava, porém ao inserir suas credenciais elas são enviadas para o atacante. Esse tipo de ataque quebra o pilar da confidencialidade e pode também afetar o da integridade, já que o atacante tem as credenciais da vítima sem o seu conhecimento.

Um outro abuso realizado por atacantes é a execução de aplicações não desejadas na máquina das vítimas, também conhecidos como *malwares*. O termo *malware* é uma abreviação para *malicious software* (do inglês, software malicioso), e é um termo geral para vírus, *worms*, trojans, *ransomwares* e outros programas que cibercriminosos utilizam para obter algum tipo de vantagem sobre a vítima (FRUHLINGER, 2019).

Com um *malware* instalado na máquina da vítima, o atacante poderia por exemplo acessar os dados que o navegador armazena, como senhas, cartões de créditos, dados de preenchimento automático como endereço, CPF, RG, entre outros dados sensíveis assim quebrando o pilar da confidencialidade. Além disso, é possível explorar o poder computacional da máquina infectada para, por exemplo, minerar criptomoedas, assim quebrando o pilar da integridade e disponibilidade.

Tabela 1 – Exemplos de *malwares*.

Tipo	O que faz	Exemplo
Ransomware	Remove o acesso da vítima aos arquivos até que um resgate seja pago	*RYUK*
Fileless Malware	Faz alterações nos arquivos nativos do sistema operacional da vítima	*Astaroth*
Spyware	Coleta dados de atividade do usuário sem que ele saiba	*DarkHotel*
Adware	Mostra anúncios indesejados	*Fireball*
Trojans	Código não desejado disfarçado de um código desejado	*Emolet*
Worms	Se espalha pela rede replicando a si mesmo	*Stuxnet*
Rootkits	Dá ao atacante acesso remoto ao dispositivo da vítima	*Zacinio*
Keyloggers	Monitora as teclas digitadas pelo usuário	*Olympic Vision*
Bots	Executa extensos ataques de inundação	*Echobot*

Um *ransomware* pode ser retratado no cenário ocorrido no STJ (Superior Tribunal de Justiça), que foi vítima do ataque de cibercriminosos, tendo seus dados e backups criptografados, ou também podendo ser considerado sequestrados, já que foi solicitado uma quantia para o resgate dos dados (PETRY, 2020). Esse ataque quebra o pilar da integridade.

Alguns dos tipos de *malwares* podem ser visualizados com mais detalhes na Tabela 1, baseada na publicação (CROWDSTRIKE, 2019). Além desses mencionados, temos os vírus, que ao infectar programas pode espalhar para outros sistemas e realizar suas próprias malícias (WOLF, 2021).

O *malware* mais comum no final de 2021, de acordo com o relatório da Arctic Wolf (WOLF, 2021), é o *Adware*. Conforme definido na Tabela 1, esses anúncios podem levar o usuário a baixar *malwares* inadvertidamente, podendo acarretar em infecções ainda mais prejudiciais.

2.4 Trabalhos Correlatos

Nesta seção, são citados alguns dos trabalhos que estão relacionados com esta monografia.

2.4.1 Uso da *Dark Web* para monitoramento de atividades maliciosas

Um artigo que ilustra bem o que é a *Dark Web* e as atividades ilegais que acontecem nela, é o (KAUR; RANDHAWA, 2020). Nele, é possível entender o funcionamento da *Dark Web* e seus conceitos, tais como as ferramentas necessárias para o acesso à rede, como

a anonimidade é garantida através da criptografia, aleatoriedade nas rotas, entre outros tópicos.

O trabalho citado anteriormente tem como objetivo analisar os crimes cibernéticos que se encontram na *Dark Web* e retrata diversos tipos de atividades criminosas que foram identificadas usando como fontes diversos outros artigos. Algumas dessas atividades são:

- **Vazamento de informações** - Grupos de cibercriminosos divulgam dados sensíveis, como cartões de créditos, credenciais de login, contas bancárias, etc.;
- **Clone de páginas** - A anonimidade da rede *TOR* faz com que seus usuários estejam vulneráveis a ataques, já que os web sites não mostram o típico *HTTPS* para indicar que um site é autêntico;
- **Fornecimento de negação de serviço** - Grupos de cibercriminosos oferecem recursos para negação de serviços ao preço de *Bitcoins*;
- **Fraudes** - É o crime mais comum que acontece na *Dark Web*, em que os fraudadores furtam os dados de cartões de crédito ou informações pessoais.

O último item citado é o alvo do desenvolvimento deste trabalho.

2.4.2 Utilização de *crawlers* para a coleta de dados

Um *crawler* é um programa designado para navegar em um website e coletar documentos HTML. Cada um deles são desenvolvidos para diferentes plataformas devido a diferença estrutural em cada uma delas. Além disso, é necessário tratar desafios como servidores não responsivos, links repetidos que podem criar uma repetição em caso de navegação por links, entre outros desafios, para que seja possível a coleta de informações relevantes. (NUNES et al., 2016).

A navegação na *Dark Web* é um tanto quanto complicada, dado que os proprietários de web sites ilegais não desejam que suas páginas sejam visualizadas por robôs. Devido a isso, é extremamente comum que para navegar no site de uma página para outra, seja necessário completar algum desafio que seria computacionalmente difícil de ser resolvido, porém resolvido em teoria com facilidade por um humano. Alguns exemplos são *Captchas* controle de requisições por segundo.

Há trabalhos que já fizeram a coleta de dados do mercado negro (NUNES et al., 2016) na *Dark Web* e obtiveram sucesso. Um deles (LAWRENCE et al., 2017) sugeriu a ferramenta *Death by Captcha* como uma opção para a resolução dos *Catphas* com um baixo custo, como uma alternativa para a resolução via API (*Application Programming Interface*) dos desafios.

O trabalho (NUNES et al., 2016) conseguiu implementar um *crawler* que coletasse informações de páginas do marcado negro e as classificassem encontrando páginas que vendem vulnerabilidades do dia zero. Utilizando a mesma técnica seria possível fazer a análise de fóruns buscando por *URLs*, e arquivos executáveis, buscando por palavras em português para filtrar os dados para o público brasileiro.

O trabalho (SOSKA; CHRISTIN, 2015) ilustra a extração de dados do *Silk Road*, um dos mercados anônimos mais conhecidos da *Dark Web*. O artigo cita algumas técnicas para evitar ser identificado como um tráfego suspeito pelo sistema de *throttle* da página, como por exemplo alternar de maneira aleatória selecionando circuitos aleatórios da rede *TOR* em cada requisição. Ainda segundo o autor, é possível detectar uma dificuldade em coletar os dados das páginas, pois além de todos os mecanismos citados, ainda existe o problema de disponibilidade da página. Em grandes sites, a coleta dos dados pode levar dias.

2.4.3 Análise dos dados para extração de informações relevantes

A execução do *crawler* retorna páginas por completo, ou seja, todo o código *HTML*, podendo também ter pedaços de código *JavaScript*, o que não é relevante para a análise dos dados posteriormente. Portanto antes de filtrar o conteúdo da busca procurando por palavras em português ou por *URLs*, é preciso extrair as informações que não tenham relevância do texto.

É possível ver em (NUNES et al., 2016) que alguns dados importantes de serem coletado nos fóruns são:

- Conteúdo do tópico - Publicações que fazem parte de um determinado tópico
- Conteúdo da publicação - Um conteúdo representando uma dúvida, opinião ou expressão de um usuário
- Autor do tópico - Usuário que criou o tópico
- Autor da publicação - Usuário que criou a publicação
- Status do autor - Equivalente ao “cargo“ de um usuário
- Reputação - Um valor que indica a relevância do usuário no fórum
- Interesse do tópico - O que cada tópico é esperado abordar

É importante ressaltar que o autor do trabalho mencionado acima não explicita o significado de cada um dos campos. O significado dado a cada um dos campos vem da interpretação do escritor deste trabalho.

Para realizar a análise dos dados para a extração de informações relevantes, é possível utilizar de técnicas como expressões regulares, e após a extração, pode-se armazenar em um banco de dados estruturado com os atributos descritos acima.

De acordo com (SOSKA; CHRISTIN, 2015), as heurísticas definidas para a extração dos dados podem se tornar bastante complicadas, dado que os sites podem estar constantemente modificando o formato da página. Por isso foi uma decisão importante separar as duas etapas.

Os trabalhos descritos nessa seção possuem como objetivo analisar os produtos do mercado anônimo com base nos dados coletados por execução de *crawlers* nos sites. Dados como métricas de volume de vendas, categoria de produtos e dados de vendedores puderam ser coletados de maneira automática, mostrando ser uma técnica funcional em (SOSKA; CHRISTIN, 2015).

3 Desenvolvimento

O desenvolvimento deste trabalho foi realizado em seis etapas. O entendimento da navegação na *darknet*, entender o processo de indexação e busca nela, o desenvolvimento de uma ferramenta para indexar os links de maneira automática, identificar um web site relevante, a extração do conteúdo de tal web site e a estruturação dos dados coletados para fácil análise a posteriori.

3.1 Metodologia

A metodologia pode ser dividida em seis passos:

- A primeira etapa é compreender o funcionamento da *darknet*, em especial, a rede *TOR*. É de conhecimento que para o acesso à rede, é necessário a utilização de um navegador específico, o *TOR Browser*.
- Um grande desafio da navegação na *Dark Web* é a difícil localização de links que permitam acesso aos sites desejados, e para isso, é necessário encontrar uma maneira de indexá-los.
- Desenvolver uma ferramenta que faça a navegação de forma automatizada na *darknet*, coletando novos links de outros serviços que estão acessíveis na rede *TOR* através de um ponto de entrada.
- Tendo uma fonte de web sites indexados na *darknet*, é necessário investigá-los com o intuito de encontrar fóruns em que cibercriminosos publiquem informações que possam ser úteis de alguma maneira, sejam recursos binários, links (não somente os maliciosos mas que possam também servir de ponte para outros fóruns), IPs, etc.
- Desenvolver e testar um *crawler* para a coleta dos dados nos fóruns, de modo que não haja execução de códigos maliciosos durante o processo de extração dos dados, sem o consenso do usuário do navegador.
- Transformar o conteúdo da página HTML (*HyperText Markup Language*) em dados estruturados, de modo que facilite a sua usabilidade e análise posteriormente.

3.2 Navegação na rede *TOR*

A navegação na rede *TOR* se diferencia da navegação na Internet de algumas maneiras. A primeira delas é que não é possível acessar um serviço da rede sem um

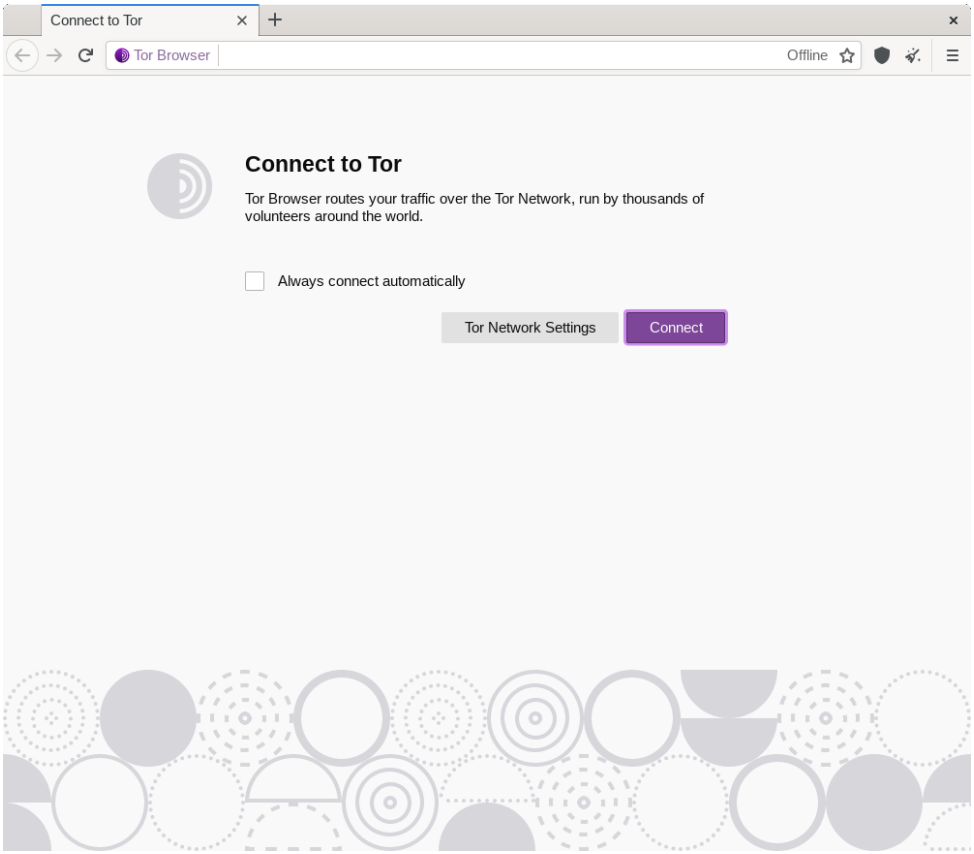

Figura 2 – Tela inicial do *TOR browser*.

navegador específico. Para isso, deve ser utilizado o *TOR Browser*, que permite acesso aos nós que fazem parte da rede *TOR*.

Além disso, como é de conhecimento, não existem motores de busca na *Deep Web*, o que significa que para obter links de acessos aos *websites*, é necessário alguma outra fonte alternativa.

Pode-se ver na Figura 2 que por padrão não há a conexão imediata à rede *TOR* ao iniciar o *Tor Browser*. Para isso, é necessário que haja explicitamente a conexão manual, conforme a Figura 3. Então, por fim, temos o navegador pronto para uso e acesso de serviços disponíveis na rede *Tor*, conforme a Figura 4.

3.3 Busca de links *onion*

Os serviços na *Dark Web* possuem um formato diferente dos da *World Wide Web*, sendo os seguintes formatos:

- *http://www.example.com/* para a *World Wide Web*
- *http://abc1defgh2ijkl34.onion/* para a *Dark Web*

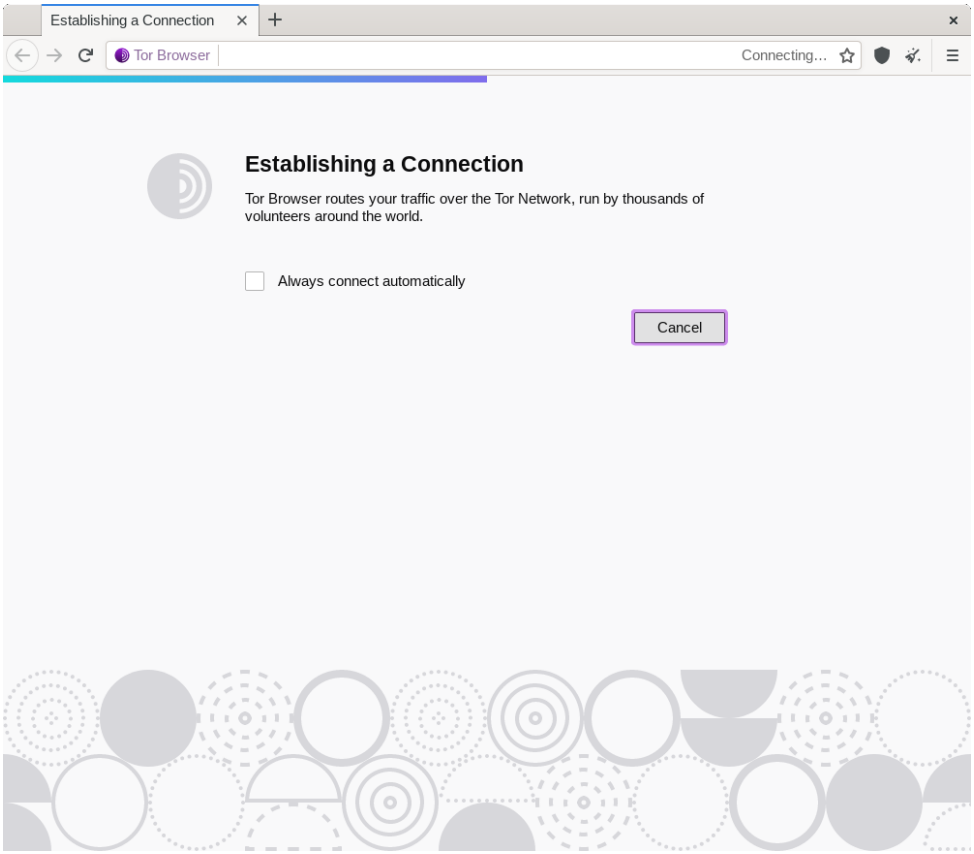

Figura 3 – Etapa de conexão à rede *TOR*.

Um usuário do *GitHub* disponibilizou uma lista com centenas de links de serviços *onion* (5KY1AR, 2020), entre eles serviços separados por categorias, como outras listas de links dentro da própria *Dark Web*, comércios ilegais, e-mails, etc.

3.4 Automação da coleta de links *onion*

Um grande problema com os links encontrados, é que grande parte deles não eram mais pertencentes a serviços ativos. Os links *onion* geralmente possuem uma volatilidade relativamente alta, tendo uma vida útil baixa dependendo do serviço, o que dificulta encontrar serviços desejados.

Durante a navegação, foi possível imaginar um grafo em que um *website* se conecta com o outros através dos links no seu conteúdo, principalmente uma lista de links, que possui arestas para diversos outros serviços. Na Figura 5 é possível ver a representação do exemplo do grafo citado.

3.4.1 Utilizando o *Python* para navegar na *Dark Web*

Tomando como base o artigo (RODRIGUEZ, 2020), foi possível realizar requisições HTTP para a rede *TOR* utilizando o *Python*. Desse modo, é possível baixar o conteúdo

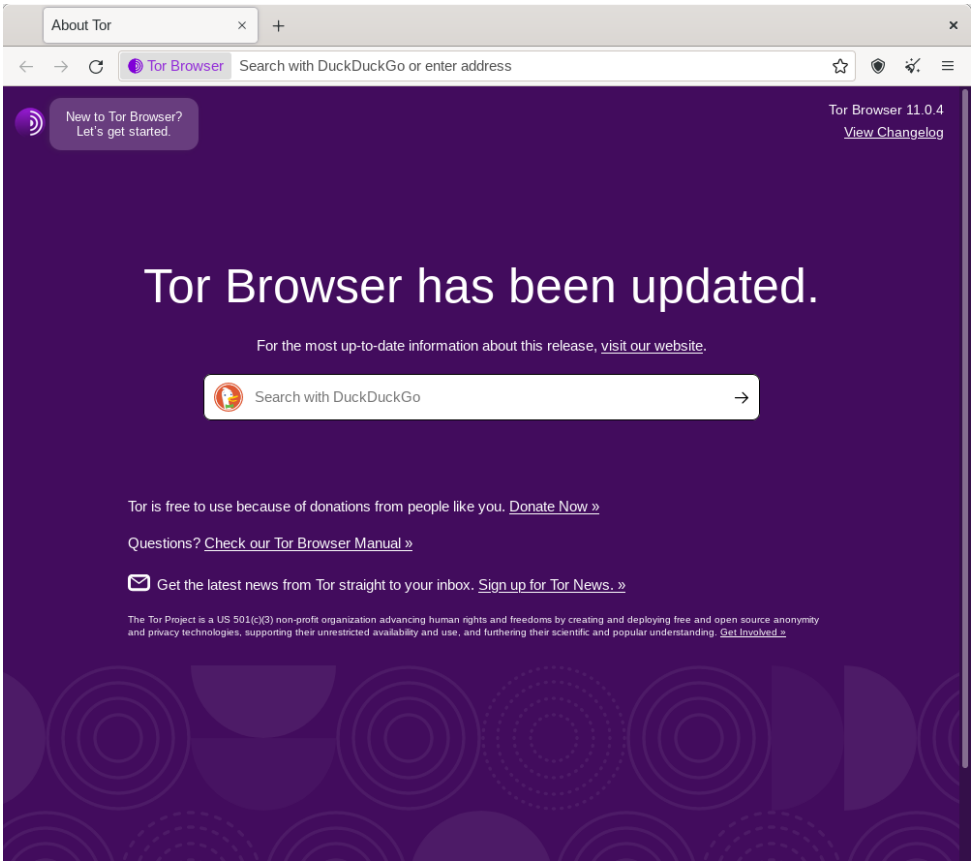

Figura 4 – *TOR browser* pronto para uso.

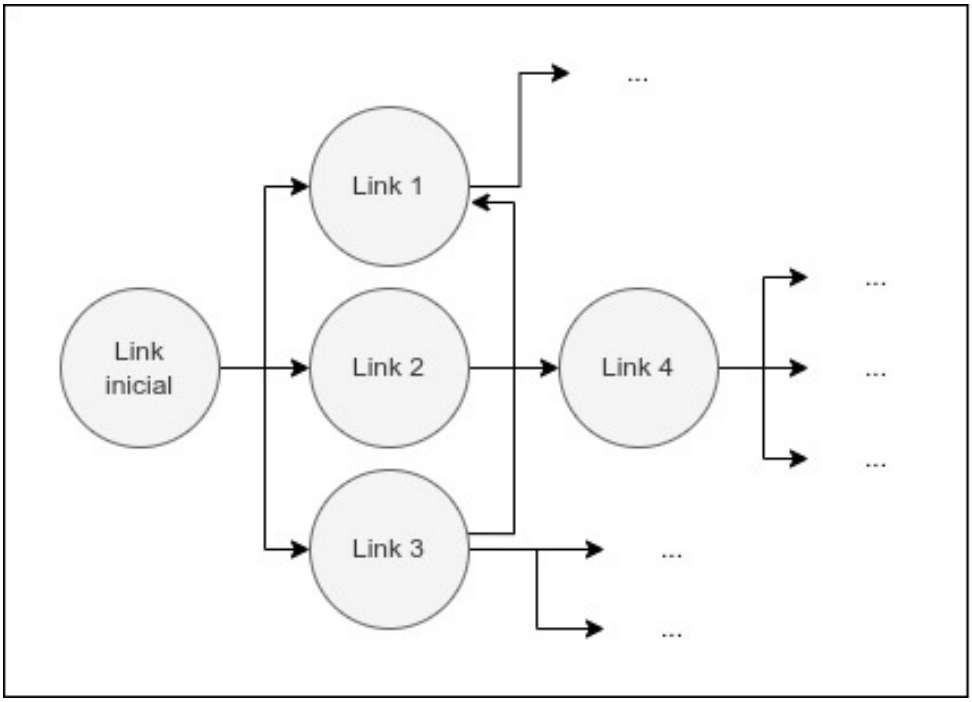

Figura 5 – Diagrama do grafo de indexação da *darknet*. - Fonte: Autoria própria

HTML das páginas e o analisar de maneira automatizada.

Para requisições na rede *TOR*, foi utilizado a requisição de rota aleatória em cada nova requisição, e além disso, um cabeçalho *User-Agent* falso, este que representa a ferramenta utilizada para a criação da requisição HTTP. Para que a requisição consiga adentrar a rede *TOR*, um *proxy* foi utilizado após o controlador do serviço estar de pé.

O serviço pode ser utilizado após instalado, utilizando o seguinte comando:

```
sudo service tor start
```

Para a randomização do cabeçalho de ferramenta, a biblioteca *Python* "fake_useragent" foi utilizado através do seguinte código:

```
from fake_useragent import UserAgent

headers = {
    "User-Agent": UserAgent().random
}
```

Já para a requisição de uma nova rota na rede *TOR* a cada requisição, foi utilizada a biblioteca "stem" através da seguinte função:

```
from stem.control import Controller

def new_tor_id():
    with Controller.from_port(port=9051) as controller:
        controller.authenticate(password=TOR_PASS)
        controller.signal(Signal.NEWNYM)
```

Nota-se que o controlador do serviço encontra-se na porta 9051, porta padrão para a comunicação com o controlador.

E por fim a requisição é realizada utilizando a biblioteca "requests", fornecendo o caminho do proxy da seguinte maneira:

```
def tor_get(url):
    new_tor_id()
    response = requests.get(url, headers=headers,
```

```
                    proxies=tor_proxy,
                    timeout=TIMEOUT)
```

O parâmetro "TIMEOUT" é utilizado para definir o tempo máximo de espera até a decisão que o serviço não está disponível e retornar um erro. Esse parâmetro é importante neste cenário, pois o tempo de resposta de serviços na *Dark Web* é consideravelmente maior do que os presentes na *World Wide Web*.

3.4.2 Desenvolvimento do indexador

A análise desejada é em busca de links de serviços *onion*, e para isso foi utilizado uma expressão regular dentro do conteúdo de cada página. A expressão utilizada foi a seguinte:

```
((?:https?://)?\w+\.onion)
```

Observe que essa expressão regular não coleta qual arquivo está sendo acessado dentro daquele serviço, isto é, quando a requisição para o site for realizada, o próprio serviço fará o redirecionamento para a página raiz do *website*. Tendo isso em mente, é de conhecimento que a coleta dos links poderia ser ainda mais eficiente, porém devido ao tempo necessário para estas requisições, a decisão tomada foi não navegar pelas páginas de um mesmo site.

A ferramenta desenvolvida demanda um ponto de entrada para aplicar a expressão regular mencionada. Neste caso, foi utilizado a lista encontrada no *GitHub* (5KY1AR, 2020), que contém centenas de links. Os links detectados durante a execução da ferramenta são adicionados a uma lista que é persistida em disco, devido ao longo tempo de execução para evitar perda de progresso, e somente caso o link já não esteja nessa lista.

A paralelização de requisições para uma execução mais rápida não teve o desempenho esperado, pois o controlador do serviço *TOR* não consegue gerenciar uma grande vazão de entrada e saída para a rede. As melhores configurações encontradas para a execução do indexador foram:

- Número máximo de *threads*: 3
- Intervalo entre requisições: 5 segundos
- Tempo de espera máximo para resposta do serviço: 5 minutos

Figura 6 – Tela inicial do fórum *Respostas Ocultas*.

Uma vazão maior que a mencionada acima causou a indisponibilidade do controlador, fazendo com que todas as requisições esgotassem o tempo máximo de espera por resposta, independente se o serviço estiver disponível ou não. A paralelização foi optada pois as requisições sequenciais tomariam um tempo muito grande, já que a lista de links identificados é consideravelmente grande.

3.5 Busca de um link relevante

A busca por um link relevante foi complicada, já que para conseguir extrair dados úteis de maneira automatizada, o serviço deve permitir o acesso automatizado na página, isso significa, não ter nenhum limite de requisições máximas por intervalo de tempo, não ser necessário resolver desafios para provar que é humano e além disso, ter um padrão bem definido na estrutura HTML.

Foram encontrados alguns fóruns, porém voltados a públicos gerais, como por exemplo *Hidden Answers*, *Dread*, *Respostas Ocultas*, entre outros. O *website* que mais chamou atenção dos mencionados foi o fórum *Respostas Ocultas*, conforme Figura 6, que apesar de ser voltado à assuntos gerais, possui uma seção específica para debater sobre *Hacking*, e além disso, os debates são em português.

Um obstáculo encontrado foi a necessidade de resolver um desafio para provar ser humano para acessar o serviço, mas felizmente este desafio só deve ser realizado no primeiro acesso, o que significa que de alguma maneira o serviço precisa armazenar algum identificador para não solicitar este desafio novamente.

Após analisar o armazenamento do *website* no navegador, foi observado um *cookie* que aparentemente identificava que o desafio já havia resolvido, e ao adicionar esse *cookie* na requisição automatizada, foi confirmado que não havia a necessidade de resolução do desafio tendo posse deste identificador.

3.6 Extração do conteúdo do fórum

A estrutura do fórum *Respostas Ocultas* possui um padrão bem definido, semelhante a de outros fóruns de debate, no qual tem-se a página inicial com sugestões de tópicos sendo debatidos no momento, uma lista de categorias para listar postagens de um determinado assunto e também a possibilidade de listar todas as postagens.

Os elementos HTML da página de uma postagem possuem classes que seguem um padrão conforme o exemplo do formato de uma pergunta a seguir:

- *qa-q-view*: Área da página que contém um conjunto de informações sobre uma questão
- *qa-q-view-content qa-post-content*: Texto descritivo da postagem
- *qa-q-view-stats*: Seção que possui dados sobre votação da postagem
- *qa-upvote-count-data*: Elementos que contabiliza os votos positivos da postagem
- *qa-downvote-count-data*: Elementos que contabiliza os votos negativos da postagem
- *qa-q-view-tag-list*: Categorias da postagem inseridas arbitrariamente pelo autor

O restante das estruturas seguem o mesmo padrão, variando algumas partes como por exemplo, após o primeiro hífen (-) tem-se um "q" para perguntas, enquanto que para respostas tem-se um "a", que pode ser visto na classe do elemento contendo o nome de usuário, "*qa-a-item-who*".

Para a extração do conteúdo das páginas, foi utilizado a mesma ideia de requisição para a rede *TOR* utilizada para a coleta dos links, com o adicional da biblioteca *BeautifulSoup*, utilizada para gerar a estrutura lógica da árvore de elementos HTML.

```
from bs4 import BeautifulSoup
from src.tor_requests import tor_get

def get_post(link):
    response = tor_get(link)
    content = response.text

    # A variável "page" armazena a estrutura de objetos da página
    page = BeautifulSoup(content, 'html.parser')
```

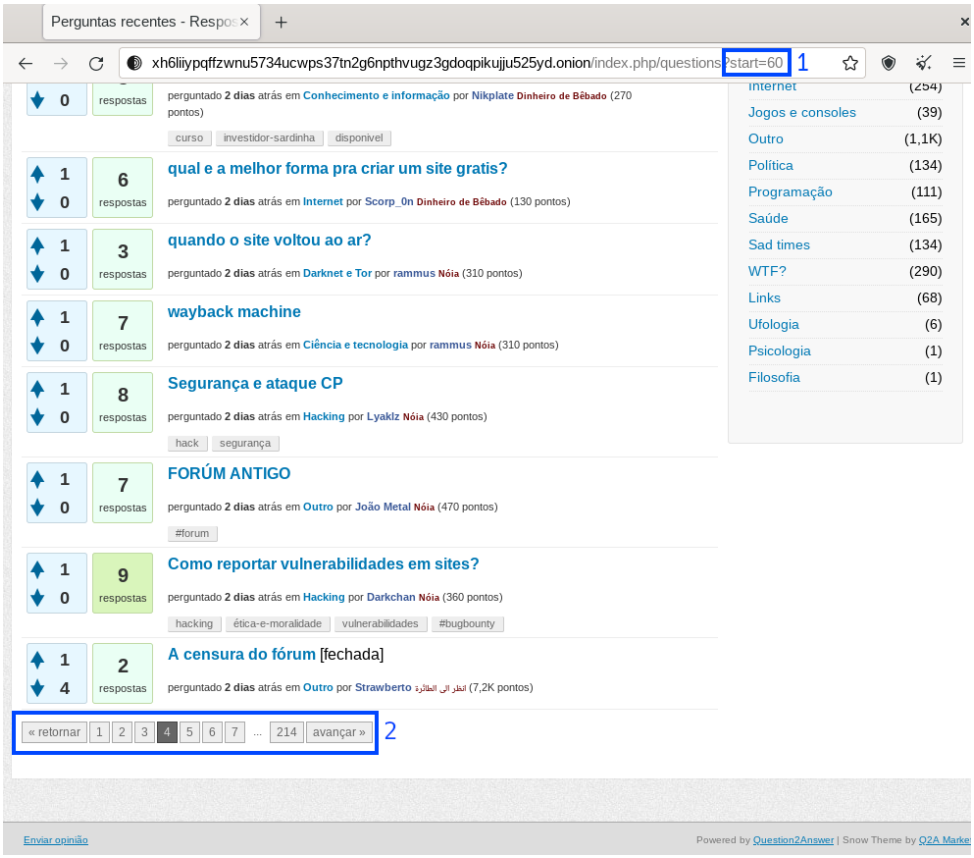

Figura 7 – Referência da página de listagem de todas as perguntas.

De posse dos objetos da página de maneira estruturada, é possível buscar por elementos específicos como por exemplo coletar os links de postagens disponíveis no atributo *href* de cada *div*, filtrados pelo nome da classe:

```
posts = []
divs = page.find_all('div')
for div in divs:
    try:
        if('qa-q-item-title' in div['class']):
            posts.append(div.a['href'])
    except KeyError:
        pass
```

A mesma lógica é aplicada para todas as páginas, percorrendo todas as publicações listadas na página contendo todas as publicações, conforme a Figura 7.

Na Figura 7 são visíveis 2 seções grifadas. A seção 1 representa a pergunta que estará na primeira posição da lista de 20 perguntas. A seção 2 é a barra de paginação das perguntas. Isso significa que na página 1, a primeira postagem vai ser a pergunta 0. Já na

página 2, a primeira pergunta será a de número 20, e assim por diante, até que a última página seja atingida.

Um possível algoritmo para coletar cada página seria:

```
pages = []
for i in range(0, LAST_PAGE):
    next_page = BASE_URL + '/questions?start=' + str(i*20)
    page_content = get_post(next_page)
    pages.append(page_content)
```

Cada página coletada tem uma lista de perguntas, em que cada um dos links é extraído, baixado e seu conteúdo armazenado. Após a extração das páginas HTML, elementos como *tags* de *script*, *style* entre outros, podem ser removidos, já que não possuem um efeito direto na página quando sendo analisada por uma automação. Desse modo, o tamanho dos arquivos a ser analisados diminuem drasticamente, e além disso, tem-se uma segurança maior para abrir uma página no navegador, já que não teremos nenhum script sendo executado sem permissão.

3.7 Estruturação dos dados

Após a análise dos dados coletados em formato HTML, é possível identificar um formato bem definido e um relacionamento entre entidades que nos permite formular a seguinte estrutura para uma postagem (ou *Post*):

```
{
    "title": string,
    "content": string,
    "category": string,
    "up_votes": integer,
    "down_votes": integer,
    "created_at": date,
    "author": User,
    "best_answer": Answer,
    "comments": [Comment],
    "answers": [Answer],
    "tags": [string]
}
```

Definindo as estruturas utilizadas acima, começando pela estrutura de uma resposta (ou *Answer*):

```
{
    "up_votes": integer,
    "down_votes": integer,
    "content": string,
    "created_at": date,
    "user": User,
    "comments": [Comment]
}
```

Definição de um comentário (ou *Comment*):

```
{
    "content": string,
    "created_at": date,
    "author": User
}
```

Definição de um usuário (ou *User*):

```
{
    "name": string,
    "title": string,
    "score": integer
}
```

3.8 Experimentos

Os experimentos levantados neste trabalho são relacionados à otimização da extração de conteúdos de páginas na *darknet*, seja de uma página específica, ou de páginas com uma estrutura não necessariamente bem definida. Uma das atividades mais importantes abordadas neste trabalho é a efetividade da paralelização da extração dos dados da página.

3.8.1 Ambiente de execução

A execução da indexação dos links foi realizada em um *container Docker* com as seguintes configurações:

- Versão do cliente: 20.10.12
- Versão da *engine*: 20.10.12
- Versão do *Go*: go1.17.5
- *Git commit* do cliente: e91ed5707e
- *Git commit* da *engine*: 459d0dfbbb
- Versão do *containerd*: v1.5.8

As especificações da máquina hospedeira são:

- CPU: Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz
- Memória RAM: 16 GB, DDR4, 2133 MHz
- Placa de rede: Intel Corporation Wireless 7265
- Sistema operacional: Arch Linux 5.15.12-arch1-1 x86_64

Nos momentos em que o controlador do serviço *TOR* é sobrecarregado, é necessário reiniciar o *container* para que ele volte a responder as requisições. Apenas reiniciar o serviço não é o suficiente, acredita-se que seja devido ao seu tempo máximo de espera, que também é atingido independente de ele estar disponível ou não.

3.8.2 Coleta de links

A Tabela 2 ilustra o resultado do processo de indexação dos links, categorizando-os como serviços ativos, indisponíveis, ou desconhecido. Serviços categorizados como desconhecidos são serviços que atingiram o tempo máximo de espera, ou que referenciavam uma máquina inacessível.

Nesses links, encontra-se diversas páginas que solicitam autenticação, listas de outros links, vendas ilegais como por exemplo mercado negro, contas e cartões clonados, entre outros. A aplicação utilizada para a coleta destes dados foi desenvolvida pelo autor desta monografia e está disponível em (LUCASJRT, 2021).

É importante notar que durante o processo de coleta dos links em páginas acessadas pelo *crawler*, as páginas não precisam de uma estrutura bem definida. O corpo da página

Tabela 2 – Resultados da indexação na *darknet*.

Status	Quantia
Ativo	7.046
Indisponível	59
Desconhecido	3.527
Total	10.632

é tratado como texto plano, isto é, texto não estruturado, e a árvore de elementos não é relevante para as detecções através da expressão regular.

3.8.3 Extração de publicações

Do fórum *Respostas Ocultas* foi possível extrair 19.654 publicações sendo a mais antiga de 26/11/2016. Em cada publicação é possível colocar *tags* personalizadas pelo usuário, e como resultado tem-se 7.408 combinações de *tags*.

Cada publicação, além das *tags*, possuem também categorias, podendo essas ser 21 valores diferentes, como por exemplo *Hacking*, *Money*, *Cryptocurrencies*, *Cooking*, etc.

Teve-se como resultado 2.689 publicações categorizadas como *Hacking* e 7.123 categorizadas como *Other*. Publicações categorizadas como *Other* pode estar também relacionada a *Hacking*, porém não foi devidamente categorizada.

A aplicação utilizada para a extração das páginas em HTML das publicações está disponível em (LUCASJRT, 2022b). A aplicação utilizada para a estruturação dos dados coletados do fórum *Respostas Ocultas* foi desenvolvida pelo autor da monografia e está disponível em (LUCASJRT, 2022a).

4 Conclusão

O principal objetivo deste trabalho era coletar publicações de diversos fóruns hackers, com o intuito de uma análise posterior para extração de dados sensíveis ou que pudessem ser uma ameaça para os usuários da *World Wide Web*. Durante o trabalho, ficou explícito que fóruns que publicam esse tipo de informação não são tão acessíveis para qualquer público.

Apesar disso, foi encontrado um fórum geral de perguntas e respostas voltado ao público brasileiro, que possui uma atividade relevante dos seus usuários, que nos fornece dados suficientes para uma análise de comportamento em um futuro trabalho. Esse fórum possui uma seção dedicada a publicações de *Hacking*, o que permite chegar mais próximo do objetivo inicial.

Conclui-se que os cibercriminosos estão em uma comunidade de não tão fácil acesso, assim dificultando a coleta de dados para prevenir campanhas de *phishing*, ou detectar algum tipo de indicadores de comprometimento (IoC). Entretanto, ainda sim é possível coletar publicações voltadas ao tópico de *Hacking*, que pode ou não se relacionar a tópicos de *phishing*, pirataria, entre outras atividades ilegais.

Para trabalhos futuros, sugere-se utilizar as postagens aqui coletadas do fórum *Respostas Ocultas* para análise de comportamento dos usuários, seja tempo de atividade, horários de atividade, quais categorias possuem mais discussões e em quais horários, entre outros. Também recomenda-se a busca de dados sensíveis como, endereços de e-mail, números de telefone, CPF, RG, fontes de conhecimento como por exemplo PDFs, livros, fontes de pirataria, busca por IoC como URLs, endereços IP, arquivos executáveis, etc.

Referências

5KY1AR. *Awesome Onion Links*. 2020. Disponível em: <https://github.com/5ky1ar/Awesome-Onion-Links>. Acesso em: 9 fev 2022. Citado 2 vezes nas páginas 20 e 23.

ARTHUR, C. *Tech giants may be huge, but nothing matches big data*. 2013. Disponível em: <https://www.theguardian.com/technology/2013/aug/23/tech-giants-data>. Acesso em: 6 out 2020. Citado na página 9.

BOSWORTH, S.; KABAY, M. E. *Computer security handbook*. [S.l.]: John Wiley & Sons, 2002. Citado na página 12.

Computer Hope. *Surface web*. 2017. <https://www.computerhope.com/jargon/s/surface-web.htm>. Acesso em: 8 out 2020. Citado na página 8.

CROWDSTRIKE. *THE 11 MOST COMMON TYPES OF MALWARE*. 2019. Disponível em: <https://www.crowdstrike.com/epp-101/types-of-malware/>. Acesso em: 2 dez 2020. Citado na página 14.

DESJARDINS, J. *How much data is generated each day?* 2019. Disponível em: <https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/>. Acesso em: 4 out 2020. Citado na página 9.

FRUHLINGER, J. *Malware explained: How to prevent, detect and recover from it*. 2019. Disponível em: <https://www.csoonline.com/article/3295877/what-is-malware-viruses-worms-trojans-and-beyond.html>. Acesso em: 11 nov 2020. Citado na página 13.

GATEFY. *What is a malicious URL?* 2019. Disponível em: <https://gatefy.com/blog/what-malicious-url/>. Acesso em: 6 out 2020. Citado na página 8.

GOOS, G.; HARTMANIS, J.; LEEUWEN, J. *Lecture Notes in Computer Science*. [S.l.: s.n.], 1999. 155 p. ISSN 16113349. ISBN 3540666664. Citado na página 8.

IBGE. *PNAD Contínua TIC 2018: Internet chega a 79,1país*. 2020. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/27515-pnad-continua-tic-2018-internet-chega-a-79-1-dos-domicilios-do-pais>. Acesso em: 4 out 2020. Citado na página 10.

Internet World Stats. *INTERNET GROWTH STATISTICS*. 2020. Disponível em: <https://www.internetworldstats.com/emarketing.htm>. Acesso em: 4 out 2020. Citado na página 9.

KAHN, R.; DENNIS, M. A. *Internet*. 2020. Disponível em: <https://www.britannica.com/technology/Internet>. Acesso em: 10 nov 2020. Citado na página 11.

KAUR, S.; RANDHAWA, S. Dark Web: A Web of Crimes. *Wireless Personal Communications*, Springer US, v. 112, n. 4, p. 2131–2158, 2020. ISSN 1572834X.

Disponível em: <https://doi.org/10.1007/s11277-020-07143-2>. Citado 5 vezes nas páginas 8, 11, 12, 13 e 14.

KIZZA, J. *Computer Network Security*. Springer US, 2005. ISBN 9780387204734. Disponível em: <https://books.google.com.br/books?id=lT5Pmcmr574C>. Citado na página 11.

KUMAR, S. et al. Deep in the Dark: A Novel Threat Detection System using Darknet Traffic. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, IEEE, p. 4273–4279, 2019. Citado na página 8.

LAWRENCE, H. et al. D-miner: A framework for mining, searching, visualizing, and alerting on darknet events. *2017 IEEE Conference on Communications and Network Security, CNS 2017*, v. 2017-January, 2017. Citado na página 15.

LUCASJRT. *miss_spider*. 2021. <https://github.com/lucasjrt/miss_spider>. Acesso em: 10 mar 2022. Citado na página 29.

LUCASJRT. *hidden_answers_parse*. 2022. <https://github.com/lucasjrt/hidden_answers_parser>. Acesso em: 10 mar 2022. Citado na página 30.

LUCASJRT. *hidden_answers_scraper*. 2022. <https://github.com/lucasjrt/hidden_answers_scraper>. Acesso em: 23 mar 2022. Citado na página 30.

MDN CONTRIBUTORS. *What is a URL?* 2020. <https://developer.mozilla.org/en-US/docs/Learn/Common_questions/What_is_a_URL>. Acesso em: 7 out 2020. Citado na página 8.

MOUTON, F.; LEENEN, L.; VENTER, H. Social engineering attack examples, templates and scenarios. *Computers Security*, v. 59, p. 186–209, 2016. ISSN 0167-4048. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167404816300268>. Citado na página 13.

NUNES, E. et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*, 2016. Citado 2 vezes nas páginas 15 e 16.

PETRY, G. M. *STJ é vítima de ransomware e tem seus dados e os backups criptografados*. 2020. Disponível em: <https://thehack.com.br/stj-e-vitima-de-ransomware-e-tem-seus-dados-e-os-backups-criptografados/>. Acesso em: 11 nov 2020. Citado na página 14.

PSAFE. *Relatório da Segurança Digital no Brasil*. 2018. Disponível em: <https://www.psafe.com/dfndr-lab/wp-content/uploads/2018/08/dfndr-lab-Relat%C3%B3rio-da-Seguran%C3%A7a-Digital-no-Brasil-2%C2%BA-trimestre-de-2018.pdf>. Acesso em: 4 out 2020. Citado na página 9.

RODRIGUEZ, A. *Accessing the Dark Web with Python*. 2020. Disponível em: <https://python.plainenglish.io/accessing-the-dark-web-with-python-4db013bf4d32>. Acesso em: 9 fev 2022. Citado na página 20.

SOSKA, K.; CHRISTIN, N. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In: *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, 2015. p. 33–48. ISBN 978-1-939133-11-3. Disponível em: <https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/soska>. Citado 2 vezes nas páginas 16 e 17.

STALLINGS, L. W. Segurança de computadores: Princıpios e práticas. *Rio de Janeiro*, 2014. Citado na página 12.

STEVENS, W. R.; WRIGHT, G. R. *TCP/IP Illustrated: volume 2*. [S.l.]: Addison-wesley, 1996. Citado na página 11.

TIWARI, A. *What Is The Difference Between Deep Web, Darknet, And Dark Web?* 2020. Disponível em: <https://fossbytes.com/difference-deep-web-darknet-dark-web/>. Acesso em: 2 dez 2020. Citado na página 8.

TOR. *Tor: Overview*. s.d. Disponível em: <https://2019.www.torproject.org/about/overview.html.en>. Acesso em: 6 out 2020. Citado na página 8.

WOLF, A. *8 Most Common Types of Malware Attacks*. 2021. Disponível em: <https://arcticwolf.com/resources/blog/8-types-of-malware>. Acesso em: 24 mar 2022. Citado na página 14.

WORLDOMETER. *Brazil Population*. 2019. Disponível em: <https://www.worldometers.info/world-population/brazil-population/>. Acesso em: 4 out 2020. Citado na página 10.